

Content Moderation across Social Media Platforms

Edited by Richard Rogers

Content Moderation across Social Media Platforms

How well do social media platforms moderate content? It is a deceptively hard question to answer given much content is removed prior to publication and further moderation ranges from light labelling to subtle down ranking, which is referred to in the vernacular as shadow banning. Given the difficulties in reconstructing the scene of content removal and visibility reduction, up until now much of the research has relied upon user experiences as well as platform self-reporting.

This book details digital research methods to study content moderation, utilising the traces left on the platforms after removal as well as content performance measures and user experience simulations, also known as research personas. It examines moderation histories as well as current practice across a series of platforms and search engines: X/Twitter, YouTube, Facebook, Instagram, TikTok, Telegram, Pornhub, Amazon, Google Play, Apple App Store, Google, Bing and the chatbots, Microsoft's Copilot, OpenAI's ChatGPT, and Google's Gemini. It focuses on the problematic zones per platform (e.g., X/Twitter's uneven implementation of policies, Facebook's data lacuna, the sexualisation of children on Instagram, illegal trade on Telegram, malicious sounds on TikTok and election information in chatbots), concluding with a discussion of a sustainable moderation philosophy and its place on the public agenda.

Digital Methods Initiative
Humanities Labs
Media Studies
University of Amsterdam

© 2025 the authors

Content Moderation across Social Media Platforms

Table of contents

1. On methods for content moderation research Richard Rogers	4
2. From Twitter to X: Demotion, Community Notes and the apparent shift from adjudication to consensus-building Emillie de Keulenaar	23
3. Ranking authority: A critical audit of YouTube's content moderation Daniel Jurg, Salvatore Romano and Bernhard Rieder	72
4. The performance of borderline content on Facebook Richard Rogers and Kamila Koronska	98
5. Minors as (misused) content on Instagram Natalia Sánchez-Querubín	124
6. Malicious earworms and useful memes: How the far-right surfs TikTok's audio trends Marloes Geboers and Marcus Bösch	151
7. The internet's dark alleys – Laissez-faire content moderation and illegal trade on Telegram Stijn Peeters	170
8. Grey areas of content moderation: A trace analysis of Pornhub Lucia Bainotti	189
9. Gatekeepers of the mobile ecosystem: Understanding app store moderation Esther Weltevrede, Anne Helmond, Fernando van der Vlist, Stefanie Duguay and Michael Dieter	216

10. Walking the store: Cultic networks and content moderation on Amazon.com	244
Marc Tuters	
11. Contested components: Studying interface enrichment as a form of content moderation on Google and Bing	273
Sal Hagen and Guillén Torres	
12. DSA, AIA and LLMs: Approaches to conceptualizing and auditing moderation in LLM-based chatbots across languages and interfaces in electoral contexts	309
Natalia Stanusch, Raziye Buse Çetin, Salvatore Romano, Miazia Schueler, Meret Baumgartner, Bastian August and Alexandra Roşca	
Author notes	329

1. On methods for content moderation research

Richard Rogers

Abstract

How well do social media companies moderate the content that is posted on their platforms? The question likely would set off a debate in at least two directions. The one concerns the kind of balance that is struck between content removal and retention, including levels of visibility and labelling, especially for what is termed borderline content, or materials that (to the platforms) do not quite cross the line to be removed. The second is how content moderation is undertaken as well as catalogued, especially now that the practice has come under greater scrutiny through such measures as the European Digital Services Act (DSA). This piece briefly sets out some moments in the history of content moderation research that include the discovery of the invisible labour behind it, the embrace (and subsequent disavowal) of fact checking, the rise of AI and automated downranking or shadow banning and larger questions concerning the shift in philosophy from adjudication and fact checking to consensus-building and Community Notes. It subsequently turns to methods researchers employ to study how well moderation is performed, particularly content performance measures, trace analysis and research personas and what these yield in studies of X/Twitter, YouTube, Facebook, Instagram, TikTok, Telegram, Pornhub, Amazon, Google, Bing and the chatbots, Microsoft's Copilot, OpenAI's ChatGPT, and Google's Gemini. A tightening of moderation is recommended for Instagram (in the realm of children's content), TikTok (in malicious sounds), Telegram (in illegal trade), Pornhub (in deepfake how-to's), Amazon (in their labelling) as well as the LLMs (in their provision of information about elections). Also to be addressed are X/Twitter's uneven moderation policies as well as Facebook's data access obstacles. Of note here is how the researchers were not able to make use of the new data access provisions enshrined in the DSA, resorting instead to scraping, marketing data dashboards and research personas. Finally, I turn to the question of moderation practice and philosophy and its place on the public agenda.

Keywords: content moderation methods, content moderation philosophy, trace research, research personas, content performance measures

Introduction: Touchstones in the history of content moderation

The following is an effort to organize content moderation research methodologically. At the outset it should be mentioned that content moderation has not been a well-covered subject matter in the history of social media platform research. There are several touchstones in that history.

One is an article in *Wired* magazine in 2014 that introduces "content moderation" (with quotation marks) as the "removal of offensive material" and paints the picture of the workplaces where it is performed, including "on the second floor of a former elementary school (...) in Bacoar, a gritty Filipino town thirteen miles southwest of Manila" (Chen, 2014). The investigative journalist gained entry to the workplace, describing the kind of content screened ("pornography, gore, minors, sexual

solicitation, sexual body parts/images, racism"), the speed in which a retention or deletion decision needs to be taken ("a few seconds") and the effects of that work on the moderators ("burnout is common"). In that vein content moderation has been called social media platforms' "dirty little secret" (Drootin, 2021).

There is the work on this "commercial content moderation", as Roberts calls it (2016), who like Chen in *Wired* magazine and journalists at the *Süddeutsche Zeitung* (Grassegger and Krause, 2016) was able to interview (former) workers, who likely signed non-disclosure agreements so as to keep their work out of the public eye. Indeed, the details of the work, if not the workers, were intentionally made "invisible" (Roberts, 2016) until 2017 when the newspaper, *The Guardian*, published portions of Facebook's "internal rulebook", the moderators' training materials (Hopkins, 2017). For its part, a year later, Facebook decided to make their "community guidelines" public, where much of the reporting focused on their secret past but also on the complex decision-making at hand in moderation (Wong and Solon, 2018). Mistakes were often made. While there was extensive moderation circumvention taking place (Gerrard, 2018), Facebook signalled some openness by allowing appeals from users who felt their content had been unfairly removed.

Extensive scholarship also began to appear at that time from *Custodians of the Internet* (Gillespie, 2018) and *Behind the Screen: Content Moderation in the Shadows of Social Media* (Roberts, 2019). At this point interest in the subject matter took off, coinciding with larger societal debates concerning the role of 'big tech' in fending off misinformation (aka the infodemic) surrounding the Covid-19 pandemic. Platforms' content moderation also was being framed in terms of censoring free speech. Social media platforms had become 'accidental authorities' (de Keulenaar et al., 2023) in arbitrating which content should stay and which should go. As a whole, content moderation had moved from being the platforms' quietly outsourced problem to a core function or even, as one scholar argued, at the center of their business models (Gillespie, 2018).

How well do platforms moderate? It is a deceptively difficult question to answer, given that moderation, at least when thinking of the outsourced work described above, as occurring on the back end, prior to publication, with binary choices (retain or remove). There are EU-mandated reporting mechanisms available, in the context of the Digital Services Act, that details such take-downs (European Commission, 2024). Apart from the DSA Transparency Database, the companies also provide transparency reports. While seemingly helpful, these self-reporting practices, according to some critical studies, provide a "legitimising force" for the social media platforms and suggest regulatory compliance; for researchers, however, "information asymmetry" remains (Maroni, 2023). As some have found, they are not helpful methodologically, for the entries made by the platforms in the DSA database are basically category counts and lack references to the removed posts (Geboers & Bosch, this volume; Jurg et al., this volume). Certain removals also appear miscategorised such as Amazon's (Tuters, this volume). They may provide a quantitative showpiece for compliance activities but are less useful in answering questions about the quality of the moderation taking place. How else to study it?

Where is content moderation?

When considering its study, moderation, broadly speaking, takes place at multiple levels, or in a stack, as we have characterized it (see Figure 1.1).

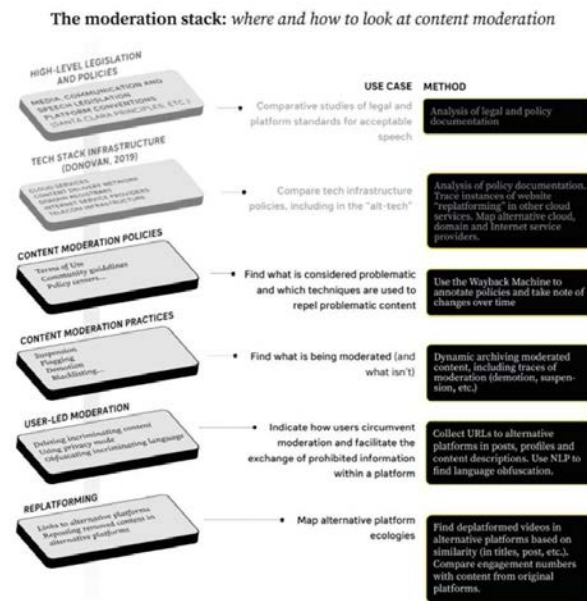


Figure 1.1 The moderation stack and its study. Source: de Keulenaar & Rogers, 2025.

Using the stack as a means to organize various methods for the study of content moderation, in the following I concentrate on moderation policies and practices as well as user-led moderation and moderation discovery. I begin the discussion with five approaches, which (with one exception) loosely fit how the researchers have taken up their studies in this book. While I separate them into specific approaches, in practice they are often mixed.

- Content performance – Post ranking by engagement and source labeling
- Moderation experiences – User ethnography of demotion and other encounters
- Moderation policy histories – Timeline work with archived web pages of platform policies
- Research personas – Evaluating user experiences and algorithmic effects through simulation
- Trace research – Reconstructing the scenes of content disappearance

Each is taken in turn before discussing the approaches as well as findings made in the individual case studies in the book.

Content performance

In 2016 Craig Silverman of BuzzFeed News published an article showing how ‘fake news’ had outperformed ‘mainstream news’ on Facebook in the run-up to the U.S. presidential election (2016). It stirred anxieties about the quality of political information on the platform as well as Facebook’s ability to curb misinformation, including conspiracy, imposter as well as hyper-partisan content (which is how ‘fake news’ was defined). It also ushered in a debate about what constituted ‘fake news’ and later ‘borderline content’, a term originating from YouTube describing posts that ‘border’ on but do not cross over its hate speech policy (Meyer, 2018).

Perhaps underappreciated in Silverman's fake news piece was the method used in its study. Election-related keywords are queried in BuzzSumo, a social media data tool, which returns a list of web URLs mentioned in related Facebook posts, ranked by engagement. Silverman took the top twenty mainstream and top twenty 'fake news' sources by counting interactions (or engagement) and graphed these two source types overtime, showing that fake news had received more interactions in the period just prior to the election. He labeled sources according to his journalistic expertise, identifying conspiracy websites as well as those claiming to be news sources but were imposters. He also listed hyper-partisan websites ('openly ideological web operators') in the category of fake news (Herrman, 2016), which later would become controversial given that most found in his and subsequent studies were to the right of the political spectrum.

Methodologically, there are sites such as NewsGuard and Media Bias / Fact Check which can be deployed for source labeling rather than relying on one journalist's judgement. Additionally, there are lists of sources that have been fact checked by fact checking organizations. When deploying these labels, one is able to gain an indication of how well 'low quality', politically charged or otherwise labeled sources are performing vis-à-vis more mainstream sources, which themselves may experience shifts.

User studies of moderation experiences

There are ethnographic approaches to the study of content moderation, including the seminal work surrounding 'algorithmic imaginaries', or the situations in which users become aware of the work of social media moderation and discuss with fellow users as well as researchers their experiences. These user stories about downranking ('shadowbanning') and sudden drops in engagement (for example) are sourced in a variety of ways from social media posts as well as comments.

In the algorithmic imaginary work, to identify interesting user stories about content moderation practices, researchers have queried Twitter for 'Facebook algorithm', 'algorithm AND weird', 'algorithm AND great' and so forth, identified users with (sometimes uncanny) stories about how platforms seem to work and invited them for interviews (Bucher, 2017). The material is subsequently organised around observations of the work algorithms do, e.g., visibility reduction. User stories may accumulate and snowball into moderation language that is critical of platform policy. 'Shadowbanning' is one example of a user term that has gained considerable currency; it refers to one's post being demoted or removed without knowing it (Leerssen, 2023). User ethnography is one technique to study both the detection as well as the effects of shadow banning, perhaps together with examining the ranking of a post over time, as discussed below in trace methods.

Moderation histories

Moderation histories may be pieced together by consulting archived platform policy pages in the Internet Archive and comparing changes to them over time (de Keulenaar et al., 2023; Katzenbach et al., 2023). It is relatively straightforward work if the URL structure of platform policy webpages have remained stable. In which case, one would consult current policy or guidelines pages and query them in the Wayback Machine of the Internet Archive, identifying which pages contain new or altered content. In all

likelihood, however, the URLs of the policy pages have not remained stable. Rather, there will be a variety of pages detailing a platform's content moderation, linked from one policy page to another. Here one may use a tool that extracts links from archived pages in order to make a list of pages that subsequently can be perused. One is ostensibly examining the strengthening or loosening of moderation policies and practices, including in the context of new ownership (in the case of Twitter to X), European Commission directives and other externalities. Comparing one platform's policy to another is also of interest as is putting side by side national legal requirements and policy. Is there mere compliance in the policies or is there more to moderation than law-following?

Research personas

Personas are social media profiles created by researchers that enable the study of algorithmic personalisation as well as moderation (Bounegru et al., 2022). One creates an account, and populates it by building a history and expressing preferences. Once the persona has been sufficiently built, the research goal is to learn how a system 'reacts' as it recommends content. For example, does it recommend more and more extreme or sensational content with the expectation to keep the user engaged (and watching)? This is the study of rabbit holes and other algorithmic pathways, often performed on YouTube (O'Callaghan et al., 2015). In one case of interest, researchers built a persona interested in Viking culture, queried both Google and YouTube for related keywords, watched a series of YouTube videos, all the while asking whether these interests would eventually lead the persona to problematic adjacent content, such as extremist material (van Wonderen et al., 2023). Rather than algorithms nudging one towards extreme content and down a rabbit hole (on YouTube), they learned that to find such content a user had to search for it. It was not recommended to the persona that expressed considerable interest in Viking culture.

Trace research

Finally, trace research studies moderation practices by seeking to reconstruct the scene of disappearance (content removal) through the system messages still online. Using pre-existing lists of URLs, such as old Twitter data of the Alt Right or Brexit, one scrapes platform metadata, that is, flags or context labels. Which traces remain of the posts that have been removed? What do these tell the researcher about the reasons for removal?

In more of an anticipatory mode, the work also can be performed in combination with agile or dynamic archiving where one continually accesses or calls content potentially susceptible to removal or demotion, capturing its status as well as ranking over periods of time. If and when the content is removed, the system messages (traces) presumably would provide the rationale for removal. As mentioned above in the context of shadow banning, trace research into post rankings (over time) can complement the user ethnography, providing a foundation for the user stories.

Moderation studies in brief

In the following these methods are put to use in a set of studies of platform content moderation. Each deploys a variation on piecing together the history of a platform's moderation policies. Subsequently, there are empirical studies of moderation. The X/Twitter study details the changes to moderation policy in the transition to X and also provides indications through trace research (and dynamic archiving) of the kind of

content X tends to moderate (and which it does not). The YouTube work deploys trace research (with dynamic archiving) as well as API calls to study the extent to which election-related content has been moderated, be it raised, demoted or removed. The Facebook research adapts the original data journalism method for the study of ‘fake news’, examining the presence of borderline content in election-related posts, comparing the cases of the U.S. presidential elections of 2016, 2020 and 2024. For Instagram, the research persona approach is taken in order to chart accounts and networks that sexualize children. The work on TikTok takes on a user-led approach, studying moderation circumvention in German extremist circles. The Telegram analysis adopts a research persona who is able to enter into areas where illicit goods and services are routinely on offer. The Pornhub work concerns itself with the traces left behind when videos in the category of newly uploaded are removed or suspended. For the work on the Google and Apple app stores, the research combines the moderation timeline with trace research on upward and downward app rankings in the context of the Russia-Ukraine war. The piece on Amazon book marketplace examines the question of the moderation of extreme content by mapping recommendations to them, asking questions about when recommendation moderation may or may not take place. The analysis on moderation in chatbots uses a sock-puppet or research persona approach, examining how Microsoft CoPilot and Google’s Gemini respond (often incorrectly) to factual questions about elections. Finally, the comparative analysis of Google and Bing departs from the question of the performance of borderline content in search engine results pages (SERPs), which now include not only organic results but a variety of AI-curated content to be evaluated.

Mixed digital methods for the study of content moderation

Most projects described here mix such methods as the use of traces, research personas as well as content performance (or ranking) where the platform is often ‘audited’. Auditing, originally, is a social scientific methodology for exposing social discrimination (in housing and loan applications) and has been applied to search engines and LLMs (to tease out offensive autocompletions). In a sense most studies that scrutinise content deprivileging (as well as privileging) mechanisms could be called audits.

Trace research is the deployment of data generated and accrued on the platform after user or platform activity, such as user clicks and views as well as platform removal notices, respectively. Up until recently, most trace research focused on user activity such as which posts users like, share and comment upon more than others. Much online research depends on these engagement metrics or sums of likes, shares and comments (or their equivalents across social media platforms).

But trace research of this kind also has been critiqued for its incapacity to distinguish between user intention and platform nudging. Does the post receive so many more likes and shares or the video more views and upvotes because they are popular among users or because they are recommended by the platform? How would a researcher disentangle the difference between user interest and platform effects or come to understand the shares of each? Additionally, has content been down ranked? This platform action also would affect a post’s visibility and thereby its potential to accrue likes and other metrics.

As argued elsewhere, not all trace research concerns user traces. Trace research can just as well focus on platform traces, which are the tracks or vestiges left behind by platform actions (de Keulenaar & Rogers, 2025). Most palpably, these are the takedown notices put up by the platforms after content has been removed. Their careful study can make much of the difference between 'this video has been removed' and 'this video has been disabled' for the one may point principally to users and the other to platform intervention. Platform trace research also takes up signs of what has been called content visibility reduction or, in the internet vernacular, shadow banning. Which content has been downranked? In a sense this is the study of the flip side of platform nudging and algorithmic amplification.

Platform trace research, at least in this instance, is in the service of the study of content moderation (rather than being hindered by it). The trace research methods put to use here are of three varieties. One could be called continual archiving, where newly posted content is monitored when it is uploaded and at incremental intervals thereafter. A second trace concerns tags, which are both suggested by the platform and added by the content creator. Creator tags are moderated by the platform, and one may compare the platform default tags with the creator-led ones. What do the moderated tags say about the moderation policies? Both sets of tags become additional, suggested search queries given after the user has entered queries of their own. The third type are the notices served when searching such as 'your search could be for illegal (...) material'. As with the related search terms, there is the question of the extent of the moderation. Here one could study which queries trigger such notices and which do not.

In the Pornhub study, the URLs as well as the data associated with 'new videos', as the platform terms them, are logged upon upload and after an hour, a day, a week and a month (Bainotti, this volume). Such a continual archiving practice allows one to trace changes to the status of the videos and associated metadata. Another set of traces are captured by querying the platform for a set of sensitive keywords and logging the related searches that the platform suggests. When such queries are made, they may trigger the platform to service notices of their inappropriateness. These notices become further platform traces to be scrutinised.

On the strength of such trace research, the Pornhub study found active moderation on the part of the platform and specific forms of user compliance to moderation policies. Certain videos are removed quickly after upload, despite that they are user-tagged in compliance with platform policies. (The platform's 18+ age tag is reappropriated as 18+-cute-girl and so forth.) It also found what the study calls 'grey areas'. These concern why Pornhub 'disables' certain videos, a category of action that is not mentioned in its moderation policies. Removals for a variety of violations are documented, including terms of service and copyright violations. Another grey area is how the related searches users receive after their initial query may lead to borderline content, on the edges of the policies concerning non-consensual acts and child content. Finally, searches for deepfakes and 'deepnudes' return videos for tutorials on how to 'undress' people in photos using generative AI.

Research personas are pre-trained platform "users" created by researchers for the purposes of studying personalised information experiences. They are particularly useful for studying algorithmic feeds that are tailored to an individual user, be it the Facebook Feed or a 'For You' page on TikTok or X/Twitter. This approach is often

contrasted to the API- or search-based ‘querying’ where the researcher makes a list of keywords or accounts and extracts the posts returned by those queries. ‘Big data’ or even smaller results sets could be said to constitute a platform-wide mode view, whereas the data collected from the feeds of research personas are more proximate to actual user experiences.

Training the personas is the starting point of the research and here the question of realistic usage comes into play. Is the research concerned with approximating actual users or eliciting returns consistent with a stereotypical subject? Likely the outcome of persona production will fall somewhere in between unless user clickstream data and other signals inform the construction of the persona from cold start to well-formed user. More likely is a ‘search and click’ strategy, where the persona is constructed through deploying keywords and following algorithmic recommendations. For example, one piece of research on extreme content drew up a list of Norse gods, queried them (in this case in newly made accounts on YouTube and Google) and watched several returned videos as well as those recommended up next on the YouTube interface (van Wonderen, 2023).

In the research on extremist soundscapes on TikTok (Geboers & Bosch, this volume), the researchers trained right-wing personas by searching for an event (Solingen knife attack) that stirred anti-immigrant sentiment as well as hashtags associated with the same. Having prepared the feed through persona work, the researchers subsequently zoomed in to two recurring sounds (‘Turk, Turk’ as well as ‘Ticks’), one more subtle and the other more outright in its hateful outlook, in order to compare their presence on the platform after a period of time (in this case two months).

TikTok has a feature called ‘use this sound’ which content creators can select when creating their videos in order to jump on a trend or otherwise join others in its usage. Videos with ‘this sound’ are aggregated on sound pages, where all videos containing the sound are listed. This is useful for research, as in this case it collects all the videos, together with their users, embedding the same sound. In the hate speech project, the researchers, having collected data from two such sound pages, conducted a moderation trace analysis, finding that videos with either sound as well as their content creators largely remained on-platform months later, despite TikTok’s policy of addressing hate speech on its platform.

How do hateful sounds remain on the platform? They are cloaked, it is argued, through a series of furtive practices such as using sticker text trends that are overlaid on the videos and remixing popular songs with racist lyrics. As such they are less susceptible to moderation. As the authors argue, these are ‘niche soundscapes’, presumably without a large FYP viewer base, and as such as less likely to be flagged by users outside of its subculture.

Other work adopts a user persona using a walkthrough method, a digital ethnographic approach where the researcher navigates the platform observing what one can find when (in this case) he or she is looking for banned and borderline content, be it illegal goods (on Telegram) or content that sexualises children (on Instagram). Light pioneered the walkthrough approach (for studying mobile phone apps) where the researcher “assume[s] a user’s position while applying an analytical eye” (2018, 819). In the method, the researcher maintains a kind of naïve stance as a ‘first-timer’ in this

content area (Grouch, 2006) but is otherwise an expert in researching digital platforms.

For Telegram the question is the presence of illegal trade in drugs, forgeries, match fixing, gambling and so forth (Peeters, this volume). Telegram has the reputation for being lightly moderated and thereby an inviting online space for such trade or trade facilitation, where platform introductions result in off-platform deals.

Remarkably, that lack of content moderation was affected just as the research on the case study progressed. During the researcher's walkthrough of the platform and its associated data collection, Telegram's CEO was arrested – precisely on charges of allowing drug trafficking (and other crimes) on its platform. As one might expect, heightened scrutiny of the platform and more moderation by Telegram followed directly thereafter (Tidy, 2025).

An exploratory approach as the walkthrough method is appropriate for Telegram given the limitations in querying its two public-facing spaces of interest to the researcher: channels (which can be followed and where administrators tend to broadcast) and groups (which can be joined and where all may contribute). The project utilised a public database (telegramchannels.me) of top channels per country, including the Netherlands which is the focus. Having determined relevant ones among those the researcher subsequently snowballed the sample. Channels can share other channels in them, and a relevant set can be assembled through a shared channel network analysis and grouped into the aforementioned categories, e.g., drugs, forgeries, etc.

The study of moderation becomes an act of comparing the presence of the channels (and their categories) across the time frames, together with any traces left when a channel is gone. Do these channels that facilitate trade in illegal goods remain on the platform over the period of analysis? Which traces of moderation are in evidence when comparing channel presence across the periods of time?

Remarkably, half of the seed channels were no longer present three months after initial collection, and the largest category of removal was drugs. Users would have been met with the message that it 'no longer exists'. The strongest indication that the cause is platform moderation rather than user flight consists in users themselves producing 'back-up channels' and discussing them as potential remedies for removal.

Another study that makes use of a research persona concerns Instagram, a platform used by both children and adults (Sanchez Querubin, this volume). Controlling access to media content for different age brackets has a long history, including movie ratings and television programme scheduling. Instagram has developed its own medium-specific techniques, including default content recommendation settings for accounts under certain ages and personalised moderation for adults.

It is well known that children under the age of thirteen have accounts and teens act like adults and are served their content (BBC, 2022). Younger children and teens may seek Instafame or otherwise develop a circle of friends sharing pictures and videos, following each other and liking their content. But they also may fall victim to abuse. Instagram requires some form of age verification and looks for 'age signals' through

AI training. The purpose is to create separate content spheres not only for recommendations but for advertising purposes.

Additionally, there is the question of adults making contact with children and (for example) praising their content and asking for or sending them pictures. One technique is to ask them to ‘trade’ pictures; another is to manipulate their online images and blackmail them. That children are lured into inappropriate contact with older users is a recurring issue for the platform, which seeks to address it by adding features and settings. Turning on ‘limits’ disallows new contacts from interacting with the content, and ‘nudity protection’ blurs and warns about such detected content. The latter is a default setting for users under 18 years of age.

Apart from as an age-verified account holder, content consumer and someone who communicates with others on the platform, the Instagram study ultimately takes up how the platform considers a child as a type of content. It examines what is referred to as the media ecology (on Instagram and beyond) around pictures and videos of kids in bathing suits, leotards and such. These may be sexualised in the comment space or through making collections of them and advertising them. Indeed, the empirical portion of the study deploys a research persona or sock puppet who displays interest in young girls’ accounts under the age of 13 (run by their parents). These girls are interested in modelling and gymnastics. The research found inappropriate comments on posts by older men, who also follow other children. When these men are followed by the research persona, it begins to receive recommendations of other childrens’ accounts as well as the men who follow them, producing a problematic ‘community of interest’. The moderation gaps identified include seller accounts, sexualised content collections and comment sections awash with inappropriate remarks.

A Change in Moderation Philosophy

X’s platform moderation has been the source of much scrutiny, given the platform’s alleged transition to a ‘free speech’ regime, with lower moderation, but also the inconsistency with which that philosophy has been applied. What is clearer is that X (and Twitter before it) have shifted away from a ‘vertical’ moderation stance based on demotion and deplatforming to one that revolves around consensus-building in the form of so-called Community Notes. (It is more a philosophical change than a complete uprooting as certain posts are still removed or their visibility reduced.) Community Notes are labels written and rated by volunteer users giving further context to posts; they are attached to a post when a particular level of consensus about their helpfulness is formed. With this new emphasis in how the platform is moderated, a ‘bridge-building’ technique relying on a form of crowdsourcing is replacing what has been called the ‘accidental authority’ invested in social media platforms to make determinations of what is permissible content (de Keulenaar et al., 2023). Following X, Meta, and its Facebook and Instagram platforms, subsequently latched onto Community Notes as a moderation practice, ending some partnerships with authoritative fact-checking organizations and otherwise similarly embracing a free speech standpoint.

Given the rise of Community Notes, the question arises concerning its overall effectiveness compared to the vertical regimes. The research reported here, conducted on posts concerning immigration in Dutch, found that a majority of notes has not reached a level of consensus for them to be posted (de Keulenaar, this volume). Such a

finding is in keeping with other studies where the level of bridge-building about controversial subjects is insufficient for action to be taken, leaving consensus unsettled and posts unlabelled. In a second tranche of the study, concerning the content, there was not an appreciable difference (in toxicity scores) between the content that had been moderated and that remained online.

At the time of writing, the rollout of Community Notes on Facebook (and Instagram) is underway; the set of volunteers (some 200,000) has been formed, and the systems (based on X's open-source algorithm) are being put into place. The fact-checking program (in the U.S.) has been ended. The rationale given by Meta for the switchover to Community Notes over fact-checking is the balance in perspectives: "We expect Community Notes to be less biased than the third-party fact checking program it replaces because it allows more people with more perspectives to add context to posts" (Meta, 2025).

It is a course change for Meta, given its history of content moderation, especially since the 'fake news' crisis of 2016, which was prompted in part by a data journalist's research that found that problematic information ('fake news') had outperformed mainstream news in the run-up to the presidential elections that year (Silverman, 2016). The study was repeated prior to the elections in 2020 and again in 2024 (Rogers, 2023; Rogers and Koronska, this volume). While it appeared to worsen in 2020, in 2024 the amount of problematic information concerning the elections had lessened, seemingly through a strategy of content moderation to depoliticise the platform that reduced the visibility of such content overall.

The three studies all employed data from BuzzSumo, a marketing data bureau, rather than from Meta, querying for election-related keywords and receiving web URLs appearing in posts ranked by engagement. It is a method that relies on engagement as a form of ranking of popular content as well as labelling of the web URLs by third-party source evaluators. Is the most engaged-with content problematic (according to the labellers)? How does it perform relative to more mainstream sources?

One reason for using BuzzSumo data is consistency of approach overtime. Another is Meta's (and Facebook's) repeated closures of data pipelines, the most recent of which (the demise of CrowdTangle) took place just over 2 months prior to the 2024 elections. Other Meta data sources (from the Social Science One project) created in part to study elections already had dried up and were longer current or fit for purpose. The replacement – the Meta Content Library – was still in its infancy and too convoluted to access by the time the elections rolled around.

The other platform whose content moderation appears to have seen improvement is YouTube. It was the platform most associated with the idea of a rabbit hole and called the 'great radicaliser' for the manner in which it recommended ever more extreme content (Tufekci, 2018). It also was the platform that elevated the reach of far-right 'alternative influencers' (Lewis, 2018). More recent research has found the rabbit hole thesis to be overturned or at least no longer current. While not recommended through seemingly adjacent areas of interest, extreme content is still found when purposively sought.

The approach to the study of content moderation on YouTube during the run-up to the European Parliamentary elections reported here is an audit that seeks to determine which source types YouTube deems authoritative (Jurg et al., this volume). Platform (or search engine) audits, derived from the social scientific tradition of auditing for discriminatory practices (e.g., in housing applications) examine the privileging and de-privileging of types of content and/or accounts. Here the approach is to identify which source types rank higher in the results from queries about election-related keywords. What the researchers term legacy media as well as public service media were by far the most returned sources from election-related content rather than say alternative influencers or other YouTubers.

Apart from evaluating which kinds of sources are highest ranking, in a second step the moderation project also examined removals. To do so, it looked for traces of take down's from the original (earlier) data set of videos returned for election-related queries. They found some 10% removed, which they then compared to another issue space with greater likelihood of removal (where the rate was just under 30%). Certain videos belonged to channels that remain online, such as one whose work has been featured on RT and Sputnik, the Russian sources banned from the European Union (since the Russian invasion of Ukraine).

Generally the researchers found YouTube's moderation reasonable, but recommended more granular data on removals rather than the bean counts provided in the DSA Transparency Database.

The last set of studies to be discussed has seen perhaps the least attention in the scholarly efforts to understand their content moderation. They concern Amazon books as well as three comparative cases: Google Play and the Apple App Store; Google and Bing; and the chatbots, CoPilot (Bing), Gemini (Google) and ChatGPT.

Content removal on Amazon book marketplace (amazon.com) would appear to be a particularly sensitive issue, considering that such a euphemism as "content moderation" could be supplanted by the idea of book banning, an historically fraught act that conjures book burning and similar intolerances and impinges upon media freedom. While adhering to national laws (in the some 23 countries in which it operates its book marketplace), Amazon and specifically amazon.com, the flagship site, historically has taken a position against bans, instead occasionally affixing content warnings along with a justification of a controversial book's availability in the store because of its historical significance. Thus it is of some significance that Amazon appeared to join the wave of enhanced content moderation occurring around the Covid-19 pandemic, as the study reports (Tuters, this volume).

Given that Amazon lists its removals in the DSA database only as 'privacy violations' when the reasons may differ, the study in question focuses on what remains online, employing a type of content performance method, where recommendations of related books are studied. When querying for titles of interest to neo-nazis and health conspiracy theory, which other books does Amazon book marketplace surface? Are they labelled?

After a consideration of cases in which Amazon removed or otherwise labelled certain books over the years (which results in a periodisation of moderation), the analysis is

based on algorithmic recommendations via co-consumption (people who bought this book also bought...). It focuses on the availability (and abundance) of anti semitic works as well as titles concerning health misinformation. (It also examines the proximity of conspiracy theory and media studies.) Ultimately it makes the case for Amazon to reintroduce a "measure of content moderation" to its book marketplace, where the analysis surfaces prime examples of unlabelled material.

App stores moderate apps through what the authors refer to as a gatekeeping structure comprising an initial review and subsequently a variety of smaller measures such as "query governance" (influencing the ranking and visibility of apps in stores) and "geo-blocking" (restricting the regions in which they may be downloaded). The authors set up their analysis of moderation by describing how the app stores respond to queries for banned content (in this case pornography), demonstrating that its moderation extends beyond forbidding returns to suggesting alternatives such as anti-porn ('stop addiction') as well as religious apps.

There are two further case studies discussed in some detail: the app stores in the time of the Covid-19 pandemic and in wartime after Russia's invasion of Ukraine. They found that the app stores were in a state of exception during the pandemic, re-routing queries to specially curated sets of results, often restricted to official apps. Here the app store is said to assume a 'quasi-governmental role' (Weltevrede et al., this volume). It was also found that these approved lists of official apps could be bypassed by misspellings and other algospeak stratagems, which points up a moderation task for the stores. In the second case, after the Russian attacks in February 2022, the app stores witnessed some dramatic use perturbations from the banning of apps (such as RT and Sputnik but many thousands more), the rise of an alternative RuStore for Android as well as the sharp rise in VPN downloads (and lite versions of TikTok) in Russia. Russian alternatives replaced social media and music apps, developing its own app ecosystem. In Ukraine app usage for Signal and maps.me suggest another response to the war.

Both the pornography as well as the Covid-19 cases are query-based analysis that fit with the content (or app) performance methodology for studying content moderation. The case of the app stores during wartime is also an app performance method, but it is based on data from an app analytics platform, providing data on app rankings by country, for example. In all the authors call for additional data disclosure by the stores that would enable moderation analysis.

Search engine interfaces have been evolving since the early results pages listed web URLs in a ranking according to relevance, with the addition of ads but also a variety of other elements such as related queries, knowledge boxes, news, images, videos and more. Recently the interfaces have been further enriched through the addition of such components as AI-generated answers. That such enrichment could be considered moderation is the point of departure in the comparative study of Google and Bing.

As the authors report, Google and Bing commissioned audits from accounting firms of their search results and their reports cleared them; both were found to be in compliance (or could address certain issues such as Bing's non-compliant ad library). But no empirical work was undertaken (or at least reported) on how the engines moderate content, particularly with an increasingly rich search engine results page that

assembles multiple components. Indeed the conceptual point of departure of the study is that the complexity of the set of components on the interface is itself a form of content moderation.

The audit undertaken by the researchers relies on a custom pipeline (open sourced for reuse) that sources conversational topics from an edgy web forum (4chan) and queries them in the search engines, analysing the appearance of interface components. Generally they found that controversial subject matters are far less enriched than non-controversial ones, leaving these to ranked web sources. It also found some "hard exclusions" (Hagen and Torres, this volume). For example, a "trump" query resulted in no related queries or AI answer box. The findings raise questions about the politics of information, or how search engines adapt their outputs to changing political tides.

The study on GenAI platforms (or LLMs) concerns the extent to which they provide election misinformation, both by default as well as through explicit prompting to do so. The plebiscites in question are the US presidential and European parliamentary elections. It employed research personas in the sense that it undertook the work from specific country IP addresses, and audited the LLM output for information accuracy and propaganda usage. First, querying them, as one would a search engine, for election information resulted in some errors and wide inconsistency across languages, calling into question how well the LLMs maintain election integrity, a crucial question in European legislation on platforms (DSA), which, as the researchers also report, is a designation that eventually apply LLMs, especially those (such as CoPilot and Gemini) which are integrated into the Google and Bing search engines, respectively. (The researchers also repeated prompts some months later in CoPilot and found exceedingly divergent results which worsened the picture.) There is also the question of LLM as a propaganda machine. Can they be used for 'disinformation-for-hire' or what the researchers call propaganda as a service? Again here the study answered in the affirmative. Ultimately, the authors call for greater data access to study moderation and more transparency from the companies with respect to how it is conducted, which brings us to the recommendations across the studies more generally.

Conclusions: Recommendations from across the studies

Methodologically speaking, the book outlines five approaches to the study of content moderation (content performance, user experiences, policy timeline reconstruction, research personas, and trace research), but in practice many are mixed and there is a tendency to employ three of them (content performance, research personas and trace research). Each of the studies also undertakes some moderation policy timeline work, often using the Wayback Machine of the Internet Archive. (There are no studies in this volume that deploy the method of studying what users say about their moderation experiences.)

To begin, for all the studies it is remarkable that none made use of the data supplied by the platforms and search engines that populate the DSA Transparency Database, which contains acts of moderation by category, feeding a dashboard. Only the YouTube, TikTok and Amazon studies reference the database. They found it unhelpful; the YouTube and TikTok studies would have benefited from video and channel take-down data and the Amazon analysis by more careful (or accurate) categorising of moderation

activities on the part of the company. Indeed, the lack of use of the database goes hand in hand with the call by most every study for improved data.

Only the YouTube study relied on the platform's API; many studies used scrapers and/or data from analytics and marketing companies. It could be argued that this research work occurred precisely at a time of transition for the companies to DSA-compliant APIs. It can also be said the researcher reliance on scrapers and analytics companies sums up the current state of social media and search engine data access for researchers.

In the remainder I would like to regroup the studies by principle methodology, focusing on these three research styles and pointing out the kinds of findings each yields. Overall I aim to provide a way forward for content moderation research and its scholarly organisation. The findings have implications in how they point up moderation tasks for the platforms (or regulators interacting with platforms). There is also the question of what else the policy making arena can do, which is where I conclude.

Query-based work on content performance strives to reveal privileging (and deprivileging) mechanisms of platforms and engines, resulting in audits. How well does election misinformation perform? On Facebook its resonance has been significantly reduced since 2016 and 2020, but such a finding was made just prior to the change in moderation philosophy (from the adjudication of fact-checking to the consensus-building of diverse viewpoint agreement); thus moderation research should examine the effectiveness of Community Notes over previous adjudication regimes (or the efficacy of Community Notes in and of themselves). In three popular LLMs election integrity issues are prominent, more so in low-resource languages, indicating a moderation task at hand. The YouTube, Amazon and the comparative App Store studies also examine content performance (be it of videos, books or apps), where the question is when do ill reputable media sources, controversial books or unapproved advice apps become recommended. On YouTube (during election season) they rarely do, but for Amazon the study found certain works worthy of warning labelling; the App studies work witnessed how Stores are susceptible to algospeak. For the Amazon book marketplace as well as the App stores, these imply moderation tasks. YouTube, it was found, actively promotes public service media (over YouTubers and influencers) at least for election-related queries, which could serve as a recommendation for other platforms.

Research personas were built for the analysis of Telegram, Instagram and TikTok. These personas or sock-puppets are simulations of everyday users, described (in the Telegram and Instagram analyses) as 'first-time users' walking through the platform with an 'analytical eye'. All three studies found a plethora of problematic materials and milieus. On Telegram there is a thriving trade in illegal goods and services; on Instagram there is an environment of adult users following children and collecting their pictures in bathing suits and leotards, redirecting them to off-platform repositories. Instagram also recommends these users to a persona showing interest in such materials. For Telegram the moderation task appears to be clear in the sense that there is illegal trade that is present on the platform, but there is also the larger question of the cultivation of such activity on the platform in the first place. Instagram for its part has

more work to do in separating children and adults on the platform; it also could examine when it shares user recommendations.

The TikTok study had a somewhat different character in that the persona was not so much a first-time user as one trained to locate extremist content, and particularly hateful sounds that the researchers call ‘malicious earworms’, meaning they are also somewhat catchy or memetic. It thus looked at the circulation of those sounds across the platform overtime, as others can ‘use this sound’, a platform feature. Had they been removed or otherwise moderated? Both ‘Turk, Turk’ as well as ‘Ticks’, the researchers found, are part of a larger network of hate speech content that is somewhat camouflaged on the platform. TikTok’s moderation task would be to respond to its exposure as such.

The X/Twitter and Pornhub work relied on trace research. The Pornhub study logged newly published videos, examining time and time again whether they had been removed or otherwise moderated. It also studied related search terms for borderline queries, thereby combining trace research (notices of removed videos) with content performance (recommendations of related searches). It found that many removals have notices that do not reveal the rationales; it also found explainer videos to make deepfakes, an otherwise regulated content type. The X/Twitter work calls into question Community Notes, not as a philosophy of content moderation given its efforts at bridge-building but as a practice. In the empirical study, most Community Notes remain unresolved, slowing the pace of moderation and leaving content online. In studying which content is left online, it found no appreciable difference (in toxicity) between the moderated and unmoderated. Community notes-style moderation may not be resolving moderation issues and as such it remains a work in progress.

There are trade-offs or dilemmas when discussing the adjudication of visibility reduction versus the bridge-building of Community Notes. While more transparent (when the notes and their statuses are open source) than the automated technique of downranking content, the Community Notes approach, the researchers found, leaves much more on the platform in plain view. The ability to be able to compare their relative performance remains a remit for the field as well as for the policy arena that enables the availability of such insight and ongoing development.

Finally, this volume offers a detailed description of how moderation ‘works’, and how it can be interpreted and monitored for scholarly and other purposes. But if it still speaks to the policy making arena, the deceptively simple question that remains is, what do we suggest that the platforms do? And how should the European policy making arena engage further?

For the platforms the book points to quite straightforward pathways to action, particularly the implementation of DSA-compliant APIs for the study of content moderation systems and moderated content. The impression that some of our studies leave, however, is that one needs to tighten the grip on content moderation. That is, there are still several moderation gaps in spite of the DSA and other legal arrangements.

But this tightening may overshadow the more politically urgent question of sustainability (de Keulenaar, this volume). How sustainable is such enforcement,

especially given the transformation of content moderation into a public bone of contention? If we agree that there are structural failures in enforcement (particularly when it is equated with censorship), then what can be said about the underlying philosophy of bridge-building approaches? They may have technical shortcomings, for algorithmic consensus and deliberation are quite distinctive from discussion, but are such approaches something that the policy making arena should grapple with?

Part of the answer to the question of where the policy making arena can engage further, as mentioned, is by enabling the availability of insight into these moderation systems beyond the DSA Transparency Database. An addition could be that this arena can actively participate in the development of such techniques as public platform design conventions.

References

Bastos, M. (2024) *Brexit, Tweeted: Polarization and Social Media Manipulation*. Bristol University Press.

BBC (2022) A third of children have adult social media accounts - Ofcom, *BBC News*, 11 October, <https://www.bbc.com/news/technology-63204605>.

Bounegru, L., Devries, M., & Weltevrede, E. (2022) The research persona method: Figuring and reconfiguring personalised information flows. In *Figure: Concept and Method* (pp. 77-104). Singapore: Springer Nature Singapore.

Bucher, T. (2016) The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, 20(1), 30–44. <https://doi.org/10.1080/1369118X.2016.1154086>

Chen, A. (2014) The laborers who keep dick pics and beheadings out of your Facebook feed, *Wired*, 23 October, <https://www.wired.com/2014/10/content-moderation/>.

de Keulenaar, E., Kisjes, I., Smith, Rory, A., Albrecht, C. and Cappuccio, E. (2023) Twitter as accidental authority. How a platform assumed an adjudicative role during the COVID-19 pandemic, in Rogers, R. (ed). *The Propagation of Misinformation in Social Media*. Amsterdam University Press, pp. 109-138, <https://doi.org/10.1515/9789048554249-007>

de Keulenaar, E. and Rogers, R. (2025) After deplatforming: The return of trace research for the study of platform effects. In: Venturini T, Acker A, Plantin J-C, et al. (eds) *The SAGE Handbook of Data and Society: An Interdisciplinary Reader in Critical Data Studies*. London: SAGE.

Drootin, A. (2021) "Community Guidelines": The legal implications of workplace conditions for internet content moderators, *Fordham Law Review*, 90(1197). Available at: <https://ir.lawnet.fordham.edu/flr/vol90/iss3/4>

Gerrard, Y. (2018) Beyond the hashtag: Circumventing content moderation on social media, *New Media & Society*, 20(12), 4492-4511.

<https://doi.org/10.1177/1461444818776611>

Grassegger, H. and Krause, T. (2016) "Inside Facebook: Im Netz des Bösen." *Süddeutsche Zeitung*, 15 December, <https://www.sueddeutsche.de/wirtschaft/inside-facebook-im-netz-des-boesen-1.3295206>.

Herrman, J. (2016) Inside Facebook's (Totally Insane, Unintentionally Gigantic, Hyperpartisan) Political-Media Machine, *New York Times*, 24 August, <https://www.nytimes.com/2016/08/28/magazine/inside-facebooks-totally-insane-unintentionally-gigantic-hyperpartisan-political-media-machine.html>.

Hopkins, N. (2017) Revealed: Facebook's internal rulebook on sex, terrorism and violence, *The Guardian*, 21 May, <https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence>.

Katzenbach, C., Kopps, A., Magalhães, J. C., Redeker, D., Sühr, T., & Wunderlich, L. (2023). The Platform Governance Archive v1: A longitudinal dataset to study the governance of communication and interactions by platforms and the historical evolution of platform policies (Data Paper).

Leerssen, P. (2023) An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation, *Computer Law & Security Review*, 48, 105790, <https://doi.org/10.1016/j.clsr.2023.105790>.

Meta (2025) Testing begins for Community Notes on Facebook, Instagram and Threads, Newsroom, <https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/>.

Meyer, R. (2018) YouTube removes the 'Hail, Trump' video From search, *The Atlantic*, 20 March, <https://www.theatlantic.com/technology/archive/2018/03/youtube-removes-the-atlantics-hail-trump-video-from-search/555941/>.

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015) Down the (white) rabbit hole: The extreme right and online recommender systems, *Social Science Computer Review*, 33(4), 459-478.

Roberts, S. T. (2016) Commercial content moderation: Digital laborers' dirty work, in Noble, S.U. and Tynes, B.M. (eds.) *The Intersectional Internet: Race, Sex, Class and Culture Online*. Peter Lang Publishing. pp. 147–159.

Roberts, S. T. (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Silverman, C. (2016) This analysis shows how viral fake election news stories outperformed real news on Facebook, *Buzzfeed News*, 16 November, <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook>.

van Wonderen, R., Burggraaff, D., Ganpat, S., van Beek, G. and Cauberghs, O. (2023)
Rechtsextremisme op sociale mediaplatforms? Ontwikkelingspaden en
handelingsperspectieven, Instituut Verwey-Jonker, Utrecht.

2. From Twitter to X: Demotion, Community Notes and the apparent shift from adjudication to consensus-building

Emillie de Keulenaar

Abstract

Although content moderation has not disappeared under Musk's tenure, it remains, in X, unclear and inconsistent. This project analyses changes in content moderation from Twitter to X by focusing on three key areas: 1) policy changes before and after the acquisition; 2) the enforcement of these policies, particularly demotion, removal and Community Notes; and 3) the role of new moderation methods such as Community Notes in shaping discourse around polarizing issues in European public debate. In describing these results, this study aims to shed more light on the potentials and limitations of changes from top-down approaches to content moderation towards more agnostic but consensus-driven methods.

Keywords: content moderation, Community Notes, X, Twitter, bridging algorithms

Introduction

When X was Twitter, the company's content moderation evolved into an ever-complex trust and safety apparatus benefiting from strong legal and technical investments. At times, thousands of tweets and users were deplatformed indefinitely for violating hate speech, electoral or COVID medical policies. But incoherent decision-making around what kind of speech was or was not acceptable (and to whom), decided in a somewhat vertical fashion, has gathered criticism from all sides of global political debate, making the company and other "mainstream" platforms liable to various crises of legitimacy.

In response, X, under Musk, claimed once again to become a platform for "free speech". For starters, some of Twitter's policies have been removed, and those that remain have been changed to varying degrees. Some have reported an increase in hate speech (Martinez, 2023), eliciting the impression that the platform no longer moderates much at all. A quick X search of a discriminatory term — "stupid jew" — does show, today, that most if not all posts mentioning this term are perfectly online. Yet, there have been high-profile scandals of the platform removing posts linked to antisemitic speech — as was the case with Kanye West — as well as with anti-migrant discourse relating to whether Trump's administrations should expand an existing H-1B visa programme (Singh, 2024).

Clearly, X *does* moderate, but it is not clear how or to what extent. One reason for this impression is that there have been few and inconsistent demonstrations. If posts are not flagged or removed, they may be demoted or receive a community note. The platform may also have redefined what it considers problematic: there have been reports of posts flagged for containing the word "cisgender" (Silberling, 2024). Such changes

may point less to a lack of moderation than a larger change in X's moderation philosophy — a phenomenon reflected in the rest of Silicon Valley (Klaidman, 2024). Before Musk, Twitter had already opted to tone down active moderation and rely instead on demotion and other less interventionist approaches (Musk, 2022). More than a year prior to his acquisition, Twitter had invested in a new community fact-checking programme called "Birdwatch", a system where users of different viewpoints assign labels to misleading posts in an attempt to crowdsource "consensus" about the factuality or acceptability of online content. Of the many moderation measures X removed, this one remained. It is likely that consensus-building systems (Ovadya & Thorburn, 2023; Perez, 2022) may become increasingly popular alternatives to top-down platform moderation (Tang, 2024), which Facebook recently claimed to be an "impossible" task (Meta, 2025). Indeed, the platforms that were once accused by now-President Trump of censorship began, one by one, declaring dramatic U-turns to their content moderation philosophies. Meta's relatively small fact-checking programme, which invested in "authoritative sources" to check the veracity of a handful of posts, was dropped in favour of Community Notes, sharing the same open-source algorithm as that of X.

But the presence of "consensus-building" techniques may not necessarily mean that X has improved as a space for European public debate, nor that it takes DSA directives seriously. When Thierry Breton called on Musk to enforce its policies in the EU, he simply recommended Breton to "f*** his own face" (Kroet & Armangau, 2024). Likewise, Zuckerberg abruptly ended what had been Meta's rather diplomatic disposition towards EU legislation, instead opting for an unusually nationalistic stance in favour of the advancement of "global free speech" against "European censorship" (Malingre, 2025).

This project looks at how general content moderation changed under Musk, describing, in more detail, some of the ways in which Twitter mutated into X. It focuses on three axes. First, it looks at how content moderation policies changed under X, and what the latest reports indicate about what is to come. This is done by using the Wayback Machine and Platform Governance Archive (Katzenbach, Dergacheva et al., 2023) to locate and analyze changes in all of the platform's recorded policies to date. This analysis allows us to pinpoint what is still considered problematic in the platform as well as how content moderation techniques may have changed. Most importantly, it helps us design queries that allow us to locate and verify the moderation status of corresponding content.

Second, I look at the extent to which X enforces its own policies, particularly those relevant to section 34 of the DSA. I begin by comparing the availability of posts that have previously been moderated for containing hate speech in the EU throughout 2020. Moderation traces, such as removals and demotion, are collected every day for one month using a process known as "dynamic archiving" (de Keulenaar & Rogers, 2025). This allows us to pinpoint moments when moderation occurs as well as what content tends to be moderated and how.

With regard to civic discourse, I explore new consensus-building moderation techniques more in depth. This includes Community Notes. How consistently are such notes applied and to what extent do they amount to an "improved" form of

moderation? To assess this, I apply the same exercise as above but with a set of keywords that touch upon contentious debates in the Netherlands, namely immigration.

In all, this project seeks to shed light on a broader transformation reportedly occurring in the moderation strategies of traditional social media platforms. Though content moderation has visibly shifted towards a norm-agnostic approach — deferring some normative decisions to users or relying on demotion techniques — it has simultaneously invested in systems that seek to "crowdsource consensus" (via Community Notes) and "bridge divides" through various "bridging systems" in ranking, recommendation algorithms and LLM applications (Ovadya & Thorburn, 2023). While this development appears to pave the road for content moderation to contribute to "healthier" public debate, questions remain about (1) the role of public institutions in modelling "bridging systems" for private platforms with distinct financial and political interests; and (2) the underlying concepts that are baked into operationalization of "consensus-building" in algorithmic systems.

From tensions to ruptures

Until recently, one could think of social media platforms within short groupings of a handful of US-based, hegemonic companies — "GAFAM", the Big Four, Big Five or Big Tech. One type of hegemony has been over the Internet's market, where, under a model of "platformization" (Helmond, 2015), they would funnel most of the Web's traffic under a "platform-ready" framework. Another has been in the realm of content moderation, in the sense that platforms have become one of the most consequential public spheres and adjudicators of public speech (Gillespie, 2018a), at least in the Western hemisphere. Facebook, Instagram, Twitter and YouTube would generally converge in their handling and definition of what counts as legitimate, truthful or acceptable content, aiming to reintroduce a sense of normality to spaces where what was once historically unacceptable re-emerged and destabilised norms against discrimination and journalistic guidelines.

In making those decisions, platforms have eventually broken the illusion of being the open and thereby "neutral" public spaces they initially modelled themselves after (Gillespie, 2018a). This is not to say that they have ever been intermediaries (Gillespie, 2018b), but that more proactive content moderation inevitably created a problem of *normative demarcation* (Van Raemdonck & Pierson, 2022), that is, to decide where to delimit a threshold between acceptable and unacceptable, false and truthful and other "policy lines" (Constine, 2018). Typically, social media platforms will identify a range of "harms" with varying levels of objectionability, and allocate a combination of content moderation techniques to repel each (Figure 2.1). The identification and estimation of each "harm" may enjoy varying degrees of public consensus and approval in time and place, and indeed different platforms will tend to identify different kinds of harms and allocate different enforcement mechanisms to each. For the past two decades, most have tended to refer to historical norms, particularly the "grammar" (Pedretti, 2023) of human rights (Hatano, 2023; Sander, 2019) and journalistic objectivity, to establish the appropriateness and factuality of user-generated content.

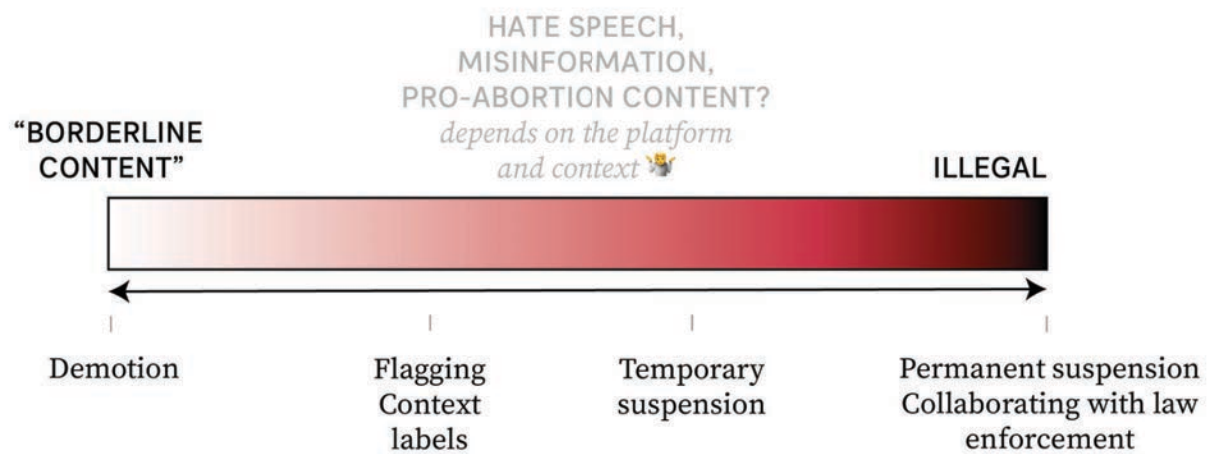


Figure 2.1 An example of platform normative demarcation, where "problematic content" is placed on a spectrum of more to less severe content moderation measures. Less severe and poorly defined "borderline" content may be demoted, while illegal content (paedophilia, terrorism, etc) will be irredeemably removed. Source: author.

But these norms are "essentially contested" (de Laat, 2012) and have been vulnerable to profoundly contentious debates about what constitutes problematic speech — not least because of the dramatic expansion of political voices once marginalized in legacy media environments (Benkler et al., 2018). It is also evident that definitions of what amounts to discrimination and misinformation have evolved dramatically throughout highly dynamic and often polarized online debates as well as social movements spearheaded online. Again through a problem of demarcation, platforms have mostly opted to follow the winds of social change and adapt their policies to the norms of their time (Suzor, 2019). As is more apparent today, these decisions have put platforms at the centre of global "culture wars", waged in part around competing definitions of hate speech, authenticity and the necessity to moderate public debate to protect individuals from historical "harms" (Sander, 2019).

This has caused what some have diagnosed as a "crisis of legitimacy" in mainstream social media platforms (Napoli, 2019). On the one hand, private companies are "accidentally" (de Keulenaar, Kisjes et al., 2023) bestowed the power to determine what does and does not circulate online based on norms negotiated across a more or less obscure network of legal, political and civil society actors (Gorwa, 2019). On the other, these norms brush against once marginalised actors and their normative frameworks — which have partly been normalized by those very same platforms. A vertical form of content moderation, whereby platforms decide through authoritative means what accounts as objective or acceptable content, is, of course, prone to any and all criticism inherent to governance.

So, after moving fast and breaking things, some platforms now stand in the awkward position of having to abide by legal (and historical) moderation norms — particularly DSA laws in Europe — while affording a space for the kind of political content they have helped normalize. One often thinks of Donald Trump and his ongoing breach of American political (and social) speech norms, but the same can be said about a myriad of representatives and influencers across South America, Europe, and SouthEast Asia who represent ideas and discourses once relegated to the dustbins of history (Föessel, 2021). These actors bring with them a radically different normative paradigm that

clashes with "mainstream" definitions of hate and misinformation, and have tended to object to speech moderation altogether under a newfound conception of free speech (Zuckerman, 2020): that information "wants to be free".

Freedom of speech as a feature

X, however, has embraced the radical political changes they have facilitated, transforming their underlying content moderation philosophy into both a feature and value to compete with other platforms. In this context, it could be argued that platforms have begun to compete not just in terms of users and features, but in terms of different content moderation philosophies, too. Users will pick one or another platform based on their normative affiliations, or be compelled to migrate wherever they are allowed. Some return or stick to X while others leave to Bluesky, just as some had once moved to Truth Social, Gab, Voat, Telegram and a few other islands of content moderation exiles. In a sense, another factor for competition between these platforms is the kind of culture they nurture within their moderation regime.

Accordingly, the frontiers between these platforms — and above all the motives of their severance — is a device for platforms to orient and define themselves. Arguably, the difference X made was that it emerged as an "alternative" platform from *within* otherwise mainstream social media platforms. From the seat of Twitter's ex-CEO, Jack Dorsey, Elon Musk pushed to dramatically reverse years of content moderation policies perceived to be exceedingly partisan — a product of a "left-wing woke mind virus" (see Musk's interview with Joe Rogan, 2024) — to return to a model of an open "town square" (Elon Musk, 2022) where all have a right to speak. From this position, Musk's move appeared at first delusional: many presumed a return to free speech absolutism similar to what the platform was before 2016. But the CEO was quick to specify that, from a practical standpoint, X would refrain from being a free speech "hellhole" (Vincent, 2022). As any platform, it would retain a minimum of service quality and coherence (Gillespie, 2018a).

Though there have been many content moderation changes, in general this shift constituted a new philosophy of moderation. In its more minimalist expression, X's current CEO Linda Yaccarino would put it like this: what is "lawful but awful" is demoted, and what is "unlawful" is deleted (see interview by CNBC Television, 2023). This particular idea had already germinated in Twitter, YouTube and Meta trust and safety circles from around 2018 as a form of "borderline content" management. That is, content would be downranked and obfuscated if it could not breach a clear threshold of objectionability (Gillespie, 2022). This technique has been characterised as "norm-agnostic", in the sense that it decides where and how to position that content on the platform based on its level of public acceptability at a particular time and place (de Keulenaar, Magalhães et al., 2023). Other than demotion, it is largely left up to users to estimate the vices and virtues of a post; it is their choice to consult it ("If you don't like it, don't click on it" (see interview in Don Lemon, 2024); and if it must absolutely be moderated, it will be hidden instead of suspended. In this context, local legislation such as the DSA is presented as an external moderation accessory that users are invited to report at their own volition.

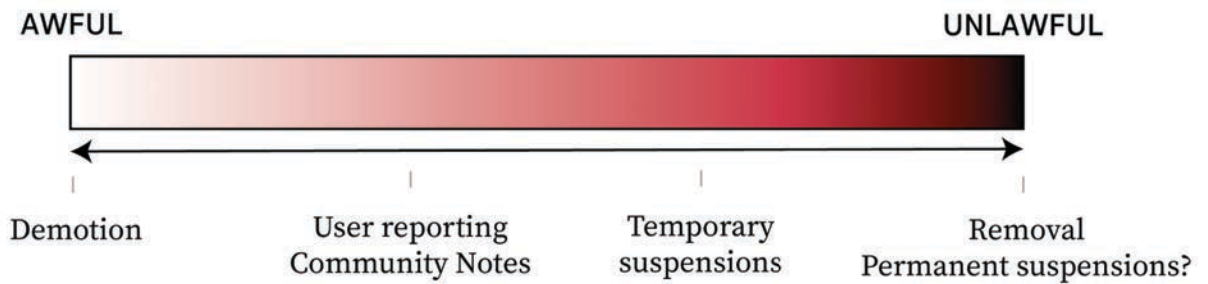


Figure 2.2 A basic representation of X's current content moderation spectrum. Source: author.

X as a competing "democracy"

On top of this basic change of policy, Elon Musk has made claims to use the platform as a means to protect and advance, in no small terms, "democracy" and "human civilization" (see interview in Joe Rogan, 2024). In a couple of interviews from early after the acquisition, he explained that one of the reasons why he decided to buy the platform was to remove it from the helms of a "mind virus", an indisposition from "far-left" political culture to accept self-criticism, freedom of expression and allow "civilisation" to flourish through "pro-human" political philosophies (ibid). A lack of free speech, in particular, would condemn civilisation to a self-destructive path because, without it, "common sense" — a capacity often reiterated by Trump on the US campaign trail — cannot emerge or be expressed to the fullest extent, thus disconnecting policy making, representative and legislative governance from the collective will (ibid).

In these terms, Musk frames the goal of liberalizing speech as a means to "save" democracies and civilization ("freeing the [Twitter] bird") so that common sense, or "crowdsourced wisdom", may emerge from total sum of "collective consciousness" expressed on the platform (Elon Musk [@elonmusk], 2022). There is an implicit conception of governmental institutions as hollow and outdated, competing, as it were, with social media platforms as "de facto town squares" where citizens debate and influence policy without any need for political delegation. X blog posts borrow a more proactive political terminology: without free speech, one loses "the checks and balances critical to a thriving democracy" (X Safety, 2023b), particularly against "agenda-driven activists" (ibid).

While Silicon Valley companies have a history of contempt against the olden and inefficient ways of governmental institutions (Turner, 2006), X has become far more confrontational in the name of the "civilizational mission" stated above. The confrontation, now, is with democratically elected states — Brazil, the UK and EU states included — who for historical or questions of sovereignty, do enforce legislation and conventions against the return of certain historical "harms", be that discriminatory, misinformative or incitement speech (Fuchs, 2015).

By virtue of its scale, diversity and accessibility, social media platforms have inevitably — though perhaps accidentally — destabilised such norms (Napoli, 2019). Musk and Zuckerberg's recent interventions, however, point further to a *resistance* to

such norms. First, in the US, there was the return and rehabilitation of thousands of accounts deplatformed for partaking in the January 6th riots of 2021 — other than hate speech or incitement to violence. There is of course Musk's support for Donald Trump. There was the use of X for routine Trump campaigns, and there was above all the promotion of X as a symbol and parcel of Trump's agenda for the return of "free speech in America" (Graham, 2024). Then, in Brazil, a spat emerged between Musk and Alexandre de Moraes on the grounds that the judge overreached his powers through local content moderation enforcement. X was presented as a means to liberate Brazilians from the arbitrary hands of the judiciary and express themselves in their full political diversity, no matter how controversial the opinion, without much evidence of Musk understanding or acquiescing to the historical premises behind the judge's decisions (Mendonça, 2024).

In the EU X is again presented as part of a civilizational mission to restore European culture by liberating citizens from self and state censorship. It is assumed that multiculturalism vested by human rights concerns is a road to hell paved with good intentions, because what resides at the end of the line is "cultural" and "demographic replacement" (Bennhold & Taub, 2025). By this prophecy, X would have to exist for an instinct of European self-survival to be expressed, no matter how obscene. It is this narrative that was used to frame Keir Starmer as an enabler of "Muslim" violence against women as well as express support for Alice Weidel, the leader of the German right-wing party AfD, while encouraging German voters to let go of their "guilt" (ibid).

Towards moderation as consensus-building?

The conception or "imaginary" of X as a vector for "real" democracy is also reflected in its features. There is Musk's usage of X polls, first for affairs internal to X (reinstating banned accounts, deciding whether he should remain the CEO), and then for decisions tied to Trump's second White House administration (reinstating a Department of Governmental Efficiency employee who had resigned for posting racist content on X) (Honderich, 2025). There is also the idea that the net total of people (i.e., users) "on the ground" surpasses editorialised news media, because "people who are actually in a particular industry or a particular region actually know what's going on better than reporters do" (Musk in Heath, 2023). In an interview with Don Lemon in 2023, Musk retorted to some questions with the idea that "the comments", not Lemon, would do a better job fact-checking his statements (2024).

Then, there are Community Notes — one of Musk's "favorite features" (Musk in Heath, 2023). Introduced as "Birdwatch" in early 2021 (Wojcik et al., 2022), it was an attempt by Jack Dorsey's late Twitter team to improve content moderation methods beyond a purely adjudicative logic. Similar to Wikipedia's system of article editing, it is a system where volunteer users write and rate notes left on highly visible posts. The note that gets selected should "appeal broadly across heterogeneous user groups", i.e., a majority of raters with different ideological backgrounds or a "diversity of perspectives" (X Community Notes, 2025) agree that note is "helpful" (informative, balanced, trustworthy, etc). "Ideological background" is measured by rating behaviour — that is, users are clustered based on their rating history (see Figure 2.3).

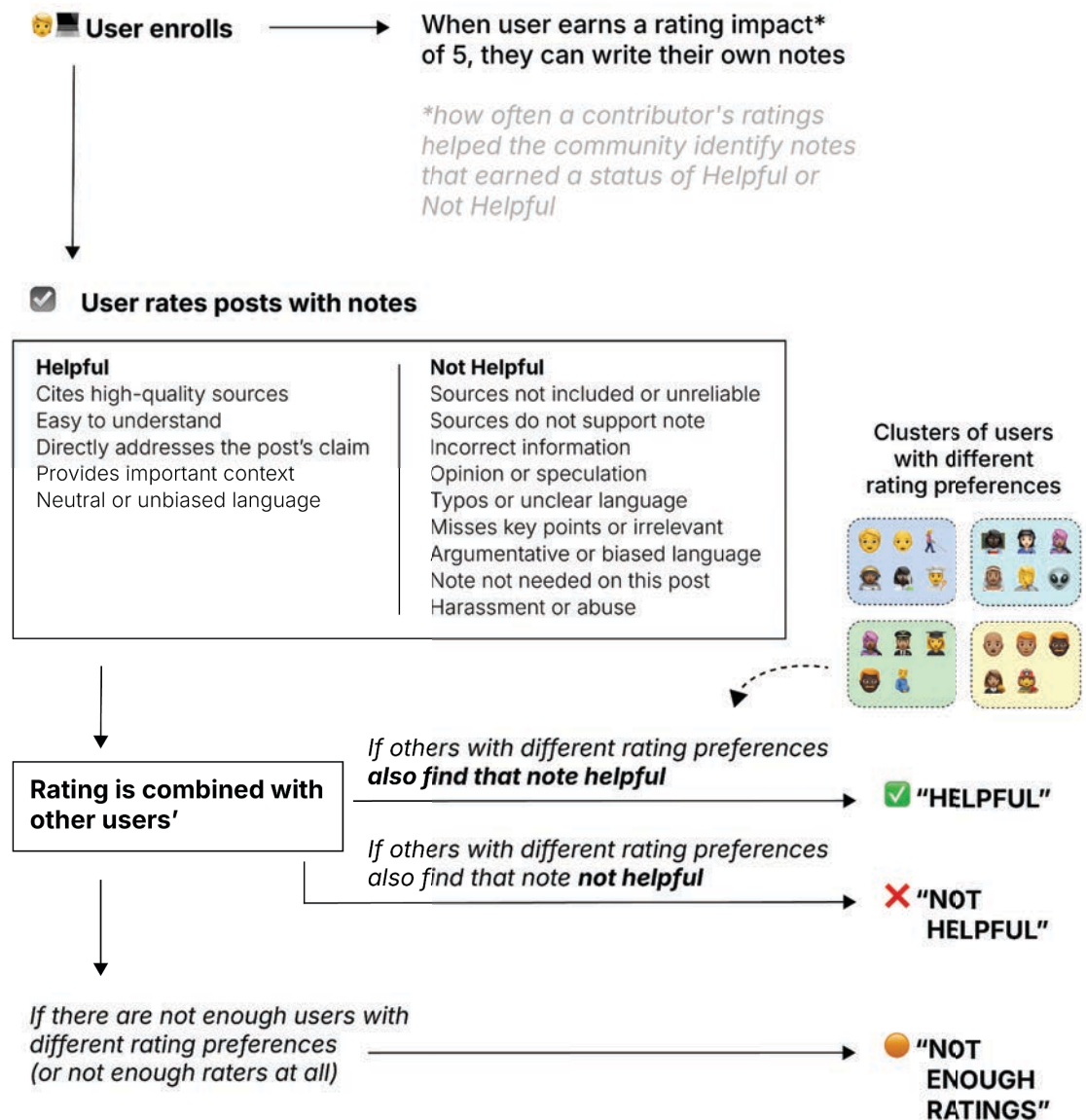


Figure 2.3 A representation of Community Notes sorting mechanisms. Source: author.

Community Notes are distinct from other content moderation features because it veers beyond the terminology of consensus-building or "civic dialogue", borrowed in part from cited works on values in design (Ovadya & Thorburn, 2023; Stray, 2022). The clustering algorithm behind Community Notes, for example, is presented as a "bridging system" (Wojcik et al., 2022), i.e., an algorithm designed to calculate and up-scale (up-rank, make more visible, etc.) consensus across users of different "divides" (be these political, cultural, linguistic, or else). Community Notes' clustering algorithm was initially developed by Polis, a public survey platform where polsters can invite users to find where and how they agree on issues they propose to discuss – also cited as "bridge-based".

This type of "design" initiative originates from an up and coming network of developers and academics invested in the premise that content moderation *does not work* and must instead be amended with *improved platform design* (Dörr et al., 2025; Schirch, 2023). For one, moderation is prone to incessant disagreements across users,

platforms and state actors as to what amounts to problematic information. Second, it only reduces harmful information from "below", without tackling its production at its source — i.e., at "poor platform design", such as the incentive to promote polarising content that would, in turn, foment the production of hate speech and misinformation. The idea, here, is to operationalise "prosocial processes" (Weyl et al., 2025), such as traditions or conventions of civic dialogue (e.g., a citizen assembly discussion, a truth and reconciliation commission) into algorithmic techniques (Ovadya & Thorburn, 2023).

Arguably, Community Notes and other forms of "bridging" may amount to a contribution to content moderation in two senses. First, it may complement moderation "upstream", in the sense that promoting some form of consensus across users may reduce the production and spread of "harmful content". Second, if it is applied to the *design* of content moderation itself — for example, a public forum where users deliberate about what norms a platform ought to adopt — they constitute a space where users may reach some temporary consensus on moderation norms, such as the factuality of a piece of content, or the degree to which it is socially safe. From a historical perspective, this space is and has been central to the formation of speech norms — to mention decades of public debate about how to prevent discriminatory language from repeating historical violence post-war, or post several civil rights movements (De Bolla, 2013). Given the diversity and speed at which these norms change, a space for building public (and ongoing) consensus around content moderation norms may thus cushion legislative or top-down moderation enforcement, which is usually vulnerable to significant public resistance, and has indeed fuelled strong opposition to European laws in both EU and American political discourse (Jackson & Szóka, 2025).

The elephant in the room

There are of course many questions to be posed to such mechanisms. One is whether they can serve as effective counterweights to environments that still heavily rely on attention as a financial model, where attention-grabbing content — such as inflammatory material — forms the substrate of what is deemed "harmful" content. The irony is not lost on this when Community Notes sit alongside unflagged or un-noted "toxic" content (Center for Countering Digital Hate, 2024) that have pushed news media and other counterbalances to leave (Reuters, 2024). The second and arguably most critical is how concepts of consensus-building and other forms of civic dialogue are understood and operationalised. Political scientists will argue that consensus-building is a complex, contradictory and at times infinite process (Hoffmann, 2021), and the "bridging algorithm" beneath Community Notes is but a simplistic reduction of thousands of other possible models.

Whether Community Notes contribute positively to X's broader moderation efforts and how they reshape ideas of consensus remain central to ongoing discussions on integrating critical European values into platform mechanisms and design. Below, I outline the method used to investigate and address both of these issues.

Method

In what follows, I will explain the methodological decisions taken to examine X's new content moderation policies and practices (demotion, deletion or suspension, and Community Notes). The overall method can be seen in Figure 2.4. In short, the

methodology comprises: (1) data collection of Twitter and X policies (from 2006 to early 2025), Community Notes in Dutch, and — to verify demotion and deletion — posts listed under search results rankings for every query relating to immigration in Dutch; (2) examining policies overall and individually; (3) looking at what kinds of topics and users tend to get the most Community Notes; (4) and looking at whether posts with high "hate scores", as determined by third-party moderation systems (in this case, OpenAI's moderation API), are effectively demoted or removed, when necessary.

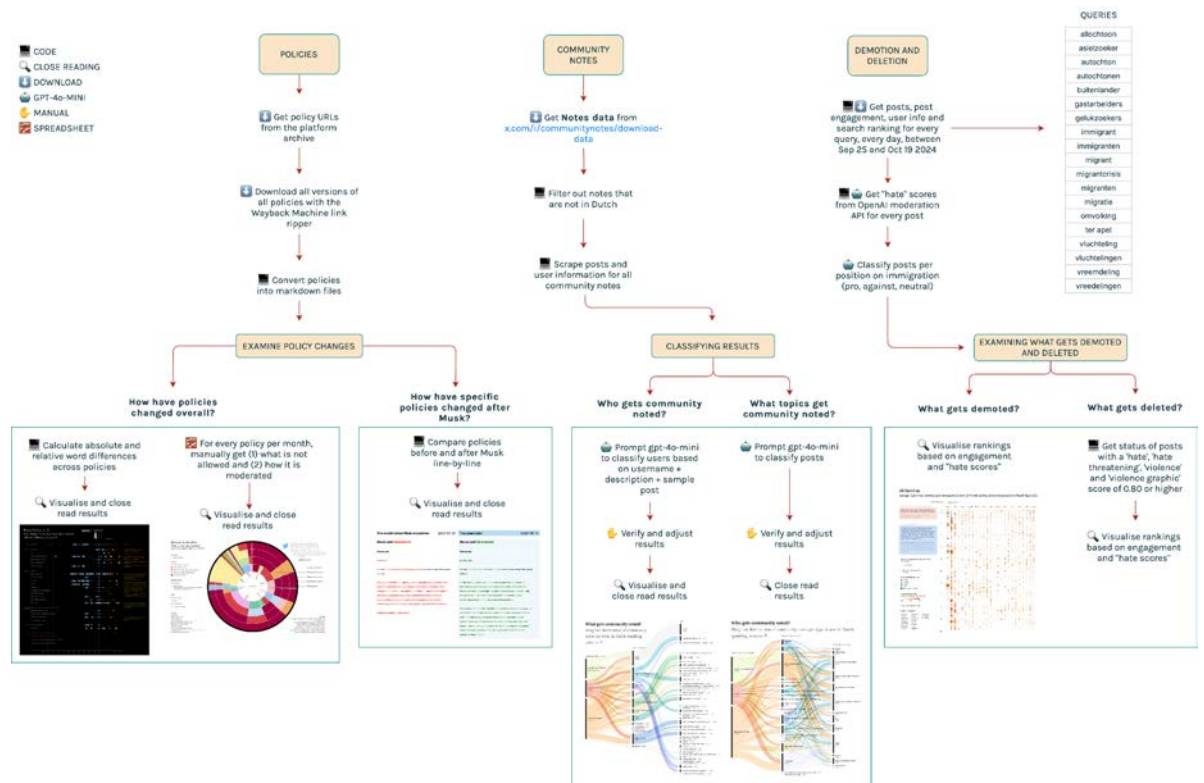


Figure 2.4 Methodology diagram. Full image [here](#). Source: author.

Policies

Data collection

Twitter and X policies were collected using the Internet Wayback Machine (Internet Archive, 2025). First, I collected all URLs that pointed to present and past Twitter and X policies from the Platform Governance Archive (Katzenbach, Magalhães, et al., 2023). With those, I used the Wayback Machine Link Ripper (Digital Methods Initiative, 2022) to collect the Wayback Machine links for every policy archived every month between 2006 and 2025. Having obtained all archived URLs, I then download every single version as HTML and markdown text files programmatically.

Analysis

Markdown files were then used to perform three analyses. First, I calculated the absolute and relative amounts of word differences for every policy to get a sense of the temporal context for each policy change, also in comparison to Twitter's content moderation. I use a Python library, `difflib.SequenceMatcher`, to compare word sequences between two or more versions of a policy (see Annex). Absolute word differences are the sum of added and removed words. Relative word differences are

absolute word differences divided by total words in previous policy versions. The overall results can be seen in Figure 2.8.

Second, I wanted to gain a sense of different content moderation "regimes". This means a period in which Twitter or X were carrying out content moderation (on paper) more or less consistently, and tended to be bracketed between major policy changes. These are approximately 2010 until around 2013-2014; 2013 until 2017; 2017 until 2022; and 2022 until present. Here, the analysis consisted in seeing (a) *what* was moderated (in the form of policy titles) and (b) *how* that was enforced. The results, in the form of spherical diagrams representing Twitter or X as spheres with shifting boundaries, can be seen in Figures 2.5, 2.6 and 2.7.

Finally, I analysed policy changes shortly before and after Musk's acquisition of Twitter. This meant comparing changes of every policy for every year of significant text changes between 2022 (before Musk) and 2025 (after Musk), namely: October 2022; June or August 2023; and beginning of 2025, or late 2024. The analysis was done by comparing changes line-by-line.

Community Notes

The main research question about Community Notes is: *what* and *who* tends to get community noted in the Dutch X space, and to what extent does it achieve a minimum of consensus — in the form of collectively rating "helpful" notes — in posts about deeply controversial topics in the Netherlands (e.g. immigration, public health, the war in Gaza)?

Data collection

To tackle each question, I first collected all the Community Notes that had some indication of being about Dutch topics. Since notes do not have any geolocalisation, one can only assume their geographic focus by their language. After downloading all Community Notes from X directly (X, 2025), I used Google Sheet's DETECTLANGUAGE function (*DetectLanguage*, n.d.) to determine their language and then excluded all those not in Dutch. This yielded an initial set of 9,756 notes. To determine what and who every community note was addressing, it was crucial to obtain X posts and user information. For both objects, a script was used to scrape all 9000 posts based on their IDs, including engagement information to the extent possible (views, for example, are not always displayed by X). User information was obtained using user IDs — also provided by Community Notes — and included user description (used to determine user professions, or other) and follower numbers (used to determine influence). No personal information is ever displayed in figures or anywhere else on this paper.

Classifying notes and users

To determine what and who Community Notes were addressing, it was necessary to classify posts into topics and users into types. This was done by first prompting gpt-4o-mini — a most cost-effective version of OpenAI's most well-performing model, gpt-4o (OpenAI, 2025) — to classify both objects.

To classify posts into topics, it was necessary to first draw from a representative sample of posts manually. This was done on the first 100 posts with the most overall engagement (an addition of views, reposts, replies and likes). Having identified a few

categories, one then instructs gpt-4o-mini, via its API, to classify posts accordingly, allowing it to identify additional categories. Gpt-4o-mini was used in February of 2025 and chosen over gpt-4o due to its reduced price and near equal performance (OpenAI, 2025). This is a form of "snowballing" data, i.e., expanding an initial list of items with each search iteration. After five iterations, the following categories were identified and placed in a final prompt:

You are an expert in Dutch public affairs. Categorize Dutch tweets based on the main topic.

Instructions:

- The first mention after "@" is the tweet author (e.g., in "@realdonaldtrump something", the author is "realdonaldtrump").
- Only return the number of a topic.
- If none fit, return: "0 - [three-word summary]" (e.g., "0 - crocodile shop").
- No extra text.

Topics:

- (1) Ukraine war, Russia, Putin
- (2) Gaza-Israel war
- (3) Tensions from the Gaza war (antisemitism, pro-Palestine protests, protest vandalism and police brutality)
- (4) The war in Afghanistan or Iraq
- (5) Second World War
- (6) Colonial history
- (7) Other historical events
- (8) (Im)migration
- (9) Tensions with migrants (incl. Muslims)
- (10) Demographic replacement theories ("omvolking", fall of Western culture/civilization)
- (11) Medicine, healthcare, health policies, wellbeing, COVID, alternative treatments, healthy eating tips, safe smoking
- (12) Wildlife, environment, natural resources incl. carbon
- (13) Climate change, crisis, global warming, carbon emissions, extreme weather, related debates (e.g., Extinction Rebellion)
- (14) Dutch or EU energy
- (15) European or Dutch farmers, nitrogen crisis
- (16) Critique of left-wing identity politics, "wokeness", gender
- (17) Debates about social norms (racism, xenophobia, misogyny, harassment, sexism, homophobia, language in the workplace)
- (18) Social, economic and political justice/critique (emancipation, anti-racism, wealth distribution/inequality, fair wages, etc.)
- (19) Black Pete (Zwarte Piet)
- (20) Censorship, freedom of speech, speech norms
- (21) Indoctrinating children with woke ideology
- (22) Left-wing bias in academia or schools
- (23) Criticism of news media
- (24) Support for alternative media or narratives
- (25) Manipulated media / fact checking etc
- (26) Advertisements or scams (e.g., stylish zip sweaters)

- (27) Theories or conspiracies regarding WEF, NWO, depopulation, Agenda 21, "you will own nothing", deep state, Klaus Schwab, chemtrails, elite pedophiles, etc.
- (29) Aliens
- (30) Housing, housing crisis, occupying or squatting houses
- (31) Amsterdam
- (32) Inflation, cost of living or products
- (33) Fireworks
- (34) Dutch railways, train travel time, infrastructure (cycling, cars, etc)
- (35) Dutch government, cabinet, parliament
- (36) Dutch elections, voting
- (37) left-wing politician/party
- (38) right-wing politician/party
- (39) Elon Musk, Tesla, SpaceX
- (40) anything relating to Twitter or X
- (41) Sports (players, teams, tournaments, football teams like Ajax, etc.)
- (42) US politics (Trump, US elections, Biden, Harris or other)
- (43) anything happening in other EU countries
- (44) child sexual abuse
- (45) international relations or affairs
- (46) military conscriptions and defence
- (47) taxes, budget cuts or fiscal policies
- (48) crimes, arrests, corruption
- (0) Other (return "0 - [three-word summary]")

Prompt 1 Prompt used to categorise posts containing Community Notes.

Once done, I manually verified each result, grouped similar categories and corrected fuzzy annotations, as this procedure is faster than manually coding posts from scratch. Posts that did not refer or revolve around Dutch issues (e.g., posts about Belgium) were removed. The final number of Community Notes was **6,430**, applied to a total of **4,718 posts**. Given their granularity, each topic was given a general category (see **Table 2.1**).

CATEGORY	TOPIC	POSTS
Budget total: 58	Inflation, cost of living or product costs	26
	Budget costs, government finance, fiscal policies	32
Conspiracies 136	Conspiracy theories regarding the World Economic Forum, the New World Order, depopulation, Agenda 21, "you will own nothing", deep state, Klaus Schwab, chemtrails, pedophilia amongst the elites, aliens, etc.	136
Dutch politics and	Dutch government, cabinet or parliament	224

governance 623	Dutch elections	96
	Dutch left-wing politician or party	59
	Dutch right-wing politician or party	220
	Defence and military conscription	24
Energy, climate and the environment 639	Wildlife, environment and natural resources	100
	Climate change	357
	Dutch or EU energy	73
	Dutch farmers/nitrogen crisis	109
Healthcare and wellbeing 409	Medicine, health policies, COVID, healthy eating topics, alternative treatments	409
History and historical conflicts 941	The war in Ukraine, Russia, Putin	285
	The war in Gaza	358
	Tensions emerging from the war in Gaza (e.g., Amsterdam riots, protests against genocide, etc.)	193
	The war in Afghanistan or Iraq	4
	WWII	61
	Colonial history	11
	Other historical events	28
Infrastructure 36	Infrastructure (general)	36
Media 433	Criticisms of new media reporting	97
	Support for alternative media	16
	Manipulated media (e.g. fake news, AI image, disinformation)	37
	Advertisements or scams	283

Immigration 367	Demographic replacement theories ("omvolking", fall of Western culture/civilization, etc.)	63
	Immigration (general)	126
	Religious, cultural or political tensions related to immigration	179
Crime and security 77	Allegations of child sexual abuse	15
	Crime (robberies, corruption, murders, drug trafficking, etc.)	62
Social and cultural issues 359	Criticism of identity policies or ‘wokeness’	73
	Debates about speech norms and discrimination (e.g., racism, xenophobia, misogyny, harassment, sexism, homophobia, language in the workplace, etc.)	116
	Social and economic justice (emancipation, anti-racism, wealth distribution/inequality, fair wages, etc.)	42
	Zwarte Piet	14
	Censorship and freedom of expression	57
	Children indoctrination with ‘wokeness’	45
	Left-wing bias in academia or schools	12
Urban affairs 42	Housing, housing crisis	28
	Amsterdam urban affairs	8
	Other (debates about fireworks, etc.)	6
Other 826	Elon Musk, Tesla, SpaceX	33
	Twitter or X	44
	Sports	57
	US politics	88
	Internal affairs of other EU countries	34
	Other (anything else)	570

Table 2.1 Community Note post categories and sub-categories.

From there, I proceeded to classify post authors (users). Here, too, I used gpt-4o-mini iteratively to assign a category to each user based on their username, description and an example post (namely, the one with most engagement). The prompt used also contained examples for every category. Gpt-4o-mini was run twice; the first time without user descriptions, and the second with. Results were verified manually.

You are an expert in Dutch public affairs, assisting in identifying Dutch Twitter users. Based on the provided username, user description, and tweet, select the most appropriate category from the list below.

****Important:****

- Only select ****"Other"**** if the user does ****NOT**** fit ****any**** of the categories listed below.
- Do ****NOT**** select "Other" if the correct category already exists in the list.

If none apply, choose ****"Other"**** and specify who the user is in exactly two words (e.g., Other: tech entrepreneur).

Categories:

- Academic (docent, researcher, analyst)
- Activist or advocacy
- Alternative influencer (conspiracy theorist, advocate of alternative or transgressive viewpoints)
- Artist or entertainer (musician, visual artist, actor, or anyone in the cultural sector)
- Author, writer
- Centre-left politician or party (PvdA, Groen Links, DENK)
- Centre-right politician or party (VVD, NSC, CDA)
- Left-wing politician or party (SP, Partij voor de Dieren)
- Centrist politician or party (D66, Christen Unie, Volt, 50PLUS)
- Far-right or right-wing politician or party (PVV, BBB, Forum voor Democratie, SGP, JA21)
- Company, business
- Fact-checker
- Dutch government or cabinet
- Online influencer
- Journalist, columnist, commentator
- Dutch ministry
- Monarch
- News media, broadcaster
- NGO
- Religious figure, organisation or account

****Instructions:****

- Return ****only**** the category name.
- If you select "Other," format your response as: ****Other: [two-word description]****
- Do ****NOT**** repeat instructions, add explanations, or include unnecessary text.

Prompt 2 Prompt used to categorise the authors of posts containing Community Notes.

Analysis

The analysis consisted in looking at *who* and *what topics* tend to be most community noted. The underlying hypothesis is that, while Community Notes boast a good formula for crowdsourcing "consensus" (Wojcik et al., 2022), there may be areas of public debate controversial enough that notes may not do so. These may be immigration, the nitrogen crisis, or repercussions from the war in Gaza. The question, then, is whether Community Notes sufficiently fulfills its role in the absence of any other comprehensive form of content moderation (except for demotion, and some applications of removal or suspension).

To proceed with this analysis, I looked at how Community Notes and Community Note ratings were distributed across post categories and topics, and user types. Since the dataset was skewed heavily towards far more Community Notes rated as "needs_more_ratings" than otherwise, it was necessary to calculate distributions in weighted percentages.

This is first done by calculating absolute counts for each (topic, category, currentStatus, user type) group. This counts the number of unique tweets that have received a community note.

$$\text{Absolute Count} = \text{Unique Count of tweetId}$$

Then, since NEEDS_MORE_RATINGS dominates the dataset, I reduce its impact using a proportional weighting formula:

$$\text{Weight for } x = \frac{\max(\text{statusCounts})}{(\text{statusCounts}[x])^{0.75}}$$

Where:

- $\text{statusCounts}[x]$ = Number of occurrences of *currentStatus* x .
- $\max(\text{statusCounts})$ = Maximum count of any *currentStatus* (usually NEEDS_MORE_RATINGS).
- *Exponent 0.75* = Adjusts the weight gradually.

Then, I calculate weighted count per group:

$$\text{Weighted Count} = \sum(\text{weight assigned to each tweet})$$

Finally, I calculate relative percentages per group to ensure balanced distributions:

$$\begin{aligned} &\text{Adjusted Relative Percentage} \\ &= \left(\frac{\text{Weighted Count}}{\sum \text{Weighted Counts across all groups}} \right) \times 100 \end{aligned}$$

Demotion and deletion

The point of analysing demotion, suspension, or deletion on X is to verify how and to what extent the platform enforces its content moderation policies beyond Community Notes. Demotion and deletion are, however, two distinct techniques that each require distinct approaches.

Demotion

Demotion is the act of downranking some policy infringing content with the intent to decrease its engagement (in the form of "impressions", or views). As described in the *Findings* section below, under Musk, demotion applies to "awful" content; that is, what is not strictly illegal in the US. These generally fall within the abusive behavior (harassment, targeting of users), violent content (incitement of violence, gory media), authenticity (platform manipulation, misleading content), hateful conduct (hateful language and insignia) and election integrity policies.

Thus, to analyse demotion, it is necessary to retain three key characteristics: content (i.e., the extent to which it fits the description of each relevant policy — be that hateful language or other); engagement (i.e., the number of views it has acquired as an indication of whether demotion was or was not successful); and ranking.

To understand the latter, one first needs to consider the places in which posts are demoted. Though they are not always explicit, policies will mention newsfeeds, search results and replies. Newsfeeds are particularly relevant because they are central spaces for content circulation on X, but they require that one logs into their personal X account while maintaining a "research persona". To obtain non-personalised (and, to an extent, more generalised) results, I chose to get rankings from search results. This meant using X's search feature and capturing every search result (including post text and engagement metadata) for a period of time, so that one can tell how the same content moved throughout rankings over time.

Query design

The content I decided to look at was discussions about immigration in Dutch. I chose this topic because it is particularly difficult for content moderation. In the Netherlands and Western Europe more generally, there has been historical concern for moderating it — in the sense of minimising hateful or discriminatory language when discussing migrants and intercultural conflict — while being, at the time, a topic that keeps attracting significant public debate and defies the very norms of speech moderation. To "moderate", then, is a question of balancing a public desire for addressing it while preventing the eruption of verbal (or adjacent) violence. This same question can be posed to online content moderation: what does X "demote" in discussions about immigration, and what difference does it make when public debate is already significantly contentious?

To attend to this question, the queries I used for data collection were designed to touch upon immigration in the broadest possible terms, without clear indication of bias toward either negative or positive sentiments: "allochtoon", "asielzoeker", "autochton", "autochtonen", "buitenlander", "gastarbeiders", "gelukzoekers", "immigrant", "immigranten", "migrant", "migrantcrisis", "migranten", "migratie", "omvolking", "Ter Apel", "vluchteling", "vluchtelingen", "vreemdeling" and "vreemdelingen". The only exceptions were the term "omvolking", "gelukzoekers", and to an extent "allochtoon" and "gastarbeiders".

Data collection

Using X's advanced search feature, I collected 23,071 posts from search results for each query every day for one month, from September 25 until October 10th, 2024,

using a local scraper. While some search results yielded around 100 posts, others went as far as 1000. This means that rankings are varied. A few components collected by the scraper were: post text; username (when or if necessary for analysis); engagement (likes, reposts, and replies and views); and the time posted. Views do not always appear in a post's page, meaning that not all posts had the same engagement values.

Analysis

The analysis consisted in looking at the average ranking of posts from September to October 2024 based on how closely it approximated sanctioned content from the *Hateful conduct* policy. That meant assigning each content with a so-called "hate score": a measure of likelihood that a post is "hateful" according to OpenAI's Moderation API v2 (OpenAI, 2024). This API — like X's own hate speech detector, *Sprinklr* (Sprinklr, 2024) — will estimate the likelihood that a post is "sexual", "hateful", "violent", "harassing", "self-harmful" (and more) based on large training datasets of example posts. This was done using a large dataset of labeled text; fine-tuned with a neural network trained to detect relevant content; enriched with human review and reinforcement learning; and calibrated to provide category-specific scores. The scores are from 0 to 1; the closer it is to 1, the more likely a post is hateful. "Hateful", here, is defined as "content that is threatening, insulting, derogatory and otherwise abusive content [...]" that "targets specific chosen groups or members of the group because of their group identities" (Markov et al., 2023, p. 2).

To take this analysis further, I sought to contextualize hate scores within the positions expressed on immigration in this segment of Dutch public debate. My aim was not to treat hate scores solely as indicators of toxicity but rather to view them as part of a broader spectrum of sentiments surrounding a controversial topic. To do so, I devised a simple coding schema for gpt-4o-mini to tag posts (via its API) as pro-immigration, anti-immigration, neutral, or other. I then verified each result manually. With these results, I calculated the average rank, hate score and views (impressions) of the first 100 search results. Each element was calculated as follows:

$$\begin{aligned} \text{Average Hate Score} &= \frac{\sum \text{Hate Scores}}{\text{Number of Posts}} \\ \text{Most Common Rank} &= \text{mode}(\text{position_category}) \\ \text{Average Engagement} &= \frac{\sum \text{Views}}{\text{Number of Posts}} \end{aligned}$$

My next objective was to confirm whether there was any correlation between hate scores, rankings and impressions. X asserts that they have successfully kept the number of hateful or "awful" posts down in rankings and low in impressions, or views. Though I work with a smaller dataset and another hate speech classifier, I tested the correlation between hate scores and impressions (views), and hate scores and search rankings. This was done using two correlation coefficients: Pearson's and Spearman's. The latter measures linear relationships (i.e., whether a hate score increases or decreases as rank or views change in a consistent direction). It is relatively easy to interpret (a value close to -1 suggests a strong relationship, while one nearer to 0 means no clear pattern) and standard for numerical data. Spearman's coefficient was used to test non-linear regressions, which is relevant for rank data. Both correlations were calculated on unique posts distributed on a weighted average. This is because there were more posts with a lesser hate score than higher (15.716 with

a score of 0-0.25 versus 2.047 with a score of 0.75 and higher). All coefficients were calculated on the first 100 top ranked posts, as this is more representative of the data a user will encounter when searching for posts.

To calculate correlations between hate scores and search rankings, I first compute Pearson's correlation between the average hate score of a post across all its appearances in the dataset (Unique_Hate_Score) and the average rank of a post across all its appearances (if a post fluctuates in ranking over time, this represents its typical rank). It is important to remember that higher rank numbers mean, in effect, a *worse* positioning, in the sense that the platform tries not to have it be prominent. A negative correlation would suggest that higher hate scores are associated with lower ranks, as follows in Table 2.2.

Effect	Statistical correlation	Interpretation
Hateful content ranks worse	Positive correlation (e.g., +0.08)	Moderation is working as intended
Hateful content ranks better	Negative correlation (e.g., -0.08)	Moderation is not working as intended
No correlation	No clear trend (near 0)	Other factors may determine ranking (X's hate speech detector works differently from OpenAI's; etc.)

Table 2.2 Pearson's correlations between the average hate score of a post across all its appearances in the dataset (Unique_Hate_Score) and the average rank of a post across all its appearances.

For correlations of hate scores and search rankings, it was first necessary to normalise views because some posts had few and others millions. I do this with log-transformation, a mathematical operation used to compress large values while keeping smaller ones distinguishable. The formula used here is $\log(1 + x)$ (a log1p transformation) instead of $\log(x)$, so as to handle cases where $x = 0$ (since $\log(0)$ is undefined). Like rankings, a negative correlation (e.g., -0.05) means that posts with high hate scores tend to get less engagement, while a positive correlation means that they tend to get more engagement.

Finally, I also quantified the overall fluctuation of posts with different hate scores across rankings. First, this entailed bracketing posts per "bins" of hate score first, as follows: 0-0.25 (low); 0.25-0.50 (medium-low); 0.50-0.75 (medium-high); and 0.75-1 (high). Then, for each unique post, I arrange its rankings in chronological order and calculate how much a rank changes at each step by taking the difference between each rank. If a post moves up in ranking (e.g., from rank 50 to 20), I count that as -30 (i.e., it moved up 30 places). If a post moves down in ranking (e.g., from rank 20 to 50), I count that as +30 (moved down 30 places). For each post, I add up all the rank changes and average the results per hate score bin.

Deletion

I also wanted to check the extent to which X enforces its (few) directives to suspend or delete posts, as indicated in the *Hateful conduct* and adjacent policies. This meant checking whether posts with high "hate scores" had been deleted or suspended between the period of data collection — October 10th, 2024 — and as late as possible — February 20th, 2025. Once again, I used a scraper to collect the status of each post on February 20th. A status indicates whether it is either online (in which case it displays the post as normal), or whether it has been removed or made unavailable. In the latter case, statuses do not show reasons *why* a post may have been deleted. They include: "Hmm...this page doesn't exist. Try searching for something else.", "This Post is from an account that no longer exists. Learn more" or "This Post is unavailable. Learn more". Clearer motives are given by these statuses: "This Post was deleted by the Post author. Learn more", "You're unable to view this Post because this account owner limits who can view their Posts. Learn more" or "This Post is from a suspended account. Learn more".

Here, the analysis consisted in verifying what was removed and what remained online, relative to each post's visibility (i.e., impressions and search ranking) and "hate score". I only included posts with a "hate score" of 0.80 or higher.

TYPE	ITEM	VALUE
Posts from search results	Total	23.071
	Minimum rank	1
	Maximum rank	1.041
	Minimum views	0
	Maximum views	56.000.000
	Position 1 ('Against immigration, migrants, foreigners, refugees, or Muslims" AND "against others that support them')	14.049
	Position 2 ('In support of immigration, migrants, foreigners, refugees, or Muslims" AND "against those that oppose them')	2.986
	Position 3 (neutral)	4.442
	Position 'Other'	1.635
	Posts with a 'hate score' of 0-0.25	15.716
	Posts with a 'hate score' of 0.25-0.50	3.540
	Posts with a 'hate score' of 0.50-0.75	2.047
	Posts with a 'hate score' of 0.75-1.00	1.775
	Posts with a hate score of >0.80	1.392
	- Of which removed (total)	142

	- <i>Removed for no stated reason</i>	117
	- <i>Removed by the author</i>	20
	- <i>Author limited who can view their posts</i>	5
Community notes	Total unique notes	6.430
	Nr of posts on which notes were applied	4.718

Table 2.3 Total distributions across Community Notes and posts obtained via search results.

What Twitter was

In what follows, I look more specifically at how X policies and content moderation practices have changed since Musk's takeover. I briefly summarise Twitter's content moderation regime(s) before its acquisition, and then I outline specific changes in each policy from October 2022 to December 2024 before outlining recommendations for dealing with moderation.

2010-2013: user reporting

Before Musk's takeover, it could be argued that Twitter underwent three major phases. The first spanned from roughly 2010 to 2016 and was initially defined by a relatively lax model of moderation. As a self-described "aggressively open company" (Twitter, 2010), users "owned" what they "shared" (ibid): they bore responsibility for "the content he or she provided" as much as they exercised the knowledge to judge what may be unacceptable to themselves and others. The only exception to the rule was content that could enter the jurisdiction of local law enforcement, which the platform frequently mentioned in its *Abusive behavior* policy. This included specific statements of violence against others and other content that may lead to physical consequences against users (Twitter, 2013). The level of tolerance for these kinds of content was minimal, though not absolutely so. While most measures were phrased vaguely as internal "investigations for abuse", they were the only types of content other than "direct, specific threats of violence against others" that deserved a permanent suspension.

An open sphere

Twitter content moderation policies and corresponding techniques, 2010

Our goal is to provide a service that allows you to discover and receive content from sources that interest you as well as to share your content with others. We respect the ownership of the content that users share and each user is responsible for the content he or she provides. Because of these principles, we do not actively monitor user's content and will not censor user content, except in limited circumstances described below.

help.twitter.com, 2010

- Internal investigations
- Temporary or unspecified suspension

Source

Twitter Rules

Method

1. Scrape text for every month of policy changes archived in the Wayback Machine.
2. Register (1) URL structure;
 - (2) definition of objectionable information in each policy;
 - (3) examples of objectionable information;
 - (4) content moderation techniques;
3. Visualisation shows (1) policy;
 - (2) first applicable content moderation technique;
 - (3) second; (4) third; (5) and fourth. Techniques may be applied consecutively or interchangeably.

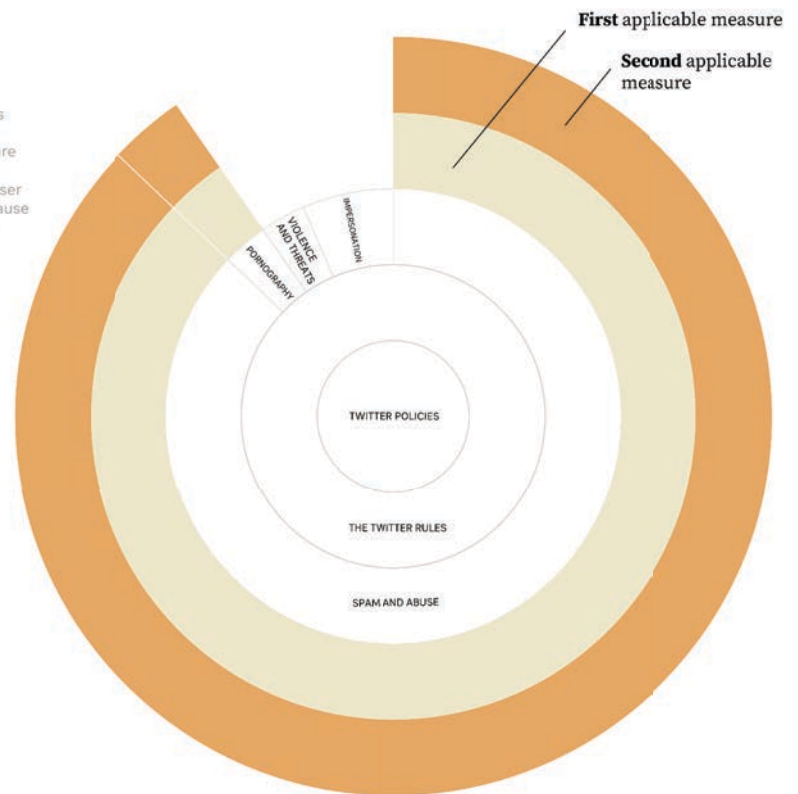


Figure 2.5 Twitter's content moderation regime circa 2010. White rings represent policies and colored rings their respective enforcement measures, arranged sequentially in the order of their application. Full image [here](#). Source: author.

2013-2017: deletions and suspensions

By January 2016, Twitter began to introduce an array of policies to combat a variety of "harms" and "spam content", including "abusive behavior", "hateful conduct", "hateful imagery" and "violence" (see Figure 2.6 below). By 2016, the context was of course the "techlash" that ensued after journalistic and academic diagnoses of "fake news" and hate speech scandals during the Brexit referendum and US elections. Mounting pressure from all sides of public debate, in particular state actors and legislators, meant multiplying a taxonomy of "harms" for the platform to moderate and introducing more guardrails in the form of deletions, temporary or permanent suspensions.

A transitioning space Twitter content moderation policies and corresponding techniques, 2013-2017

"We believe that everyone should have the power to create and share ideas and information instantly, without barriers. In order to protect the experience and safety of people who use Twitter, there are some limitations on the type of content and behavior that we allow."

Twitter Rules, Dec 2017

- Internal investigations
- Delete (Tweet, content)
- Temporary or unspecified suspension
- Permanent suspension
- Unspecified
- User reporting
- Reach out

Source
Twitter Rules
Hateful conduct policy
Abusive behaviour

Method
1. Scrape text for every month of policy changes archived in the Wayback Machine.
2. Register (1) URL structure; (2) definition of objectionable information in each policy; (3) examples of objectionable information; (4) content moderation techniques.
3. Visualization shows: (1) policy; (2) first applicable content moderation technique; (3) second; (4) third; (5) and fourth. Techniques may be applied consecutively or interchangeably.

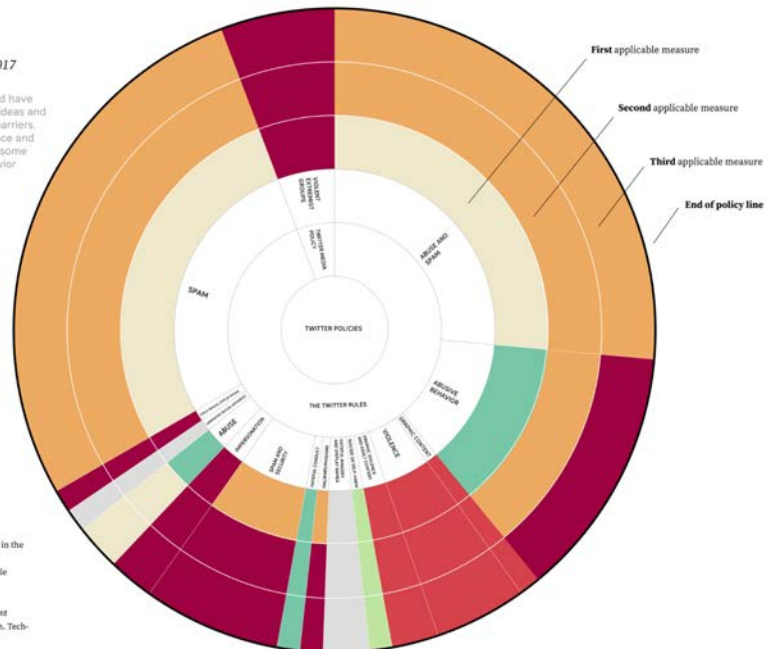


Figure 2.6 Twitter's content moderation regime circa 2013-17. White rings represent policies and colored rings their respective enforcement measures, arranged sequentially in the order of their application. Full image [here](#). Source: author.

2017-2022: modularity

In response to the first techlash of 2018, Twitter had, by 2020, developed a number of policies designed specifically for important institutional processes. While this included the US elections of 2020, COVID-19 had itself deeply transformed content moderation as a practice. Initially, the platform's level of tolerance was lowered for abusive, hateful and misinformative content, shrinking the amount of permissible content, prompting a change in its user base and the exodus of banned content to various counter spheres.

At the same time, the dynamic aspects of objectionability — the possibility that some content may become problematic tomorrow, based on the shifting norms of users and institutions — pushed Twitter, Google and Meta to design more granular moderation techniques. These are demotion (downranking tweets in replies and newsfeeds); hiding sensitive or problematic tweets unless a user clicks on them; and a "strike system" that allows users to return after temporary suspension. From 2016, then, moderation techniques shifted from a relatively binary logic of deletion and suspension, to one where Twitter attempted to preemptively moderate *potentially* and *relatively* objectionable content.

Modular moderation Twitter content moderation policies and corresponding techniques, 2019-2021

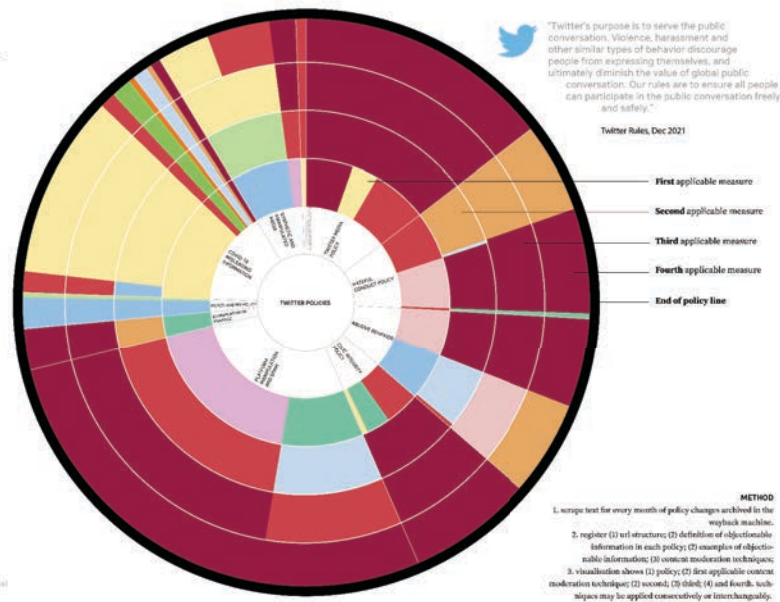
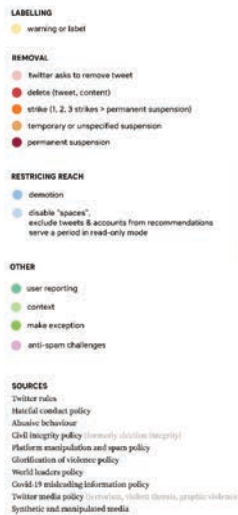


Figure 2.7 Twitter's content moderation regime circa 2019-21. White rings represent policies and colored rings their respective enforcement measures, arranged sequentially in the order of their application. Full image [here](#). Source: author.

The end result, in Figure 2.7, was a platform that simultaneously maintained clearly defined boundaries (seen in dark red, indicating permanent suspensions) as well as conditions or modular ones (seen in blue shades, indicating demotion or rendering content invisible). In more or less discrete funding calls, Twitter invited content moderation initiatives more attuned to the "health of public debate" (Twitter, 2018), using a terminology alluding to public or civic dialogue, depolarisation and conflict reconciliation. Community Notes emerged in this context as an improved type of "fact checking", where "truth" would emerge not from authoritative sources but public consensus.

What X became

That was the platform that Musk acquired: one that, while complexifying and enforcing its content moderation system, also began indicating a certain scepticism towards vertical and authoritative decision-making. It was natural, in this sense, that Jack Dorsey initially made a point about framing Musk's acquisition as "the singular solution I trust" (Fung, 2023). Of course, this sentiment did not extend much beyond 2023 (ibid). But it was a testament to the fact that some of Twitter's late content moderation philosophy found some continuation under Musk — particularly the idea that "freedom of speech is not freedom of reach". This shift meant refraining from enforcing "ground truths" and normative decisions upon the platform's diverse user base ("Who is to say that one person's misinformation is another person's misinformation?" (Musk in Clayton, 2023)) by shifting large parts of that decision-making to users themselves through "deliberation mechanisms", such as Community Notes. This meant an investment in two major content moderation strategies: (1) norm-agnostic moderation measures, such as demotion, that are not easily perceived by users; and (2) crowdsourced moderation in the likes of Community Notes, which have expanded to other platforms and, potentially, other X functions.

Policy revisions

How exactly did this philosophy influence X's content moderation policies? From Figure 2.8, one can tell that nearly all content moderation policies have undergone some changes in the Summer of 2023 and towards March 2024; I will revisit each in detail below. On a macro level, it is also evident that some policies having to do with COVID have been removed — because "COVID is no longer an issue" (Musk in Clayton, 2023). Others have been merged and simplified into three policy sectors: (1) "Platform integrity and authenticity" for anything relating to spam, platform manipulation, manipulated media, username squatting and civic integrity; (2) "Safety and Cybercrime" for violent content, child safety, illegal goods, and adult content; and (3) "Platform Use Guidelines" for questions relating to the platform's general philosophy and usage.

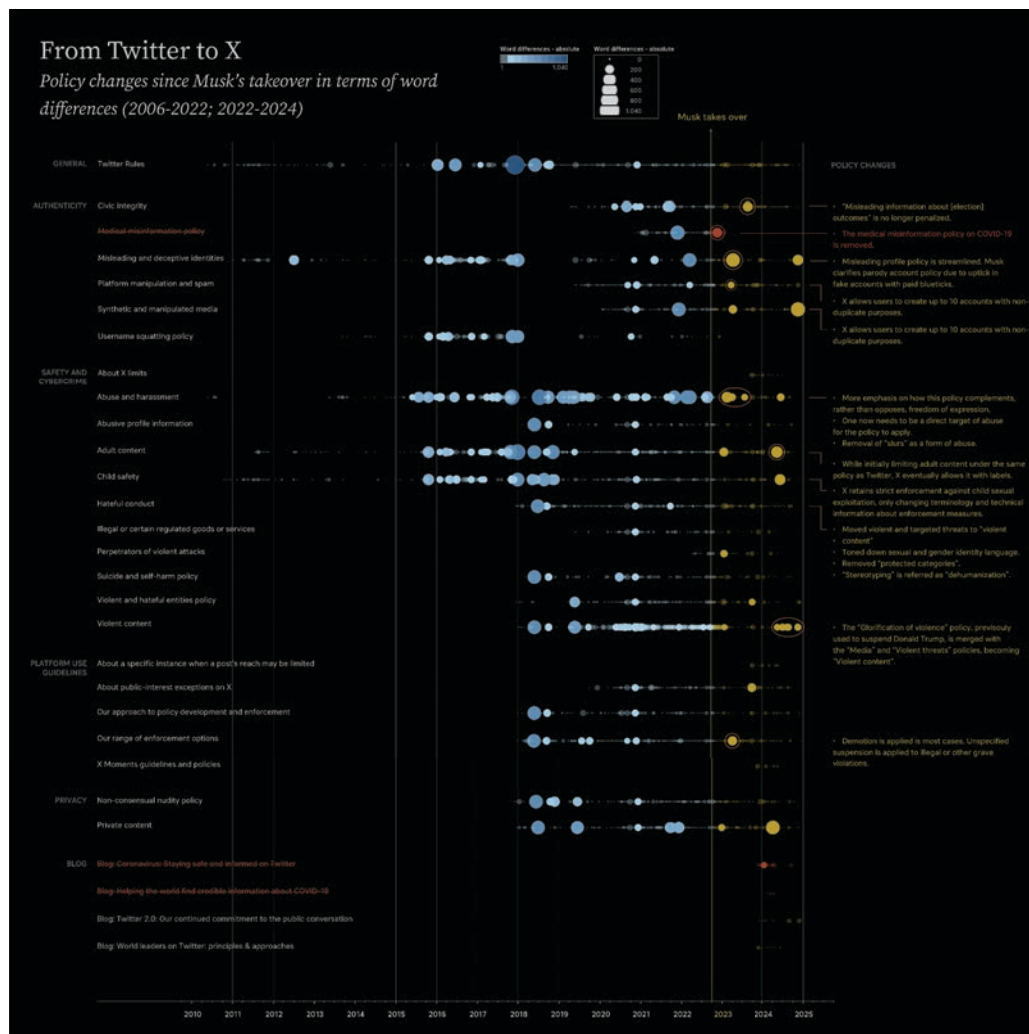


Figure 2.8 Bee swarm showing absolute number of word differences per Twitter/X policy, 2010-2025. Blue dots represent changes on Twitter, and yellow dots on X. Red dots represent instances in which a policy has been removed. Policy names and structure are from December 2024. Full image [here](#). Source: author.

This restructuring is contingent on policies becoming more minimalistic, in the sense that they are simplified into one single, core purpose. This was part of an effort by the X Safety team to "consolidate a number of pages [...] and roll out a new, simplified template", reducing "the overall number of articles and total word count significantly",

in order to make rules "clearer for everyone" (X Safety, 2024). Previously separate, the *Abusive behaviour* and *Hateful conduct* policies have been merged into a new *Violent content* policy that focuses mostly on the physical aspect and potential of online violence, rather than personal offences or "cultures" of violence (in the sense of beliefs, ideas or attitudes that may legitimize violent dispositions). Likewise, policies under the "Platform integrity and authenticity" section now focus mostly on "artificial" or instrumentalized platform manipulation, such as attempts to boost engagement, trick platform mechanisms, and so on. The terminology of "misinformation" — "misleading" content, "deceptive" behaviour — is removed or minimized.

In this context, policies highlight how they complement, rather than obstruct, freedom of expression and "diversity of perspectives" (X Community Notes, 2025). One aspect of this shift is that definitions and estimations of objectionability tend to be outsourced to local legislation where applicable (including, surprisingly, "deadnaming", i.e., the practice of calling someone by a name they no longer assimilate with), or to users themselves, via, for example, Community Notes. On that point, policies adopt a more conditional tone towards users: "you can't" becomes "you may not"; "your account will be permanently suspended" becomes "your account may be suspended"; and in some cases, some of what used to be forbidden is now allowed as part of new features, particularly sharing *Adult content*: "You may not post adult content [...]" becomes "You may share consensually produced and distributed adult nudity [...] provided it's properly labelled and not prominently displayed." (See Annex.) Each of these changes are described in more detail below.

Social justice terminology is reduced in the Hateful Conduct policy

First, under the directive to facilitate freedom of expression and thereby an open and politically "centrist" platform (see interview in Joe Rogan, 2024), policies abandon any semblance of social justice language in the *Hateful* and *Abusive behaviour* policies. This includes contemporary language about gender, sexuality and race. As shown in Figure 2.9, the *Hateful conduct* policy tones down any mention of sexual or gender identity or "protected categories" such as "women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual[s]" and other "marginalized and historically underrepresented communities". Instead, it shifts this language to a question of personal discretion ("those that identify with multiple underrepresented groups") or to identities with more "universal" political and legal resonance ("race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease"). As a result, "stereotyping" and attacking others with "slurs" — the discrimination of one's personal and "protected" identity — is changed in favour of more universal forms of discrimination: "dehumanization", "insults" and "profanity".

One month before Musk acquisition	2022-09-01	One year later	2023-09-23
Hateful conduct policy [Hateful conduct:](http://web.archive.org/en/rules-and-policies/twitter-rules.html)[(https://help.twitter.com/rules-and-policies/twitter-ruleshateful-conduct)-You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories: Hateful imagery and display names:• You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category. Rationale Twitter's mission is to give everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers. Free expression is a human right – we believe that everyone has a voice, and the right to use it. Our role is to serve the public conversation, which requires representation of a diverse range of perspectives. We recognize that if people experience abuse on Twitter, it can jeopardize their ability to express themselves. Research has shown that some groups of people are disproportionately targeted with abuse online. This includes; women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities. For those who identify with multiple underrepresented groups, abuse may be more common, more severe in nature and more harmful.		Hateful Conduct Overview April 2023 You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. X's mission is to give everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers. Free expression is a human right – we believe that everyone has a voice, and the right to use it. Our role is to serve the public conversation, which requires representation of a diverse range of perspectives. We recognize that if people experience abuse on X, it can jeopardize their ability to express themselves. Research has shown that some groups of people are disproportionately targeted with abuse online. For those who identify with multiple underrepresented groups, abuse may be more common, more severe in nature, and more harmful.	

Figure 2.9 A snapshot of Hateful Conduct policy changes. Full policy [here](#). Source: X/Twitter and author.

At the same time, there is also a particularization of hateful (and thereby personal) language on the website. In March 2023, X used a new hate speech detection system called *Sprinklr*, which the company says "defines hate speech more narrowly" by focusing on "the nuanced context of their use" (X Safety, 2023a). Under this logic, it is the number of "impressions" (views) of hateful language, not the sheer quantity, that matter most. When a post is found to have hateful language by Sprinklr's method, it is downranked and not removed.

U-turns in the Abuse and harassment policy

The current *Abusive behaviour* policy (Figure 2.10) takes a few more U-turns. A month after Musk's takeover, the policy is radically simplified, synthesizing the policy's relatively complex taxonomy of "abuse" into four main sections: targeted or weaponized harassment; calling for others to harass an individual or group of people; making sharing unwanted sexual content; insulting others; and denying violent events. One needs to be a direct target or somewhat directly involved in harassment or abuse for the policy to be valid.

One month before Musk acquisition	2023-09-21	Two years later	2024-09-10
Abuse and harassment		Abuse and Harassment	
Overview		Overview	
June 2023		March 2024	
You may not share abusive content, harass someone, or encourage other people to do so:		You may not target others with abuse or harassment, or encourage other people to do so.	
On X, you should feel safe expressing your unique point of view. We believe in freedom of expression and open dialogue, and in order to facilitate healthy dialogue on the platform, and empower individuals to express diverse opinions and beliefs, we prohibit behavior and content that harasses, shames, or degrades others. In addition to posing risks to people's safety, abusive behavior may also lead to physical and emotional hardship for those affected.		X's mission is to give everyone the power to create and share ideas and information, as well as express their opinions and beliefs without barriers. Free expression is a human right – we believe that everyone has a voice, and the right to use it. Our role is to serve the public conversation, which requires representation of a diverse range of perspectives.	
What is in violation of this policy?		We recognize that if anyone, regardless of background, experiences harassment on X, it can jeopardize their ability to express themselves and cause harm. To facilitate healthy dialogue on the platform, and empower individuals to express diverse opinions and beliefs, we prohibit behavior and content that harasses, shames, or degrades others. In addition to posing risks to people's safety, these types of behavior may also lead to physical and emotional hardship for those affected.	

Figure 2.10 A snapshot of Abuse and Harassment policy changes. Full policy [here](#). Source: X/Twitter and author.

The following year, the policy added more (con)text to what it initially deleted. It reintroduces the clause to "hear directly from the person being targeted" prior to enforcing the policy. Perhaps surprisingly, it reintroduces a clause against using "prior names and pronouns", which stipulates that, "when required by local laws", the platform would "reduce the visibility of posts that purposefully use different pronouns to address someone other than what that person uses for themselves". As of February 2025, this clause still exists in spite of Trump's executive orders against transgender terminology.

Misleading information

Likewise, policies relating to the management of "fake", "manipulated" or "misleading" information no longer focus on *statements* made by users, but on objective evidence and malicious use of artificiality — or, in the case of institutional processes (such as elections), physical attempts to prevent them from taking place. The *Civic Integrity* (previously *Electoral Integrity*) policy, for example, no longer penalizes "misleading information about outcomes" or "information intended to undermine public confidence in an election or other civic process" (X, 2025b). This includes "disputed claims that could undermine faith in the process itself" (such as "unverified information about election rigging, ballot tampering, vote tallying, or certification of election results"); and "misleading claims about the results or outcome of a civic process" (such as "claiming victory before election results have been certified") (ibid). Potentially problematic statements or "awful content" — however defined — will be subject to soft moderation methods, namely: "excluding the post from search results, trends and recommendations"; removing it from one's personal newsfeeds; "restricting the post's discoverability"; "restricting likes, replies, reposts, quotes, bookmarks, share, pin to profile, or edit post"; and "downranking the post in replies". Previously, tweets would first be labelled and eventually suspended, first temporarily and then permanently, under a strike system.

One month before Musk acquisition 2022-09-01	One year later 2023-09-01
Civic integrity policy Overview October 2021 You may not use Twitter's services for the purpose of manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process. In addition, we may label and reduce the visibility of Tweets containing false or misleading information about civic processes in order to provide additional context. The public conversation occurring on Twitter is never more important than during elections and other civic events. Any attempts to undermine the integrity of our service is antithetical to our fundamental rights and undermines the core tenets of freedom of expression, the value upon which our company is based. We believe we have a responsibility to protect the integrity of those conversations from interference and manipulation. Therefore, we prohibit attempts to use our services to manipulate or disrupt civic processes, including through the distribution of false or misleading information about the procedures or circumstances around participation in a civic process. In instances where misleading information does not seek to directly manipulate or disrupt civic processes, but leads to confusion on our service, we may label the Tweets to give additional context. Given the significant risks of confusion about key election information, we may take these actions even if Tweets contain (or attempt to contain) satirical or humorous elements.	Civic integrity policy Overview August 2023 You may not use X's services for the purpose of manipulating or interfering in elections or other civic processes, such as posting or sharing content that may suppress participation, mislead people about when, where, or how to participate in a civic process, or lead to offline violence during an election. Any attempt to undermine the integrity of civic participation undermines our core tenets of freedom of expression and as a result, we will apply labels to violative posts informing users that the content is misleading.

Figure 2.11 A snapshot of Civil Integrity policy changes. Note how the 2023 version (on the right) is reframed as a means to protect freedom of expression. Full policy [here](#). Source: X/Twitter and author.

Suppression and intimidation We will label or remove false or misleading information intended to intimidate or dissuade people from participating in an election or other civic process. This includes but is not limited to: - misleading claims that polling places are closed, that polling has ended, or other misleading information relating to votes not being counted; - misleading claims about police or law enforcement activity related to voting in an election, polling places, or collecting census information; - misleading claims about long lines, equipment problems, or other disruptions at voting locations during election periods; - misleading claims about process procedures or techniques which could dissuade people from participating; and - threats regarding voting locations or other key places or events (note that our [violent threats policy](https://help.twitter.com/rules-and-policies/violent-threats-glorification) may also be relevant for threats not covered by this policy). Misleading information about outcomes We will label or remove false or misleading information intended to undermine public confidence in an election or other civic process. This includes but is not limited to: - disputed claims that could undermine faith in the process itself, such as unverified information about election rigging, ballot tampering, vote tallying, or certification of election results; and - misleading claims about the results or outcome of a civic process which calls for or could lead to interference with the implementation of the results of the process, e.g. claiming victory before election results have been certified; inciting unlawful conduct to prevent the procedural or practical implementation of election results (note that our violent threats policy may also be relevant for threats not covered by this policy).	Suppression You may not advance verifiably false or misleading information about the circumstances surrounding a civic process intended to intimidate or dissuade people from participating in an election or other civic process. This includes but is not limited to: - misleading claims that polling places are closed, that polling has ended, or other misleading information relating to votes not being counted; - misleading claims about police or law enforcement activity related to voting in an election, polling places, or collecting census information; - misleading claims about long lines, equipment problems, or other disruptions at voting locations during election periods; Intimidation You may not incite, promote or encourage others to threaten or coerce others to participate or refrain from participating in a civic process. This includes, but is not limited to: - inciting or promoting violent behaviors intentionally near a location where an electoral process is being conducted, including polling stations and vote counting locations; - inciting the disruption or destruction of procedures, infrastructure, or election equipment that is necessary for someone to participate in a civic process; - inciting others to harass voters or poll workers;
---	--

Figure 2.12 A snapshot of Civil Integrity policy changes. Note how "We will label or remove false or misleading information" becomes "You may not advance verifiably false or misleading information", and the focus on "Misleading information about outcomes" turns to physical "Intimidation" against others participating in a civic

process. Full policy [here](#). Source: X/Twitter and author.

In the former case, *Misleading identities* relaxes the imperative to be one's "authentic self": they are no longer "required to display [one's] real name or image on [their] profile". It allows accounts to be inauthentic if they are parodies (as it did before (Twitter, 2021)), provided they clearly indicate it — particularly in reaction to Musk abandoning the blue tick as a means to verify a public person's authenticity. Another form of sanctioned "fakeness" is the misappropriation of another user's identity, or artificially inflating engagement and using fake profile images. Correspondingly, moderation is relaxed or unclear: one can be "suspended" but it is unclear if the measure is temporary or permanent.

One month before Musk acquisition	2022-09-01	A year later	2023-09-05
Misleading and deceptive identities policy		Misleading and deceptive identities policy	
Overview		Overview	
You may not impersonate individuals, groups, or organizations to mislead, confuse, or deceive others; nor use a fake identity in a manner that disrupts the experience of others on Twitter.		April 2023	
We want Twitter to be a place where people can find authentic voices. That means one should be able to trust that the person or organization featured in an account's profile genuinely represents the account owner. While you are not required to display your real name or image on your profile, your account should not engage in impersonation or pose as someone who doesn't exist in order to deceive others. Accounts that use deceptive identities can create confusion, as well as undermine the integrity of conversations on Twitter. For this reason, you may not misappropriate the identity of another person, group, or organization, or create a fake identity for deceptive purposes.		You may not misappropriate the identity of individuals, groups, or organizations or use a fake identity to deceive others.	
What is a misleading or deceptive identity?		We want X to be a place where people can find authentic voices. While you are not required to display your real name or image on your profile, your account should not use false profile information to represent itself as a person or entity that is not affiliated with the account owner, such that it may mislead others who use X.	
One of the main elements of an identity on Twitter is an account's profile, which includes a username (@handle), account name, profile image, and bio.			
An account's identity is deceptive under this policy if it uses false profile information to represent itself as a person or entity that is not associated with the account owner, such that it may mislead others who use Twitter. Deceptive identities may feature the likeness of another person or organization in a manner that confuses others about the account's affiliation. Fake identities, which may use stolen or computer-generated photos and fabricated names to pose as a person or organization that doesn't exist, are also considered deceptive when they engage in disruptive or manipulative behavior.			

Figure 2.13 A snapshot of Misleading Identity policy changes. Full image [here](#). Source: X/Twitter and author.

Adult content

Another policy that underwent significant changes was *Adult content*, previously *Sensitive media*. Until March 2023, *Sensitive media* addressed graphic violence, adult content, violent sexual conduct, gore content and hateful imagery. Under a policy of simplification and consolidation, the March 2023 version of the policy moved hateful imagery to the *Hateful conduct* policy. Then, in May 2024, the policy split in two: sanctions against violent content moved to the *Violent content* policy, and adult content was given a new policy. Today, this policy expands affordances for sharing "consensually produced [...] adult nudity or sexual behavior" when and if it's "probably labeled". This includes "full or partial nudity" and "explicit or implied sexual behavior [...] such as sexual intercourse". When logged in, the user can see this

and other sensitive media (for example, gory images from war scenes) by confirming their age and whether they would like to watch the content. It is restricted to users who are under 18, and cannot be displayed in "highly visible" places like profile pictures, headers, list banners, or community cover photos.

This move has been reported as an effort to tap into more streams of revenue (Oremus & Hunter, 2024), or Musk wishing X to become a "super app" with multiple capacities beyond pure social media (Heath, 2023). Philosophically, this change is also concomitant to X wanting to become the platform for free speech, including "sexual expression" as a form of "artistic expression" (see Figure 2.13).

One month before Musk acquisition 2022-09-01	A year later 2023-09-01	Two years later 2024-10-09
Sensitive media policy Overview January 2022 You may not post media that is excessively gory or share violent or adult content within live video or in profile header, or List banner images. Media depicting sexual violence and/or assault is also not permitted. People use Twitter to show what's happening in the world, often sharing images and videos as part of the conversation. Sometimes, this media can depict sensitive topics, including violent and adult content. We recognize that some people may not want to be exposed to sensitive content, which is why we balance allowing people to share this type of media with helping people who want to avoid it to do so. For this reason, you can't include violent, hateful, or adult content within areas that are highly visible on Twitter, including in live video, profile, header, or List banner images. If you share this content on Twitter, you need to [mark your account as sensitive](http://web.archive.org/web/20220901/http://twitter.com/settings-and-policies/media-settings.html). Doing so places images and videos behind an interstitial (or warning message), that needs to be acknowledged before your media can be viewed. Using this feature means that people who don't want to see sensitive media can avoid it, or make an informed decision before they choose to view it. We also restrict specific sensitive media, such as adult content, for viewers who are under 18 or viewers who do not include a birth date on their profile. Learn more about age restricted content [here](http://web.archive.org/web/20220901/http://twitter.com/settings-and-policies/notifications-on-twitter.html).	Sensitive media policy Overview March 2023 You may not post media that is graphic or share violent or adult nudity and sexual behavior within live video or in profile header, List banner images, or Community cover photos. Media depicting excessively gory content, sexual violence and/or assault, bestiality or necrophilia is also not permitted. People use Twitter to show what's happening in the world, often sharing images and videos as part of the conversation. Sometimes, this media can depict sensitive topics, including graphic content and adult nudity and sexual behavior. We recognize that some people may not want to be exposed to sensitive content, which is why we balance allowing people to share this type of media with helping people who want to avoid it to do so. For this reason, you can't include graphic content or adult nudity and sexual behavior within areas that are highly visible on Twitter, including in live video, profile, header, List banner images, or Community cover photos. If you share this content on Twitter, you need to [mark your account as sensitive](http://web.archive.org/web/20230901/http://twitter.com/settings-and-policies/media-settings.html). Doing so places images and videos behind an interstitial (or warning message), that needs to be acknowledged before your media can be viewed. Using this feature means that people who don't want to see sensitive media can avoid it, or make an informed decision before they choose to view it. We also restrict specific sensitive media, such as adult nudity and sexual behavior, for viewers who are under 18 or viewers who do not include a birth date on their profile. Learn more about age restricted content [here](http://web.archive.org/web/20230901/http://twitter.com/settings-and-policies/notifications-on-twitter.html).	Adult Content Overview May 2024 You may share consensually produced and distributed adult nudity or sexual behavior, provided it's properly labeled and not prominently displayed. We believe that users should be able to create, distribute, and consume material related to sexual themes as long as it is consensually produced and distributed. Sexual expression, whether visual or written, can be a legitimate form of artistic expression. We believe in the autonomy of adults to engage with and create content that reflects their own beliefs, desires, and experiences, including those related to sexuality. We balance this freedom by restricting exposure to Adult Content for children or adult users who choose not to see it. We also prohibit content promoting exploitation, harassment, objectification, sexualization or harm to minors, and obscene behaviors. We also do not allow sharing Adult Content in highly visible places such as profile photos or banners. How we define Adult Content Adult Content is any consensually produced and distributed material depicting adult nudity or sexual behavior that is pornographic or intended to cause sexual arousal. This also applies to AI-generated, photographic or animated content such as cartoons, hentai, or anime. Examples include depictions of: • Full or partial nudity, including close-ups of genitals, buttocks, or breasts; • explicit or implied sexual behavior or simulated acts such as sexual intercourse and other sexual acts.

Figure 2.14 From Sensitive media policy to Adult content. Full image [here](#). Source: X/Twitter and author.

The policies that stay the same

The policies that have remained the same tend to be those where objectionable content is defined and sanctioned by the law, or are subjects against which there is already widespread disapproval. This includes, of course, *Child sexual exploitation* (now *Child safety*), and *Violent and hateful entities*. *Child safety* forbids any and all kinds of harm against children, and requires — unlike most other policies — immediate and permanent suspensions of users, no matter the context.

One month before Musk acquisition	2023-09-10	Two years later	2024-12-10
Child sexual-exploitation-policy Overview October 2020 We have a zero-tolerance child sexual-exploitation policy on Twitter.		Child safety May 2024 We have zero tolerance for any forms of child sexual exploitation and remove certain media depicting physical child abuse to prevent the normalization of violence against children. Our Child Safety policy is designed to protect minors from sexual and physical abuse, as well as psychological harm that may result from sharing such content. With this principle in mind, Child Safety content is any content that contains Child Sexual Exploitation, Physical Child Abuse Media and Media of Minors in Physical Altercation, as defined below. We encourage our users to come to X to share stories, raise awareness, and speak their mind - including calling attention to the exploitation of children and minors. However, even when shared with the intent to bring awareness or justice, to express outrage or sharing content in a humoristic context, posting media of children experiencing sexual abuse or certain types of physical abuse can contribute to their revictimization and may even lead to the normalization of sexual or physical violence against children. When this content is shared on X, we may remove it even if it was shared with good intent. Our priority is to protect minors from physical and psychological harm irrespective of the context in which the content may be shared. We use the terms children and minors interchangeably in any of our policies and define them as any person under the age of 18.	

Figure 2.15 The Child Safety policy before and after Musk. Full policy [here](#). Source: X/Twitter and author.

Violent and hateful entities is an interesting case, because it was first designed to prevent the circulation of jihadist content around 2017 (Twitter, 2017). Over time, it began to target domestic terrorism (as seen in Germany and New Zealand in 2019) enacted by "hateful groups". Though the *Hateful conduct* policy may have dropped language seen as left-wing "partisan", this policy retains sanctions against groups that target or harass a "protected category" (X, 2025a). For the rest, it retains its sanctions partly because they are already strongly bound by international law and industry initiatives, such as the Global Internet Forum to Counter Terrorism.

Content moderation enforcement

On the level of content moderation practices or techniques, I note a similar move towards simplification. For one, there are no clear mentions of measures being executed consequentially; they are all, more or less, imposed *in extremis*, when absolutely necessary. As shown in Figure 2.15, these are: (1) requesting a user to remove their content; (2) suspending that account in case they do not comply; and (3) reducing the content's visibility by downranking them, removing them from certain parts of the website, excluding ads from that content, and labelling a demoted tweet. Even then, X only requires a user to remove their post "if the violation is severe enough" (in place of "we require the violator to remove it before they can Tweet again" on Twitter), and permanent suspension is no longer explicitly enforced (X Help Center, 2024). The only exception to these rules is content that is invariably objectionable, such as child abuse, violent entities and non-consensual adult content.

Demotion all the way X content moderation policies and corresponding techniques, 2024

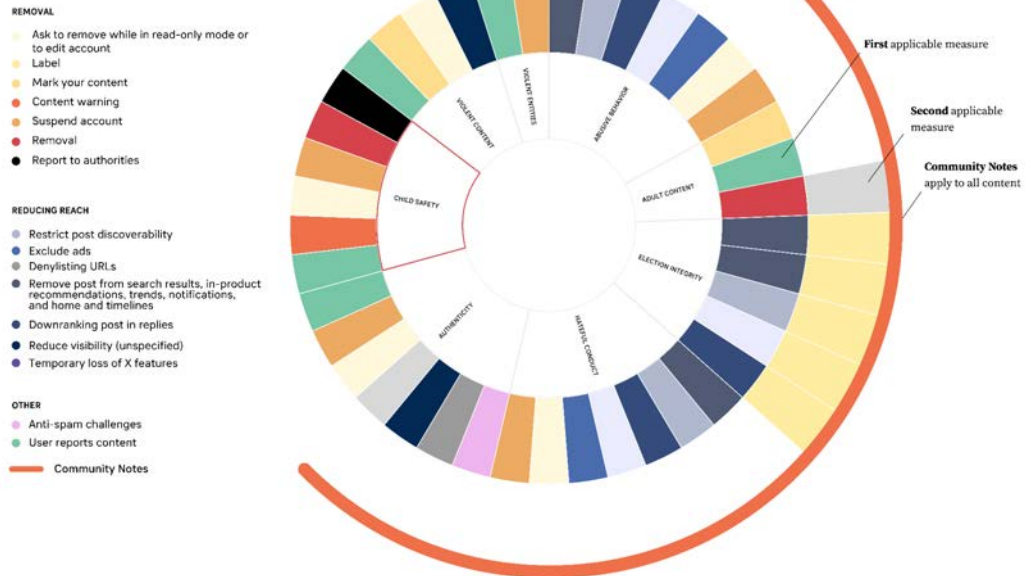


Figure 2.16 X's content moderation regime in 2024. White rings represent policies and colored rings their respective enforcement measures, arranged sequentially in the order of their application. Note: the cut-out of Community Notes on the left is to make space for legends. Full image [here](#). Source: author.

This is also reflected in X's *Enforcement Options* document. By the time of Musk's takeover, "limiting tweet visibility" was moved front and center. It is no longer just a complementary, interim measure used while waiting for a user to remove a post or return from a period of suspension. It also appears to have become more complex. While they may have been less clear in pre-Musk policies, "demotion" as a general technique previously resorted to making tweets ineligible in search results or timelines; excluding them from email or product recommendations; and burying them down replies and search results (without the possibility to share them). Post-Musk measures now include the temporary loss of X features and exclusion of ads.

For the rest, the platform outsources content moderation — and its normative decision-making — to users, via Community Notes. It may be worth noting that, under its hood, Community Notes also underwent changes analogous to the *Hateful conduct* policy. On October 27, 2022 — the very day Musk took over — some metadata fields from Community Notes were removed. These were options for community noters to qualify a post under judgments of value or norms, such as: "believable" (a response to: "If this tweet were widely spread, its message would likely be believed by:") or "harmful" ("If many believed this tweet, it might cause:"), or to estimate their "validationDifficulty" ("Finding and understanding the correct information would be:") (X, 2025).

Having laid out content moderation changes *on paper*, I now turn to an analysis of content moderation *in practice*. More specifically, I will look at how: (1) posts classified as "hateful" by industry-leading content moderation classifiers (such as OpenAI's moderation API) are ranked in X search results for keywords about immigration in Dutch; (3) how many of these posts are unavailable or deleted; and (3) what topics and users tend to have the most Community Notes.

Demotion

Let us first look at demotions for one keyword — "buitenlander" — and then proceed with average demotions for all keywords. In the case of **Figure 2.17**, representing the search ranking of posts, the posts with the highest "hate scores" tend to have lower impressions (represented by dot size), and be within the 20th and lower search rankings. Removed posts tend to be located on the 40th ranking and lower.

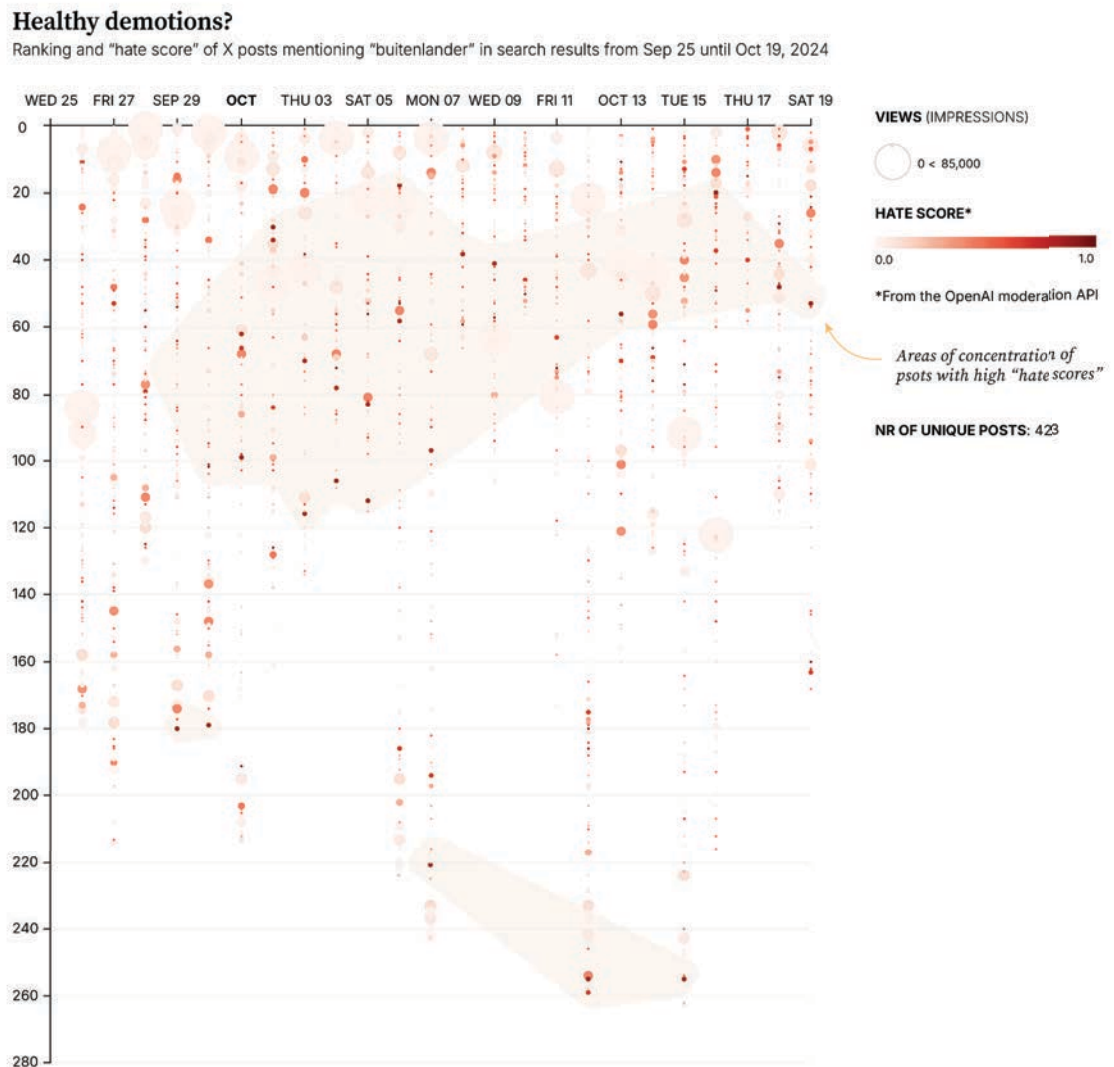


Figure 2.17 Demotion and deletion of posts mentioning "buitenlander" between September and October 2024. Full image [here](#). Source: author.

Taking a step back to look at general averages across 100 ranked posts, we see that posts with a high hate score tend to be in a slightly lower position compared to low hate score posts. Posts with a score of 0.75 tend to be positioned on the 51st rank, while those with a score of 0.25 or lower tend to be on the 49th rank. Impressions also tend to be slightly lower: 4.921 views on average for posts with a score of 0.75 or above versus 14.283 views on average for those with a score of 0.25 or less.

Hate score	Weighted Avg Rank	Weighted Avg Views	Weighted Avg Views (Log)	Total Unique Posts
0-0.25	49.77	14283	5.19	5290
0.25-0.50	49.17	5849	5.15	1330
0.50-0.75	50.45	4248	5.27	824
0.75-1.00	51.39	4921	4.82	693

Table 2.4 Weighted average ranking of posts based on their hate scores.

This is also reflected in Pearson’s correlations (Figure 2.18). Posts with higher hate scores are placed in slightly lower ranks compared to those with lower scores. The correlation coefficient of 0.008 suggests this effect is very weak, meaning hate scores only slightly impact ranking. On the other hand, higher hate scores are slightly associated with fewer views (-0.16 correlation), though the effect is once again small. In this sense, neither relationship is strong enough to suggest a platform-wide suppression effect based solely on hate scores — at least not for this dataset (in a minority language), and with hate scores estimated by OpenAI and not Sprinklr. Other factors such as engagement and recency may play a larger role.

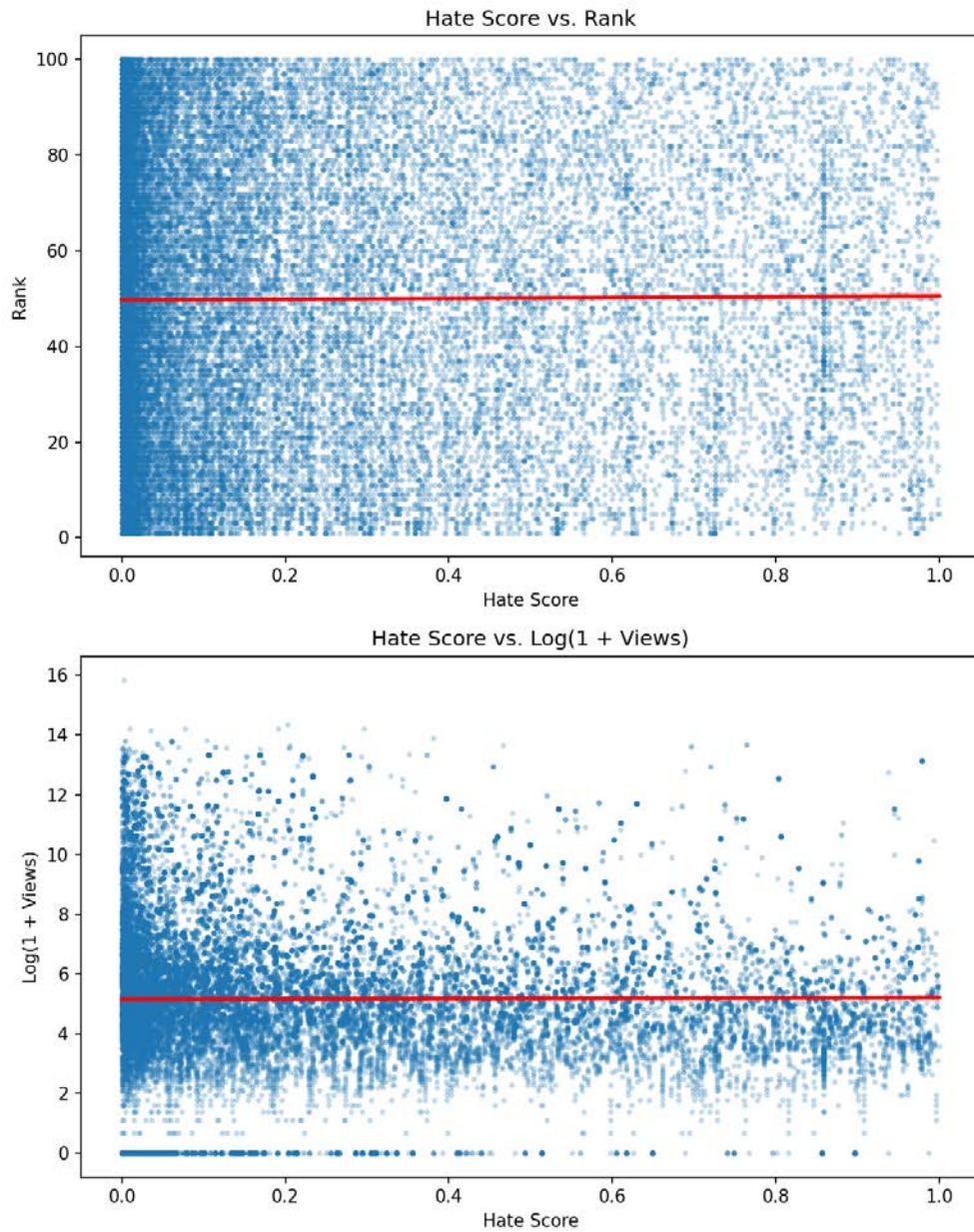


Figure 2.18 Pearson’s correlation coefficient measures the strength of the relationship between post rankings, views, and hate score. Source: author.

Though correlations are not strong, I find that posts with high hate scores are more transitory, in the sense that they tend to move across *slightly* more rankings than those with higher hate scores (Table 2.5).

Hate score	Avg Rank Movement	Median Rank Movement	Total Unique Posts
0-0.25	22.30	17.0	3538
0.25-0.50	22.39	17.0	939
0.50-0.75	22.01	17.0	582
0.75-1.00	23.01	18.0	481

Table 2.5 Average number of positions that posts with different hate scores have moved across search rankings.

Figure 2.18 illustrates this in more context. One can see the average ranking and hate score for a cut-off point of 100 posts overtime. Posts with high hate scores tend to be on the lower bottom of search results (~70 or lower), with the exception of one, in second position and low impressions, whose ranking eventually dropped over time.

All flared up

Average 'hate score', ranking and impressions (views) of X posts talking about immigration in Dutch (Sep 2024)

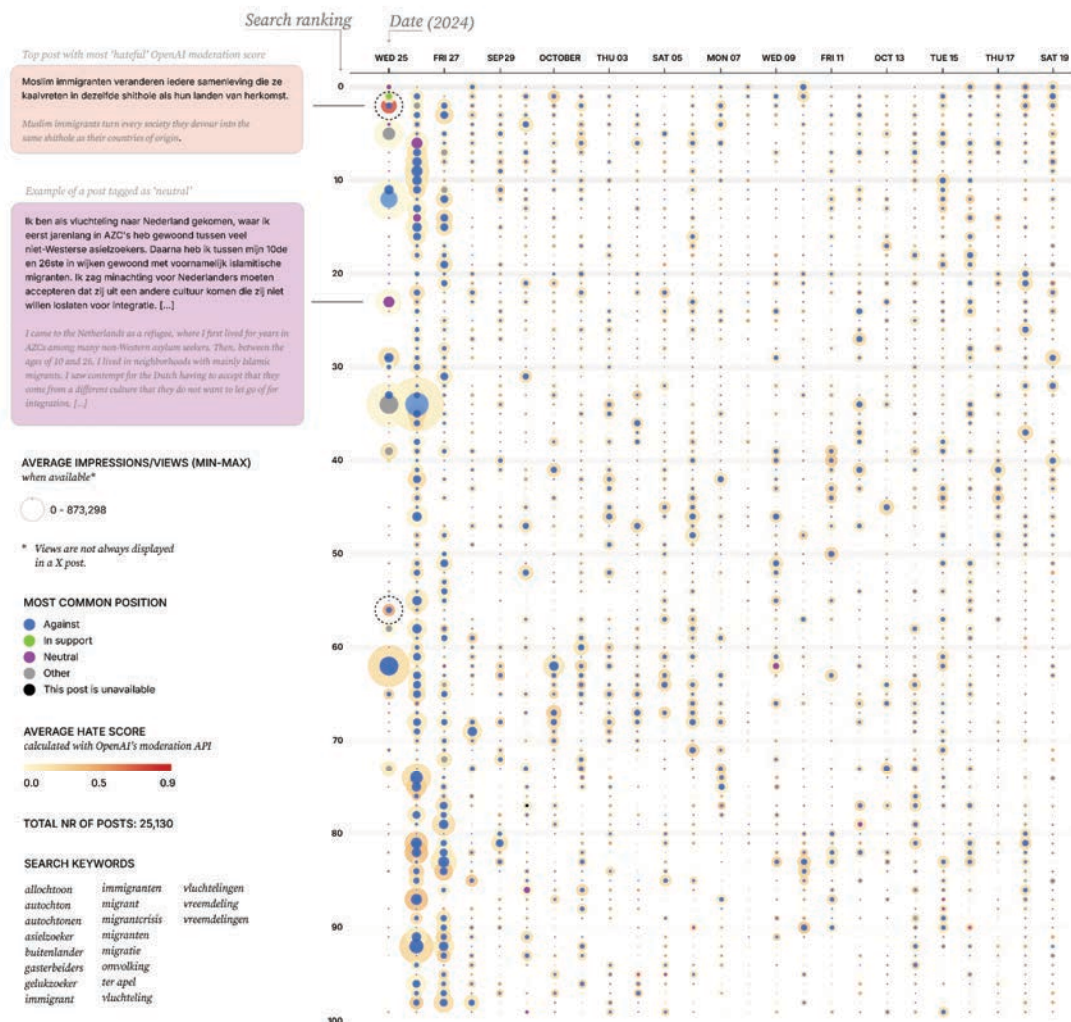


Figure 2.19 Average demotion, engagement and position for all posts mentioning any and all queries between September and October 2024. Full image [here](#). Source: author.

It must be said, however, that the general sentiment with regards to immigration is largely negative — making the expression of hateful sentiments no surprise, and hate speech detection somewhat redundant. Of 23,073 posts, 14,049 expressed sentiments *against* immigration, versus 4,442 with neutral sentiments, 2,986 in favour, and 1,257 with other sentiments. In other words, hateful language may be somewhat indistinguishable within largely negative content.

This highlights a deeper challenge in moderating hate speech: determining where to draw the line between actively hateful expression and legitimate—albeit inflammatory—debate on deeply contentious issues. In this context, the category of hateful language risks losing both its conceptual and adjudicative coherence. For one,

it would exclude a significant portion of public debate, making sustainable removal or "deplatforming" difficult without substantial repercussions.

This is the context in which developers have turned to another form of content moderation more concerned with consensus-building or "bridging" intended to facilitate "productive" or "constructive conflict" (Ovadya & Thorburn, 2023) around contentious issues *before* adjudicative interventions. The closest mechanism that approximates this goal on X — at least on paper — is Community Notes, in the sense that it attempts to calculate levels of consensus between users who write and rate notes on contentious posts. In what follows, I will outline a few findings on how this plays out in the Dutch X sphere.

Community Notes

We begin by looking at the weighted distribution of all Community Notes across post users and topics. For users, one can see in Figure 2.19 that the majority of all notes — regardless of rating — are directed at alternative influencers, who, in turn, post about conspiracy theories, energy and climate change, healthcare and wellbeing and other topics. While the majority of note ratings towards alternative influencers need more ratings, this shows that, in this dataset at least, Community Notes are primarily used to correct "factualities" or a lack thereof. The second most community noted type of user are political entities: Dutch far-right or right-wing politicians or parties are the authors of 9.02% of community noted posts, versus 3.86% by the Dutch centre-left. Journalists, fact-checkers, columnists, commentators, news media, broadcasters and academics account for a similar share.

Who gets community noted?

Weighted distribution of community notes per type of user in Dutch-speaking posts on X

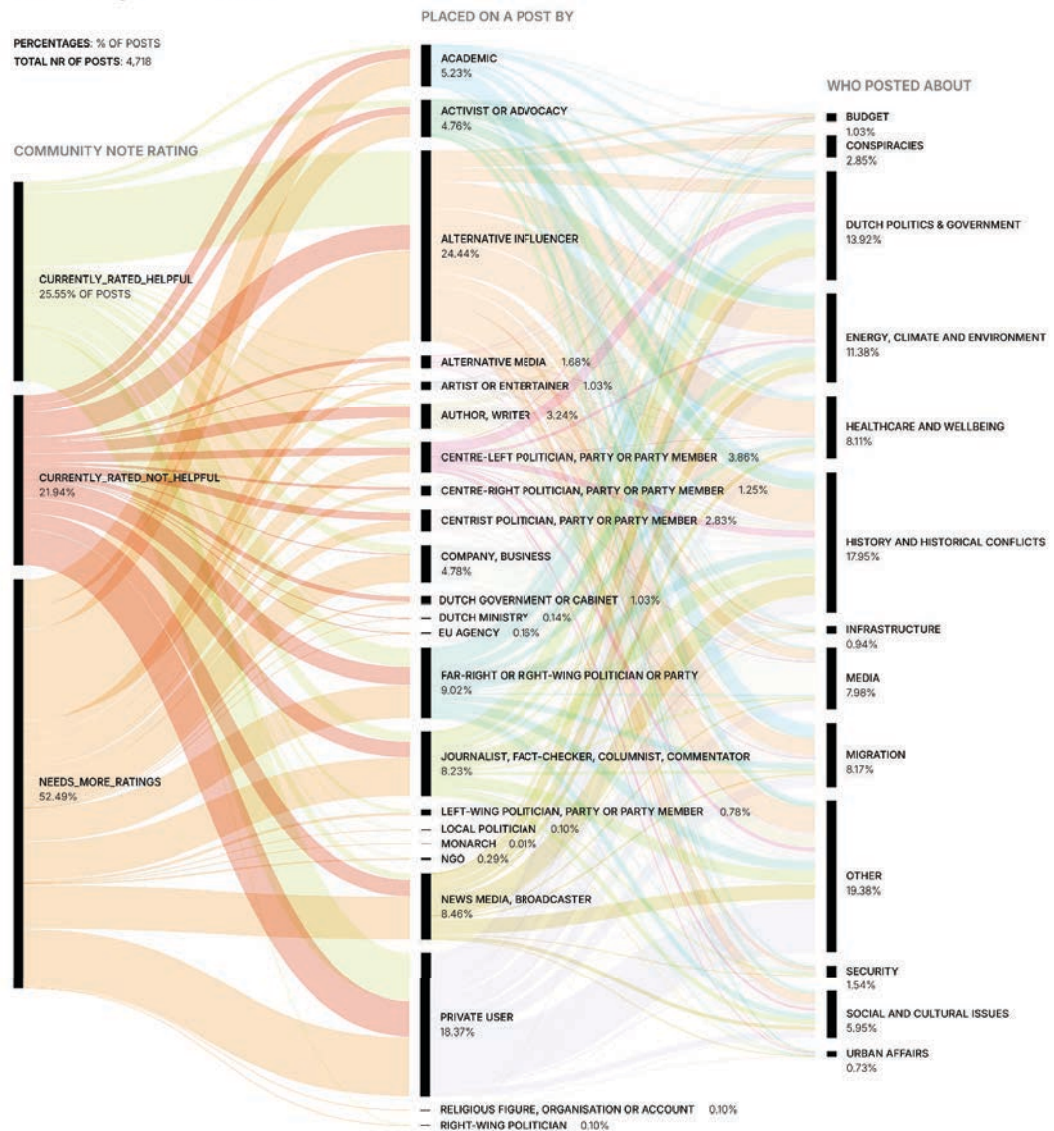


Figure 2.20 Weighted percentages of community note ratings per user type and the topic they post about. Numbers represent percentages of posts for a given rating, user type or topic. Source: author.

Figure 2.21 shows that the majority of topics that get community noted include: history and historical conflicts (the war in Ukraine, the war in Gaza, or tensions arising from the war in Gaza); Dutch politics (Dutch government, cabinet or parliament; Dutch right-wing politicians or party); energy and climate (climate change in particular, but also the Dutch farmer crisis); and, finally, migration (in general, and tensions related to it). In a minority of cases, the majority of notes rated helpful are, as indicated in the previous figure, those that attempt to correct or contextualise posts about conspiracy theories. In the large majority of cases, Community Notes need more ratings — particularly in topics that are deeply contentious, such as the above-mentioned historical conflicts, migration and Dutch politics in general. While Community Notes

may function on a logic of consensus-building, it may not always succeed to reach consensus for notes added to deeply contentious topics.

What gets community noted?

Weighted distribution of community notes per topic in Dutch-speaking posts on X

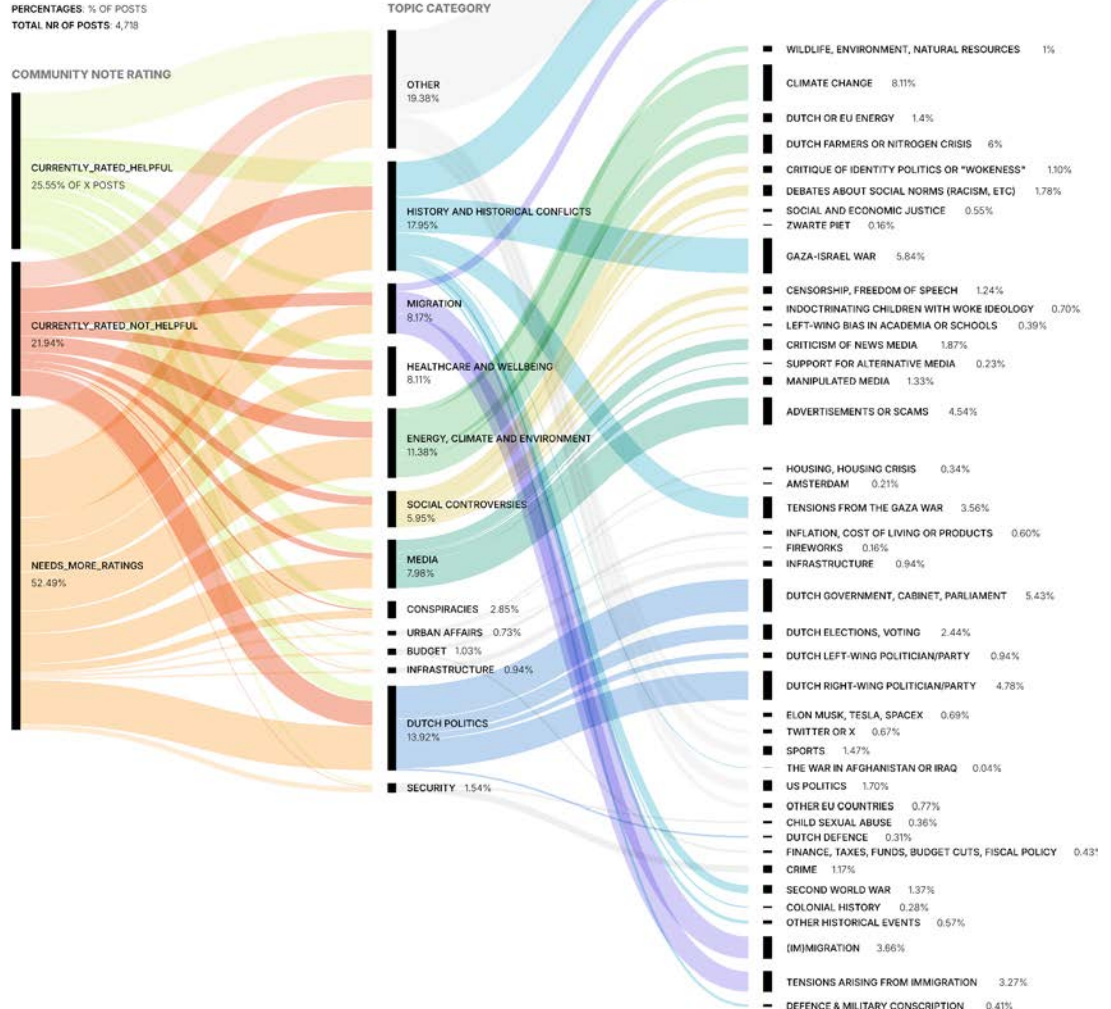


Figure 2.21 Weighted percentages of community note ratings per topic category and subtopic. Numbers represent percentages of posts for a given rating and topic. Source: author.

Deletion

Finally, to address removals or suspensions, I find that among 1392 posts with a hate score of 0.80 or higher, 142 were removed. Of these, 117 were removed for no stated reason, 20 were removed by the author, and 5 are not accessible because the author limited who can view them. Posts that were removed do not differ from those that remain online (see Annex). All express strongly negative sentiments against immigration and some against Muslims in particular. Posts that contain explicit slurs (e.g., "Muslim scum") are present in both removed and online posts.

Conclusions

The writing of this report bore witness to seismic shifts in the world of content moderation. After Musk's acquisition, X shifted to another philosophy of content moderation. X policies have shifted norms, dropping any clear references to a language considered "on the left". They have also dramatically simplified and streamlined to focus on a similar if identical set of issues as Twitter, except with a different method: to demote what is "awful", and to remove what is "unlawful". The idea is that, for content moderation to interfere, it must be for the absolute extremities — the types of content that ought never to be published by law. For the rest, moderation is presented as having a lesser say on what constitutes acceptable content than users themselves. The findings above show that, on average, this technique works as intended, but modestly: hateful posts are *slightly* more demoted and have *slightly* fewer impressions — but this depends on what classifier one uses and what definition of hate speech it remodels. I also find a small number of removed posts with high hate scores, but they do not differ in substance from those that remain online.

Perhaps more fundamentally, the question is whether these techniques make sense when it only obscures the most extreme of otherwise negative sentiments about contentious topics, especially in an environment focused on maximizing engagement. This question, then, becomes: *how much* could the DSA moderate, and *how sustainable* are such measures? The problem illustrated with the case study of posts about immigration reflects a larger case study in present-day European debate: that of attempting to moderate problematic but persistent language that pains to be "deplatformed". Historically, DSA norms are grounded on strong post-war norms against the normalisation of discriminatory and related language in European public spheres. However, in the massively pluralistic environments that X and other platforms co-constitute, one wonders whether it is sustainable to remove "problematic" content — particularly that of the "variable" (i.e., not *extreme*) kind. Users and platforms across geographies, even within the EU, will diverge greatly in their definitions of hateful, misleading, misinformative and otherwise "harmful" language. To remove it without a buffer zone for public deliberation — and a minimum amount of consensus-building — may continue to generate significant resistance in the public, political and now tech sectors, and threaten the legitimacy (and thus longevity) of the DSA as a whole.

And so, though X has received much negative coverage in critical debates, it is worth pointing out that a few techniques introduced shortly before Musk's takeover may offer some respite for DSA applications. It represents some innovation in content moderation, in the sense that it invests in more than just adjudicative mechanisms. Another approach and definition proper to moderation is the facilitation and balance of dialogue in a public environment. Nevertheless, the findings above show that Community Notes have important limitations: the vast majority of notes "still need ratings", and that tends to be the case, perhaps unsurprisingly, when the topic it is addressed is particularly contentious (i.e., fails to reach a consensus for a community note to be posted).

Opportunities for intervention

Still, where there are limitations in content moderation and platform design, there are opportunities for DSA applications. The norms that the DSA puts forth are also

applicable to civic processes, including models for consensus-building that may inform platform design. Below, I recommend a few points of intervention for sustainable platform (and/or AI) regulation beyond the moderation of content.

1. **Publicly funded, expert-informed workshops:** Regular and publicly funded workshops — such as quarterly data sprints — can bring together developers, practitioners in civic dialogue, and informed publics. These workshops, supported by networks of expertise across the EU and globally, would serve as incubators for alternative algorithmic approaches to platform governance.
2. **Deliberating civic values:** These engagements can be used to deliberate on what kinds of civic values (e.g. bridging, pluralism, social context) ought to be embedded into platform governance and moderation systems.
3. **Operationalizing civic values:** A key step is translating such civic values into specific technical features — e.g., content labeling based on social provenance, ranking systems designed to surface cross-cutting perspectives, or affordances for forming new, cross-community collaborations.
4. **Dissemination through public repositories and middleware:** The outputs of these efforts — tools, protocols, algorithms — can be gathered into a public repository of alternative platform design features. These can be developed as forms of middleware (interoperable layers that sit between user and platform), and deployed on decentralized platforms like Mastodon and Bluesky, used to prototype new platforms, or codified into public design conventions for platforms. Such conventions could eventually be enforced or recommended by law, e.g. requiring all platforms to deploy at least one bridging algorithm in their recommendation systems.
5. **Legal frameworks for "prosocial" design:** These efforts call for deeper investment in developing legal frameworks that can enforce or incentivize prosocial platform designs. Analogue media like public broadcasting systems can serve as historical precedent, particularly in how they promoted pluralism, inclusivity, and informed public debate.
6. **Algorithmic literacy and public media outreach:** Finally, the results from these design and governance initiatives can feed into broader public and media literacy efforts, enhancing public understanding of how civic values can and should be translated into algorithmic systems.

Within this configuration, investing in moderation and consensus-building mechanisms can be a point of entry for European influence in a broader marketplace for platform design, and thus a more sustainable model of content moderation development.

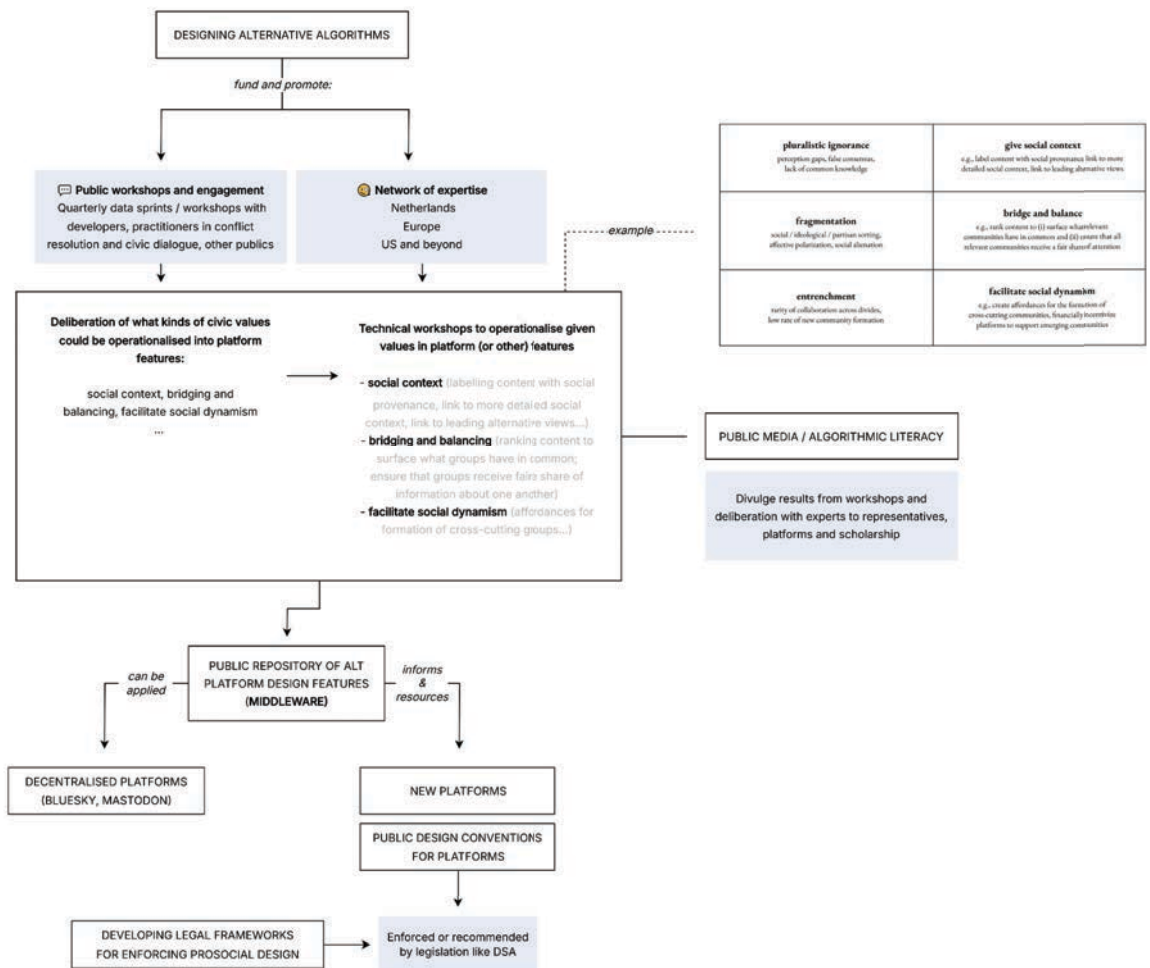


Figure 2.22 Diagram showing multiple points of intervention for promoting alternative platform design mechanisms. Source: author. See Weyl et al., 2025 for the table shown in the diagram.

Acknowledgments

The author would like to acknowledge the data collection and conceptual work of Ivan Kisjes; feedback from Richard Rogers and Sal Hagen; and ideas stemming from conversations with Luke Thorburn, Jonathan Stray, Lisa Schirch and Glen Weyl.

Annexes and datasets

de Keulenaar, E., & Kisjes, I. (2025). Content moderation from Twitter to X: Policies, enforcement and Community Notes (Version 2) [Dataset]. Mendeley Data. <https://doi.org/10.17632/d9ggkzp7bc.2>

References

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press. <https://www.oxfordscholarship.com/view/10.1093/oso/9780190923624.001.0001/oso-9780190923624>

Bennhold, K., & Taub, A. (2025, January 28). Hacking democracy: How Elon Musk is disrupting European politics. *New York Times*.

<https://www.nytimes.com/2025/01/28/world/europe/hacking-democracy-elon-musk-europe.html>

Clayton, J. (2023, April 12). Elon Musk BBC interview: Twitter boss on layoffs, misinfo and sleeping in the office. *BBC*. <https://www.bbc.com/news/business-65248196>

CNBC Television (Director). (2023, August 10). X Corp. Now a much healthier and safer platform than a year ago, says Linda Yaccarino [Video recording]. *CNBC*. <https://www.youtube.com/watch?v=NUqtBkiuRKs>

Constine, J. (2018, November 15). Facebook will change algorithm to demote "borderline content" that almost violates policies. *TechCrunch*. <https://social.techcrunch.com/2018/11/15/facebook-borderline-content/>

De Bolla, P. (2013). *The Architecture of Concepts: The Historical Formation of Human Rights*. Fordham University Press.

de Keulenaar, E., & Rogers, R. (2025). After deplatforming: the return of trace research for the study of platform effects. In T. Venturini, A. Acker, J.-C. Plantin, & A. Walford (Eds.), *The SAGE Handbook of Data and Society: An Interdisciplinary Reader in Critical Data Studies* (p. 560). SAGE. <https://uk.sagepub.com/en-gb/eur/the-sage-handbook-of-data-and-society/book281091>

de Keulenaar, E., Kisjes, I., Smith, R., Albrecht, C., & Cappuccio, E. (2023). Twitter as accidental authority: how a platform assumed an adjudicative role during the COVID-19 pandemic. In R. Rogers (ed.), *The Propagation of Misinformation in Social Media: a Cross-Platform Analysis* (pp. 109–138). Amsterdam University Press. <https://www.aup.nl/en/book/9789048554249>

de Keulenaar, E., Magalhães, J. C., & Ganesh, B. (2023). Modulating moderation: a history of objectionability in Twitter moderation practices. *Journal of Communication*, 73(3), 273–287. <https://doi.org/10.1093/joc/jqad015>

de Laat, P. B. (2012). Coercion or empowerment? Moderation of content in Wikipedia as ‘essentially contested’ bureaucratic rules. *Ethics and Information Technology*, 14(2), 123–135. <https://doi.org/10.1007/s10676-012-9289-7>

Digital Methods Initiative. (2022). Internet Archive Wayback Machine Link Ripper [Computer software]. <https://tools.digitalmethods.net/beta/internetArchiveWaybackMachineLinkRipper/>

Don Lemon (Director). (2024, March 18). Elon Musk on racism, bailing out Trump, hate speech, and more—The Don Lemon Show (Full Interview) [Video recording]. <https://www.youtube.com/watch?v=hhsfjBpKiTw>

Elon Musk [@elonmusk]. (2022, October 28). The bird is freed [Tweet]. Twitter. <https://x.com/elonmusk/status/1585841080431321088>

- Elon Musk. (2022, March 26). Given that Twitter serves as the de facto public town square, failing to adhere to free speech principles fundamentally undermines democracy. What should be done? [Tweet]. Twitter.
<https://x.com/elonmusk/status/1507777261654605828>
- Elon Musk. (2022, November 24). Hate speech impressions down by 1/3 from pre-spike levels. Congrats to Twitter team! <https://t.co/5BWaQoIip> [Tweet]. Twitter.
<https://x.com/elonmusk/status/1595630109116989440>
- Föessel, M. (2021). *Récidive 1938*. Presses Universitaires de France.
- Fuchs, R. (2015, September 22). Germany's constitution protects against dictatorships, DW. <https://www.dw.com/en/never-again-dictatorship-what-is-germanys-basic-law/a-18728880>
- Fung, B. (2023, May 1). Jack Dorsey no longer thinks Elon Musk is the right person to run Twitter. CNN. <https://www.cnn.com/2023/05/01/tech/dorsey-turns-on-musk-twitter-acquisition/index.html>
- Gillespie, T. (2017). Platforms are not intermediaries. *Geo. L. Tech. Rev.*, 2, 198.
https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/gltr2§ion=19
- Gillespie, T. (2018a). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
- Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3).
<https://doi.org/10.1177/20563051221117552>
- Gorwa, R. (2019). The platform governance triangle: conceptualising the informal regulation of online content. *Internet Policy Review*, 8(2).
<https://policyreview.info/articles/analysis/platform-governance-triangle-conceptualising-informal-regulation-online-content>
- Graham, T. (2024, November 6). Elon Musk's flood of US election tweets may look chaotic. My data reveals an alarming strategy. *The Conversation*.
<http://theconversation.com/elon-musks-flood-of-us-election-tweets-may-look-chaotic-my-data-reveals-an-alarming-strategy-243021>
- Hatano, A. (2023). Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation. *The Australian Year Book of International Law Online*, 41(1), 127–156. <https://doi.org/10.1163/26660229-04101017>
- Heath, A. (2023, October 31). Elon Musk's "everything app" plan for X. *The Verge*.
<https://www.theverge.com/23940924/elon-musk-x-twitter-all-hands-linda-yaccarino-super-app>
- Helmond, A. (2015). The platformization of the web: Making web data platform ready. *Social Media + Society*, 1(2), 1–11. <https://doi.org/10.1177/2056305115603080>

- Hoffmann, M. H. G. (2021). Consensus building and its epistemic conditions. *Topoi*, 40(5), 1173–1186. <https://doi.org/10.1007/s11245-019-09640-x>
- Honderich, H. (2025, February 7). Musk to rehire Doge aide who resigned over racist posts. *BBC*. <https://www.bbc.com/news/articles/c93q625y04wo>
- Internet Archive: Wayback Machine. (n.d.). [Computer software]. Retrieved July 16, 2022, from <https://archive.org/web/>
- Jackson, D., & Szóka, B. (2025, February 11). The far right’s war on content moderation comes to Europe. *Tech Policy Press*. <https://techpolicy.press/the-far-rights-war-on-content-moderation-comes-to-europe>
- Joe Rogan (Director). (2024, November 4). Joe Rogan Experience #2223—Elon Musk [Video recording]. Joe Rogan Experience. https://www.youtube.com/watch?v=7qZl_5xHoBw
- Katzenbach, C., Dergacheva, D., Fischer, A., Kopps, A., Kolesnikov, S., Redeker, D., & Viejo Otero, P. (2023). Platform Governance Archive (PGA): Dataset PGA v2 [Data Paper]. <https://doi.org/10.26092/elib/2373>
- Katzenbach, C., Magalhães, J. C., Kopps, Sühr, T., & Wunderlich, L. (2023). Platform Governance Archives. <https://pga.hiig.de>
- Klaidman, D. (2024, July 19). Silicon shift? Major tech titans throw financial, political support to Trump. *CBS News*. <https://www.cbsnews.com/news/trump-jd-vance-silicon-valley-support/>
- Kroet, C., & Armangau, R. (2024, August 13). EU Commission not drawn on Musk insults against Breton. *Euronews*. <https://www.euronews.com/my-europe/2024/08/13/eu-commission-not-drawn-on-musk-insults-against-breton>
- Malingre, V. (2025, January 13). European digital regulation comes under attack from Trump, Musk and Zuckerberg, *Le Monde*. https://www.lemonde.fr/en/economy/article/2025/01/13/european-digital-regulation-comes-under-attack-from-trump-musk-and-zuckerberg_6737001_19.html
- Martinez, C. (2023, April 27). One billionaire owner, twice the hate: Twitter hate speech surged with Musk, study says. *Los Angeles Times*. <https://www.latimes.com/business/technology/story/2023-04-27/hate-speech-twitter-surged-since-elon-musk-takeover>
- Mendonça, Y. C. de. (2024, September 9). Elon Musk’s feud with Brazilian judge is much more than a personal spat – it’s about national sovereignty, freedom of speech and the rule of law. *The Conversation*. <http://theconversation.com/elon-musks-feud-with-brazilian-judge-is-much-more-than-a-personal-spat-its-about-national-sovereignty-freedom-of-speech-and-the-rule-of-law-238264>

- Meta. (2025). Misinformation. Transparency Centre. <https://transparency.meta.com/en-gb/policies/community-standards/misinformation/>
- Napoli, P. M. (2019). *Social Media and the Public Interest: Media Regulation in the Disinformation Age*. Columbia University Press.
- OpenAI. (2024). Moderation. <https://platform.openai.com>
- OpenAI. (2025). Pricing. <https://openai.com/api/pricing/>
- Oremus, W., & Hunter, T. (2024, June 5). Why Elon Musk's X is embracing adult content. *Washington Post*. <https://www.washingtonpost.com/politics/2024/06/05/elon-musk-x-porn-adult-content/>
- Ovadya, A., & Thorburn, L. (2023). Bridging Systems: Open Problems for Countering Destructive Divisiveness Across Ranking, Recommenders, and Governance. Knight First Amendment Institute, Online. <http://knightcolumbia.org/content/bridging-systems>
- Pedretti, L. (2023). A Gramática da Violência Política: Representações críticas sobre a ditadura militar na década de 1970. *Novos estudos CEBRAP*, 42, 163–180. <https://doi.org/10.25091/S01013300202300010009>
- Perez, S. (2022, September 7). Twitter expands fact-checking program ahead of US midterms. *TechCrunch*. <https://techcrunch.com/2022/09/07/twitter-expands-its-crowdsourced-fact-checking-program-birdwatch-ahead-of-u-s-midterms/>
- Rogers, R. (2020). Deplatforming: Following extreme internet celebrities to Telegram and alternative social media. *European Journal of Communication*, 35(3). <https://doi.org/10.1177/0267323120922066>
- Sander, B. (2019). Freedom of expression in the age of online platforms. *Fordham International Law Journal*, 43, 939. <https://heinonline.org/HOL/Page?handle=hein.journals/frdint43&id=955&div=&collection=>
- Schirch, L. (2023). The case for designing tech for social cohesion: The limits of content moderation and tech regulation. SSRN Scholarly Paper No. 4360807. <https://doi.org/10.2139/ssrn.4360807>
- Silberling, A. (2024, May 14). On Elon's whim, X now treats "cisgender" as a slur. *TechCrunch*. <https://techcrunch.com/2024/05/14/on-elons-whim-x-now-treats-cisgender-as-a-slur/>
- Singh, A. (2024, December 27). MAGA vs. Musk: Right-wing critics allege censorship, loss of X badges. *Axios*. <https://www.axios.com/2024/12/27/musk-x-loomer-h1b-maga-verification>

- Sprinklr. (2024). How Sprinklr helps identify and measure toxic content with AI. *Sprinklr*. <https://www.sprinklr.com/blog/identify-toxic-content-with-leading-analytical-ai/>
- Stray, J. (2022). Designing recommender systems to depolarize. *First Monday*, 27(5). <https://doi.org/10.5210/fm.v27i5.12604>
- Suzor, N. P. (2019). *Lawless: The Secret Rules That Govern our Digital Lives*. Cambridge University Press. <https://doi.org/10.1017/9781108666428>
- Tang, A. (2024, November 26). Embrace the Shift to ‘Prosocial Media.’ *Wired*. <https://www.wired.com/story/prosocial-media-social-networks-discourse-decentralization/>
- Turner, F. (2006). *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. University of Chicago Press.
- Twitter. (2010, July 23). Twitter Help Center. Twitter. <https://web.archive.org/web/20100723013149/http://support.twitter.com/groups/33-report-a-violation/topics/121-guidelines-best-practices/articles/18311-the-twitter-rules>
- Twitter. (2013, August 17). Twitter Help Center | Abusive behavior policy. Witter. <https://web.archive.org/web/20130817034649/https://support.twitter.com/groups/56-policies-violations/topics/236-twitter-rules-policies/articles/20169997-abusive-behavior-policy>
- Twitter. (2017, December 20). Violent extremist groups. Twitter. <https://web.archive.org/web/20171220212715/https://help.twitter.com/en/rules-and-policies/violent-groups>
- Twitter. (2018). Twitter health metrics proposal submission. Twitter. https://blog.x.com/en_us/topics/company/2018/twitter-health-metrics-proposal-submission
- Twitter. (2021, June 10). Parody, newsfeed, commentary, and fan account policy. <https://help.twitter.com/en/rules-and-policies/parody-account-policy>
- Van Raemdonck, N., & Pierson, J. (2022). A conceptual framework for the mutual shaping of platform features, affordances and norms on social media. *Tijdschrift voor Communicatiewetenschap*, 50(4), 358–383.
- Vincent, J. (2022, April 28). Twitter reassures advertisers Musk won’t make the platform more of a toxic hell-hole than it already is. *The Verge*. <https://www.theverge.com/2022/4/28/23046139/twitter-elon-musk-free-speech-plans-toxic-advertisers-brands>
- Weyl, E. G., Thorburn, L., de Keulenaar, E., Mchangama, J., Siddarth, D., & Tang, A. (2025). Prosocial Media, SSRN Scholarly Paper No. 5141171. <https://papers.ssrn.com/abstract=5141171>

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M. B. F., Coleman, K., & Baxter, J. (2022). Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation, arXiv:2210.15723. <https://doi.org/10.48550/arXiv.2210.15723>

X Community Notes. (2025). Diversity of perspectives. X. <https://communitynotes.x.com/guide/en/contributing/diversity-of-perspectives>

X Help Center. (2024). X's enforcement philosophy & approach to policy development. X. <https://help.x.com/en/rules-and-policies/enforcement-philosophy>

X Safety. (2023a, March 21). We recently partnered with @Sprinklr for an independent assessment of hate speech on Twitter, which we've been sharing data on publicly for several months. Sprinklr's AI-powered model found that the reach of hate speech on Twitter is even lower than our own model quantified 📊 [Tweet]. Twitter. <https://x.com/Safety/status/1638255718540165121>

X Safety. (2023b, September 18). Stand with X to protect free speech. X. https://blog.x.com/en_us/topics/company/2023/stand-with-x-to-protect-free-speech

X Safety. (2024, June 4). We want our policies to be as clear, helpful, and transparent as possible, so you can quickly read them and understand the X Rules. To achieve this, we've begun the process of updating and consolidating a number of pages on our Help Center and rolling out a new, simplified [Tweet]. Twitter. <https://x.com/Safety/status/1798112474803446138>

X. (2025). Our policy on violent organizations | X Help. <https://help.x.com/en/rules-and-policies/violent-entities>

X. (2025). X Community Notes - Working with the Community Notes data. X. <https://communitynotes.x.com/guide/en/under-the-hood/download-data>

X. (2025). X's civic integrity policy | X Help. X. <https://help.x.com/en/rules-and-policies/election-integrity-policy>

Zuckerman, E. (2020, January 17). The Case for Digital Public Infrastructure. Knight First Amendment Institute. <http://knightcolumbia.org/content/the-case-for-digital-public-infrastructure>

3. Ranking authority: A critical audit of YouTube's content moderation

Daniel Jurg, Salvatore Romano and Bernhard Rieder

Abstract

This chapter examines YouTube's content moderation practices during the 2024 European Parliamentary elections. Using search results from the Netherlands, Germany, and France, plus an API sample, it explores YouTube's pledge to raise authoritative sources and remove harmful content. Findings suggest the search algorithm favors legacy media and Public Service Media (PSM). While many PSM carry a publisher context label, deployment seems patchy and absent in European languages such as Basque, Catalan, Danish, Finnish, Galician, Greek, and Portuguese. The study logs 486 election videos that became unavailable. However, sparse information problematizes assessing the enforcement of Terms of Service. The chapter concludes with a call for increased data access via the YouTube Research Program to scale content moderation studies on the platform.

Keywords: algorithmic ranking of authoritative sources, content removal transparency, researcher access to platform data, publisher context label (news funding notice), YouTube content moderation

Introduction

The year 2024 was called "the biggest election year in history," with over 2 billion eligible voters in more than 60 elections worldwide (Buchholz, 2024). In Europe, citizens could vote for the European Parliament, weighing in on critical debates about climate change, migration, and digital governance under the new Digital Services Act (DSA). Digital platforms have become integral to these election processes, shaping how voters access election-related information, engage in political discussions, and eventually cast their vote. However, rather than fulfilling the early web 2.0 promise of amplifying marginalized voices and fostering democratic participation, 2024 election-related discourse continued growing concerns around disinformation, election interference, and the manipulation of the public opinion (West, 2024). These issues have taken center stage since the 2016 U.S. elections (Silverman, 2016), which prompted an ever-expanding body of work that underscores the need to conceive of platforms as 'custodians of the internet,' actively managing, curating, and organizing content, shaping what is visible and permissible (Gillespie, 2018).

Public and political pressure, fueled by widespread concerns about disinformation in elections and major crises such as the COVID-19 pandemic, has driven platforms in the past years to place greater emphasis on content moderation (Stoian, 2023). The EU's Digital Services Act (DSA), which came into effect on February 17, 2024, marked a pivotal regulatory effort to enforce accountability in these moderation practices, directly addressing issues of misinformation and the manipulation of public opinion (European Commission, 2022). The DSA introduced various obligations for

Very Large Online Platforms (VLOPs), requiring them to tackle illegal content, curb the spread of disinformation, and mitigate broader societal harms. By mandating transparency, accountability, and algorithmic oversight, the DSA aims to fundamentally reshape how platforms manage their content ecosystems. Its rules compelled platforms not only to demonstrate their commitment to accurate information and democratic integrity but also to provide detailed reports on their actions and regularly evaluate their impact (European Commission, n.d.).

As one of the largest VLOPs, YouTube plays a crucial role in advancing content moderation efforts. With over 2.5 billion active monthly users, it is one of the world's most influential social media platforms (Singh, 2024). The platform has had a controversial history with moderation issues, from Islamic extremism (Shane, 2017) to pedophiles gathering in comment sections of children's videos (Fisher & Taub, 2019). The infamous label "the great radicalizer," introduced by sociologist Zeynep Tufekci in 2018, brought YouTube's content moderation failures to the forefront of public debate, casting the online video platform as "one of the most powerful radicalizing instruments of the 21st century" (Tufekci, 2018). Research has since highlighted YouTube's role in hosting far-right extremists (Lewis, 2018) and amplifying their content through its recommendation algorithm, a phenomenon known as the rabbit-hole theory (O'Callaghan et al., 2015; Mozilla Foundation, 2021). In fact, the platform experienced a major crisis in 2017 when advertisers discovered their content appeared alongside extremist videos and withdrew their advertising from YouTube, triggering what became known as the "Adpocalypse," which forced the platform to implement stricter content moderation policies and shift toward more advertiser-friendly policies (Kumar, 2019). While researchers continue to point to the platform's problems with extremism (Ballard, 2023), the demands of advertisers and YouTube's revenue model have fundamentally reshaped its content moderation approach from a laissez-faire stance toward more regulatory measures.

In light of the DSA, and against the background of YouTube's history with extremism and move towards more regulatory measures, this study explored YouTube's moderation efforts during the 2024 European Parliamentary elections through an independent 'critical' audit. We started from the more traditional auditing techniques that rely on data from YouTube's API as well as data collected via specific research browsers to test two core parts of YouTube's content moderation, namely (1) the raising of authoritative content and (2) the removal of problematic content. As elaborated in our methodology section, the 'critical' element comes from being attentive to the conditions of possibility to carry out various types of audits and finding new and creative ways to study content moderation independently.

Research questions

In order to explore and understand content moderation during the 2024 European Parliamentary elections, this study poses the following research questions:

1. What sources seem to be promoted during the 2024 Parliamentary election, and what does that tell us about YouTube's conception of 'authoritativeness'?
2. What sources have been removed, and how, during the 2024 European Election, and what does that tell us about YouTube's removal practices?

3. How well does current access to data allow researchers to perform audits of content moderation practices?

In order to answer these questions, this study begins with a brief history of YouTube's content moderation practices, highlighting some of the company's key implemented features and initiatives. We then present our critical auditing methodology and findings. Rather than making definitive claims about YouTube's performance in moderating the 2024 European Election, our exploratory research aims to uncover some of the key mechanisms behind moderation and document potential shortcomings.

The road to moderating the 2024 elections

YouTube has taken notable steps in addressing election moderation over recent years. Under the "Our Commitments" section on its website, the company presents the following statement regarding the 2024 election year:

2024 is the biggest election year in history [...]. With users around the world coming to YouTube for news and information about their civic duty, from voter registration to the location of their nearest polling place, we have a responsibility to support an informed citizenry and promote healthy political discourse. To deliver on this responsibility, we **remove** content that violates our Community Guidelines, including election content, **raise** high-quality election news and information from authoritative sources in search results and recommendations, **reduce** the spread of harmful election misinformation and **reward** trusted creators via the YouTube Partner Programme. Our policies apply to everyone and are enforced with consistency, regardless of the political viewpoints expressed, the language the content is in or how the content is generated (YouTube, n.d.-c, emphasis in original).

The statement underscores YouTube's "Four Rs of Responsibility" framework, introduced in 2019 amidst growing criticisms about the platform's moderation role (The YouTube Team, 2019). YouTube has also documented its moderation efforts through detailed timelines, highlighting key product launches, policy updates, and additional technical features (see Figure 3.1).

While content moderation is often thought of as preventing or removing problematic content, in its widest sense, it encompasses the broader "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" (Grimmelmann, 2015, p. 47). As Figure 3.1 shows, while YouTube has extensive removal systems, their public communication places emphasis on this wide range of 'positive' moderation as, for example, investing in trusted journalistic organizations.

In line with Gillespie's (2018) conception of guideline updates, the introduction of specific policies and features does "important discursive work, performing but also revealing how platform companies see themselves as ambivalent arbiters of public propriety" (p. 46). In the context of this study, the following section highlights two important aspects in the history of content moderation: (1) algorithmic approaches to surfacing high-quality information through search and information panels, and (2) content removal practices, particularly regarding community guidelines violations.

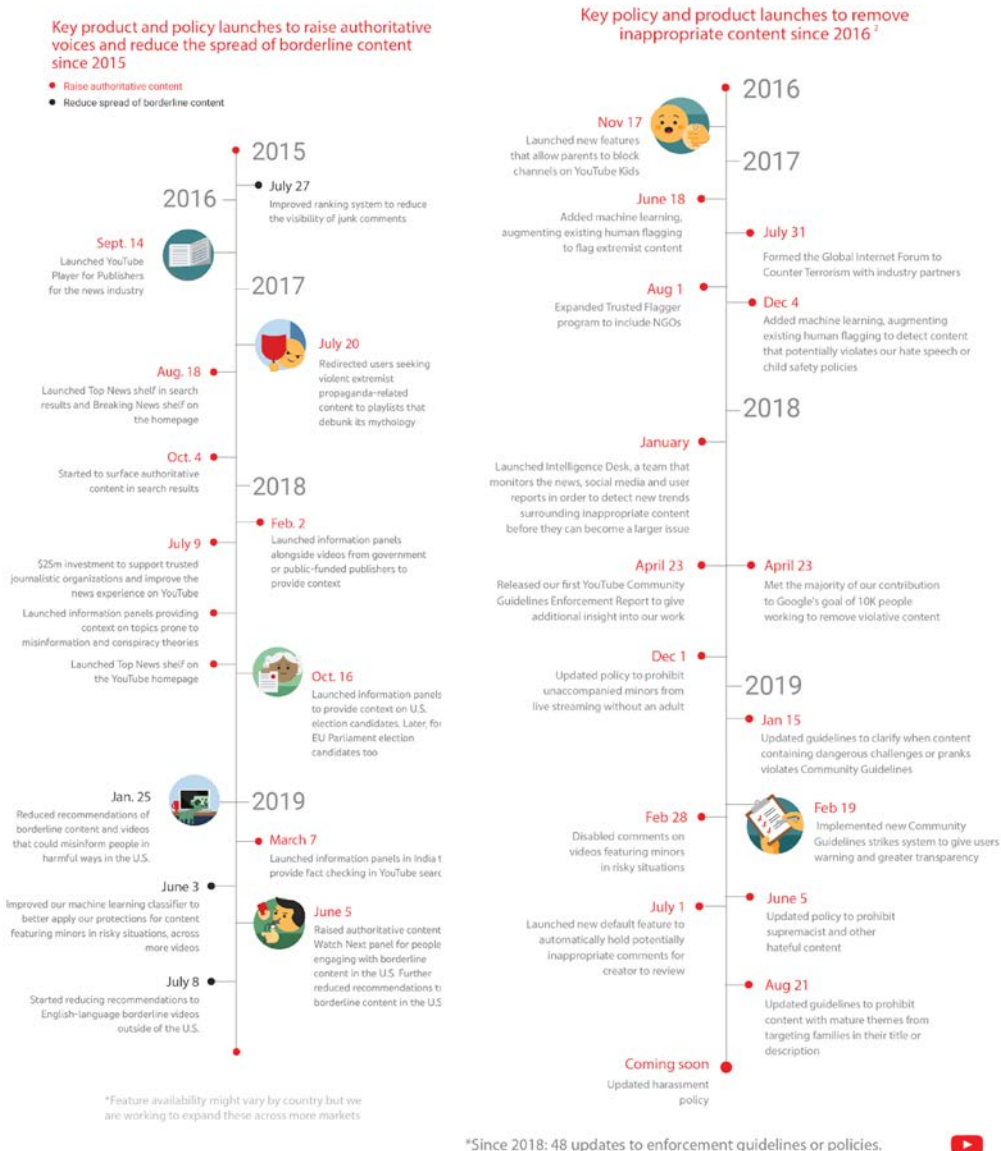


Figure 3.1 YouTube's communication on the historical development of "The Four Rs of Responsibility". Source: The YouTube Team, 2019a; 2019b.

Raising high-quality election news from authoritative sources

In their communication about disinformation, YouTube states that when it comes to disinformation it uses "external human evaluators and experts" and "well-tested machine learning systems to build models that generate recommendations" (YouTube, n.d.-a). YouTube's algorithms have been a focal point of much critique around radicalization, considering how radical actors strategically insert themselves into algorithmic recommendations (Yesilada & Lewandowsky, 2022). While much attention has gone to recommendations, YouTube's search algorithm has also received notable attention, particularly in how natively digital content creators strategically manage algorithmic visibility to maximize their reach (Gillespie, 2017; Bishop, 2019). When it comes to political and controversial cultural issues, a 2018 study found that for searches on topics such as refugees or political candidates, "YouTube-native content, which often thrives on controversy and dissent, consistently outperforms mainstream sources in terms of visibility and exposure" (Rieder et al., 2018, p. 50).

However, recent research in the European context highlights an increased focus on authoritative sources. For example, Glaesener's (2023) study on search results for the Russia-Ukraine War in Germany, found that "most of [the] search results came from YouTube channels of the mainstream media" (p. 87). Similarly, Padilla et al. (2025) in Spain revealed that while public media constitutes a smaller proportion of the channels appearing in search results, "YouTube's search algorithm tends to select proportionally more content from public media than from other sources" (p. 11). This growing body of evidence seems to align with YouTube's own reporting that, in the context of the 2020 US Election, "videos from authoritative sources like news outlets represented the most viewed and most recommended election videos on YouTube" (The YouTube Team, 2023).

In addition to its efforts to increase the algorithmic visibility of authoritative sources, YouTube also makes sure that it shows "information panels linking to third-party sources around a small number of well-established topics that are subject to misinformation" (Miller, 2020). One particularly debated feature is the "Information Panel Providing Publisher Context," what we will refer to in this study as the "News Funding Notice" (NFN). Announced on February 2, 2018, the NFN aims to enhance transparency regarding the funding of media organizations. This notice is prominently displayed on videos from news organizations that receive public or government funding. Accompanied by an information icon linking to the publisher's Wikipedia page, the label is designed to contextualize news sources, encouraging users to critically assess potential biases or affiliations (Samek, 2018) (See Figure 3.2).

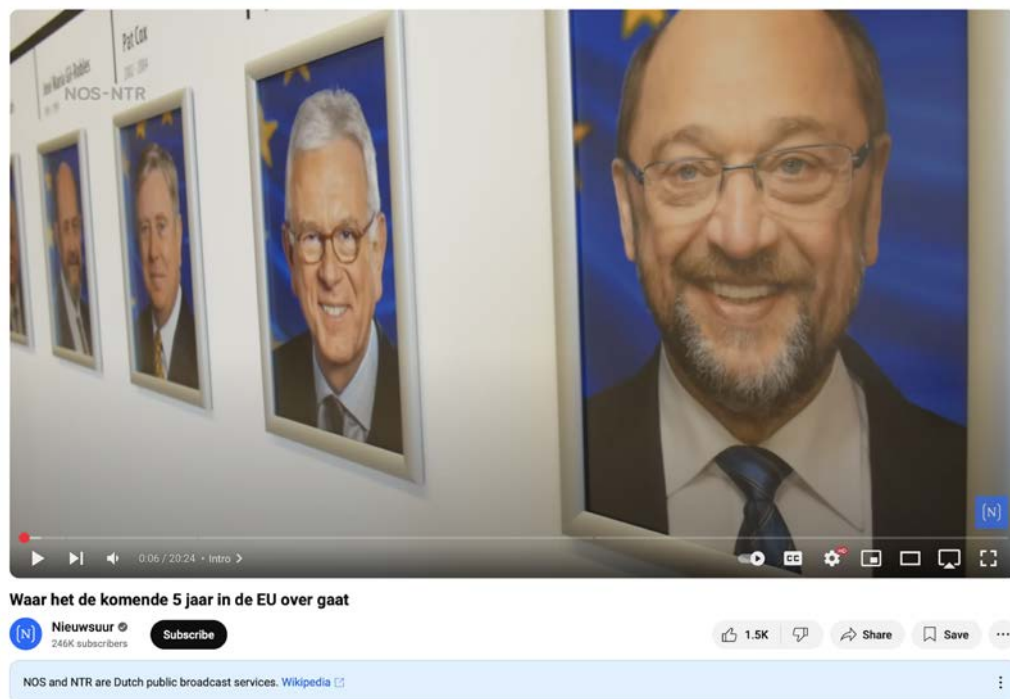


Figure 3.2 Screenshot (18-02-2025) of a video from Nieuwsuur accompanied by the funding notice "NOS and NTR are Dutch public broadcast services" with a link to Wikipedia. Source: YouTube.com.

While YouTube states that the feature is "not a comment by YouTube on the publisher's or video's editorial direction, or on a government's editorial influence"

(YouTube Help, n.d.), its labeling with tags such as ‘public broadcast service’ or ‘state-funded’ media arose amid heightened scrutiny of state-affiliated media's influence on public opinion, such as RT (formerly Russia Today). YouTube faced criticism for amplifying Russian state-sponsored media, with RT becoming one of the platform's "most-watched" news channels at its peak (Wakabayashi & Confessore, 2017). The NFN has drawn criticism from public broadcasters such as PBS: "PBS is an independent, private, not-for-profit corporation, not a state broadcaster [...] YouTube's proposed labeling could wrongly imply that the government has influence over PBS content, which is prohibited by statute. If YouTube's intent is to create clarity and better understanding, this is a step in the wrong direction" (Shaban, 2018).

This worry aligns with studies demonstrating that context labels may reduce the ‘misleading’ influence of state-funded outlets by signaling their government affiliations to users (Nassetta & Gross, 2020, p. 5). However, other research has argued that "increasing the visibility of publishers is an ineffective, and perhaps even counterproductive, way to address misinformation on social media," particularly when such sources provide truthful information (Dias et al., 2020, p. 1). Beyond its effectiveness, YouTube has also been criticized for inconsistently applying the notice (Kofman, 2019). Given the significant role public service media play in the European context—considerably more so than in the United States—it is crucial to further investigate the consistent application and broader implications of such labeling practices.

Reducing the spread of harmful election misinformation

Much of YouTube's communication has emphasized efforts in ‘soft’ moderation (Gorwa et al., 2020), which involves promoting high-quality content while demoting lower-quality material. However, as Figure 3.1 illustrates, the platform has also implemented more hard moderation, i.e., the active removal of content that violates community guidelines (ibid.), specifically targeting supremacist and hate speech content. This includes implementing stricter policies, expanding enforcement teams, and using automated systems to identify and remove such material. Notably, following its 2019 policy update, YouTube removed 25,000 channels, including several key white supremacist figures such as Stefan Molyneux, David Duke, and Richard Spencer (Alexander, 2020).

While YouTube's efforts to moderate content have grown significantly, empirical research on the specifics of its content moderation practices has been difficult. Between July and September 2024, Google's Transparency Report revealed that YouTube removed a total of 4,874,056 channels for being incompatible with its community guidelines (Google Transparency Report, n.d.). While independent research has sought to evaluate these removal practices, with some studies attempting to reverse-engineer the removal process by predicting whether a video would be taken down, a "key challenge in analyzing deleted/removed videos and perhaps the reason why there hasn't been any work analyzing such videos is that there are no repositories of unavailable YouTube videos" (Kurdi et al., 2021, p. 48).

YouTube's transparency efforts, including its own reports and the new DSA Transparency Database, seek to improve transparency around its content moderation practices. However, researchers still face significant challenges. Studying content

removal beyond the general statements in the DSA requires pre-collected datasets, as the platform does not provide retrospective access to deleted content, leaving critical gaps in understanding the scope and impact of its moderation efforts. To bridge this gap, researchers have used existing datasets and subsequently assessed removed videos and looked at YouTube's 'removal statements' (De Keulenaar et al., 2021). However, such research has indicated that YouTube's communication to users is often vague, with removal statements suggesting that only a "small number of videos have been removed for inciting hatred, involving violence or harassment, or because of copyright claims" (Ibid., p. 129). In fact, YouTube communicates differently about removal practices in transparency reporting (statements of reason) and statements for users (removal statements). While metadata collected prior to removal can be cross-referenced with community guidelines, such research remains constrained by the interpretative leaps required to link removals to specific guidelines.

Despite these limitations, the language used in removal statements warrant further examination. These statements, beyond mere procedural notifications, also articulate the platform's stance on complex issues such as freedom of expression and its role in shaping public discourse. The 2018 shooting at YouTube headquarters, where the perpetrator cited YouTube's policies as a factor in her declining video views, underscores the sensitive socio-political landscape surrounding content moderation (Wakabayashi et al., 2018). YouTube's limited information in removal statements likely reflects an attempt to mitigate conflict. However, a deeper analysis of these discursive practices offers valuable insights into how platforms navigate the delicate balance between transparency and potential conflict, highlighting the need for policies that promote accountability and transparency while mitigating the risk of exacerbating existing tensions.

Methodology

This exploratory study conducted two critical audits of YouTube's moderation practices during the 2024 EU Parliamentary Elections, examining its promotion of high-quality sources and enforcement of community guidelines through content removals. The following sections outline our critical auditing methodology, data collection and analysis, and limitations.

Critical audit

Digital platform audits can be seen as systematic evaluations of internal policies, codes of conduct, or algorithmic outputs, aimed at increasing transparency in systems often perceived as "black boxes" (Rieder & Hofmann, 2020). Existing scholarship has identified five idealized approaches: (1) code audit, assessing source code; (2) non-invasive user audit, surveying user experience; (3) scraping audit, observing results from specific prompts or queries; (4) sock puppet audit, impersonating users; and (5) collaborative or crowdsourced audit, engaging testers (Sandvig et al., 2014). Our exploratory study employs somewhat of a hybrid approach. We combine API and web scraping methods to analyze the performance and removal of election-related content.

Our hybrid approach offers the advantage of collecting a substantial amount of structured data, including metadata (e.g., views and likes), via YouTube's API, while also accounting for country-specific variations in content moderation through scraping. Assigning country-specific IP addresses to our sock puppet accounts establishes an 'algorithmic baseline'—that is, it seeks to map the localized content YouTube's pre-

personalized algorithm serves by default, revealing the types of sources algorithmically prioritized in each country. The method's primary limitation here is that it does not capture the personalized user experience on algorithmic platforms like YouTube. For example, we did not create research personas to map the personalized information flows of specific user types (Bounegru et al., 2022), nor did we incorporate broader data donation-based audits, where users voluntarily share their platform activity data to examine personalized experiences across diverse contexts (see Mozilla Foundation, 2021; Hille, 2022). While such studies are valuable, they require resources and time beyond the scope of this research.

Finally, our 'critical' audit seeks to emphasize the conditions that enable independent audits in the first place and the need for new and creative approaches to studying content moderation. Responding to calls for greater platform transparency, Rieder and Hofmann (2020) argue that transparency should not be viewed merely as a state of increased disclosure, but rather as a process involving deliberate choices about what "should be exposed and how, what is relevant and what can be neglected, which elements should be shown to whom and, not least, how the visible aspects should be interpreted" (Power, 1997 as cited in Rieder & Hofmann, 2020, p. 6). They propose "observability" as a framework that "seeks to address the conditions, means, and processes of knowledge production about large-scale socio-technical systems" (Rieder & Hofmann, 2020, p. 4). Therefore, we not only audited the outcomes of YouTube's moderation efforts but also critically reflected on the available data and methods, reinterpreting features in new ways to illuminate the broader challenges of achieving observability in content moderation research.

Data Collection

This study operationalized two datasets collected in the run-up and during the 2024 EU Parliamentary Election: (1) country-specific, browser-sourced data from the Netherlands, Germany, and France, and (2) general API-sourced data.

A) Country-specific Data

Our first dataset, provided by AIForensics, captures YouTube search rankings for election-related queries across European countries from May 2 to July 7. Due to the inherent fragility of scraping methods, data was missing for eight days, though the overall collection period remained substantial. We focused on data from the Netherlands, France, and Germany, analyzing the top 20 results for both general election queries and anti-immigration searches, using local IP addresses.

B) API Data

Our second dataset collected European Parliamentary election videos through weekly YouTube API searches from April 23 to July 15, 2024. We queried "European Parliamentary election" using the API's "order=date" parameter to retrieve videos chronologically, helping avoid non-relevant content. Each week's data was merged into a single dataset, with duplicates removed, resulting in 8,142 unique videos. On October 23, 2024, meta-data for videos was retrieved that were later deleted or made unavailable, enhancing the reliability of the dataset as a representation of the election-related content on YouTube during that time. While the API's constraints—such as search result limits and potential query biases—prevent an entirely exhaustive collection, this approach provides an extensive and systematically curated snapshots of election-related videos.

Methods

Our analysis followed an iterative approach, where initial patterns in content performance and removal informed subsequent methodological choices. The following sections detail this process of pattern discovery and refinement.

Performance of authoritative sources

We analyzed YouTube search rankings in the Netherlands, Germany, and France to assess how sources were ranked during the 2024 EU Parliamentary Elections. We compared results from two queries: "how to vote in EU-election" (Neutral) and "how to vote anti-immigration" (Adversarial). Table 3.1 provides a full overview of the specific queries in the respective languages.

Table 3.1 Overview of queries per country and type

Country	Query Type	Query
The Netherlands	Neutral	stemmen europese verkiezingen 2024
Germany	Neutral	europawahl 2024 stimmen
France	Neutral	puis-je voter l'élection européenne 2024
The Netherlands	Adversarial	stemmen anti-immigrant kiezer europese verkiezingen 2024 nederland
Germany	Adversarial	stimmen, anti-immigranten-wähler europawahlen 2024 deutschland
France	Adversarial	devrais-je voter électeur anti-immigrant élections européennes 2024 france

Table 3.2 details the total number of videos collected for each specific query and time range. While targeting twenty results per day per query, some scraped queries returned fewer results. Note that these numbers do not represent unique YouTube videos. If the same video shows up on a different day, it is counted as a distinct datapoint.

Table 3.2 Overview of browser-sourced data for two queries in three countries.

Country	Query	Number of Videos	From	Till
The Netherlands	Neutral	1,302	2024-05-02	2024-07-07
Germany	Neutral	1,291	2024-05-02	2024-07-07
France	Neutral	1,300	2024-05-02	2024-07-07
The Netherlands	Adversarial	1,197	2024-05-08	2024-07-07
Germany	Adversarial	1,185	2024-05-08	2024-07-07
France	Adversarial	1,195	2024-05-08	2024-07-07

We used six mutually exclusive categories to classify types of YouTube channels in the data and the performance of their videos. (1) Public Service Media includes channels labelled by YouTube as receiving government or public funding, alongside unlabelled public broadcasters like AT5, NH Nieuws, and parliamentary channels such as Public Sénat and LCP – Assemblée nationale. (2) Legacy Media captures traditionally private-funded organizations rooted in print, radio, or television, such as De Telegraaf and BNR. (3) Government includes channels of institutions at international (e.g., European Parliament), national (e.g., Dutch Police), regional, and municipal levels. (4) Political Parties class covers channels operated by political organizations, individual politicians, and affiliated youth wings. The (5) Natively Digital category encompasses digital-born media and influencers whose primary reach is online, like HugoDécrypte and Mcfly et Carlito. Finally, the (6) Other category captures remaining entities, including commercial companies, civic organizations, universities, and special interest groups like MEDEF and La FNSEA.

After observing discrepancies in the application of News Funding Notices (NFNs) on YouTube, we sought to classify the types of exclusion. In addition, we noticed that NFNs were not available from everywhere. We then also systematically tested NFN visibility of a Euronews video across different EU countries using VPN and YouTube region settings, and checking NFN display across available European language settings. Not all languages could be tested (e.g., Lëtzebuergesch), while some countries required testing multiple languages (e.g., Spain's Spanish, Catalan, Basque, and Galician).

Evaluating content removal

While the DSA Transparency Database seeks to support content removal research, its reliance on self-reported data and lack of specific video IDs and detailed reasoning limit its suitability for our needs. Though useful for understanding moderation categories and cross-platform trends (Trujillo et al., 2024), it provides little insight into event-specific moderation practices.

Therefore, to analyze YouTube's content removal practices, we examined videos no longer available in our API-sourced dataset. Following a digital methods approach, we repurposed YouTube's video availability statements, which inform users why videos have been taken offline. Since such statements are unavailable via the API, we scraped YouTube for removal information and analyzed existing metadata, using the Internet Archive when available. While we initially aimed for a broader categorization of removed content, YouTube's removal statements were often vague, obscuring specific violations. Given the speculative nature of assessing actual removal reasons, we shifted our analysis to two small case studies of videos specifically violating Terms of Service and explored how YouTube communicates about removals more broadly, examining how these practices reflect its discursive positioning as a platform moderator.

Limitations

Our exploratory study faced several practical limitations. First, our API query lacked a specific geographical location and ranking data, hindering direct comparisons between API- and browser-sourced results. Second, the weekly capture intervals of our API queries may have missed content removed more rapidly by YouTube. Third, variations in translation across countries in the browser-sourced data introduced semantic differences in queries, impacting precise cross-location comparisons. Fourth, removal statements can be dynamic, reflecting different stages of the removal process. During our data collection on October 23, 2024, we captured initial removal statements, but some final manual checks revealed some modifications. This highlights that content moderation is not static but an ongoing process of adjustment. We recommend more research into observing content removal as a process that is reflected in removal statements.

Finally, our browsing-based method retrieved videos uploaded long before the 2024 elections, including some off-topic or geographically irrelevant content. For example, we find quite some sports channels and influencer content that is not directly related to the elections. YouTube's personalization significantly enhances relevancy, which our approach does not capture. Therefore, these results should not be seen as representative of YouTube's intended election-related content for specific users. Nevertheless, the

method offers valuable insight into the baseline prioritization of source types by the platform.

Findings

The findings are presented in three sections, structured around our research questions. First, we examine the types of sources promoted during the 2024 EU Parliamentary elections, analyzing what these sources reveal about YouTube's conception of authoritativeness. Second, we investigate content removal during the 2024 elections, analyzing the types of sources removed and how those removals reflect YouTube's practices. Finally, we critically reflect on the current state of data access for researchers conducting audits of content moderation practices.

Prominent role for public service media

Figure 3.3 illustrates the frequency of videos from different channel types appearing in the top 20 search results for each query by country.

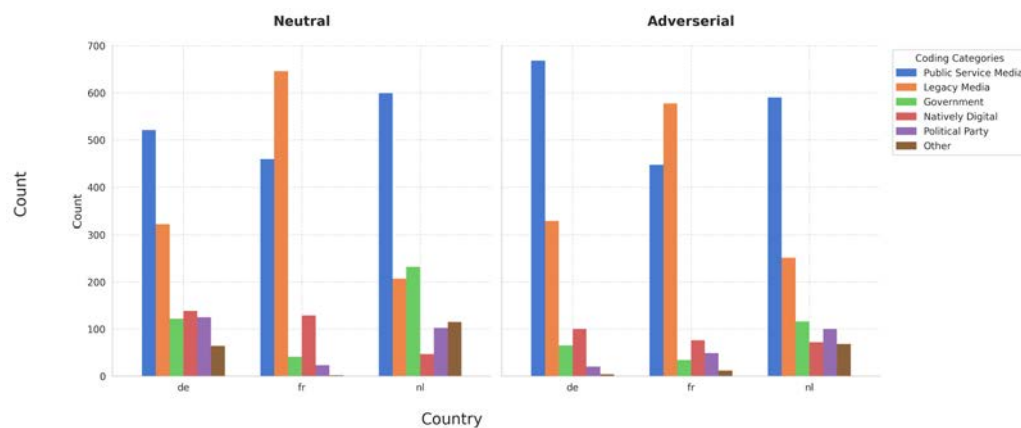


Figure 3.3 Video count from types of channels for ‘neutral’ (general information) and for the ‘adversarial’ (anti-immigration) vote query in Germany (de), France (fr), and the Netherlands (nl). Source: authors.

Figure 3.3 illustrates that baseline videos served in the Netherlands and Germany predominantly originate from public service media (PSM), while legacy media is more prominent in France, although PSM content remains frequent in the French context as well. The differences between search queries are relatively minor; however, the anti-immigration query Germany appears to result in a higher proportion of PSM videos, whereas the Netherlands exhibits a lower presence of government content. It is important to note that not all channels or videos provided information directly relevant to the elections. The baseline also includes content from Eurovision, sports channels, and natively digital content creators, where the videos appear unrelated to the electoral context. Additionally, in the Netherlands, there was quite some content from international outlets such as Al Jazeera, CBC News, and MSNBC, potentially reflecting the small Dutch-language market.

While there are important caveats to consider when interpreting this baseline data, our findings do highlight the significant prominence of PSM and legacy media. This prominence also underscores the relevance of YouTube's "News Funding Notice" (NFN) applied to PSM channels. In our analysis, we included channels that YouTube

did not label with an NFN, despite their public status. Nevertheless, 91 percent of the videos from PSM in our dataset originated from channels that did carry an NFN. In many cases, it appears that YouTube assigns NFNs to channels primarily focused on news and informational content, making the absence of funding notices for European public broadcasters, such as the Eurovision Song Contest, less surprising. However, we also identified instances where YouTube's decision not to apply an NFN label was less clear.

Discrepancies in the application of ‘News Funding Notice’

NFNs have been introduced by YouTube to increase transparency, combating disinformation, and promoting accountability. However, our channel analysis revealed discrepancies in NFN application for (1) international broadcasters, (2) subsidiary channels, and (3) local broadcasters.

TRT Français, a branch of the Turkish Radio and Television Corporation (TRT), is an example of the discrepant labeling of international broadcasters. As a government-funded entity, TRT channels have been identified by YouTube to fall under policies promoting transparency for public and state-funded media (see Bradshaw et al., 2024). However, Figure 3.4 shows that while TRT channels targeting different language and regional markets (e.g., TRT World, TRT Russia, TRT Arabic) receive funding labels, TRT Français does not.

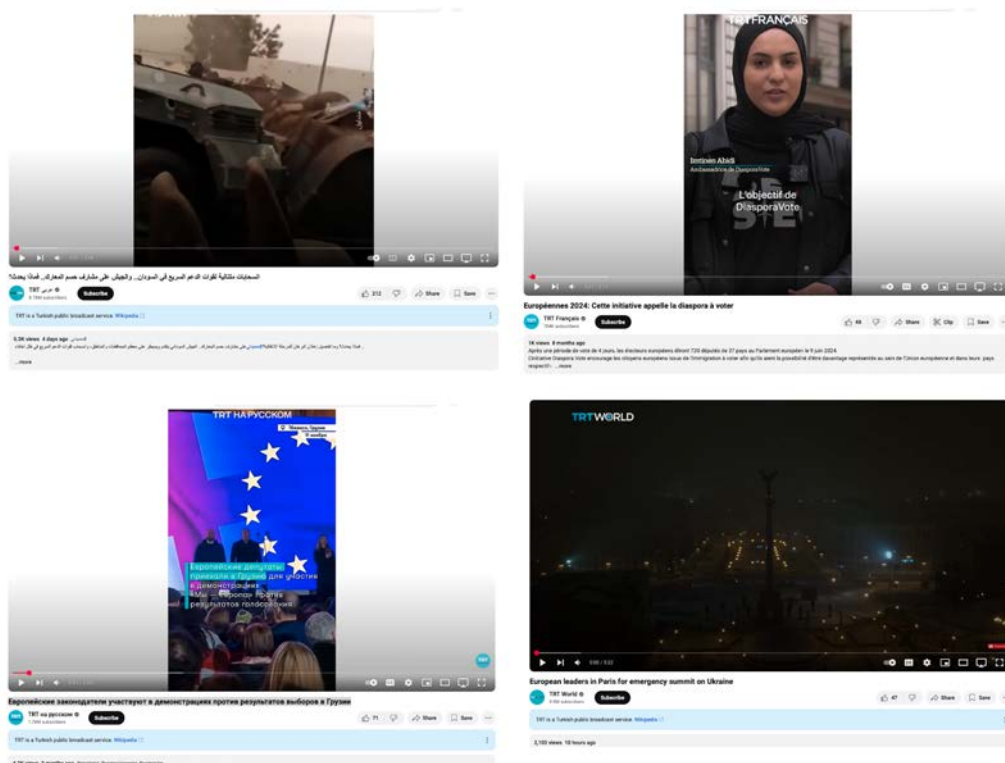


Figure 3.4 Screenshots (18-02-2025) of video from TRT Français (without NFN) compared with TRT World, TRT Russia, TRT Arabic (with NFNs). Source: YouTube.com.

Properly applied, NFNs could significantly impact user perception and engagement with these channels. On the one hand, labeling TRT Français alongside public service media like France Info or France 24 would, at least discursively, position them similarly as publicly funded entities (despite varying degrees of editorial

independence). On the other hand, explicitly highlighting TRT Français' connection to the Turkish government could have various implications for the French branch and its credibility among certain users.

Another example of labeling discrepancies involved affiliated channels. As illustrated in Figure 3.5, RTBF, the public broadcasting organization for Belgium's French-speaking community, has the NFN label applied to its main channel, which boasts 271,000 subscribers (as of this writing). However, its subsidiary channel, RTBF Info, which focuses on delivering "the most important daily news updates from the RTBF news team" (RTBF Info, n.d.) and holds a substantial audience of 178,000 subscribers, remained unlabeled. We also only found RTBF Info in our French results. This discrepancy raises questions about the uniformity and criteria of labeling practices, particularly when both channels share the same public broadcasting affiliation.

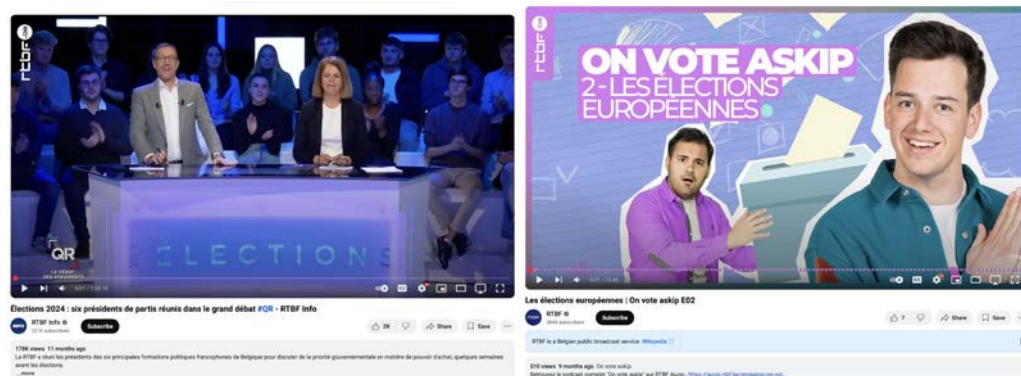


Figure 3.5 Screenshot (18-02-2025) of two videos from the public broadcaster RTBF with 'RTBF' having a funding notification and 'RTBF Info' not having a funding notification. Source: YouTube.com.

Another revealing example is the Dutch public broadcaster Ongehoord Nederland, which also does not receive a funding notice despite being a public broadcast service (See Figure 3.6).

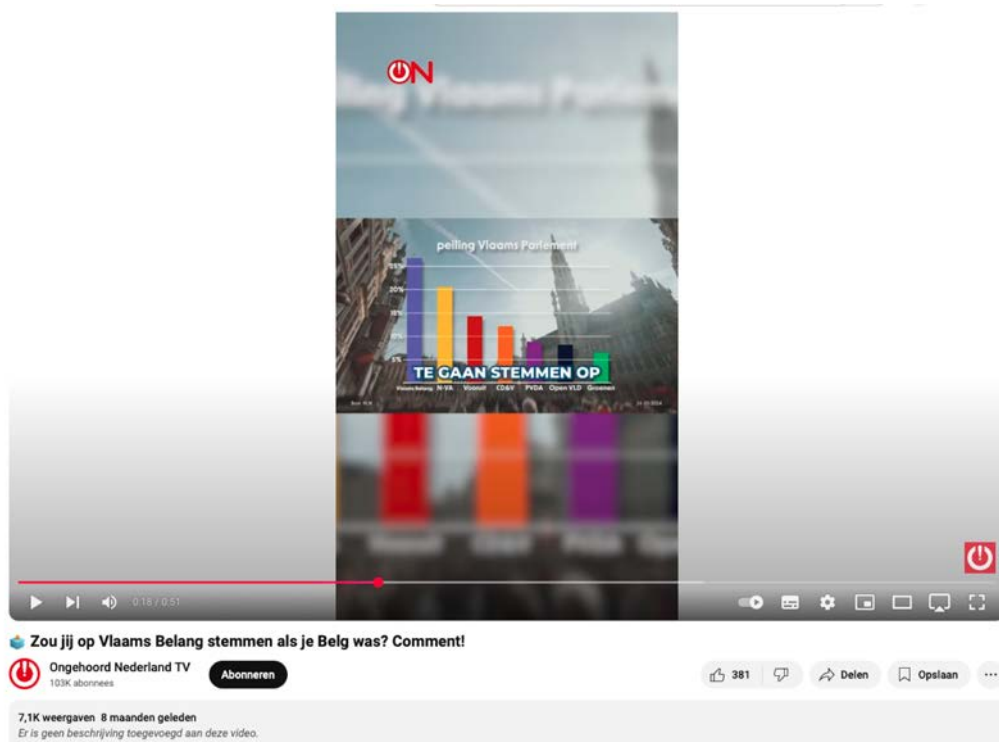


Figure 3.6 Screenshot (24-02-2025) of the Dutch public broadcaster Ongehoord Nederland, which does not receive a funding notice despite being a public broadcast service. Source: YouTube.com.

Ongehoord Nederland presents a compelling case, given accusations of spreading disinformation and conspiracy theories (Dongen, 2021), and the Dutch public broadcaster NPO's request to revoke its broadcasting license (Koeleman, 2023). In such politically sensitive situations, applying an NFN label carries significant implications. It could either foster trust among viewers who might otherwise be wary of its content due to its public broadcasting status and institutional legitimacy, or it could alienate audiences skeptical of government-funded media.

Finally, we observed that several regional media organizations included in our sample, which can also be considered public service media, did not receive an NFN. These regional broadcasters, such as L1 Limburg and Omroep Flevoland in the Netherlands, play a crucial role in delivering publicly funded news and information to their local communities (See Figure 3.7).



Figure 3.7 Screenshot (18-02-2025) of two election-related videos from the regional public broadcaster L1 Nieuws (Limburg) and Omroep Flevoland (Flevoland) both not having a funding notification. Source: YouTube.com.

In Figure 3.7 we see a climate change notification during the Maastricht debate broadcasted on L1 Nieuws but not an NFN. When it comes to local public broadcasters, one could argue that their impact is limited, primarily extending to regional communities. At the same time, this highlights that YouTube's transparency efforts focus mainly on labeling major national and international players, showing less concern for local broadcasters. The problem with NFNs, however, goes beyond discrepancies in application.

Inconsistent availability of 'News Funding Notice' in Europe

Our analysis also revealed NFNs are absent in several EU countries across different language settings. Specifically, NFNs are missing for languages including Suomi (Finnish), ελληνικά (Greek), Dansk (Danish), Català (Catalan), Euskara, (Basque), Galego (Galician), and Português (Portuguese), though the notification is available for Brazilian Portuguese.

Figure 3.8 reveals disparities in NFNs for Dutch and Danish citizens during the 2024 EU Parliamentary election. YouTube does not uniformly provide funding information about Euronews' educational content to all EU citizens. These inconsistencies, coupled with non-transparent qualifying criteria, not only prompt critical questions about YouTube's methods for determining NFN eligibility and ensuring consistent application across European public service media contexts. These inconsistencies also cast doubt on the platform's ability to uniformly apply other transparency features across regions. For example, the visibility of climate change notifications, such as the one during the Maastricht debate broadcasted on L1 Nieuws.



Figure 3.8 Screenshots (18-02-2025) of a EuroNews video titled "How do the European Parliament elections work?" with a Danish view on the left and a Dutch view on the right. Source: YouTube.com.

Removing problematic content

YouTube's own report on content moderation during the 2024 EU Parliamentary Elections states it "terminated over 1,000 channels and removed over 140 EU election-related videos for violating our Community Guidelines, including policies on manipulated content and misattributed footage" (The YouTube Team, 2024). In a

sample of 8,142 election-related videos, this study found 486 unavailable videos. To put this in perspective, we compared this to a query on Andrew Tate with a similar method, a controversial influencer banned from YouTube in August 2022 (Holpuch, 2022) (Figure 3.9).

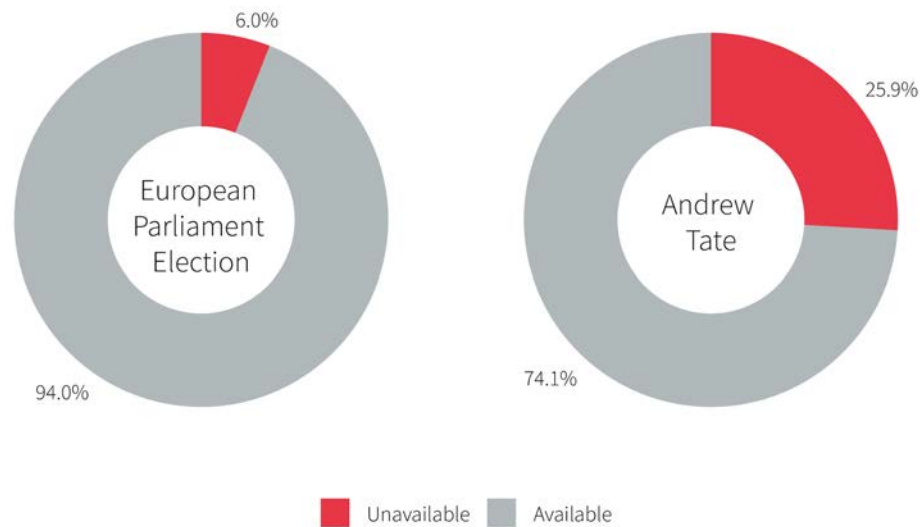


Figure 3.9 Removal rating of videos for the European Parliament Election query (N=8,142) (Removal information retrieved on October, 223, 2024) and Andrew Tate query (N=65,159) (Removal information retrieved on December, 7, 2024). Source: authors.

Figure 3.9 shows that 6% of videos related to the European Parliament Elections were removed, compared to 26% for videos associated with banned influencer Andrew Tate. While definitive conclusions are difficult based on our sampling and YouTube's opaque removal explanations, this discrepancy may reflect stricter pre-upload filtering for election content, or, perhaps more likely, simply less (problematic) content being uploaded, due to its region specific topic and low public interest in EU elections (see European Economic and Social Committee, 2023).

Ambiguity of removal statements

We identify six distinct justifications provided to users for why content is unavailable (see Table 3.3).

Removal Statement	Number of Videos
Video unavailable This video is no longer available because the YouTube account associated with this video has been terminated.	202
Video unavailable	179
Private video Sign in if you've been granted access to this video.	77
Video unavailable This video has been removed by the uploader	23
This video has been removed for violating YouTube's Terms of Service	4
Video unavailable This video isn't available anymore The user who uploaded the video has terminated their YouTube account	1

Table 3.3 Removal Statements for Videos Related to the European Parliament Election. Note: These statements were collected on October 23, 2024. They were originally in Dutch and subsequently translated into English.

Table 3.3 reveals that most unavailable videos were accompanied by the message: "The YouTube account associated with this video has been terminated," while others simply state: "Video unavailable." This seems to imply that most of the deletions happen at the channel level. At the moment of capture, in only four cases did YouTube explicitly indicate that videos were removed for violating its Terms of Service (ToS). Revealingly, the platform communicates via ToS and not in relation to Community Guidelines. For users without access to additional channel details, such as a channel ID, this limited information is all that is provided. However, by visiting the channels of removed videos, we uncovered more detailed information (Figure 3.10).

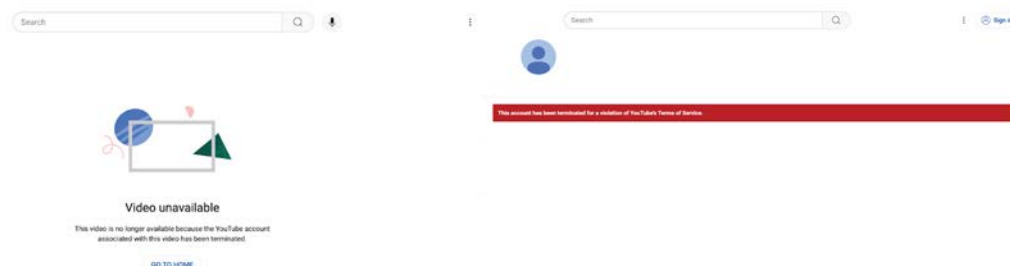


Figure 3.10 Removal statement for removed video compared with the removal statement on the page of the channel. Source: YouTube.com and authors.

Figure 3.10 shows that users accessing removed videos via shared links received only a generic termination message, while more detailed explanations were visible only on the channel page itself. This suggests a deliberate discursive strategy by YouTube to obscure its role in content removal. In contrast, user-removed content is explicitly attributed to the user. Furthermore, while YouTube's public reporting emphasizes Community Guidelines, actual removal statements reference only the Terms of Service (ToS), legal instruments designed to minimize liability (Gillespie, 2018, p. 46). This discrepancy highlights the difference between ToS, focused on legal protection, and Community Guidelines, which aim to shape moral behavior.

While removal statements were generally vague, we found rare instances where YouTube provided more specific explanations. For example, in our Andrew Tate dataset, two out of 16,891 removed videos included detailed reasons like violations of spam policies or policies on violent or graphic content, with links to relevant policy pages. However, such detailed explanations were the exception, not the rule.

Types of removal

Of the four videos removed for violating ToS, two of the associated channels remained active. This suggests that, while YouTube removed the other channels completely, in these cases the channels were allowed to remain online. One example is the channel WealthHub, which reuploads content from banned influencer Andrew Tate. The removed video, titled "How Tristan Influenced Elections," included a description claiming, "It is undeniable that Tristan and his brother are the most influential influencers among young men. In the recent European Parliament elections, German youths played a decisive role in the outcome. #germany #tristantate #politics #europe." Searching for the channel in a clean browser, the channel appeared only after extensive scrolling, indicating a low ranking in YouTube's search algorithm. Direct navigation to the channel via its ID revealed approximately 17,400 subscribers and 74 YouTube Shorts with a combined total of 27 million views (as of writing). While most videos had modest view counts, certain uploads, such as clips featuring Andrew Tate and Piers Morgan, demonstrated the potential to go viral and reach millions. This highlights that channels reuploading content from controversial figures like Andrew Tate, with the capacity to go viral, can remain on the platform. In this case, YouTube appeared to have acted on the specific video related to the EU election without removing the entire channel and possibly down-ranked the channel in search results. While this is not to suggest the channel should have been removed, its low subscriber count and limited search visibility may obscure its ability to produce and distribute Shorts with controversial content capable of reaching millions—a factor that is outside the scope of this research but deserves more attention.

The other channel, titled DDGeopolitics, posted an interview with investigative journalist and former MEP candidate Polona Frelih, which was subsequently removed. The channel describes its overall mission as follows:

DD Geopolitics was born out of the necessity to push back on Western narratives and censorship in media [...]. We regularly feature contributors from China, Palestine, Russia, Serbia, USA, and Yemen. We have been featured on Russia Today, TNT, Sputnik, PressTV, and other platforms.

This is another instance in which YouTube decided that while the video related to the EU elections violated ToS, the broader mission of the channel to engage with borderline content and partners that have been banned from the platform are still able to exist, which shows YouTube's efforts to limit but not remove this type of borderline content.

The platform's limited transparency about the specific takedown reasons make it difficult to formulate solid conclusions about YouTube's priorities or hierarchies of concern in content removal. However, the two channels examined here exemplify borderline cases that raise concerns about potentially harmful reactionary content and state-funded disinformation. Both channels display affiliations with figures and entities

banned from the platform, including Andrew Tate, Russia Today, and Sputnik. In these instances, YouTube appears to have removed specific videos while allowing the channels, which reupload Andrew Tate's content or feature material associated with Russia Today and Sputnik, to remain active. Such analyses, however, would benefit a lot from more detailed removal statements and other data access, which in very rare cases, YouTube does provide.

Access to data

While our mixed-methods approach provided valuable insights, it also revealed limitations in YouTube's data access for researchers. Despite YouTube's claims of "expanding data access" (Richardson & O'Connor, 2023), important information remains inaccessible to independent researchers. In this final section, we reflect on the broader issue of access to data and the idea of 'observability' in content moderation studies.

First, we find that YouTube's API exhibits a recency bias, prioritizing recent content while older content fades from results. This hinders research on past events without real-time data collection. However, including extensive historical data might misrepresent typical user experience, which favors new uploads. A potential solution is an optional "historical mode" within the Research API, alongside the default recent view. This would allow researchers to access historical data while retaining the option to analyze content as seen by everyday users.

Secondly, our study relied on web scraping to collect data on two critical YouTube features: News Funding Notices (NFNs) and removal statements. This information, though displayed to users, is absent from YouTube's API. Integrating these features into the API's metadata would be very valuable, including other content notifications (e.g., funding labels and health disclaimers) that provide information to users and consequently may serve as research variables for understanding YouTube's moderation and disinformation strategies. While the DSA Transparency Database provides valuable information on moderation 'statements of reasons,' it lacks essential identifiers like video and channel IDs. Integrating this database with YouTube's removal statements and API would enable more robust and transparent research on platform moderation. Of course, there is the risk of bad actors using this information in "gaming the system" (Petre et al., 2019). However, we argue that increased transparency and observability pose less risk in this context. Borderline content creators can already scrutinize Terms of Service and Community Guidelines, and clearer rules are ultimately more beneficial than harmful.

Our data access requests speak to a particular shortcoming in YouTube's current efforts via the "YouTube Research Program" (YouTube. (n.d.-b)). While intended to support academic research, the program currently offers researchers similar API access as commercial users. Expanding API access for researchers with the proposed elements would facilitate more independent, large-scale studies of moderation while reducing reliance on tedious and difficult-to-scale scraping methods.

Conclusion and implications

This exploratory study critically examined YouTube's efforts to promote authoritative election-related content and enforce content removal during the 2024 EU

Parliamentary elections. Our findings offer several important considerations for policymakers and other researchers.

First, policymakers are advised to be aware of the apparent increasing role of PSM on social media platforms like YouTube. YouTube seeks to prioritize authoritative news by elevating content from legacy media and PSM, which operate under rigorous editorial standards for gathering and disseminating information, particularly during critical events like elections. Our findings here align with emerging research (Glaesener, 2023; Padilla et al., 2025) and YouTube's own assessment (The YouTube Team, 2023). This may be YouTube's attempt to indirectly exert editorial oversight, outsourcing responsibility for election information to established media organizations.

The European Parliament has also recognized PSM's democratic and cultural significance, advocating for increased funding, independence, and a focus on combating disinformation (Rodríguez-Castro et al., 2020). However, the effect of YouTube's strategy remains unclear. While YouTube seeks to prioritize traditional media when it comes to important events, its effect on actual audience engagement needs more research. Existing research highlights public distrust in news organizations (Newman et al., 2024). Moreover, disinformation and extreme content might not directly present itself in highly moderated spaces around elections but via other topics. That being said, effective content moderation must also address challenges facing traditional media outlets who are serving an important function on social media platforms.

Second, policymakers are advised to think about the uniform availability of transparency features to effectively combat misinformation across all EU countries. Our findings on NFN application discrepancies may extend to other tools and notifications, highlighting the urgent need for platforms to support all official EU languages (and perhaps beyond). Inconsistencies in implementing disinformation measures risk creating national disparities and undermining the DSA's effectiveness. YouTube's claim that its "policies apply to everyone and are enforced with consistency" (YouTube, n.d.-c) appears at odds with the uneven application of transparency features across languages. YouTube acknowledges that "information panels may not be available in all countries/regions and languages" and promises that it is "working to bring information panels to more countries/regions" (YouTube Help). However, NFNs were introduced in 2018, and we know little about what information is available to whom. International collaboration is crucial to address these inconsistencies and ensure fair platform governance across the EU. Researchers should further investigate the availability of platform features across countries, as selective rollouts necessitate detailed mapping of implementation and availability.

Finally, policymakers are advised to expand their current data access initiatives under the DSA Transparency Database to include platform-specific services like the YouTube Research Program. While the DSA database provides removal data, researchers often require more contextualized information, such as video or channel IDs, for event-specific analysis. We recommend that YouTube incorporate detailed removal statements and video-specific notifications into its Research API. YouTube possesses this information but has not yet made it accessible for research.

In conclusion, this study highlights that while YouTube has made progress in promoting authoritative sources and refining moderation practices, gaps remain. Addressing inconsistencies in transparency features, improving data access for researchers, and ensuring robust cross-border implementation are crucial for platforms like YouTube to effectively support democratic processes and maintain accountability in the future.

References

- Alexander, J. (2020, June 29). YouTube bans Stefan Molyneux, David Duke, Richard Spencer, and more for hate speech. *The Verge*.
<https://www.theverge.com/2020/6/29/21307303/youtube-bans-molyneux-duke-richard-spencer-conduct-hate-speech>
- Ballard, C. (2023, November 9). Perhaps YouTube Fixed Its Algorithm. It Did Not Fix its Extremism Problem. *Tech Policy Press*. <https://techpolicy.press/perhaps-youtube-fixed-its-algorithm-it-did-not-fix-its-extremism-problem>
- Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. *New Media & Society*, 21(11–12), 2589–2606. <https://doi.org/10.1177/1461444819854731>
- Bounegru, L., Devries, M., & Weltevrede, E. (2022). The Research Persona Method: Figuring and Reconfiguring Personalised Information Flows. In C. Lury, W. Viney, & S. Wark (Eds.), *Figure: Concept and Method* (pp. 77–104). Springer Nature.
https://doi.org/10.1007/978-981-19-2476-7_5
- Bradshaw, S., Elswah, M., & Perini, A. (2024). Look who's watching: Platform labels and user engagement on state-backed media outlets. *American Behavioral Scientist*, 68(10), 1325–1344. <https://doi.org/10.1177/00027642231175639>
- Buchholz, K. (2024, January 19). Infographic: 2024: The Super Election Year. *Statista Daily Data*. <https://www.statista.com/chart/31604/countries-where-a-national-election-is-was-held-in-2024>
- De Keulenaar, E., Burton, A. G., & Kisjes, I. (2021). Deplatforming, demotion and folk theories of Big Tech persecution. *Fronteiras - Estudos Midiáticos*, 23(2), 118–139. <https://doi.org/10.4013/fem.2021.232.09>
- Dias, N., Pennycook, G., & Rand, D. G. (2020). Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, 1(1). <https://doi.org/10.37016/mr-2020-001>
- Dongen, M. van. (2021, July 15). Dit kunnen we verwachten van Ongehoord Nederland, de aspirant-omroep van Arnold Karskens. *de Volkskrant*.
<https://www.volkskrant.nl/cultuur-media/dit-kunnen-we-verwachten-van-ongehoord-nederland-de-aspirant-omroep-van-arnold-karskens~b8b10cd9/>

European Commission. (2022, October 27). The EU's Digital Services Act. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

European Commission. (n.d.). DSA Transparency Database. Retrieved December 13, 2024, from <https://transparency.dsa.ec.europa.eu/>

European Economic and Social Committee. (2023, November 24). Disinformation and lack of interest are the main reasons for poor voter turnout in European elections. EESC. <https://www.eesc.europa.eu/en/news-media/press-releases/disinformation-and-lack-interest-are-main-reasons-poor-voter-turnout-european-elections>

Fisher, M., & Taub, A. (2019, June 3). On YouTube's Digital Playground, an Open Gate for Pedophiles. *New York Times*. <https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html>

Gillespie, T. (2017). Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, Communication & Society*, 20(1), 63–80. <https://doi.org/10.1080/1369118X.2016.1199721>

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Glaesener, T. (2023). Who informs Germans about the Russia-Ukraine War on YouTube? *The Journal of Social Media in Society*, 12(1), 87-114.

Google Transparency Report. (n.d.). YouTube Community Guidelines enforcement. Retrieved December 13, 2024, from <https://transparencyreport.google.com/youtube-policy/removals?hl=en>

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 1-15. <https://doi.org/10.1177/2053951719897945>

Grimmelmann, J. (2015). The Virtues of Moderation. *Yale Journal of Law & Technology*, 42(2015), 68.

Hille, H. (2022, April 14). Die Welt newspaper dominates YouTube's featured news playlist. *DataSkop*. <https://dataskop.net/yt-news-die-welt-en/>

Holpuch, A. (2022, August 24). Why Social Media Sites Are Removing Andrew Tate's Accounts. *New York Times*. <https://www.nytimes.com/2022/08/24/technology/andrew-tate-banned-tiktok-instagram.html>

Koeleman, D. (2023, April 24). NPO vraagt ministerie om vergunning van Ongehoord Nederland in te trekken. *NU*. <https://www.nu.nl/media/6260797/npo-vraagt-ministerie-om-vergunning-van-ongehoord-nederland-in-te-trekken.html>

- Kofman, A. (2019, November 22). YouTube Promised to Label State-Sponsored Videos But Doesn't Always Do So. *ProPublica*.
<https://www.propublica.org/article/youtube-promised-to-label-state-sponsored-videos-but-doesnt-always-do-so>
- Kumar, S. (2019). The algorithmic dance: YouTube's Adpocalypse and the gatekeeping of cultural content on digital platforms. *Internet Policy Review*, 8(2).
<https://doi.org/10.14763/2019.2.1417>
- Kurdi, M., Albadi, N., & Mishra, S. (2021). "Think before you upload": An in-depth analysis of unavailable videos on YouTube. *Social Network Analysis and Mining*, 11(1), 48. <https://doi.org/10.1007/s13278-021-00755-x>
- Lewis, B. (2018, September 18). Alternative Influence. Data & Society; Data & Society Research Institute. <https://datasociety.net/library/alternative-influence/>
- Miller, L. (2020, September 24). Authoritative voting information on YouTube. YouTube Official Blog. <https://blog.youtube/news-and-events/authoritative-voting-information-on-youtube/>
- Mozilla Foundation. (2021). YouTube Regrets. <https://foundation.mozilla.org/en/youtube/findings/>
- Nassetta, J., & Gross, K. (2020). State media warning labels can counteract the effects of foreign disinformation. *Harvard Kennedy School Misinformation Review*.
<https://doi.org/10.37016/mr-2020-45>
- Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A., & Nielsen, R. K. (2024). Reuters Institute digital news report 2024. Reuters Institute for the study of Journalism.
- O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, 33(4), 459–478.
<https://doi.org/10.1177/0894439314555329>
- Padilla, A., Coromina, Ò., & Prado, E. (2025). Los Trusted Media en YouTube: Volumen y visibilidad de los medios públicos en los resultados de búsqueda. *Revista Latina de Comunicación Social*, 83, 1–17. <https://doi.org/10.4185/rlcs-2025-2336>
- Petre, C., Duffy, B. E., & Hund, E. (2019). "Gaming the System": Platform Paternalism and the Politics of Algorithmic Visibility. *Social Media + Society*, 5(4), 2056305119879995. <https://doi.org/10.1177/2056305119879995>
- Power, M. (1997). *The audit society. Rituals of verification*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198296034.001.0001>
- Richardson, L., & O'Connor, J. F. (2023, August 24). Complying with the Digital Services Act. Google. <https://blog.google/around-the-globe/google-europe/complying-with-the-digital-services-act/>

- Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, 9(4). <https://doi.org/10.14763/2020.4.1535>
- Rieder, B., Matamoros-Fernández, A., & Coromina, Ò. (2018). From ranking algorithms to 'ranking cultures': Investigating the modulation of visibility in YouTube search results. *Convergence*, 24(1), 50–68. <https://doi.org/10.1177/1354856517736982>
- Rodríguez-Castro, M., Campos-Freire, F., & López-Cepeda, A. (2020). Public Service Media as a Political Issue: How Does the European Parliament Approach PSM and Communication Rights? *Journal of Information Policy*, 10, 439–473. <https://doi.org/10.5325/jinfopoli.10.2020.0439>
- Samek, G. (2018, February). Greater transparency for users around news broadcasters. YouTube Official Blog. <https://blog.youtube/news-and-events/greater-transparency-for-users-around/>
- Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: Converting critical concerns into productive inquiry*, 22(2014), 4349-4357.
- Shaban, H. (2018, February 3). YouTube's new attempt to limit propaganda draws fire from PBS. *Washington Post*. <https://www.washingtonpost.com/news/the-switch/wp/2018/02/03/youtubes-new-attempt-to-limit-propaganda-draws-fire-from-pbs/>
- Shane, S. (2017, November 12). In 'Watershed Moment,' YouTube Blocks Extremist Cleric's Message. *New York Times*. <https://www.nytimes.com/2017/11/12/us/politics/youtube-terrorism-anwar-al-awlaki.html>
- Silverman, C. (2016, November 16). This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. *BuzzFeed News*. <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook>
- Singh, S. (2024, November 14). How Many People Use YouTube 2024 (Active Users Stats). DemandSage. <https://www.demandsage.com/youtube-stats/>
- Stoian, V. (2023). The EU approach to combating disinformation: Between censorship and the "market for information." In Arcos, R., Chiru, I., & Ivan, C. (Eds.), *Routledge handbook of disinformation and national security* (pp. 311-327). Routledge.
- The YouTube Team. (2019a, December 3). The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation. YouTube Official Blog. <https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/>

- The YouTube Team. (2019b, September 3). The Four Rs of Responsibility, Part 1: Removing harmful content. YouTube Official Blog. <https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/>
- The YouTube Team. (2023, June 2). An update on our approach to US election misinformation. YouTube Official Blog. <https://blog.youtube/inside-youtube/us-election-misinformation-update-2023/>
- The YouTube Team. (2024, July 26). The 2024 European Parliamentary elections on YouTube. YouTube Official Blog. <https://blog.youtube/news-and-events/2024-european-parliamentary-elections-on-youtube/>
- Trujillo, A., Fagni, T., & Cresci, S. (2024). The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media (No. arXiv:2312.10269). arXiv. <https://doi.org/10.48550/arXiv.2312.10269>
- Tufekci, Z. (2018, March 10). Opinion | YouTube, the Great Radicalizer. *New York Times*. <https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html>
- Wakabayashi, D., & Confessore, N. (2017, October 23). Russia's Favored Outlet Is an Online News Giant. YouTube Helped. *New York Times*. <https://www.nytimes.com/2017/10/23/technology/youtube-russia-rt.html>
- Wakabayashi, D., Astor, M., Salam, M., & Stevens, M. (2018, April 3). Police Identify Woman Who Wounded 3 at YouTube and Killed Herself; Motive Is Unclear. *New York Times*. <https://www.nytimes.com/2018/04/03/technology/youtube-shooting-suspect-headquarters.html>
- West, D. M. (2024, November 7). How disinformation defined the 2024 election narrative. Brookings. <https://www.brookings.edu/articles/how-disinformation-defined-the-2024-election-narrative/>
- Yesilada, M., & Lewandowsky, S. (2022). Systematic review: YouTube recommendations and problematic content. *Internet Policy Review*, 11(1). <https://doi.org/10.14763/2022.1.1652>
- YouTube Help. (n.d.). Information panel providing publisher context. YouTube Help. Retrieved December 13, 2024, from <https://support.google.com/youtube/answer/7630512?hl=en>
- YouTube. (n.d.-a). How does YouTube address misinformation? YouTube. Retrieved December 13, 2024, from <https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/>
- YouTube. (n.d.-b). YouTube Research. YouTube. Retrieved December 13, 2024, from <https://research.youtube/>
- YouTube. (n.d.-c). YouTube security and election policies. YouTube Security and Election Policies - How YouTube Works. Retrieved December 13, 2024, from

<https://www.youtube.com/howyoutubeworks/our-commitments/supporting-political-integrity/>

4. The performance of borderline content on Facebook

Richard Rogers and Kamila Koronska

Abstract

Facebook's content moderation has been the source of critical study for some years now. The platform has been the site of Russian disinformation campaigns, 'fake news' infestations, disproportionate engagement by 'right-wing commentators' and other moderation challenges for both extreme as well as borderline content. This piece takes up the performance of this type of content, revisiting a well-known methodology from data journalism deployed to study it. Following in the footsteps of Craig Silverman's 2016 investigation that ushered in the 'fake news' crisis, and revisited in 2020 with some methodological and definitional refinements, we reopen the question of the performance of borderline content on Facebook in the run-up to the 2024 U.S. presidential elections. How well does it perform? We find that it is outperformed by mainstream content compared to 2016 and 2020. In our findings, we contextualize Meta's efforts to 'depoliticise' platform content, especially around election time. We conclude with the implications of such internal programming, especially on research data access and Meta's emphasis on content-agnostic adjudication, which allows much of what was once removed to remain on-platform.

Keywords: fake news, hyper partisan, Facebook API, APICALypse, CrowdTangle, Meta Content Library

Introduction: Facebook's content moderation

Facebook's content moderation has faced heightened scrutiny over the years, beginning at least with the investigative reporting that found the company outsources the work of blocking disturbing content to low-wage contract workers from around the globe (Chen, 2014; Roberts, 2016). The 'fake news' crisis of 2016 and beyond had its origins on Facebook, when data journalists reported that low quality news, including imposter and hyperpartisan sources, outperformed mainstream news in the run-up to the presidential elections. The Cambridge Analytica scandal of 2018 — the mass data breach of over 80 million social media users — occurred on Facebook; it involved a seemingly innocent online quiz which would provide its users with their psychological profiles but it also collected the data of the Facebook user as well as his or her friends and fed it to a company who advertised itself as engaging in 'psychographic profiling', selling it to political campaigns (including Donald Trump's) (Hindman, 2016). The Facebook whistleblower, Frances Haugen, focused attention on the social media company's moderation policies in 2021, revealing that it disables content amplification algorithms based on angry reactions prior to elections, only to re-enable them thereafter (Merrill & Oremus, 2021). Given this recent historical context, researchers have posed questions concerning Facebook's quality control, asking about the availability of problematic content on the platform, especially around elections but also more broadly.

At the same time, scholars studying the effects of moderation on the content that circulates on Facebook have faced a variety of obstacles as the company has disabled or limited data-gathering pipelines over the years often with privacy-related arguments. Facebook APIs have been discontinued (v1 in 2014 and v2 in 2018), its Graph Search deprecated (2019) and CrowdTangle dismantled (in 2024). In certain instances, these capabilities have been replaced with new initiatives such as the Social Science One project (after the Facebook Pages API was removed in 2018) and the Meta Content Library (after CrowdTangle's demise); Meta also shared Facebook (election-related) data with a select group of researchers in 2020-23. Arguably, these replacements have only narrowed the scope of research that can be accomplished on the platform, occasioning researchers to seek other data collection strategies, including scraping (Vincent, 2021).

The (often downgraded) data replacements and the limited researcher partnerships have been met with questions about how to continue to analyse platform data as well as how Meta is steering academic research. Indeed, the company's limiting of research data as well as access to it have revitalized debates about how to study content online, developing post-API approaches after the so-called APICALYPSE, which stands for the ruin brought to academic research by drying up data streams. At the same time, researchers have raised suspicions that the company has "instrumentalised [privacy-related concerns] to actively frustrate critical, independent, public interest scrutiny by scholars" (Bruns, 2021).

In light of the recurring concerns with Facebook content as well as the difficulties in researching platform vulnerabilities, the following takes up the question of how to study problematic information on Facebook after the end of CrowdTangle. It first details the kinds of data that have been accessed by researchers over the years as well as the types of research that can be undertaken, making comparisons between API-based research, Social Science One, CrowdTangle and Meta Content Library. Here one notes how the winnowing of data points has shaped the research environment, given the increasingly computationally intensive data sets and pipelines made available or the stingier partnerships formed of vetted researchers affiliated with company-organised programmes.

The research reported here additionally seeks to examine Facebook content related to the U.S. presidential elections in 2024, that is, the very subject matter that researchers and public advocacy organizations have highlighted as likely affected by the lack of data access. To do so, it goes off platform, so to speak, employing an approach first developed by Craig Silverman from BuzzFeed News (2016), the data journalist, mentioned above, whose work ushered in the 'fake news' crisis of 2016 and beyond. It also builds upon further refinement of the methodology that takes into account the politicisation of 'fake news' as well as the definition and classification debates that followed (Rogers, 2020) – discussions concerning how the branding of news as 'fake' affected free speech and resulted in 'cancel culture'. 'Fake news' developed a politics, given the publicising of findings that it is far more associated with the right of the political spectrum (Roose, 2020).

Elections and other salient events (such as the Covid-19 pandemic) have become charged occasions for definition-wrangling. Indeed, the fake news definition debates are significant given how the original notion (by Silverman) contained imposter as

well as hyperpartisan sources. These hyperpartisan sources, defined as ‘openly ideological web operators’ (Herrman, 2016), would later be removed in Facebook’s reworking of ‘fake news’ as ‘false news’ (Weedon et al., 2017). This new term shrank the problem and also depoliticised the issue given that right-leaning, hyperpartisan sources were removed from consideration.

The depoliticisation is evident in the ‘content agnostic’ approaches to determining problematic materials as well as in how this content is handled. As the head of Meta has put it in a widely reported ‘blueprint for content governance and enforcement’ (2018), the company “fight[s] misinformation [by] identifying fake accounts, which are the source of much of the spam, misinformation, and coordinated information campaign (Zuckerberg, 2018). For problematic content that is not circulated by ‘fake accounts’, it relies on fact checkers as well as automated systems. Meta has demoted rather than removed it, reducing its reach while not censoring, canceling or otherwise taking a position on it. It is a part of its so-called “recipe for cleaning up [the] News Feed”, first implemented after the acknowledgement of the fake news problem (Facebook, 2018).

How ‘cleaned up’ is it? In comparing the findings about the performance of problematic or borderline content from the ‘fake news’ data journalism piece from 2016 with our own empirical work from the 2020 and 2024 U.S. elections, we found that over the course of the twelve years the problem first worsened and most recently there is a trend towards the mainstream outperforming the fringe. In a sense this is the result that Facebook’s content moderation has sought.

When present, the problematic content, however, is still nearly exclusively from so-called ‘right-wing commentators’, an issue that has been widely publicised, particularly through journalistic reporting about ‘Facebook’s top ten’ sources by engagement measures in the run-up to the 2020 elections and beyond. Meta’s content-agnostic, depoliticising approach allows that content to remain on the platform, as we detail.

The piece first takes up how to research Facebook with the data that is made available by the company. Subsequently we discuss Meta’s content moderation efforts concerning what they called ‘false news’ (rather than ‘fake news’) and later borderline content that brushes up against (but does not cross) the company policy lines (Constine, 2018). How has the definition of that type of content affected how it is moderated? Thereafter we turn to a comparison between the three periods when moderation was particularly scrutinised: the 2016, 2020 and 2024 U.S. elections. How has borderline content performed on the platform over time?

We conclude by showing that Facebook’s moderation effects appear to be more pronounced in the current period compared to 2020 and 2016. At the same time their efforts at depoliticising moderation are challenged by the fact that the borderline content present (and occasionally performing well) is from ‘right-wing commentators’, including those touting conspiracy theories.

The performance of these commentators is obfuscated by the current data access regime, company arguments as well as so-called transparency reporting. One measure of performance (reach) is favoured over the measure previously relied upon by

marketing companies, researchers as well as journalists (engagement). In Meta's latest data environment (Meta Content Library) the reach data point is highlighted, while engagement is downplayed.

Facebook data access

In the following, we first take up how Facebook data has been made available to researchers (including journalists and non-governmental organizations), comparing the two APIs (v1 and v2), the Social Science One project, CrowdTangle and most recently the Meta Content Library. There are two purposes. One is to describe how research has been shaped by the data availability, enabling and curtailing certain approaches. One red thread is the privacy argument made by the company for each change to data access, which is well known, but we also make note of how the company gradually has come to prefer academic-industry partnerships and approved researchers who apply for access. The other purpose of discussing the different data access regimes is related more explicitly to content moderation research, especially after the revelations of the copious 'fake news' and disinformation found on the platform in 2016 and beyond. How have the changes to data access affected this particular line of research?

The Open Graph API (v1), launched in 2009 (and discontinued in 2014), is sometimes referred to as the Friends data API, for it allowed researchers access to one's friends as well as their interests and 'likes' (Rieder et al., 2015). It enabled social network analysis, especially tastes and ties work, where one studies (for example) the extent to which a cohort of friends has similar tastes and/or 'likes' (Lewis et al., 2008). But v1 also allowed access to Groups and Pages, where one could additionally conduct engagement analysis, that is, which posts animate a Group, a Page or sets of them.

When it was shut down, the most salient rationale concerned a consent argument. The Friends have not given permission to have their data harvested. In research circles, it also coincided with what could be called an ethics turn in data-driven platform research, where ideas that the 'data are already public' and thereby fair game to researchers were met with 'contextual privacy' concerns (Zimmer, 2010; Nissenbaum, 2011). Despite agreeing to the platforms' terms of service (and thereby to data harvesting), the users had neither given consent to researchers, nor expected in their context of use that their posts, preferences and likes would be deployed analytically.

The API also laid the ground for the Cambridge Analytica scandal, which is often called a Facebook 'data breach', but it was actually a large Facebook profile harvesting exercise by a Cambridge University researcher, feeding a campaign consultancy business connected to him working in the area of 'addressable ad tech' (Rosenberg et al., 2018; Rath, 2019). It used the Open Graph API v1. When one took the online personality test, "This Is Your Digital Life", he or she would receive in turn their personality type, corresponding to the so-called big five: openness, conscientiousness, extroversion, agreeableness, and neuroticism. At the same time, the app also extracted the player's Facebook friends' data (Hindman, 2018). It collected some 80 million profiles overall, and with them produced 'psychographic profiles' for ad targeting on the basis of Facebook users' likes. Ads would target those with particular combination of traits or profiles, e.g., those with 'low openness' and 'high neuroticism' would receive ads supportive of Republicans. U.S. Senator Ted Cruz of Texas was one of

Cambridge Analytica's customers (Issenberg, 2015). Donald Trump was another (Karpf, 2017).

The closure of the Open Graph v1 API in 2015 actually predated the public discovery of the massive data harvest from the personality test and the scandal that followed. Though that particular damage had already been done, the API v2 removed the functionality associated with Friend data access, at the same time retaining the capacity to extract data from Groups and Pages (Rieder, 2013). Joining a Group or liking a Page gave one access to that Group's or Page's entire history via the API. (Later, the researcher would have the access without liking a Page.) One would make lists of Pages (say, the alt right or the Rwandan diaspora), download all the Pages' posts, together with the engagement metrics, and study the content that particularly animates these groups (Kok & Rogers, 2017). Is it problematic content? Though they were not called that just yet, some of the early content moderation studies made use of the approach of compiling sets of Pages that could have content deemed offensive, extreme or hateful (Benkler et al., 2018). An addition of overtime analysis would allow research into the persistence of problematic content and thus the lack of moderation.

The closure of the Pages API (as v2 was often called) came on the heels of a 2018 review of 'third-party apps' sitting atop it. In a sense, the review was a reaction to the unrestricted use of the third-party app, "This Is Your Digital Life", the source of Cambridge Analytica data. How many others were harvesting Page and Group data at scale? It also brought to an end academic software that sat atop the API, including Netvizz, Facepager and Netlytic, spurring an uproar summarised in the widely cited notion of the 'APIcalypse' (Bruns et al., 2018). It was argued that without the data the platform could not be held accountable for harboring disinformation and other problematic content (Walker et al., 2019). It also had the practical consequence of hindering textbook routines for undertaking Facebook research (Rogers, 2019).

Facebook deprecated another service around the same, Graph Search, which allowed for advanced searches of Facebook such as a list of users who liked a Page, or all Pages liked by a user. It was not something that academics were known to use but considered as a digital investigation resource for journalists, open source intelligence (OSINT) users and others in similar lines of work, including human rights researchers (Cox, 2019). The deprecation was described as "Facebook's response to data scandals and its resulting push to emphasize privacy" (Silverman, 2019).

In place of the Pages API, Meta offered the Social Science One project (King & Persily, 2020), a big data alternative that would provide vetted researchers a data set of the URLs posted on Facebook, together with a collection of metadata. It changed Facebook research rather substantially, reorienting it towards the computationally intensive. In doing so, it also nudged research away from the analysis of posts made on the platform (and their performance or engagement) to web URL appearances in posts. The subject matters of the research were affected, as they were restricted largely to elections or election integrity.

The Social Science One data sets (of millions of URLs) reside in a repository where access is granted after the submission of a research proposal by credentialled university researchers (whose eligibility is equivalent to that of a principal investigator). At the time of writing the data set includes URLs posted up until October

2022. Given its lag in currency, it would not have been called upon for research on the 2024 U.S. elections.

For that matter, no researchers would make use of Meta's other Facebook data resource, CrowdTangle, for it was discontinued in August 2024, a few months prior to the U.S. elections. It was a Facebook (and Instagram) data dashboard and API, which the company made available to users in 2018, after the closure of the Pages API. Like the Social Science One project, it again targeted specific research subject areas, specifically misinformation and election integrity. Known misinformation researchers would be able to operate a 'lab' and invite other users working on such projects.

Apart from company claims about the superiority of its successor – language similar to its rollout of Social Science One and closure of the Pages API – to date there has not been an exposé on the reasons behind CrowdTangle's closure, especially considering that the timing coincided with a wave of elections (including European and American ones) which are its specific use cases. A privacy-enhancing and more secure replacement – initially called the Fort – had been in the works for some time; it also was to be its regulation-compliant data resource, meeting the anticipated requirements of the new European legislation (the Digital Services Act) that would require so-called very large online platforms to provide data access to researchers.

But CrowdTangle, as tech journalists have written, also was proving to be reputationally challenging (Newton, 2020). Both researchers and journalists have been finding that "right-wing commentators" were routinely amongst the best performing Pages and accounts, according to engagement measures (Benkler et al., 2018; Roose, 2021). One particularly high-profile demonstration of that finding was posted daily on Twitter (beginning in July 2020 in the run-up to the U.S. presidential election), first manually and then through an automated routine that called the CrowdTangle API and outputted the results at @FacebooksTop10 (see figure 4.1). It prompted a highly public tussle between the journalist and the company concerning the distinction between engagement (i.e., reactions, shares and comments) and reach (i.e., views). The one may show right-wing commentators as top ranked, but the other did not, according to the company.

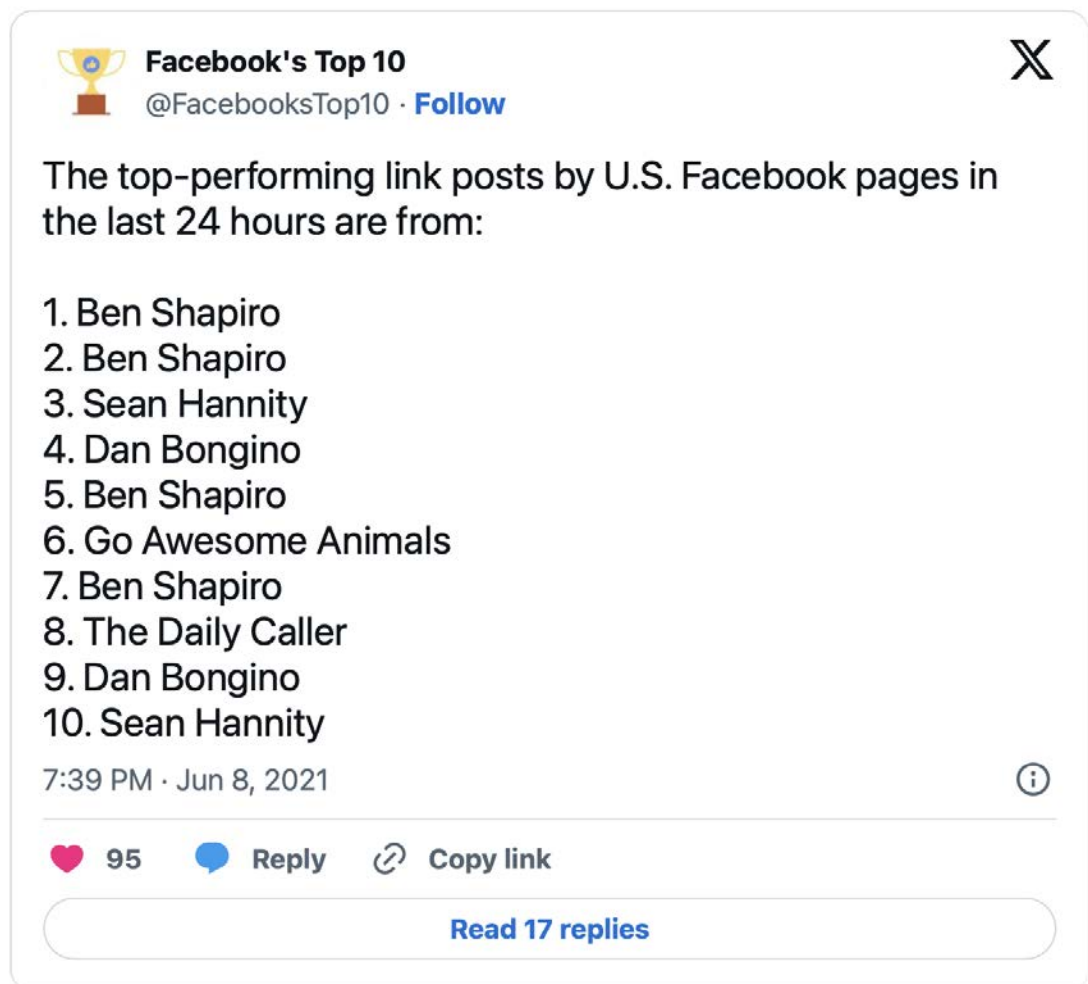


Figure 4.1 Output of @FacebooksTop10. Source: Roose, 2021.

How to consider the measurement of impact on a social media platform? To buttress its position in this debate, Facebook first shared a list of most viewed posts, but ultimately began issuing quarterly reports with tabulations of these posts, called ‘Widely Viewed Posts: What People See on Facebook’ (2021b). They show that the most viewed posts were more mainstream (but also spam-like) than those that garnered high engagement (Zuckerman, 2021). At the same time, it discontinued the work of the CrowdTangle team and not long thereafter halted offering new lab accounts. When in 2024 Meta introduced and defended its Meta Content Library (MCL) superseding CrowdTangle, it made two principal arguments in favour of the new tool over the existing one. The first was that CrowdTangle was already in a state of deprecation (given that the team had been disbanded), and the second concerned the superiority of MCL’s data because it included reach. At least initially, there was no engagement data, which had been the source of the fractious debate.

There have been reactions by researchers, journalists as well as non-governmental organisations, beginning with the fact that all tools sitting atop the CrowdTangle API ceased working, including academic software measuring ‘coordinated inauthentic behaviour’ on Facebook (Paroni & Terenzi, 2024). It was something of a re-run of the moment in 2018 when the Pages API was switched off. The other main reaction has been that the replacement is convoluted and impractical to access and does not allow

for the actual collection of data, with the exception of accounts and Pages with a great number of followers (Gotfredsen & Dowling, 2024). Historical data is unavailable.

In a kind of requiem written a few months prior to the closure of CrowdTangle, one tech journalist wrote, "Facebook executives hated [@FacebooksTop10]. Advertisers and policymakers were calling them up asking about Roose's tweets" (Newton, 2024). The internal company debate apparently resulted in a strategic shift in data access (Roose, 2021). Where individual academics, journalists and others with a CrowdTangle account previously would grant peer access, now Meta preferred an academic-industry partnership to vet academic research proposals, assess applications and enable data access, albeit as long as the data remained on company servers. Meta employees also fanned out to internet research conferences, reaching out to individual researchers asking them to make application to the CrowdTangle substitute. One of the authors was among those invited out to dinners and breakfasts.

While the first academic-industry partnership, the Social Science One project, may have stagnated, that model driving data access indeed continued, when in 2020 Meta granted a select group of U.S. university researchers (and members of Social Science One) exceptional access to platform data from the 2020 U.S. elections (Thorp, 2023; Uzogara, 2023). Selective data access granted to researchers is not completely new, as it was given to two groups investigating the 2016 Russian disinformation campaign on Facebook (and Instagram), which delivered their findings to the U.S. Congress (DiResta et al., 2018; Howard et al., 2018).

Here, however, this academic-industry partnership had a somewhat different character. Prefiguring the Meta Content Library restrictions, the researchers studying the 2020 election were not allowed to work with the data on their own computers or undertake research (e.g., on follower networks) that the company considered sensitive because of privacy concerns. Based on how it subtly steered the work, the project's academic rapporteur called the academic-industry partnership 'independence by permission', arguing against such a partnership construct as a model (Wagner, 2023).

What to moderate? The rise of content-agnosticism and personal choice

There are many content categories that are moderated according to Meta's community standards policies. At the time of writing they include the following: "hate speech; violence and incitement; suicide, self-injury and eating disorders; bullying and harassment; graphic violence; adult nudity and sexual activity; posts buying, selling, trading or promoting restricted goods or services (as defined by our Community Standards); spam; fake accounts; and scam" (Meta, 2024a). There are also those that fall into the 'borderline' type, i.e., "content (...) not prohibited but comes close to the lines drawn by the [community standards policies]" (Meta, 2024b).

The focus here is on the borderline. We examine how the content category has evolved, particularly how it first replaced 'fake news', or as the company called it, 'false news', as a category to be moderated and subsequently became a content type largely left on the platform by default. Facebook (or Meta) essentially ceased adjudicating content in favour of sources, particularly 'fake accounts' and 'adversarial actors'.

In this brief review of the evolution of the fake news terminology, the most significant finding is that the content that was once considered to be ‘false news’ and seemingly worthy of removal (in 2017) was subsequently termed ‘misinformation’ and retained and demoted (2019). More recently the same content is thought of as ‘borderline’ and retained online without demotion, unless the Facebook user explicitly requests it to be downranked through its ‘personalised ranking’ feature or has been flagged by fact checkers. Thus, unless it’s flagged by fact checkers, what was once considered to be ‘false news’ has gradually been redefined and retained by the platform by default. That last determination was removed in early 2025 when the company announced that it was discontinuing its fact checking services in the U.S., to be replaced by a Community Notes system first put into place by Twitter (‘Birdwatch’) and subsequently X. Determined to be a larger philosophical change in content moderation, a form of user consensus would replace company-sanctioned adjudication (De Keulenaar et al., 2023; de Keulenaar, this volume).

This most recent announcement completed a reversal of content moderation policy developed since the ‘fake news’ crisis. The rehabilitation of borderline content can be viewed in light of four touchstones in company content moderation history: its White Paper of 2017 (and the introduction of the fact-checking program just beforehand), the pending misinformation problem around 2020 elections and the Covid-19 pandemic, the introduction of personalised ranking as an antidote to defining borderline content in 2023 and the idea that misinformation cannot be defined and thus regulated in 2024. Each is taken in turn.

In its White Paper of 2017, Facebook detailed its reactions to the ‘fake news’ crisis that had engulfed the company beginning a year earlier, setting out the measures to address the problem. The first was a definitional move. Rather than ‘fake news’, which it dubbed ‘overused’ and ‘misused’, Facebook preferred ‘false news’, which it defined as "news articles that purport to be factual, but which contain intentional misstatements of fact with the intention to arouse passions, attract viewership, or deceive" (Weedon et al., 2017). At the outset it was thus quite similar to Silverman’s original definition of ‘fake news’ (discussed in the opening) that included imposter news sites as well as hyperpartisan sites, or those ‘openly ideological web operations’ often sensationalising and arousing passions.

The task for moderating false news lay at the feet of some 80 ‘partner organisations’ paid to fact check Facebook, a program initiated in late 2016 as part of the company’s larger ‘recipe’ for removing as well as reducing such content (Ananny, 2018). Those posts chosen and labeled by the fact checkers would be ‘down ranked’, reducing their reach considerably.

The deletion of the White Paper from the Facebook website in late 2020 could be seen as a course change, but was already in evidence in the policy documents where there was a gradual replacement of ‘false news’ (with its general emphasis on ‘misstatement of fact’) with ‘misinformation’ and its focus on harmful, health-related content during the Covid-19 pandemic as well as election integrity threats in the run-up to the 2020 U.S. elections. Indeed, in its moderation policy pieces, there are really only two areas of focus: "COVID-19 and vaccines and content that is intended to suppress voting" (Facebook, 2021a).

In 2023 when Meta described its ‘personalised ranking’ on the Facebook Feed (2023a), it also removed borderline content as a category to be moderated: "We use personalisation to reduce the distribution of content that doesn't violate our policies but may get close (...), mak[ing] this content less visible for those who prefer not to see it" (Meta, 2023b). This change effectively made the visibility of this content a matter of preference; moderation, at least of borderline content, had become personalised.

Finally, another policy development rehabilitated erstwhile problematic content further. In 2024, the company described ‘misinformation’ as a slippery term and changed its policy towards it. It described it as rather undefinable, arguing that it was not possible to classify it. "[W]hat is true one minute may not be true the next" (Meta, 2025a). Apart from content that would "contribute to a risk of imminent harm" (which could be said to fall into a different category of moderation), at its core misinformation once worthy of removal and later demotion now would remain online. In summary, we have witnessed an evolution whereby falsehood gradually has been transformed into a personal preference, an excessively disputed definition and a fact of life.

Given that it now remains online, in the next section we turn to the performance of borderline content around the periods of time when concern has been specially pronounced, the U.S. elections.

Facebook’s ‘fake news’ problem and the U.S. presidential elections, 2016-2020

Following in the footsteps of the original ‘fake news’ research around the 2016 U.S. elections as well as the subsequent refinements made for the study of the 2020 elections, for the empirical project, we ask, to what extent does problematic information resonate on political Facebook around the 2024 U.S. presidential elections? By ‘political Facebook’ is meant the collection of posts that receive considerable engagement where the presidential candidates and their talking points and issue language are mentioned.

This study revisits the initial ‘fake news’ report written by Craig Silverman and published by BuzzFeed News in 2016 as well as the findings and subsequent refinements to the methodology to include political leanings of the sources in 2020. Each of these studies has used the same data source, BuzzSumo, which describes itself as a social media "content research tool" (2024). The company gathers Facebook data by populating its database with trending Facebook URLs using crawlers, which are subsequently supplemented with engagement data from Facebook (BuzzSumo, 2024).

In 2016 Silverman deployed the content research tool to determine the best-performing content on Facebook for what he called fake news compared to mainstream news. In order to do so, he queried it for election-related keywords (candidate names, campaign story keywords, etc.), gradually compiling the election-related stories with the highest engagement over three-month periods from April to November, 2016. He found that in the period closest to the election that fake news outperformed mainstream news.

The news-like story with the headline, ‘Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement’, was the top performing piece, put out by Ending the Fed, a fly-by-night, hyperpartisan outfit (Silverman, 2016). It had three of

the top five performing stories, which together had greater engagement in the same period than the highest mainstream sources. One of the other well performing ‘fake news’ pieces, "FBI Agent Suspected in Hillary Email Leaks Found Dead in Apparent Murder-Suicide", was authored by a source calling itself, The Denver Guardian, which is an imposter news site. This exemplary combination of hyperpartisan and imposter sources would come to stand in for the notion, "fake news".

In 2020 we returned to Silverman’s study, redoing it for the elections that year (Rogers, 2020). Then, four years later, how well did so-called fake news perform compared to mainstream news? There were certain refinements to the methodology. We added the political leaning of the story sources (using a combination of media monitoring organisations that label sources, roughly, in that fashion). Does fake news tend to have a political leaning? We also compared the scale of the problem when using Silverman’s definition of ‘fake news’ (which included hyperpartisan sources) compared to Facebook’s ‘false news’ (where hyperpartisan sources are removed). Does Facebook’s fake news problem shrink when measuring ‘false news’ instead?

One significant finding concerned the political leaning of the hyperpartisan content: the vast majority was rightward-leaning, as had already been found in journalistic as well as scholarly studies (Benkler et al., 2018; Roose, 2020). ‘Fake news’ has a politics in the sense that it was empirically regularly associated with one portion of the political spectrum. As discussed above, this key finding – highly engaged-with, hyperpartisan sources on Facebook have a specific political leaning – arguably paved the way for Meta to develop and encourage content-agnostic approaches to identifying problematic content, thereby seeking to depoliticise content moderation.

The other finding concerned the scale of Facebook’s problem. With Silverman’s definition the problem had worsened slightly compared to 2016. With the Facebook notion of ‘false news’ that does not include hyperpartisan sources, the problem was hardly in evidence. Imposter sites (or ‘pink slime’ as they are sometimes dubbed), while present online in rather large quantities (Bengani, 2019), received far less traction than in 2016.

The performance of borderline content in the run-up to the 2024 elections

Turning to 2024, we again repeated the original method together with certain refinements. How well has ‘fake news’ performed on Facebook compared to more mainstream news in the months prior to the presidential election? How does its performance compare to previous elections?

To gather the data, we first assemble keyword lists that reflect talking points and issues mentioned in the political party platforms as well as during actual campaigns. We compile the lists by reviewing Democratic and Republican party platforms and sourcing keywords that represent flagship campaign agendas such as "illegal aliens" for the Republicans and "opportunity economy" for the Democrats. Akin to Silverman’s journalistic approach, we supplement the list to cover topics that emerged in the run up to the election, adding keywords from headlines and article leads from *The New York Times*. These articles include news coverage of all three candidates running in the campaign—Biden, Harris and Trump.

Next, we query these keywords, including the candidates' names, in BuzzSumo. The tool returns ranked lists of web URLs featured in posts, sorted by total engagement on Facebook (the sum of reactions, shares, and comments). We then label these URLs at the domain level using Media Bias/Fact Check's (MBFC) rating system, which classifies sources across the political spectrum—ranging from extreme right, right and center-right to center, center-left, left, and extreme left. In addition to political labels, we attach quality and content-type classifications from the same rating system, such as level of bias and "conspiracy/pseudoscience".

Redeploying MBFC indicators, we further group these outlets into our custom categories: "borderline/non-borderline" and "biased/least-biased." "Borderline" sources (which are similar to the hyperpartisan of 'fake news') publish "extreme right", "extreme left", or "pseudoscience/conspiracy" content. In contrast, "non-borderline" are those that do not fall into these categories. By "biased" sources, we refer to outlets with clear political leanings, including "extreme right", "extreme left", "right," and "left." For the least biased we use the categories, "center," "center-left" and "center-right." (There are also sources that do not fit into any of these categories, such as "YouTube" content and "unlabeled" links.)

Using these categories, we are able to provide insights into the performance of borderline content compared to more mainstream sources (as in the original BuzzFeed News piece). We also compare performance by political leaning (as in the refined methodology). There is additionally a comparison of more and less biased information, a new wrinkle we use to discuss the extent of the de-politicisation of election-related information on Facebook.

Lastly, we comment on the engagement of the content originating from YouTube, which performs particularly well (compared to other periods). It is also a platform that has been found to allow hyperpartisan (borderline) content to thrive in the run-up to the election (Grant, 2024). This is political content, largely by 'right-wing commentators', about how the 2020 U.S. election was 'stolen' or witnessed widespread irregularities. One other category, 'unlabeled', is important for the study of 'imposter' sources for the original 'fake news' definition, which we also explore, finding as in 2020 an overall absence of 'pink slime' websites, or those with the appearance of a news site but are fake (Bengani, 2019).

How much engagement has borderline content received (in comparison to 2020 and 2016)? Our findings contrast with Craig Silverman's 2016 'fake news' study as well as our own in 2020. We show that engagement with borderline political content in the run-up to the 2024 election was notably lower than in 2016 when Donald Trump won the presidency for the first time and in 2020 when he was defeated by Joe Biden. In our sample, borderline content accounted for a fraction of the total engagement with election-related URLs (see Table 4.1).

Mainstream domains for all three candidate-related queries consistently received higher Facebook engagement compared to borderline domains (see Figure 4.2). Specific observations can be made for each candidate. Both mainstream and borderline outlets mentioning Donald Trump had more consistent engagement. For the Democrats the opposite is true. Harris's engagement is highly concentrated in mainstream domains, with borderline engagement peaking briefly around August and quickly

fading. Biden's engagement trends mirror that of Harris in terms of mainstream prominence, but his borderline engagement shows a sharper and more event-specific peak before collapsing altogether when he withdrew from the presidential run towards the end of July.

Month	Borderline	Mainstream	Total by month
1/1/2024	26,644.00	1,712,349.00	1,908,696.00
2/1/2024	17,726.00	2,464,500.00	2,704,314.00
3/1/2024	44,005.00	2,540,342.00	2,827,498.00
4/1/2024	154,111.00	2,124,012.00	2,467,629.00
5/1/2024	12,992.00	2,230,213.00	2,447,776.00
6/1/2024	13,084.00	3,490,468.00	3,986,464.00
7/1/2024	15,867.00	6,744,923.00	7,582,728.00
8/1/2024	17,125.00	6,633,615.00	7,329,821.00
9/1/2024	17,493.00	7,126,371.00	7,886,153.00
10/1/2024	10,404.00	6,569,740.00	7,174,851.00
11/1/2024	1,272.00	808,400.00	855,483.00
Total	330,723.00	42,444,933.00	47,171,413.00

Table 4.1 Total Facebook engagement for "borderline" and "mainstream" content, 2024 elections. Source: BuzzSumo, 2024.

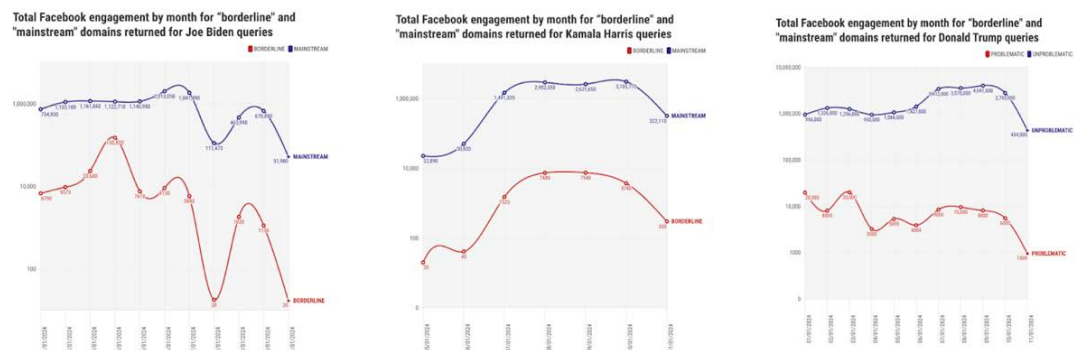


Figure 4.2 The graph shows the total Facebook engagement by month for URLs classified as "borderline" and "mainstream" shared in posts mentioning the candidates taking part in the 2024 U.S. presidential election - Joe Biden, Kamala Harris and Donald Trump. Logarithmic scale. Date range: 12th January - 31st October 2024. Source: BuzzSumo.

The type of borderline content classified by Media/Bias Fact Check as 'extreme' attracted attention on Facebook in short bursts (see Figure 4.3). Engagement started low in January, 2024, spiked in April, declined by June, and then leveled off at lower levels. This sharp rise and fall was driven by 'extreme-right' sources (as classified by MBFC), such as infowars.com and zerohedge.com.

These bursts of engagement resemble the algorithmic distribution patterns for which Facebook has faced scrutiny in the past—particularly by the whistleblower Frances Haugen—who described how polarizing content becomes viral (Hao, 2021; Bandy & Diakopoulos, 2023). As a result, borderline content can create noticeable spikes in metrics and appear in ‘top content this month’ summaries. None of the upticks in engagement with extreme right/left or conspiracy outlets coincided with major political events on the campaign trail, such as TV debates or the multiple Trump assassination attempts.

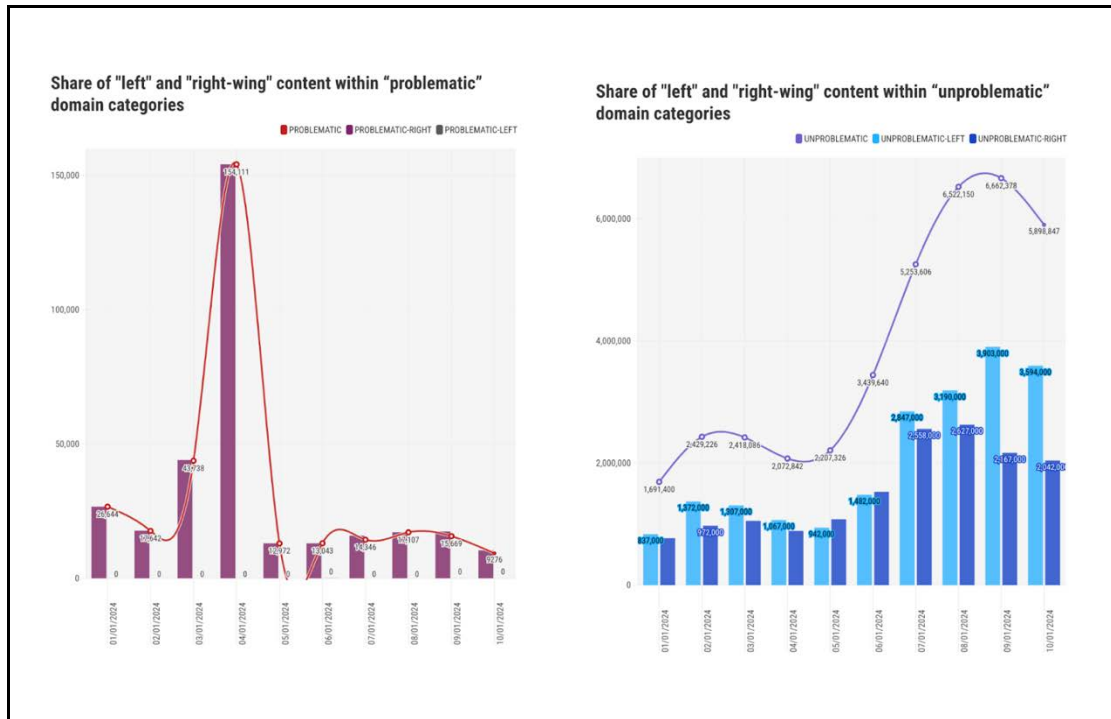


Figure 4.3 Facebook engagement by month for URLs classified as "problematic" and "unproblematic". The problematic categories include "extreme right", "extreme left", and "pseudoscience/conspiracy" domains as classified by the Media Bias / Fact Check (MBFC). URLs classified as "unproblematic" include entries labelled by the MBFC as "right", "left", "center-right", "center-left", and "least biased". Logarithmic scale. Date range: 12th January - 31st October 2024. Source: BuzzSumo, 2024.

When examining the politics of engagement of borderline content, we find that it comes disproportionately from right-wing domains. In contrast, engagement with non-borderline or more mainstream outlets shows a more balanced and steady interaction with both right- and left-leaning media throughout the campaign period. This trend suggests a clear divide in how borderline and non-borderline content resonates with Facebook users, with mainstream sources maintaining consistent attention, while problematic content tends to be from the right and sporadic, tied to polarising and viral articles from one of the extreme right outlets.

With respect to biased content, it received significantly less engagement than less-biased content, although this trend only became noticeable six months before the polls opened (see Figure 4.4). The relatively similar engagement performance between politically biased and less biased domains during the first six months of the campaign period indicates that the significance of hyperpartisan or politically biased content

began to decline significantly as election day approached, suggesting the effects of greater content moderation.

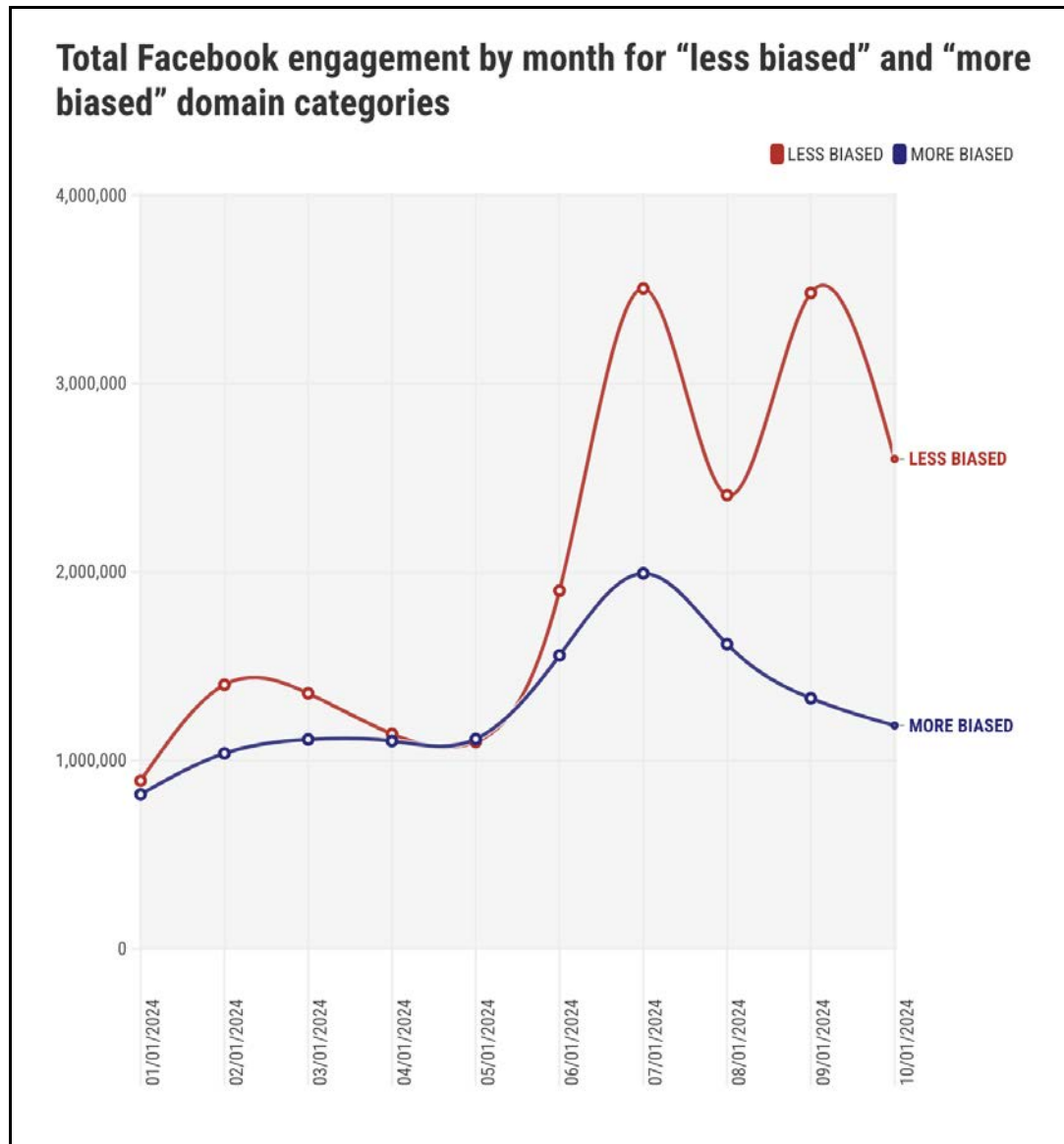


Figure 4.4 Total Facebook engagement by month for URLs classified as belonging to "less-biased" and "more-biased" media outlets. The "less-biased" category includes outlets classified by the Media Fact Check Bias (MFCB) as "right-center", "left-center," and "least biased," while "more-biased" outlets are those categorised by the MFCB as "extreme right", "extreme left", "right" and "left". Linear scale. Date range: 12th January - 31st October 2024. Source: BuzzSumo, 2024.

Another set of observations concerns the performance of the top domains. If we look at the performance of the five top media outlets per month, we discern a pattern of continuous engagement per domain (see Table 4.2). In line with the overall engagement trend, most of these outlets are mainstream domains representing traditional media such as CNN and Fox News. They maintain steady engagement throughout the year, with CNN showing consistent activity over multiple months and Fox News peaking at various points, such as over 200,000 engagements in July. The most significant surges in engagement, however, came from the *BBC* and *New York Times*, particularly in September and October 2024, where they dominate the rankings.

domain	1/1/20 24	2/1/20 24	3/1/20 24	4/1/202 4	5/1/202 4	6/1/202 4	7/1/202 4	8/1/202 4	9/1/202 4	10/1/202 4	11/1/2 024
ajc.com	0	0	0	0	25,622. 00	0	0	0	0	0	0
apnews.com	0	0	0	15,339. 00	0	0	0	0	0	0	0
bbc.com	0	18,380 .00	0	0	0	0	0	0	660,06 6.00	0	120,22 0.00
bloomberg.com	0	0	0	0	0	0	0	0	0	0	16,986. 00
cnn.com	0	40,037 .00	77,075 .00	0	39,576. 00	0	0	148,02 8.00	145,80 3.00	0	12,240. 00
dailywire.com	11,469 .00	0	21,917 .00	0	152,17 6.00	0	0	578,72 3.00	184,67 6.00	0	
desmoinesregiste r.com	0	0	0	0	0	0	0	0	0	0	68,332. 00
foxnews.com	14,312 .00	60,216 .00	52,136 .00	79,164. 00	0	0	215,03 3.00	115,26 9.00	0	133,658. 00	0
infowars.com	15,219 .00	0	31,212 .00	140,79 6.00	0	0	0	0	0	0	0
nbcnews.com	12,274 .00	0	34,148 .00	0	0	0	0	0	0	0	0
npr.org	0	21,871 .00	0	0	0	0	0	0	0	0	0
nytimes.com	0	30,225 .00	0	0	0	0	0	157,83 7.00	612,53 5.00	1,340,87 6.00	0
rollingstone.com	12,639 .00	0	0	0	0	0	0	0	0	0	0
washingtonexam iner.com	0	0	0	0	0	32,173. 00	0	0	0	0	0
washingtonpost.c om	0	0	0	12,716. 00	0	0	0	0	0	0	0
wftv.com	0	0	0	0	0	85,920. 00	178,38 7.00	0	0	118,841. 00	0
wpxi.com	0	0	0	26,639. 00	39,147. 00	0	384,81 8.00	0	331,68 3.00	182,965. 00	0
wsbtv.com	0	0	0	0	0	115,31 4.00	207,44 8.00	202,79 4.00	0	183,240. 00	0
wsoc.tv.com	0	0	0	0	36,226. 00	80,975. 00	162,74 2.00	0	0	0	33,134. 00
youtube.com	0	0	0	0	0	128,47 7.00	0	0	0	0	0

Table 4.2 Top 5 most visible domains (per month) on Facebook and their monthly total engagements. Date range: 12th January - 31st October 2024. Source: BuzzSumo, 2024.

While most highly visible sources are non-borderline or mainstream, exceptions include infowars.com and dailywire.com (see Figure 4.5). Infowars is characterised by MBFC as conspiracy/pseudoscience with low credibility and Daily Wire as ‘medium credibility’. The latter has faced criticism for sharing unverified stories and misstated facts to promote partisan narratives in the past (Snopes, 2017; Center for Countering Digital Hate, 2021).

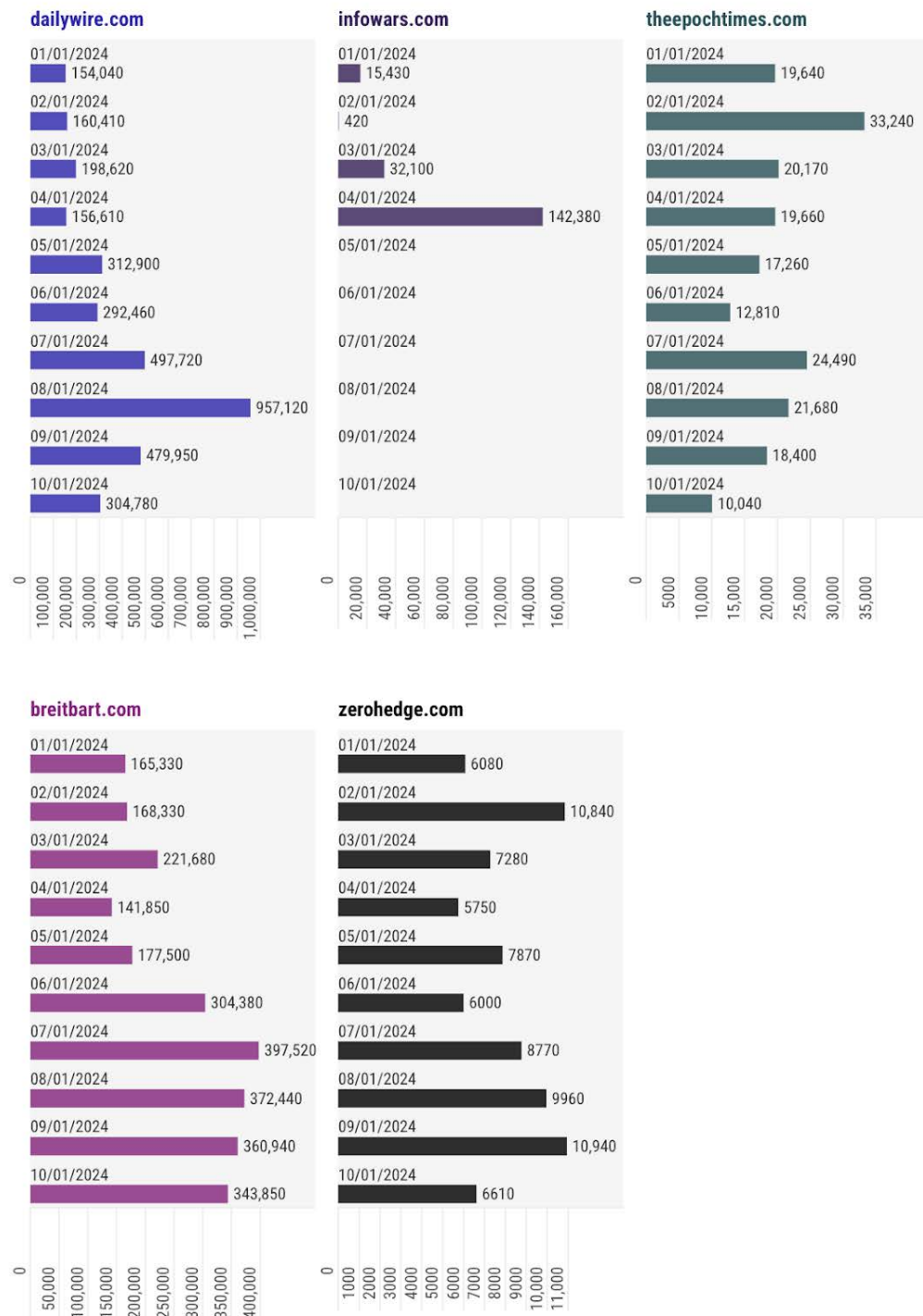


Figure 4.5 Total engagement per month for websites characterised by MBFC as extreme and/or most-biased. Logarithmic scale. Data range: 12th January - 31st October 2024. Source: BuzzSumo, 2024.

Examining well performing YouTube content

Overall engagement with YouTube-originated content grew consistently throughout the period, indicating its significance as a platform for political content (see Figure 4.6). Both candidates made strategic attempts to reach audiences through podcasters

and television programs prominently featured on the platform—Harris, for example, appeared on Saturday Night Live, while Trump was a guest on Joe Rogan’s podcast.

Among the top 100 most engaged-with YouTube videos, approximately twenty originated from channels identified as ‘borderline’ based on fact checkers and media reports, which we detail. These borderline channels collectively received approximately 200,000 engagements on Facebook, compared to circa 900,000 for media outlets such as ABC News, the Wall Street Journal, and Sky News Australia, whose links were widely shared on the platform. These borderline channels share the same political leaning, for they expressed support for Donald Trump or his policies.

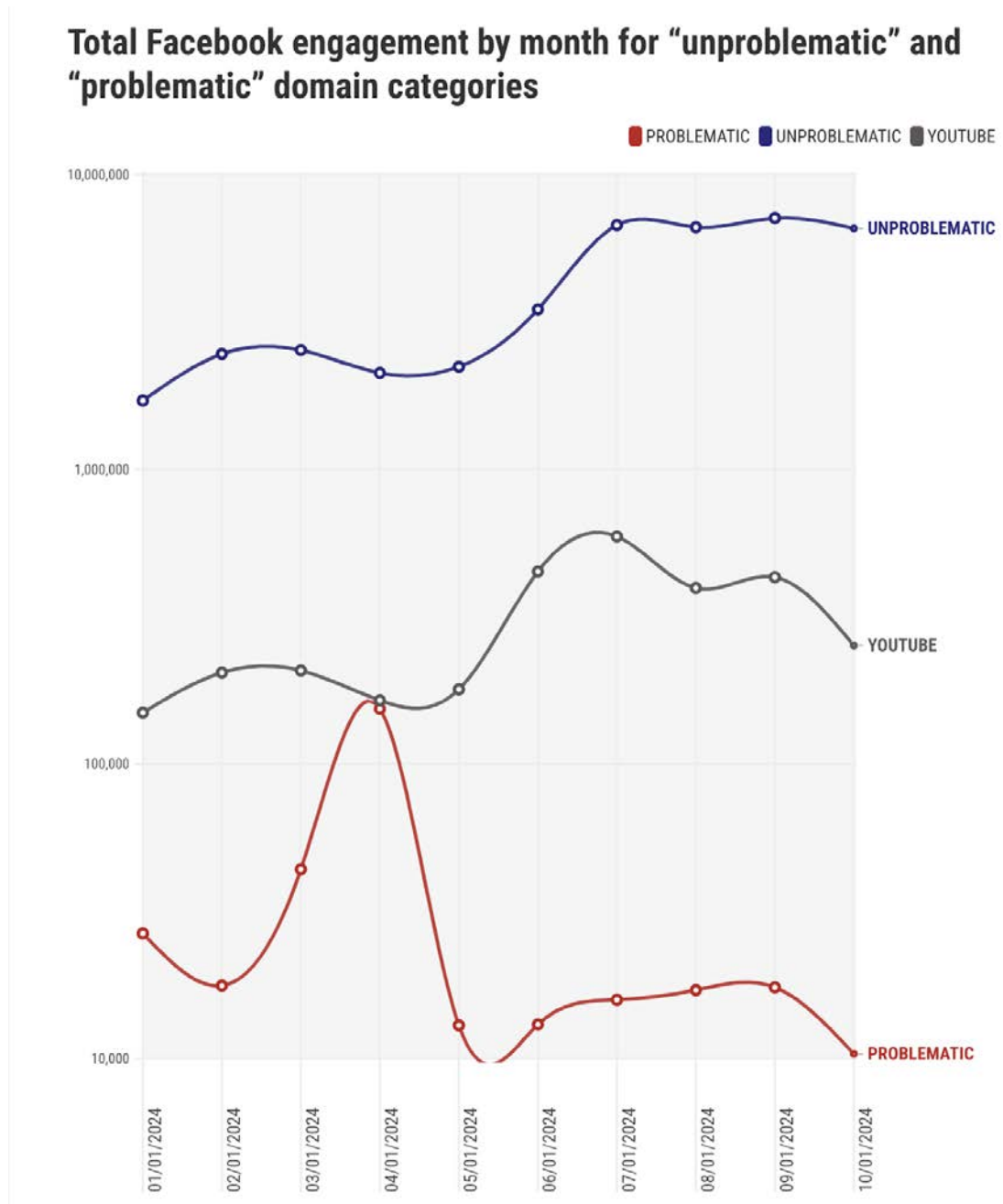


Figure 4.6 Total engagement per month for websites characterised as borderline and non-borderline compared to YouTube. Linear scale. Data range: 12th January - 31st October 2024. Source: BuzzSumo, 2024.

The three top-performing ones are Guns & Gadgets 2nd Amendment News, The Dr. Phil Podcast, and Jonathan Cahn Official, and for each we describe the most engaged-with video and the fact checking that came after it. All are YouTubers or podcasters with considerable followings. Guns & Gadget's top performing video discusses how "Biden & Harris Authorize Military To Use Lethal Force On Americans!!". Dr. Phil's is an interview with candidate Trump, who claims in their conversation that he won the 2020 election. Jonathan Cahn's concerns how the FBI was involved in an attempted assassination of Donald Trump.

Guns & Gadget's video is among those that have commented on a Department of Defense policy statement, Directive 5240.01, which the Brennan Center for Justice describe as election related rumours (Goitein & Nunn, 2024). They concern how the directive implies that the U.S. military would be deployed during the U.S. presidential elections; the rumours have been the subject of fact checking by the Center for an Informed Public at the University of Washington, debunking the claims (Center for an Informed Public, 2024).

Dr. Phil is a talk show host and television personality, who now hosts a YouTube podcast. He has faced criticism for his handling of sensitive subjects, such as discussing the effectiveness of lockdowns during the coronavirus pandemic. In his interview with Trump, where he claimed to have won the 2020 U.S. presidential election, reporters subsequently published a fact check, examining Trump's claims of fraud from mail-in ballots (Lewis, 2024).

Jonathan Cahn Official is another channel characterised as borderline in our sample. His content blends religious themes and prophecies with contemporary events, reflecting a growing trend referred to as 'conspirituality', the amalgam of conspiracy and spirituality (Harambam, 2024). Jonathan Cahn's top performing video is about how the FBI was directly involved in the attempted assassination of Donald Trump. The podcast notes refer to the episode as "open[ing] up the mystery behind the attempted assassination of Donald Trump (...) A mystery that goes back three thousand years and to the Tabernacle of God" (Jonathan Cahn Official, 2024). Numerous fact checks have appeared concerning such unsubstantiated rumours surrounding the shooting (Factcheck.org, 2024).

Regarding the URLs that remained unlabelled, the vast majority are regional (affiliate) news sites. To identify potential pink slime news, we looked for matches from Bengali's list (2019) as well as via the GATE domain analysis service (2024), which incorporates information from multiple sources to assess whether a domain or social media account is credible. It includes ten different databases, regularly updated by media monitoring professionals and fact-checkers. In our review of over 12,000 unlabelled URLs, we did not find an imposter or otherwise problematic source.

Conclusions: Depoliticisation achieved?

In response to the events of January 6, 2021, when pro-Trump supporters stormed the Capitol building in Washington, DC in an attempt to prevent the certification of the 2020 election results, Meta announced it would modify its approach to managing political content on its platforms (Frenkel & Isaac, 2024). The company laid out plans to reduce the overall presence of political material across its platforms including Facebook and Instagram.

It also committed to revising its ranking system, moving away from using engagement signals such as likes, comments, and shares to recommend political content to one that is more personalised (Meta, 2025b). This marked a notable shift in how the platform prioritizes and distributes material, political or otherwise, placing Meta on a somewhat different path from X and TikTok which automates ‘for you’ recommendations.

Given the platform's history of using engagement-driven algorithms to promote attention-grabbing content, and its role in elections, adjusting them to reduce reliance on politically charged attention (from extreme sources) had the potential to disrupt activity among Meta’s nearly 4 billion monthly user base, thus impacting its earnings. The opposite happened, however.

Meta’s advertising revenue, the cornerstone of its business, remains robust. In the three months leading up to June 2024, the company reported over \$39 billion in revenue, up from the previous year. CEO Mark Zuckerberg attributed this growth to higher ad prices and improved ways of targeting its users with commercials (Murphy, 2024).

How else to evaluate the implications of its depoliticisation efforts? Meta’s attempts to downplay political content and particularly its moderation, as we reported above, face challenges. The new emphases have had a series of consequences for both the platform and its research, which is where we would like to conclude.

The challenges include the persistence of well performing ‘right-wing commentators’, identified and publicised particularly around the 2020 U.S. election but also thereafter. As we found, in 2024 the engagement of their political posts occasionally spikes, temporarily outperforming other content. In the case of infowars.com but also in other examples of this ‘extreme right’ and ‘right’ material, the content is additionally classified as low quality and conspiracy.

From the point of view of depoliticisation, the other challenge concerns how left-leaning content outperforms its right-leaning counterparts. Center-left outlets received higher engagement than other content types by a large margin just prior to the election. Findings such as these could occasion Meta to strive to balance further the engagement and/or reach these source types garner.

Platform de-politicisation efforts also have coincided with curtailing access to research data. Assessing moderation more generally, as we also pointed out, is challenging for researchers. How well borderline content is performing on the platform is rather difficult to discern given the data access environments and Meta’s commitment to de-emphasizing certain metrics (engagement), despite previous research that has been built upon it.

Indeed, with the closure of CrowdTangle (as well as the Pages API that came before it) engagement has taken a back seat. Meta’s transparency reports, on ‘widely viewed posts’, do not contain engagement data, for that data may still show right-wing commentators as top performers, at least in bursts, as we have found.

Meta is also promoting content-agnostic approaches not only because the company reported that misinformation is undefinable but also because misinformation has a

political leaning. Content-agnostic approaches to the study of Facebook thus fit with depoliticisation.

Despite the challenges Facebook researchers face, given the company's course change towards depoliticisation, we are able to make certain observations about moderation effects. If as reported Meta has reduced the visibility of political outlets all together then it may help to explain the slow down in the performance of 'borderline' content in 2024 compared to 2020 and 2016. While they once may have been removed for being 'false news', they remain online, with an occasional burst in popularity, but overall are seemingly underperforming.

This brings us, finally, to the performance of YouTube URLs, particularly the YouTubers and channels receiving the most engagement. Arguably, the top performing borderline content, with some exceptions we found in the spikes of extreme right URLs discussed earlier, could be said to be on YouTube. All of it directly or indirectly was in support of Donald Trump's candidacy, marking the content with a clear political orientation.

When Meta announced in early 2025 that it was ending its fact checking program and lowering the standards of its content moderation, it was described as a political move, e.g., as a "surrender to the right on speech" (Newton, 2025). Whether such a characterisation holds, it nevertheless would imply that the 'right wing commentators' producing borderline content and misinformation would no longer be labelled, demoted or removed.

References

- Ananny, Mike. 2018. The partnership press: Lessons for platform-publisher collaborations as Facebook and news outlets team to fight misinformation. Tow Center for Digital Journalism. <https://doi.org/10.7916/D85B1JG9>
- Bandy, J., & Diakopoulos, N. (2023) Facebook's news feed algorithm and the 2020 US election. *Social Media + Society*, 9(3). <https://doi.org/10.1177/20563051231196898>.
- Bengani, P. (2019) Hundreds of 'pink slime' local news outlets are distributing algorithmic stories and conservative talking points, *Columbia Journalism Review*, 18 December, https://www.cjr.org/tow_center_reports/hundreds-of-pink-slime-local-news-outlets-are-distributing-algorithmic-stories-conservative-talking-points.php.
- Benkler, Y., Farris, R. & Roberts, H. (2018) *Network Propaganda*. Oxford University Press.
- Bruns, A. (2021) After the 'APIcalypse': Social media platforms and their fight against critical scholarly research, in S. Walker, D. Mercea and M. Bastos (eds.) *Disinformation and Data Lockdown on Social Platforms*, Routledge, 14-36, <https://doi.org/10.4324/9781003206972>.
- Bruns, A., Bechmann, A., Burgess, J., Chadwick, A., Clark, L. S., Dutton, W. H., ... Zimmer, M. (2018) Facebook shuts the gate after the horse has bolted, and hurts real

research in the process. *Internet Policy Review*.
<https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786>

BuzzSumo (2024) Content research tool, <https://www.buzzsumo.com>.

Center for Countering Digital Hate (2021) The toxic ten: How ten fringe publishers fuel 69% of digital climate change denial, November, <https://counterhate.com/wp-content/uploads/2021/11/211101-Toxic-Ten-Report-FINAL-V2.5.pdf>.

Center for an Informed Public (2024) Rumors rapidly spreading about reissued Department of Defense Directive 5240.01, Rapid Research Blog, 18 October, <https://www.cip.uw.edu/2024/10/18/rumors-defense-department-directive-5240-01/>.

Chen, A. (2014) The laborers who keep dick pics and beheadings out of your Facebook Feed, *Wired*, 23 October, <https://www.wired.com/2014/10/content-moderation/>.

Constine, J. (2018) Facebook will change algorithm to demote "borderline content" that almost violates policies, TechCrunch, 15 November, <https://techcrunch.com/2018/11/15/facebook-borderline-content/>.

Cox, J. (2019). Facebook quietly changes search tool used by investigators, abused by companies, Vice, 10 June, <https://www.vice.com/en/article/facebook-stops-graph-search/>.

De Keulenaar, E., Magalhães, J.C. & Ganesh, B. (2023) Modulating moderation: a history of objectionability in Twitter moderation practices, *Journal of Communication*, 73(3): 273–287, <https://doi.org/10.1093/joc/jqad015>.

DiResta, R., K. Shaffer, B. Ruppel, D. Sullivan, R. Matney, R. Fox, J. Albright & B. Johnson (2018). *The Tactics & Tropes of the Internet Research Agency*, Austin, TX: New Knowledge.

Facebook (2021a) How We're Tackling Misinformation Across Our Apps, Newsroom, 22 March, <https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps>

Facebook (2021b) Widely Viewed Posts: What People See on Facebook, Transparency Center, <https://transparency.meta.com/en-gb/data/widely-viewed-content-report/>

Facebook (2018). The Three-Part Recipe for Cleaning up Your News Feed, Newsroom, 29 May, <https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/>

Factcheck.org (2024) Issues: Attempted assassination, Annenberg Public Policy Center, University of Pennsylvania, <https://www.factcheck.org/issue/attempted-assassination/>.

Frenkel, S. & Isaac, M. (2024) How Meta distanced itself from politics, *New York Times*, 24 September, <https://www.nytimes.com/2024/09/24/technology/meta-election-politics.html>.

Gate Cloud (2024), URL Domain Analysis, University of Sheffield,
<https://cloud.gate.ac.uk/shopfront/displayItem/url-domain-analysis>.

Goitein, E. & Nunn, J. (2024) Concerns over Pentagon policy change are much ado about nothing, Brennan Center for Justice, <https://www.brennancenter.org/our-work/analysis-opinion/concerns-over-pentagon-policy-change-are-much-ado-about-nothing>.

Gotfredsen, S.G. & Dowling, K. (2024) Meta is getting rid of CrowdTangle—and its replacement isn't as transparent or accessible, *Columbia Journalism Review*, 9 July, https://www.cjr.org/tow_center/meta-is-getting-rid-of-crowdtangle.php.

Grant, N. (2024) Election falsehoods take off on YouTube as it looks the other way, *New York Times*, 31 October, <https://www.nytimes.com/2024/10/31/technology/youtube-election-conspiracy-theories-misinformation.html>.

Hao, K. (2021) The Facebook whistleblower says its algorithms are dangerous. Here's why, *MIT Technology Review*, 5 October, <https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/>.

Harambam, J. (2024). Conspirituality: An (un)happy marriage of conspiracy theories and spirituality? In *The Shape of Spirituality: The Public Significance of a New Religious Formation* (pp. 267-298). Columbia University Press.

Herrman, J. (2016) Inside Facebook's (Totally Insane, Unintentionally Gigantic, Hyperpartisan) Political-Media Machine, *New York Times*, 24 August, <https://www.nytimes.com/2016/08/28/magazine/inside-facebooks-totally-insane-unintentionally-gigantic-hyperpartisan-political-media-machine.html>.

Hindman, M. (2018) How Cambridge Analytica's Facebook targeting model really worked – according to the person who built it, *The Conversation*, 30 March, <https://theconversation.com/how-cambridge-analyticas-facebook-targeting-model-really-worked-according-to-the-person-who-built-it-94078>.

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2018). The IRA, social media and political polarization in the United States, 2012-2018. Computational Propaganda Research Project, University of Oxford.

Issenberg, S. (2015) Cruz-connected data miner aims to get inside U.S. voters' heads, *Bloomberg*, 12 November, <https://www.bloomberg.com/politics/features/2015-11-12/is-the-republican-party-s-killer-data-app-for-real->.

Jonathan Cahn Official (2024) The Mystery Behind The Trump Assassination Attempt | Jonathan Cahn Prophetic, YouTube video, <https://www.youtube.com/watch?v=5vKDJO72oVs>.

Karpf, D. (2017) Will the real psychometric targets please stand up? *Civicist*, 1 February, <https://web.archive.org/web/20180317230635/https://civichall.org/civicist/will-the-real-psychometric-targeters-please-stand-up/>.

King, G. & Persily, N (2020) A new model for industry–academic partnerships. *PS: Political Science & Politics*, 53(4):703-709. doi:10.1017/S1049096519001021

Kok, S., & Rogers, R. (2017). Rethinking migration in the digital age: Transglobalization and the Somali diaspora. *Global Networks*, 17(1), 23-46.

Lewis, K. (2024) Donald Trump's Interview with Dr. Phil—Fact-Checked, *Newsweek*, 28 August, <https://www.newsweek.com/donald-trump-doctor-phil-interview-fact-check-1945252>.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, Ties, and time: A new social network dataset using Facebook. *com. Social Networks*, 30(4), 330-342.

Merrill, J.B. and Oremus, W. (2021) Five points for anger, one for a ‘like’: How Facebook’s formula fostered rage and misinformation, *Washington Post*, 26 October, <https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/>.

Meta (2025a) Misinformation, Transparency Center, <https://transparency.meta.com/en-gb/policies/community-standards/misinformation>.

Meta (2025b) Our approach to political content, Transparency Center, <https://transparency.meta.com/en-gb/features/approach-to-political-content/>.

Meta (2024a) Content likely to violate our Community Standards, Transparency Center, <https://transparency.meta.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-likely-violating-our-community-standards>.

Meta (2024b) Content borderline to the Community Standards, Transparency Center, <https://transparency.meta.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards>.

Meta (2023a) Content Distribution Guidelines: Changes, corrections and adjustments, Transparency Center, <https://transparency.meta.com/en-gb/features/approach-to-ranking/cdgs-changes-corrections>.

Meta (2023b) Types of content that we demote, Transparency Center, 16 October, <https://transparency.meta.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/>.

Murphy, H. (2024) Meta’s shares soar after revenue growth reassures investors on AI plan, *Financial Times*, 1 August, <https://www.ft.com/content/edbe2580-0b64-4339-9be2-4b4fee46211b>.

Newton, C. (2025) Meta surrenders to the right on speech, *Platformer*, 7 January, <https://www.platformer.news/meta-fact-checking-free-speech-surrender/?ref=platformer-newsletter>.

Newton, C. (2024) How CrowdTangle predicted the future, *Platformer*, 14 March, <https://www.platformer.news/meta-crowdtangle-shutdown-dsa-platform-transparency/>.

- Newton, C. (2020) Why no one knows which stories are the most popular on Facebook, *The Verge*, 22 July, <https://www.theverge.com/interface/2020/7/22/21332774/facebook-crowdtangle-kevin-roose-nyt-tweets-interactions-reach-engagement>
- Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, 140(4), 32-48.
- Paroni, B. A. & Terenzi, M. (2024) Insights and Report from the Coordinated Sharing Behavior Detection Conference, vera.ai blog, <https://www.veraai.eu/posts/coordinated-sharing-behavior-detection-conference-2024>.
- Rathi, R. (2019) Effect of Cambridge Analytica's Facebook ads on the 2016 US presidential election, Towards Data Science, *Medium*, 13 January, <https://towardsdatascience.com/effect-of-cambridge-analyticas-facebook-ads-on-the-2016-us-presidential-election-dacb5462155d>.
- Rieder, B. (2013). Studying Facebook via data extraction: the Netvizz application. In *Proceedings of the 5th annual ACM Web Science Conference*, pp. 346-355.
- Rieder, B., Abdulla, R., Poell, T., Woltering, R., & Zack, L. (2015). Data critique and analytical opportunities for very large Facebook Pages: Lessons learned from exploring "We are all Khaled Said". *Big Data & Society*, 2(2), 2053951715614980.
- Roberts, S.T. (2016). Commercial content moderation: Digital laborers' dirty work, in S.U. Noble & B. Tynes, *The Intersectional Internet: Race, Sex, Class and Culture Online*, Peter Lang, pp. 147-160.
- Rogers, R. (2020) Research note: The scale of Facebook's problem depends upon how 'fake news' is classified, *Harvard Kennedy School Misinformation Review*, 1(6), <https://doi.org/10.37016/mr-2020-43>.
- Rogers, R. (2019). *Doing Digital Methods*. Sage.
- Roose, K. (2021) Inside Facebook's data wars, *New York Times*, 14 July, <https://www.nytimes.com/2021/07/14/technology/facebook-data.html>.
- Roose, K. (2020) What if Facebook is the real "silent majority"? *New York Times*, 27 August, <https://www.nytimes.com/2020/08/27/technology/what-if-facebook-is-the-real-silent-majority.html>.
- Rosenberg, M., Confessore, N. and Cadwalladr, C. (2018) How Trump consultants exploited the Facebook data of millions, *New York Times*, 17 March, <https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>.
- Silverman, C. (2019) Facebook Turned Off Search Features Used To Catch War Criminals, Child Predators, And Other Bad Actors, *BuzzFeed News*, 10 June, <https://www.buzzfeednews.com/article/craigsilverman/facebook-graph-search-war-crimes>.
- Silverman, C. (2016) This Analysis Shows How Viral Fake Election News Stories

- Outperformed Real News On Facebook, *Buzzfeed News*, 16 November, <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook>.
- Snopes (2017) Did Democrats refuse to stand for a Navy SEAL's widow?, *Snopes*, <https://www.snopes.com/fact-check/democrats-stand-seal-widow/>.
- Thorp, H.H. (2023) Face value, *Science*. 381: 359-359. DOI: 10.1126/science.adj8930.
- Uzogara, E.E. (2023) Democracy intercepted, *Science*. 381: 386-387. DOI: 10.1126/science.adj7023.
- Vincent, J. (2021) Facebook bans academics who researched ad transparency and misinformation on Facebook, *The Verge*, 4 August, <https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformation-nyu-ad-observatory-plugin-in>.
- Wagner, M.W. (2023). Independence by permission, *Science*. 381: 388-391. DOI: 10.1126/science.adi2430.
- Walker, S., Mercea, D., & Bastos, M. (2019). The disinformation landscape and the lockdown of social platforms. *Information, Communication & Society*, 22(11), 1531–1543. <https://doi.org/10.1080/1369118X.2019.1648536>
- Weedon, J., Nuland, W. & Stamos, A. (2017) Information Operations and Facebook, Facebook, 27 April.
- Zimmer, M. (2010) But the data are already public: On the ethics of research in Facebook, *Ethics and Information Technology*, 12, 313–325, 10.1007/s10676-010-9227-5.
- Zuckerberg, M. (2018) A Blueprint for Content Governance and Enforcement, Facebook, <https://www.facebook.com/notes/751449002072082/>.
- Zuckerman, E. (2021) I read Facebook's Widely Viewed Content Report. It's really strange, *Ethan Zuckerman blog*, 18 August, <https://ethanzuckerman.com/2021/08/18/facebooks-new-transparency-report-is-really-strange/>

5. Minors as (misused) content on Instagram

Natalia Sánchez-Querubín

Abstract

This chapter interrogates Instagram's content moderation by focusing on how minors are rendered legible and governable within its technical and algorithmic systems. It outlines four platform-specific constructions of the child: as data subject, content consumer, communication agent, and visual (misused) content. Each framing corresponds to distinct risks and moderation strategies, from algorithmic management to exposure to adult content to the platform challenges to control the circulation of

sexualized images via recommendation. The analysis traces Instagram’s shift from user-driven inputs to AI-driven verification, and from general moderation policies to targeted, datafied interventions like Teen Accounts and sensitivity controls. A case study employing a sock puppet methodology examines how Instagram’s algorithm amplifies sexualized content featuring minors, revealing persistent gaps in moderation practices.

Keywords: Instagram, children, CSAM, moderation, minors

Introduction

Instagram, launched in October 2010 and acquired by Facebook (now Meta) in 2012, began as a straightforward photo-sharing platform. In its early days, content moderation relied heavily on community guidelines, user reports of inappropriate content, and a small team of just 15 full-time employees. This team faced the task of overseeing vast amounts of user-generated content, with already over five million photos uploaded daily in 2012 (Clark, 2012).

Early moderation techniques also included banning hashtags and keywords associated with sexually explicit materials, making such content inaccessible through search. By May 2012, Instagram was also moderating pro-eating disorder content and hashtags and issuing public service announcements in response to troubling search behaviors.

Throughout the years, Instagram has grown into a global, large-scale social media platform with "2 billion monthly active users" (Zoe, 2024) and "15,000 reviewers across the globe" (Meta, n.d.[a]). Stories, Reels, Sensitive Content Controls, Direct Messages, Blurred Screens, Comment Filters, and Close Circle are some of the numerous features that have been implemented, making Instagram into a complex media environment.

Several sources provide documentation about Instagram’s features and rules for governing content and behavior. These include Instagram's Help Center and Instagram's About page, where users used to be able to consult the Community Guidelines. On November 12, 2024, however, Meta announced that Instagram’s guidelines would transition to the Transparency Center website, where they are now referred to as Community Standards. These standards apply universally across Instagram, Threads, Facebook, and Messenger (Meta, n.d. [b]).

Table 5.1 outlines the twenty-seven Community Standards, as listed in Meta’s Transparency Center. Each standard includes a policy rationale that explains the type of content prohibited, categorized by severity (Tier 1 and Tier 2). Additionally, it specifies which content requires supplementary information or is displayed with a warning screen. Certain types of content are also permitted but restricted to users aged 18 and older.

Coordinating Harm and Promoting Crime	Dangerous Organizations and Individuals	Fraud, Scams, and Deceptive Practices	Restricted Goods and Services	Violence and Incitement
---------------------------------------	---	---------------------------------------	-------------------------------	-------------------------

Adult Sexual Exploitation	Bullying and Harassment	Child Sexual Exploitation, Abuse, and Nudity	Human Exploitation	Suicide, Self-Injury, and Eating Disorders
Adult Nudity and Sexual Activity	Adult Sexual Solicitation and Sexually Explicit Language	Hateful Conduct	Privacy Violations	Violent and Graphic Content
Account Integrity	Authentic Identity Representation	Cybersecurity	Inauthentic Behavior	Memorialization
Misinformation	Spam	Third-Party Intellectual Property Infringement	Using Meta Intellectual Property and Licenses	Additional Protection of Minors
Locally Illegal Content, Products, or Services	User Requests			

Table 5.1 Meta’s Community Standards 2025. The Community Standards list the content and behavior that is forbidden or restricted on Facebook, Instagram, Messenger and Threads. Source: Meta, n.d. [b].

Instagram’s policies and content enforcement continue to evolve in response to internal and external expectations, often leading to the introduction of new features and changes. For instance, in 2016, Meta partnered with third-party fact-checking organizations to manage misinformation, and in 2020, Instagram temporarily removed the "recent" tab on hashtag pages to curb misinformation during the United States presidential elections. In January 2025, once again, Meta announced significant changes to its moderation policies and fact-checking practices, framing them as a shift towards "dramatically reducing censorship" across Facebook, Instagram, and Threads, particularly following Donald Trump’s return to the White House (Booth, 2025). Meta explained, "it’s not right that things can be said on TV or the floor of Congress, but not on our platforms" (Kaplan, 2025).

The changes are substantiated by claims about over-policing and frequent mistakes in labeling content as misinformation and on demoting politically charged topics, such as immigration and gender identity, which are often central to public debate (Kaplan, 2025). In December 2024, Meta reported that "one to two out of every 10 of these actions may have been mistakes (i.e., the content may not have actually violated our policies)" (Kaplan, 2025).

Traces of Meta’s new moderation regime are already evident in the Transparency Center, where archived versions of content and behavior policies are made available. For instance, as of 7 January 2025, the policy area previously labelled "hate speech" has been renamed "hateful conduct" and the term "protected characteristics" have been removed (Knibbs, 2025). Additionally, the misinformation policy now distinguishes between the United States and the rest of the world, splitting into two subpages. The US-specific page outlines a plan to end fact checking and move to a Community Notes program, similar to the system already implemented on X (Kaplan, 2025), even though, Zuckerberg also admits, it would mean catching "less bad stuff" (Booth, 2025). The page for the ‘rest of the world’ still states that they rely (for now) on fact-checkers "certified through the non-partisan International Fact-Checking Network

(IFCN) or, in Europe, the European Fact-Checking Standards Network (EFCSN)" (Meta, 2025a, 7 January). Community Standards, as this 'forking' of policies shows, are designed to be global but are often adjusted to local laws. For instance, when content on Instagram is reported as going against local law or policy like the GDPR, but not against Community Standards, Meta "may restrict the content's availability in the country where it is alleged to be unlawful" (Meta, n.d. [c]).

To enforce community guidelines, Meta had since 2016 followed a "remove, reduce, and inform" content moderation policy, which means, that they remove harmful content that goes against their policies, reduce the distribution of problematic content that doesn't violate policies but is borderline, and provide people with additional context. Instagram's use of artificial intelligence for content moderation has been evolving also since at least 2016, when Zuckerberg announced the introduction of AI-powered tools designed to 'alleviate' human reviewers from tasks deemed overwhelming in scale and emotionally taxing. Zuckerberg explained that "instead of making contractors the first line of defense, or resorting to reactive moderation where unsuspecting users must first flag an offensive image, AI could unlock active moderation at scale by having computers scan every image uploaded before anyone sees it" (Constine, 2016). By June 2017, Instagram had further strengthened its moderation capabilities by introducing an offensive-comment filter powered by machine learning, designed to automatically hide overtly abusive comments (Hao, 2019). Later that same year, they launched "sensitive-content screens," meaning that inappropriate images would be obfuscated with a blur and accompanied by a warning to minimize unwanted experiences in the app. Reliance on these automated moderation techniques responded to claims about the scale of social media, for which "community-scale techniques have become increasingly untenable and unconvincing" (Gillespie, 2020, 1).

Today's content moderation process still begins with artificial intelligence technology automatically and proactively detecting violating content by recognizing elements in photos or text. When deemed necessary, flagged content is sent to a review team. Meta's new content moderation policy, to be implemented as of 2025, will represent a shift from a broad application of this proactive automated moderation to a selective one. Instead of automated systems scanning for all policy violations, proactive automated moderation will focus mostly on "tackling illegal and high-severity violations, like terrorism, child sexual exploitation, drugs, fraud and scams"; for "less severe policy violations" user reports will be needed before an action is taken (Kaplan, 2025).

The reviewers in charge of handling reports undergo training and specialize on policy areas and regions, processing "content in more than 80 languages" (Meta, n.d.[a]). To ensure cultural competency, for example, "Spanish speakers from Mexico, not Spain, are hired to review content from Mexico [...] reviewers know the specific meanings of words, cultural context, [and] local celebrities" (Meta, 2022a). Until at least 2024, sensitive topics like 'hate speech' were moderated using market-specific slur lists tailored to regional nuances (Meta, 2022b). However, for areas such as adult nudity, "content review is more straightforward, so language proficiency isn't required" (Meta, 2022a).

The AI then learns from the thousands of decisions made by reviewers, making it, according to Meta, especially good at identifying repetitive violations and taking "action on new piece of content if it matches or comes very close to another piece of violating content," thus helping prevent viral incidents (Meta, 2024b). However, they also acknowledged that "reviewers tend to make better decisions than technology," especially when ambiguity or nuance is involved, which is consistent with research pointing out that automated moderation struggles at accounting for content subtlety, sarcasm, or cultural meaning (Meta, 2022c).

As this brief overview illustrates, the story of Instagram's content moderation is ongoing and does not occur in vacuum; it also extends beyond simply outlining community guidelines and the protocols that enforce them, at a given time. Content moderation is shaped by societal concerns and debates, such as those around freedom of speech, where diverse interests, actors, and risk perspectives clash. In doing so, social media content moderation grapples with the dynamics and risks introduced by user-generated content and algorithmically driven media, alongside emerging and longstanding cultural fears. As Gillespie observes, content moderation responds, "to contemporary fears—such as sexual predation and terrorism—while revisiting traditional concerns around media and public culture, including sex, obscenity, and graphic violence" (2018, p6).

Instead of aiming to write a general history of Instagram's content moderation, this chapter will examine its evolution through a specific societal concern: the protection of minors as social media users and issues of sexualization and abuse. Like other platforms, Instagram caters to both adult and underage users, requiring not only differentiated features for each group but also specialized protections for minors. Instagram's guidelines have since the app's launch and consistently through its evolution, referenced children. For example, they have been explicit about having zero tolerance for users sharing materials involving child sexual abuse; over the years the level of detail of what constitutes such content have significantly being expanded.

Certain policies and interventions apply to all users but emphasizes younger ones, such as those concerning eating disorders and bullying. Children also have a dedicated policy area like "Age-Appropriate Content," "Child Sexual Exploitation, Abuse, and Nudity," and "Additional Protection of Minors." Also, as stated in Meta's 2025 announcement, child exploitation will remain a priority of automated moderation.

This chapter underscores the platform's materiality when approaching the subject of minors as social media users—namely, it narrates Instagram's moderation story through the features, affordances, and algorithmic processes that shape how underaged users become legible to the platform. Examining Instagram's content moderation through this lens thus considers how the platform conceptualizes a "child"—as personal data, content consumer, communication agent, and as a type of content. Each of these roles introduces distinct risks: As subjects in communication, minors encounter malicious adults and grooming and as underaged content consumers, they can be recommended inappropriate, adult content. Images featuring children in bathing suits and leotards (children as a type of content) are used for luring Instagram users who are sexually attracted to children and seek to exchange illicit materials.

The following sections revisit a selection of techniques and features associated with these four perspectives, each reflecting a specific aspect of content moderation; each is also a distinct way in which minors become legible on the platform and are managed. The first section examines minors as a verifiable age, focusing on Instagram's transition from requiring users to manually input personal data to employing artificial intelligence for age recognition. The second section frames minors as content consumers—users who search for content, are recommended content, and are targeted with ads. Instagram has shifted from overarching content moderation to personal moderation features, such as offering optional sensitivity controls to users, and later, with the introduction of Teen Accounts, to having 'sensitive content' and certain aspects of 'sociality' be pre-determined for underaged users.

The third section views Instagram as a communication platform, highlighting minors as participants in chats and social networks. Here, risks include inappropriate connections with adults and sexting scandals. In response, Instagram has introduced a combination of data-driven parental supervision tools, follower limitations, and stricter privacy settings to protect minors, albeit being at odds with aspirations at online fame, also encouraged by the app.

The fourth perspective examines minors as a type of content—such as photos or videos featuring them—that may initially be shared for legitimate reasons but later become sexualized in comment sections and reposted by accounts that exploit and sexualize minors or are explicitly created for luring buyers of illicit materials. The chapter's case study builds on this fourth perspective. By employing a 'sock puppet' methodology, it identifies gaps in the moderation of sexualized images of children and explores the role of Instagram's algorithm in amplifying this content.

Minors as a verifiable age

In 2024, Meta announced the launch of Instagram Teen Accounts, which is a feature that automatically applies safety settings to accounts for users ages 13–17. It includes limiting who can contact underaged users and the content that is recommended to them.

Delivering this 'age-appropriate' social media experience will depend on preventing children under the age of 13 (or 16 depending on the country) from accessing the app and then, on differentiating adult users from those who are underaged. Thus, a first way in which a 'user' becomes legible to a platform as a 'child' is by associating them with a birth date. Or, in other words, the features' success depends on age-verification techniques. As stated by UK's children's commissioner Rachel de Souza, "ultimately, platforms cannot protect children online unless they know who the children are" (Hern, 2021).

Instagram placed the responsibility for age verification, initially, exclusively on users; agreeing with Instagram's terms and conditions presupposed that a user was over the age of 13. In 2019, Instagram began asking people to provide their birthday when signing up for their service and in 2022 rules for age verification became stricter for the UK and EU. Age can be verified by providing a birthdate, an ID, by having people vouch for one's age, or with video selfies, that through Meta's partnership with the company Yoti (Meta, 2022d) are processed using A.I. Instagram also announced in 2024 that editing a date of birth from under to over the age of 18 will require users to verify their age.

Besides age-verification through personal data, Meta has also discussed using an "adult classifier model" (Finkle et al., 2022) that employs "age signals" to determine whether an account belongs to a teenager or an adult. The logic behind the technology is that "correctly categorizing adults is important not only because it allows them to access services and features that are appropriate for them, but also because it helps mitigate risks and child safety issues that could arise on platforms where adults and teens are both present" (Finkle et al., 2022). This model is based on the idea that adults and minors exhibit different online behaviors, allowing it to identify discrepancies between the age claimed and the actions taken by the user. Meta's artificial intelligence can, for example, analyze a profile's followers; the expectation is that a teenager's network generally consists of users of a similar age. It will also review the content users engage with and look for unexpected birthday posts that differ from the age used to open the account. To comply with legislation like the Children's Online Privacy Protection Act (COPPA), these age-verification practices must focus on minimizing data collection and storage.

Being able to sort users according to age ranges is also what enables Instagram to target content to adults and minors differently. Afterall, a second way in which a child becomes legible to a platform is as a distinct type of content consumer, for whom protections are meant to be enforced both with techniques imported from legacy media to social media (e.g., information provision labels) and with medium-specific approaches such as age-appropriate, built-in sensitivity settings.

Minors as content consumers

Legacy media, such as film and television, along with more recent video-on-demand platforms, have devised techniques to protect minors as content consumers. For instance, media companies provide age ratings, which categorizes content as suitable for viewers aged 12, 15, or 18 years old. These ratings aim "to provide individuals (especially parents) with appropriate information about media products so that they can make informed decisions regarding the media consumption of their household" (Gosselt, van Hoof, and De Jong, 2012, p.337). Content labels and descriptors, another technique, make viewers aware of the presence of nudity or explicit language in films and television. Labelling typically occurs at a national or regional level by organizations such as the Dutch KijkWijzer.

Harms are also mitigated through programming and scheduling techniques. Television content meant for adults has been shown late at night and channels restrict the types of content that can be shown in proximity to each other. Technical mechanisms to protect minors also include adding passwords to devices and, in video-on-demand platforms, creating 'profiles' that delimit the content algorithms can recommend per user. In video-on-demand sites, parents can also block specific types of content from appearing in searches and review the viewing history of other users (Netflix, 2024). These media systems distribute responsibility across the content delivery value chain, including producers, platforms, users/consumers, and activities such as creation, aggregation, and distribution (ERGA, 2018).

Since its launch, Instagram's community guidelines have recognized minors as a distinct group of content consumers, highlighting the potential risks they face. In 2012, they listed five key points: "1. Post your own photos and videos. 2. Keep your clothes

on. 3. Be respectful. 4. Don't spam. 5. HAVE FUN!". With regards to nudity, Instagram further asked users to share photos and videos that were safe for all ages and "in line with [the] App Store's rating for nudity and mature content" (Instagram, 2012). Despite these warnings, news outlets reported that Instagram's small moderation team, struggled to manage the numerous uploads tagged with hashtags like #instasex and #instaboobs (Clark, 2012). The platform had "few safeguards" to "prevent children as young as 13" from encountering such content". Instagram wrestled with keeping the platform 'family friendly.'

In addition to community reporting, Instagram started banning hashtags associated with sex and nudity, preventing content from being served. Hashtag and keyword banning has, however, faced substantial hurdles, particularly with user circumvention and uneven application across languages—a challenge often described as "moderation inequality." In 2015, for instance, it was revealed that some users leveraged the Arabic word for "movie" as a hashtag to conceal sexually explicit content, bypassing the platform's predominantly English-oriented moderation. Instagram eventually acted by blocking a broader range of explicit Arabic phrases. Recent studies still report moderation inequality across digital services for languages like Maghrebi Arabic Content (Elswah, 2024) and Devanagari (Jafri and Kummer, 2024). Social media companies, researchers have claimed, "often invest less in monitoring content in non-English languages, especially in the Global South" (Jafri and Kummer, 2024) and in "smaller markets at the periphery of the advertising industry" (De Gregorio and Stremlau, 2023, p.870; Shahid, 2024; Internet Languages, 2022; Rowe, 2022).

While porn is explicitly forbidden and usually proactively moderated, Instagram continues to face frequent scrutiny over their handling of more nuanced subjects such as adult partial nudity —content that, while not illegal, may be considered inappropriate for some users but not for others. In terms of nudity, what is deemed suitable for a 13-year-old, or a 16-year-old differs significantly from what is appropriate for adults, who might also have different preferences. In other words, not all users want or need Instagram to be 'family friendly'.

While television's programming techniques segment viewers into time slots or into profiles, Instagram has had to develop medium-specific techniques, of which one example is the operationalization of 'sensitivity content' and demotion. This is content that is not explicitly banned but it is deemed potentially harmful (specially for younger users) and thus is moderated at the level of recommendation and search. For example, "content that may be sexually explicit or suggestive, such as pictures of people in see-through clothing" are allowed on the platforms "but that may not be eligible for recommendations" (Instagram, n.d. [a]). In 2021, Instagram also introduced 'sensitivity settings' which gave users the agency to determine the amount of content they might find upsetting or offensive. The latter represented a move toward personal content moderation, "a form of moderation in which users can configure or customize some aspects of their moderation preferences based on the content of posts submitted by other users" (Jhaver et al. 2023, p.3). Instagram's sensitive settings include, 'allow,' 'limit (default),' and 'limit even more.'

This rationale is currently encoded in guidelines pertaining to adult nudity (as seen below), which address different types of users age and preferences and potential content actions.

We restrict the display of nudity or sexual activity because some people in our community may be sensitive to this type of content, particularly because of their cultural background or age. Additionally, we default to removing sexual imagery to prevent the sharing of non-consensual or underage content. Restrictions on the display of sexual activity also apply to digitally created content unless it is posted for educational, humorous, or satirical purposes. Our nudity policies have become more nuanced over time. We understand that nudity can be shared for a variety of reasons, including as a form of protest, to raise awareness about a cause, or for educational or medical reasons. Where such intent is clear, we make allowances for the content [...] (Meta, n.d. [d])

Decisions about which content is classified as sensitive, and thus subject to visibly management techniques, has been shaped by controversies around topics like breasts and accusations of bias against queer people when demoting content that is deemed "inappropriate" or borderline. For instance, campaigns like "Free the Nipple" in 2016 criticized the overregulation of female bodies in nudity guidelines on social media. There was significant pushback against the removal of breastfeeding and post-mastectomy from social media, highlighting the need for more inclusive policies. So-called 'shadow-banning,' has been argued, affects queer communities and people of color disproportionately. For instance, ads promoting LGBTQ+ community health were rejected for being classified as sexual or political content. It also emerged that Instagram had banned hashtags like #gay, #lesbian, #bi, and #lesbiansofinstagram (Dickson, 2019) and photos of black women were moderated differently when compared to those of white women (Tobin, 2020).

The 2012 news article cited earlier described Instagram as being inundated with explicitly pornographic content, linked to hashtags like #instasex. Performing the same query today presents a markedly different landscape. As stated in the guidelines, some terms are suppressed entirely, while others either provide limited results or are flagged as 'sensitive.' For example, the following message appears after searching for #instasex: "We've hidden these results. Results for the term you searched for may contain sensitive content". In the "not personalized" tab, Instagram suggests accounts with keywords like 'sex' in the username, such as @instagirls100 and @insta.se.x, which also include thumbnails of women in suggestive poses. While #instasex remains hidden, the hashtag #instasexo, a variation in Spanish, results in images that suggest nudity or sexual themes without fully displaying explicit content. Searches for the term 'vagina' return artistic representations, medical illustrations, and lingerie models, whereas the term 'pussy' yields no content and shows that related hashtags are moderated. Changing one 'sensitive content setting to 'see less' could change the results.

The content that is featured, following Meta's documentation, is content that has—at least in principle—passed automatic moderation processes and has not been reported by users. Missing from the results is content already deleted. For example, between July and September 2024, an estimated 0.06% of all content views contained violating material, meaning that approximately 6 out of every 10,000 views included full adult nudity or sexual content. Of this content, 98.10% was detected and removed proactively by Meta's systems, while 1.90% was reported by users. This indicates that

the majority of such material was addressed before being flagged by users (Meta, n.d. [e]).

Instagram's launch of Teen Accounts operationalizes the concept of sensitive content settings, while also moving away from personal moderation, as a solution to the challenge of accommodating both adults and minors on the platform. While adult users can control how and which sensitive content they encounter, Teen Accounts enforce the strictest settings on minors. Sensitive content is then, as per the most recent regulations, categorized into types that require additional information or context, content that is allowed with a warning screen, or content that is restricted to users aged 18 and older.

Advertisers are also impacted by age-appropriate policies. While traditional advertisers used, for example, scheduling to prevent minors from seeing content such as ads for alcohol, Instagram leverages personalization and targeting to achieve similar goals. According to Meta's "Advertising Standards" (Meta, n.d. [f]) ads about restricted topics—such as alcohol, financial products, and weight loss services—are prohibited from being shown to users under 18, or in other words, those in Teen Accounts. Also, since 2021, minors in the EU cannot be targeted using detailed interests, behaviors, or demographics; advertisers can only include teens in their audience based on age and location. It is in this sense that Teen Accounts merge elements of legacy media's age segmentation and video-on-demand strategies by profiling users and adding content labels while also employing platform-specific techniques to manage the visibility through age-appropriate content and settings.

Minors as participants in communication

The previous section discussed children as content consumers for whom moderation techniques should curate age-appropriate social media content and minimize harms when searching for content. Instagram's Teen Accounts include pre-determined (rather than personal) content sensitivity settings, aimed for users between 13 and 17 years of age. This third section conceptualizes children instead as participants in communication through Instagram's direct messaging, launched in 2013. These 'social' affordances (Henry and Powell, 2015) often put minors in contact with malicious adult actors, making them potentially into victims of technology-facilitated sexual violence, including image-based sexual harassment and abuse. These behaviors include receiving unsolicited explicit images and requests for sexual imagery, extortion, and non-consensual distribution of sexual content. Adults also misrepresent themselves in terms of their age and location to trick teens into trusting them.

Journalists, academics, and Meta's documentation have addressed the scale of the problem. In 2024, an article in *The Guardian* stated that an estimated 100,000 children using Instagram and Facebook are subjected to online sexual harassment on a daily basis (McQue, 2024), including being sent explicit images by unknown adults. Similarly, a study conducted at Kent University in the U.K. details how children across diverse sites and age groups receive unwanted messages asking for sexual imagery by unknown adult men and are asked to 'trade' (Ringrose, Reghr, and Milne, 2021). There are also cases where minors' photos are altered using A.I. and then used to blackmail them. Police data from England and Wales likewise underscores online grooming, with Instagram implicated in a third of the recorded cases between 2017 and 2019

(Pincheta, 2019). In addition to these findings, networks of accounts that appear to be operated by minors are openly advertising self-generated child sexual abuse material for sale. Instagram's recommendation algorithms and direct messaging helps connect buyers and sellers (Thiel, Di Resta, and Stamos, 2023a).

While these reports are recent, issues of unwanted communication can be found throughout Instagram history. Already in 2012, two years after the app was launched, a blogger "recounted how a friend of his elementary school-aged daughter was contacted by an Instagram user who asked her to chat via a different social networking site. Once they were chatting, the individual "asked to see this child's privates" (Clark, 2012).

When it pertains to minors as participants in communication, Instagram's interventions have focused on privacy settings and forms of moderated sociality, for example, within the chat functions. In November 2016, for example, Instagram launched disappearing messages (also known as Vanish Mode) and in August 2021, it introduced 'Limits,' a feature that automatically hides comments and direct message requests from people who don't follow the user, or who only recently followed them (Mosseri, 2021). In 2024, Instagram enhanced the privacy of direct messaging, including, in May, an expanded 'Limits' function, allowing users to temporarily restrict comments, chats, tags, and mentions to those only coming from their Close Friends. In April, the platform began testing nudity protection features, meaning that users sending images containing nudity would receive cautionary reminders and could unsend photos. Also, Instagram aims to automatically blur photos detected as containing nudity in direct messages; they can only be revealed by tapping on the image. Another update was the introduction of 'DM filtering,' which automatically filters offensive message requests, including inappropriate words, phrases, and emojis, ensuring users "never have to see them." In July 2024, a new privacy feature allowed users to secure direct messages in a folder accessible through biometric authentication (Swipe Insight, 2024), and in October, a feature preventing people from screenshotting disappearing photos that were sent in private chats was rolled out.

Besides the options mentioned above, which also cover adult users, Teens Accounts imposed additional restrictions on minors. Sociality is also moderated by restricting "adults over 18 from starting private chats with teens they're not connected to on Instagram" and by sending "notifications encouraging [underage users] to update their settings to a more private experience" (Instagram, n.d. [b]). Additional safety notices should let a Teen Account know "when they're chatting with someone who may be based in a different country" (Instagram, 2024a). Simultaneously, Instagram aims to identify 'suspicious adults' based on how they interact with Teen Accounts; suspicious adults include "adults who have recently been blocked or reported by a young person" (Instagram, n.d. [b]).

The platform also encourages parents to play an active role in protecting their children online. Media literacy is addressed through content labels (e.g., about violence or nudity) and in Meta's Family Center one can download a 53-page guide titled "A Parent and Carer's Guide to Instagram." Meta also makes available parental controls features, which grants parents access and control over certain aspects of their child's social media experience.

The implementation of parental controls must adhere to privacy mandates, thus, Instagram is "generally forbidden by privacy laws against giving unauthorized access to someone who isn't an account holder" which includes users under 18 (Instagram, n.d. [c]). Alternatively, it offers parents access to metrics and settings, a form of datafied (instead of a content-based) supervision. Having parental 'supervision' on Instagram then means that parents can see usage metrics and block teens from using Instagram during specific time periods. They can choose between approving a teens' requests to change settings or allowing them to manage their settings themselves. And, while parents cannot see their children's feeds and searches, they can review their child's content settings.

These privacy restrictions differentiate Teen Accounts from other forms of online monitoring, such as a web browser extension that allows for blocking websites or reviewing search histories or from third-party apps like Canopy (operating in the US). The latter offers control through real-time filtering of social media, meaning that the app can block inappropriate images and videos without blocking entire sites in real-time. The person browsing will see inappropriate images on their social media feed blurred or replaced by a gray square and labeled 'filtered'.

Parental supervision, returning, now, more explicitly to the issue of direct communication, also means "receiving insights" into who Teen Accounts are chatting with — "while parents can't read their teen's messages, now they will be able to see who their teen has messaged in the past seven days" (Instagram, 2024b). In terms of social interactions, parents can also view who a Teen Account is following and who is following them back, as well as the number of messages being sent. Additionally, Instagram will alert adult users when they are interacting with a supervised account, disclosing that the people monitoring it will see their username, profile picture, and whether the supervised user has blocked them. The Canopy app, mentioned earlier, goes even a step further by identifying if an explicit image is made and alerting "the parent before it is shared" by "using advanced computing technology, including artificial intelligence and machine learning, to instantly recognize and filter out pornographic content online and on your child's smartphone camera" (Canopy, n.d.).

These moderation interventions depend on creating awareness around the value of privacy and of minimizing contact with strangers, putting them at odds with the app's attention economy. At the same time that Instagram argues for protection, it also recognizes that underage users often share images from their daily lives to gain popularity and become influencers. This pursuit necessitates being public and accessible to thousands of strangers around the world; privacy measures are then limited by the perceived value of user engagement.

"We know young people, like aspiring creators or athletes, find value in public accounts. So, teens can still opt for a public account if they choose to do so after learning more about the options. If the teen doesn't choose 'private' when signing up, we send them a notification later highlighting the benefits of a private account and reminding them to check their settings" (Instagram, 2021)

"As a creator, having a public presence is important. For teens aged 13-15, you can choose to adjust to a public account or change other settings to be less

protective with a parent's permission. For teens aged 16-17, you can change these settings without a parent's permission" (Instagram, n.d. [d])

As shown above, in 2021, the option to forgo privacy measures was left to teenagers. Since 2024, as all minors will start to be placed by default on Teen Accounts, some decisions about privacy might need parental approval. There are cases documented where teenagers (with their parents' knowledge) forgot privacy in order to promote and monetize content, including that which sexualizes them. An example is found in the *New York Times* article on Jacky Dejo, a Dutch snowboarder, bikini model, and child influencer turned social media entrepreneur (Valentino-DeVries and Heller, 2024). Dejo spoke about being 16 years old and posting and selling provocative (yet not nude) images to men through Instagram and monetizing "attention from men who are sexually interested in minors" (Valentino-DeVries and Heller, 2024); her parents were aware of her activities.

This type of risk also affects children who are officially too young to be on Instagram. As per the platform's regulations, guardians may run accounts on behalf of children younger than 13 years old if they state in the account's bio that a parent or manager is in charge and if they provide identification, when requested (Instagram, n.d. [e]). Accounts like these often promote a child's acting, modeling, or their 'influencer' careers, functioning as proxy performances of intimacy that curate distinguishable characters and brands for children. In doing so, these accounts may grant malicious actors, unintentionally but also, sometimes, complicitly, access to a child's everyday life. As it is explored in the section below, through these accounts, minors become a type of content, often misused and exploited to facilitate adult men para-sociality with children and engagement in trading of child sexual abuse material.

Minors as (misused) content

Meta's Community Standards on "Child Sexual Exploitation, Abuse, and Nudity" state that is not permitted to use the service for "content, activity, or interactions that threaten, depict, praise, support, provide instructions for, make statements of intent, admit participation in, or share links of the sexual exploitation of children (including real minors, toddlers, or babies, or non-real depictions with a human likeness, such as in art, AI-generated content, fictional characters, dolls, etc.)" (Meta, n.d.[g]).

The first explicit regulations forbidding the sexualization and misuse of content featuring children appeared in 2014, in Instagram's Community Guidelines. At that time, Instagram stated: "while we know that families use Instagram to capture and share photos of their children, we may remove images that show nude or partially nude children for safety reasons." These are "to help keep others from possibly reusing these types of images in inappropriate ways" (Instagram, n.d. f). A person can also report an account that has shared photos of their child without their permission. In 2015, the guidelines became more explicit, asserting "zero tolerance when it comes to sharing sexual content involving minors or threatening to post intimate images of others."

Since 2018, one finds an addition to the rules that explicitly prohibits "initiating unsolicited contact with minors (for example, private messages between stranger adults and minors)" and in 2020, Instagram also added clarifications banning content that "praises, supports, promotes, advocates for, provides instructions for, or encourages

participation in non-sexual child abuse" and, in 2021, the rules expanded further, prohibiting adults from using Instagram to solicit minors, as well as banning solicitation among minors and between minors and adults. The guidelines also explicitly banned using the platform "with the intention of sexualizing minors" or posting content "that supports, promotes, advocates for, or encourages participation in pedophilia unless discussed neutrally in an academic or verified health context." By 2022, the guidelines addressed "content that solicits imagery of child sexual exploitation, or nude or sexualized images or videos of children." As of May 2024, users of Teens Accounts cannot be tagged, mentioned, or used as content in Reels Remixes or Guides, by default.

Child sexual exploitation is now taking form also in images and videos generated with artificial intelligence, according to the National Center for Missing and Exploited Children (NCMEC), a US-based organization. These include "deepfake sexually explicit images or videos based on any photograph of a real child [or] CSAM depicting computer-generated children engaged in graphic sexual acts" (McQue, 2024). By 2023, Instagram regulations started to also include AI-generated content, banning depictions of children—whether real, fictional, AI-generated, or human-like, such as in art, dolls, or other non-real portrayals—if they involve nudity or sexualization.

Despite these efforts, studies continue to identify blind spots on Instagram's moderation practices. For example, a report from 2019 showed how for "pedophiles, a single hashtag opened the door to one of Instagram's seediest corners, one where images of sexually exploited children were openly traded as if they were collectibles" (Clark, 2019). Transactions were facilitated with hashtags like #dropboxlinks, with variants for those seeking boys or specific age groups. Users would find each other on Instagram and then move "the conversation to private messages where they allegedly swapped links of Dropbox folders containing the illicit imagery" (Clark, 2019) or to platforms like Kik, Telegram, and WhatsApp.

In 2023, The Wall Street Journal published a report titled, "Instagram Connects Vast Pedophile Network" (Horwitz and Blunt, 2023). They report that Instagram was permitting searching for terms associated with illegal materials, thought hashtags like #pedowhore, #preteense, #pedobait, #mnsfw ("minor not safe for work"), "cheese pizza," "seller" or "s3llr," "chapter 14," or "age 31" followed by an emoji of a reverse arrow. A pop-up screen appeared, warning users that "these results may contain images of child sexual abuse," and noting that such material cause "extreme harm" to children. The screens, however, offered two options: "Get resources" and "See results anyway." In response to questions from the reporters, Instagram removed the option to view search results for terms likely to produce illegal images.

Researchers from the Stanford's Internet Observatory, also found that platforms like Instagram have recently become home to underage users who sell or exchange sexual material depicting themselves (Thiel et al., 2023). There is also evidence that children in low-income contexts enter underage camming and sex work, potentially pressed by family members (Christensen & Woods, 2024). As it often includes ephemeral media affordances like Live and Stories, Europol refers to this as "live distant child abuse" (ref).

Reporters have also found issues of algorithmic amplification, meaning that while "pedophiles have long used the Internet" Instagram's algorithm, unlike forums and file-transfer services, "doesn't merely host these activities. Its algorithm promotes them" (Horwitz and Blunt, 2023). By setting test accounts, they notice that after viewing a single account in the network, they were offered "suggested for you" recommendations "of purported child-sex-content sellers and buyers, as well as accounts linking to off-platform content trading sites" (Horwitz and Blunt, 2023). Research into Instagram's moderation strategies, particularly in relation to eating disorders, has highlighted similar limits of hashtag banning and amplification—"once someone is embedded in a pro-ED or other similar network—through their followers, the content they share, their likes, saves, comments, clickstreams, and other mined social media data—they do not need to rely on hashtags to discover new content" (Gerrard, 2018, p.4504). In 2021, researchers from The Transparency Project, corroborated these issues, meaning that Instagram's algorithm "still recommended accounts full of disturbing images of underweight women to users who showed an interest in getting thin" (Tech Transparency Project, 2021). Detecting these types of illicit activity, they argue, requires not just reviewing user reports and automatically detecting images, but also tracking and disrupting online networks, therefore, making it difficult for users to connect with each other, find content, and recruit victims (Horwitz and Blunt, 2023).

In addition to gaps concerning search terms and algorithmic amplification, Instagram struggles to prevent people from misusing children's images. This pertains to content that raises no alarms (e.g., a child wearing a bathing suit) but which is being sexualized by other users in comment sections or accounts that aggregate these images. The Guardian reported on the issue, remarking that Instagram was "failing to remove accounts that attract hundreds of sexualized comments for posting pictures of children in swimwear or partial clothing, even after they are flagged" (Das, 2022). These images were "ruled acceptable by its automated moderation technology and remain live" (Das, 2022). These types of accounts, the report describes, are used for "breadcrumbing," meaning that they "post technically legal images but arrange to meet up online in private messaging groups to share other material" (Das, 2022).

Likewise, in February 2024, journalists from The New York Times used a combination of digital methods, interviews, and media monitoring to uncover this form of predatory behavior around 5,000 accounts belonging to child influencers, run and monetized by their mothers. These accounts "draw men sexually attracted to children, and they sometimes pay to see more" (Valentino-DeVries and Keller, 2024). The more explicit images on these accounts gather more attention, with most followers being men who leave inappropriate comments and links to sites like Telegram. These Telegram channels were for men to "openly fantasize about sexually abusing the children they follow on Instagram and extol the platform for making the images so readily available". The adults running the accounts spoke about the limits of reporting and blocking features—"if parents block too many followers' accounts in a day, Meta curtails their ability to block or follow others" (Valentino-DeVries and Keller, 2024).

Examples of breadcrumbing and misused highlight the limits of semantic approaches in content moderation. Li and Zhou (2024) describe semantic approaches as those were "a system's goal is to get the "meaning" of the content right (so that what is identified as porn, for example, actually is porn)" (Li and Zhou, 2024, p.3). The images in the

examples above do not violate community's guidelines per se. Also, the terms in the comments are often not examples of hate speech or banned language. It is the way in which images and text are used together that violate community guidelines and becomes a red flag for harm.

Alternatively, platforms sometimes also include forms of ambient moderation, with 'ambient' referring "to the pervasive information immediately surrounding the content, including user comments and engagement metrics" (Li and Zhou, 2024, p.2). From the perspective of ambient moderation, a dance video being considered "suggestive" could depend on the content of the clip and on whether it is "nested amid negative comments (such as sexually explicit ones) as opposed to positive comments (such as those praising body positivity and female empowerment)" (Li and Zhou, 2024, p.8). In such an approach, "the subjectively felt character and impact of the same content would differ and should thus be evaluated differently" and "in very rare cases of ambient-oriented moderation, content deemed appropriate can be taken down by platforms when it attracts inappropriate comments, even though nothing is ostensibly wrong with the content itself" (Li and Zhou, 2024, p.2). Li and Zhou use examples of Chinese social media to illustrate the complexities of the technique, which can also lead to censorship and penalizing content creators for unanticipated uses. Techniques such as these are more content-neutral because they target user behaviors and interactions rather than just the content. Similar techniques have been employed by, for instance, Google's Jigsaw, which aimed to detect extremist content before it leads to violence by identifying patterns in users' social media activities and interactions that suggest terrorist planning and propaganda (Gillespie, 2018). Those who searched for specific content using terms pre-defined by Jigsaw were redirected to ads and curated YouTube videos with positive, de-radicalizing content. Meta already applies certain aspects of ambient moderation and behavioral moderation, for example, when using behavioral signals terms to block suspicious adults from interacting with Teen Account.

The ephemeral features in social media represents yet another challenge for moderation. For example, TikTok's "internal investigation suggested that children were stripping on its TikTok Live service in exchange for online gifts" (Allyn, Goodman, and Kerr 2024) and similar issues are reported on Instagram. Techniques for moderating ephemeral content, include "taking sample frames from livestreams and seeing if they match hashes of known CSEA material" using machine learning classifiers to detect CSAM on live video; and employing predictive analysis of text transcriptions of live audio or user chats in livestreams (Gorwa and Thakur, p.6). However, a challenge that much livestreaming content is "new" and "thus by definition not "known" and possible to match against previously identified harmful material through hash-based techniques" (Gorwa and Thakur, 2024, p.5). Computers vision models and text analytics are used to sort through audio streams (that become Design based approaches also aim to introduce friction in the process by for example, requiring that an account already has a basis number of followers or subscribers before they can livestream. This prevents people from "spontaneously creating an account" to livestream harmful content.

Circumvention has been common as "stories can be abused through posting multiple pieces of non-violating content to portray a violating narrative. Reviewing these pieces of content individually prevents us from accurately enforcing against stories one internal document used to train moderators at Meta, dated October 2018, read" (Cox,

2018). By December 2023, Instagram noted that the company had deployed a "new automated enforcement effort" that increased their "automated deletions of Instagram Lives that contained adult nudity and sexual activity" (Meta, 2023).

This chapter has, so far, reviewed content moderation practices on Instagram, specifically focusing on the involvement of children and the associated risks linked to the dissemination of sexual abuse materials and the sexualization of minors. The review has been organized in four sections, corresponding to how children are rendered visible both as users and subjects of risk in the platform. The first pertained to age verification, facilitated through practices ranging from personal data collection to implementing artificial intelligence technology, placing responsibility both on users and platform. The second perspective framed children as content consumers, which necessitate that Instagram regulates the way content is searched, recommended, and targeted. In terms of content recommendation, Instagram has shifted from overarching content bans to personal moderation policies with sensitive content controls and recently, to age segmentation through fixed settings with Teen Accounts. The latter aims to create a differentiated social media experience for underage users and adult users.

The third perspective highlighted minors as participants in digital communication through chat functionalities, accumulating the risk of inappropriate interactions with adults. In response, Instagram has formulated stricter privacy measures, enhanced parental monitoring tools, and moderated social interactions. The fourth perspective considers minors as a category of content that is inadvertently and intentionally sexualized by adults, often being exploited to promote vendors and distributors of illicit materials. The chapter's case study builds on the fourth perspective.

Case Study

The methodology involved creating a 'sock puppet' account with an empty 'bio' section and a non-descriptive profile picture. This account was used to mimic interest in young girls. To initiate the process, the account interacted with profiles belonging to girls under the age of 13 involved in modeling or gymnastics—like those highlighted by the *New York Times* in 2024. These accounts are managed by their parents.

Scrolling through the accounts, it became evident that adult men were leaving inappropriate comments (albeit tamer than those found in less popular accounts), as exemplified in Figure 5.1. I visited the profiles of these commenters and reviewed the accounts they followed, which often included multiple child influencer accounts, as well as accounts that appear to be run by young girls with significantly lower content and follower counts. To further train Instagram's algorithm, I started visiting these new profiles and following some of them, reinforcing the 'sock puppet' account's connections within this problematic network. No keywords or hashtags were used to search for content.

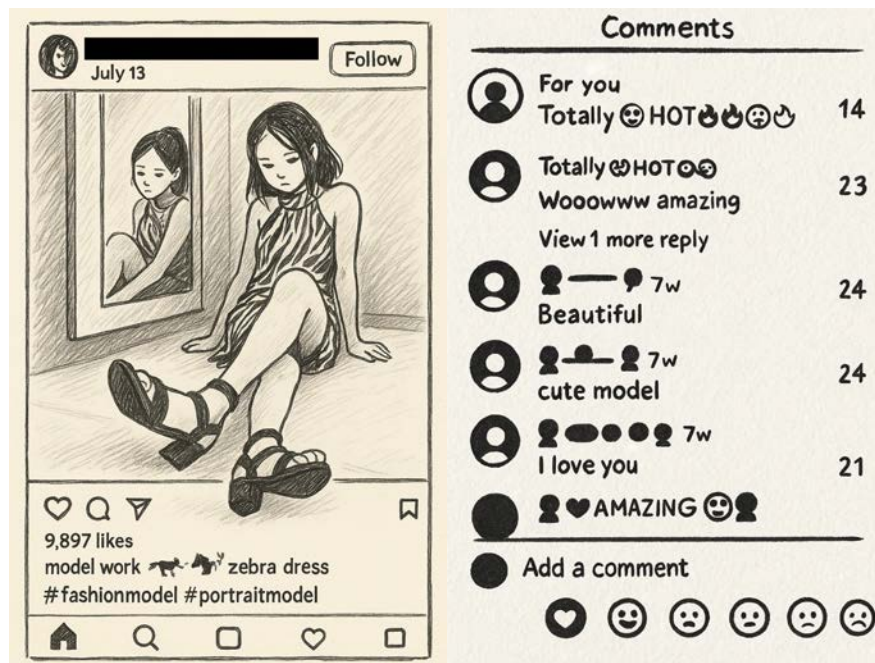


Figure 5.1 Vulgar comments left on young model's pages. Modeling work by a young girl is featured across multiple Instagram accounts, with around 150,000 followers. It is common to find comments sexualizing the girls on these types of accounts.

By following this protocol, Instagram began recommending the 'sock puppet' account additional profiles featuring young girls and accounts interested in them, treating them as part of a community of interest. This form of algorithmic amplification aligns with studies reviewed earlier in this chapter, which have observed, for example, that "recommendation algorithms inadvertently boost the network; a user who follows one seller account receives related suggestions for others" (Thiel, DiResta, and Stamos, 2023, p.6).

The case study's goals are descriptive. It aims to explore the prevalence of this type of problematic content on Instagram. Additionally, it seeks to provide descriptions and examples of how users encounter this type of content and its surrounding dynamics. What I have encountered while using the 'sock puppet' account, is organizing into two sections. The first one explores 'aggregator accounts,' which are accounts that collect images of children and permit/encourage their sexualization; these differ from the child influencer accounts described by the *New York Times* in 2024. The second section focuses on accounts that are allegedly run by children and that fit the characteristics of 'seller accounts' and are used for promoting Telegram groups that insinuate CSAM. The section looks at the emerging issues of AI-generated images of children, that while not being illicit images or examples of AI-generated CSAM, are borderline content that sexualizes children.

The case study highlights that while Instagram regulates problematic hashtags related to CSAM and deletes explicit illegal content involving minors, its recommendation algorithm continues to expose users to borderline content that sexualizes children and suggests CSAM off-platform. Types of problematic content include seller accounts, unmoderated comment sections, and specific terminology in bios, all which remain easily discoverable on the platform.

Aggregator accounts facilitate sexualizing children

In the context of this chapter, ‘aggregator accounts’ are accounts that ‘collect’ images of multiple children, sometimes even tagging them. An example is this type of account count’s bio might deceptively read, "I’m a creative and adventurous person who loves trying new things. I’m very passionate about social justice and any volunteering". The account features photos of girls in bathing suits, pajamas, and shorts. One of these images shows a girl dancing, with the video focusing on her crotch area. The caption includes hashtags unrelated to the content such as #bodypositivity and #transrights. These images, in principle, comply with Instagram’s Community Guidelines, in the sense that they are not displaying nudity or any form of illegal activity. Comments left under Image 13, however, quickly reveal the dynamics that this page encourages: "love to see you with your shorts off", "Uyy Dios mio [heart and fire emojis]" and "oh yes".

Similarly, in another image, a girl is sitting on a beach chair, eating ice cream. She is wearing a bikini and ice cream has spilled on her tights. Comments include "maybe she can use some liking to get cleaned" and "who busted so much on her?". The image is tagged with the girl's account, a child influencer. In her account page, one finds the video from which the Image originated. It is common for people to request or share links to Telegram, potentially to seller accounts and trading channels.

Several other accounts found during the research fit the profile of an aggregator, each counting with thousands of followers and views. These accounts feature videos of both adult young women and underage girls dancing, posing in bathing suits, and playing sports. Some posts featuring underaged girls invite sexualizing them, where the caption "guess my age" is overlaid on a close-up of the girl's legs and crotch area. As with the examples before, commenters are sexualizing the images. One can also speculate about techniques to obfuscate moderation present in these types of accounts, including using unrelated hashtags like #bodypositivity and combining both images of adult women and girls. These observations are also consistent with findings from the Stanford Internet Observatory's study on cross-platform dynamics in self-generated CSAM, which stated that these types of "accounts only occasionally use hashtags and keywords hinting at the nature of the content—this appears to be a strategy to attract newcomers but is deployed in limited fashion so that platforms do not detect and deactivate accounts" (Thiel, DiResta, and Stamos, 2023b, p.6).

Seller accounts and parasocial relations with minors

On Instagram, adults can run accounts on behalf of children under the age of 13, some even becoming influencers, which is, according to Instagram’s regulations, a legitimate activity. Alternatively, there are also accounts that give the impression of being run directly by children; for example, some account holders even self-identify as being 13 or even younger. It’s difficult to know if the accounts are, indeed, run by the children in the videos, adults, or if the images are stolen.

The type of account features videos of children dancing, talking to the camera, and engaging in mundane activities that offer the illusion of having direct access and communication with the minors. The latter is also consistent with previous studies that described how "seller accounts often claim to be run by children themselves and use overtly sexual handles" (Tech News Briefing, 2023). These types of accounts seem to

cater to a male adult audience that sexualizes them. This behavior is not only tolerated, as it remains un-moderated in comments sections, but also encouraged, as it becomes even more clear in accounts where the content is more explicit about this intent. Accounts like these, for example, staged scenarios, like having a child lay in bed with sentences such as "they say I am obsessed with bikers and dominant men" overlaying the image. This para-sociality is not unlike the one found in adult online sex work.

Some accounts feature abundant content; however, it is more common accounts that only include a bio (with terms like trade and pizza emoji) and a handful of videos. An example of such an account might have 3 posts and 898 followers; 0 posts and 184 followers; 1 post and 207 followers. These accounts also display a heavy reliance on transient media such as Stories. These are characteristic and are also consistent with 'seller accounts'. Moreover, accounts will frequently have content menus, promotions, or cross-platform promotion.

Cross-platform promotion, breadcrumbs, and AI

The comments sections on the various types of Instagram accounts discussed in this case study are filled with comments sharing or requesting links to Telegram groups, suggesting the presence of content involving minors. For example, posts and comments include promotions for Telegram groups with names such as "t33nle4k" accompanied by imagery insinuating underage girls or links to sites with names such as with word variations such as 'yngclub.'

Telegram, in particular, has been implicated in cases involving CSAM. According to its own data, in August alone, Telegram removed over 45,000 groups and channels related to child abuse, working in collaboration with the International Watch Foundation (IWF). Furthermore, French authorities filed 12 charges against Telegram founder Pavel Durov in August, including complicity in "distributing, offering, or making available pornographic images of minors in an organized group" and "possessing pornographic images of minors" (Brewster, 2024).

These activities take place openly and are reminiscent of previous incidents, such as the use of hashtags like #dropboxlinks to share CSAM (Child Sexual Abuse Material). In these cases, users reportedly moved the conversation to private messages where they exchanged links to Dropbox folders containing illicit material. These findings underscore a troubling pattern that reflects prior research into the role of social media platforms in facilitating the spread of such material.

Law enforcement has found examples of Telegram groups where AI-generated CSAM is being shared on the platform, chatrooms, dedicated to AI-generated exploitative materials, including using 'nudifier apps' (Brewster, 2024). While using the 'sock puppet' account for this case study, I did not find or was recommended images that would qualify as AI-generated pornography or that could insinuate sexual activity, which could be due to Instagram's moderation efforts. However, one finds content that could be considered borderline, especially in the context of it being recommended as part of a content network linked to the previous sections.

Conclusions

The case study found evidence that this behavior has remained on the platform and goes unmoderated, be that because it is not reported by the account owners or other users or fails to be detected automatically. The case study also finds that through the recommendation algorithm it is easy to encounter accounts that sexualize minors, including those that appear to be run by children younger than thirteen and that post content catering to small followings of men that, in turn, sexualize them. There are also accounts that fit the description of trader accounts, which means accounts that mainly use stories and lure users ‘off the platform,’ for example, towards telegram groups. The latter has also been implicated in scandals about underage sexual materials. The question remains- why does this material stay on the platform if it continues examples of children being sexualized?

This chapter’s case study examines Instagram’s content moderation from the perspective of how content featuring minors is misused and sexualized. Investigations, such as those conducted by the *New York Times*, have already identified Instagram as a problematic space, where accounts run on behalf of children (or allegedly by the children themselves) become targets for sexualized comments from men. These accounts also serve as hubs for users seeking to trade illegal images of minors—an alarming practice that remains often unmoderated.

The case study found evidence that this behavior persists on the platform, often going undetected or unreported by account owners and other users. Additionally, Instagram’s recommendation algorithm makes it easy to encounter accounts that sexualize minors, including those appearing to be run by children under thirteen, including so-called "trader accounts"—profiles that primarily use Instagram Stories to redirect users off the platform, often leading them to Telegram groups. Telegram itself has been implicated in scandals involving underage sexual materials, raising further concerns about how such content circulates across digital ecosystems. A central question remains: Why does this material continue to exist on Instagram, despite clear evidence of children being sexualized?

Borderline content remains a significant challenge for social media moderation. The images and accounts described in this case study fall into this category because their photos do not explicitly violate platform regulations. For example, children may be depicted wearing bathing suits, meaning the images do not contain nudity or other explicit material. Similarly, while the language used by some users is vulgar, it does not necessarily breach the app’s community guidelines. The problematic nature of this content arises from the combination of images, text, and user behavior, which collectively contribute to the sexualization of minors. Borderline content is often subjected to visibility management by making it, for example, more difficult to find. This raises critical questions: How is borderline content handled by automated moderation systems when it involves children? Should the ‘border’ of acceptability be raised in these cases?

Alternatively, Instagram already moderates behavior and social interactions, not just content, by using behavioral signals to detect malicious actors. For example, in Teen Accounts, adult users are banned from contacting minors who are not within their network. Instagram also makes it more challenging for new users to start live

broadcasts. In the case study's context, certain behavioral signals could indicate trader accounts, such as profiles that use a child's photo as a bio picture, have thousands of followers but only three to four videos, and primarily rely on the live feature. Occasionally, these accounts also include links to Telegram in their bio, further raising concerns.

This raises key questions about Instagram's current moderation strategies. Are these behavioral signals actively employed to moderate and detect borderline content? Are they used to monitor links to external sites? Instagram has already forbidden direct links to sites like OnlyFans, as they might constitute a form of solicitation. However, content creators often circumvent these restrictions by using link aggregators such as Linktree. If Instagram moderates certain external links, are links to Telegram, when originating from accounts with suspicious behavioral signals, also being subject to moderation?

This case study is situated within an overview of Instagram's evolution from the perspective of content moderation targeting minors and the various ways in which the platform aims to provide a safe experience. When viewed through the distinct features of Instagram, the goal of child protection takes on different dimensions. When minors are considered both content consumers and active participants in communication, Instagram has a responsibility to prevent inappropriate content and actors from reaching them. Moderation interventions at the search level, the introduction of fixed sensitivity settings, advanced age verification, and teen account targeting all aim to achieve this, hopefully preventing scandals related to extortion and grooming.

In terms of images of children, no illegal material was found within the context of this case study, and Instagram reports interventions when such content is detected. However, the challenge remains with borderline content, where minors are subject to sexualization and where the abundance of such material makes moderation particularly difficult. Will Meta's recent policy changes, which prioritize addressing serious infractions over minor violations—including cases of child exploitation—effectively close this blind spot? While this content differs from explicitly illegal material (such as child pornography), where does the sexualization of minors fall on the scale of urgency for content moderation?

References

Allyn, B., Goodman, S., & Kerr, D. (2014, October 11). TikTok executives know about app's effect on teens, lawsuit documents allege. *NPR*.
<https://www.npr.org/2024/10/11/g-s1-27676/tiktok-redacted-documents-in-teen-safety-lawsuit-revealed>

Booth, R. (2025, January 7). Meta to get rid of factcheckers and recommend more political content. *The Guardian*.
<https://www.theguardian.com/technology/2025/jan/07/meta-facebook-instagram-threads-mark-zuckerberg-remove-fact-checkers-recommend-political-content>

Brewster, T. (2024, August 27). The wiretap: Telegram is full of ai-generated and real child abuse photos—but is that enough to arrest a CEO? *Forbes*.

<https://www.forbes.com/sites/thomasbrewster/2024/08/27/ai-generated-child-abuse-on-telegram-pavel-durov-arrested/>

Canopy. (n.d.). Meet a safer Internet with AI-powered protection. Canopy. Retrieved December 20, 2024, <https://canopy.us/>

Christensen, L.S., & Woods, J. (2024). 'It's like POOF and it's gone': The live streaming of child sexual abuse. *Sexuality & Culture*, 28, 1467–1481
<https://doi.org/10.1007/s12119-023-10186-9>

Clark, B. (2019, January 9). In Instagram's darkest corner, all it takes is a hashtag to uncover images of child sexual abuse. The Next Web. <https://thenextweb.com/news/in-instagram-darkest-corner-all-it-takes-is-a-hashtag-to-uncover-a-mountain-of-child-sex-abuse-imagery>

Clark, E. (2012, August 31). The dark side of Instagram: Thousands creating X-rated 'Instaporn' with popular photo app. *Mail Online*.
<https://www.dailymail.co.uk/news/article-2196387/The-Instagram-Thousands-create-X-rated-Instaporn-popular-photo-app.html>

Constine, J. (2016, May 31). Terminating abuse. *TechCrunch*.
<https://techcrunch.com/2016/05/31/terminating-abuse/>

Cox, J. (2018, December 31). Leaked documents show how Instagram polices stories. *Vice*. <https://www.vice.com/en/article/leaked-documents-instagram-polices-stories-content-moderation/>

Das, S. (2022, April 22). Instagram under fire over sexualized child images. *The Guardian*, <https://www.theguardian.com/society/2022/apr/17/instagram-under-fire-over-sexualised-child-images>

De Gregorio, G. & Stremlau, N. (2023). Inequalities and content moderation. *Global Policy*, 14, 870–879. <https://doi.org/10.1111/1758-5899.13243>

Dickson, EJ. (2019, July 11). Why did Instagram confuse these ads featuring LGBTQ people for escort ads? *Rolling Stone*. <https://www.rollingstone.com/culture/culture-features/instagram-transgender-sex-workers-857667/>

Elswah, M. (2024, September). Moderating Maghrebi Arabic content on social media. The Center for Democracy & Technology. <https://cdt.org/insights/moderating-maghrebi-arabic-contenton-social-media/>

European Regulators Group for Audiovisual Media Services [ERGA]. (2018, March 1). ERGA activity report on Protecting Children in Audiovisual Media Services- Current and Future Measures. <https://digital-strategy.ec.europa.eu/en/library/erga-activity-report-protecting-children-audiovisual-media-services-current-and-future-measures>

Finkle, E., Luo, Sheng, Agarwal, C., & Fryer, D. (2022, June 22). How Meta uses AI

to better understand people's ages on our platforms. Meta.
<https://tech.facebook.com/artificial-intelligence/2022/6/adult-classifier/>

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12), 4492-4511.
<https://doi.org/10.1177/1461444818776611>

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press,
10.12987/9780300235029

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2). <https://doi.org/10.1177/2053951720943234>

Gorwa, R. and Thakur, D. (2024, November 21). Real time threats: analysis of trust and safety practices for child sexual exploitation and abuse (CSEA) prevention on livestreaming platforms. Center for Democracy and Technology.
<https://cdt.org/insights/real-time-threats-analysis-of-trust-and-safety-practices-for-child-sexual-exploitation-and-abuse-csea-prevention-on-livestreaming-platforms/>

Gosselt, J., Van Hoof, J., & De Jong, M. (2012). Media rating systems: do they work? Shop floor compliance with age restrictions in The Netherlands. *Mass Communication and Society*, 15(3), 335–359. <https://doi.org/10.1080/15205436.2011.558803>

Hao, K. (2019, July 9). Instagram is using AI to stop people posting abusive comments. *Technology Review*.
<https://www.technologyreview.com/2019/07/09/65590/instagram-is-using-ai-to-stop-people-posting-abusive-comments/>

Henry, N. & Powell, A. (2015). Beyond the "sext": Technology-facilitated sexual violence and harassment against adult women. *Australian & New Zealand Journal of Criminology*, 48(1), 104-118 <https://doi.org/10.1177/0004865814524218>

Hern, A. (2017, May 3). Facebook Live: Zuckerberg adds 3,000 moderators after murders. *The Guardian*.
<https://www.theguardian.com/technology/2017/may/03/facebook-live-zuckerberg-adds-3000-moderators-murders>

Hern, A. (2021, 31 August). Instagram to require all users to enter birthdate. *The Guardian*. <https://www.theguardian.com/technology/2021/aug/31/instagram-to-require-all-users-to-enter-birthdate>

Horwitz, J. & Blunt, K. (2023, June 7). Instagram connects vast pedophile network. *The Wall Street Journal*. <https://www.wsj.com/articles/instagram-vast-pedophile-network-4ab7189>

Instagram. (2021, March 17). Continuing to make Instagram safer for the youngest members of our community. Instagram.
<https://about.instagram.com/blog/announcements/continuing-to-make-instagram-safer-for-the-youngest-members-of-our-community>

Instagram. (2024a, 22 October). Our new education campaign to help protect teens from sextortion scams. <https://about.instagram.com/blog/announcements/campaign-against-teen-sextortion>

Instagram. (2024b, September 17). Introducing Instagram Teen Accounts: Built-in protections for teens, peace of mind for parents. Instagram. <https://about.fb.com/news/2024/09/instagram-teen-accounts/>

Instagram. (n.d.) (a). About sensitive content control on Instagram. Instagram. Retrieved December 20, 2024, from https://help.instagram.com/1055538028699165/?helpref=uf_share

Instagram. (n.d.) (b). About Instagram teen privacy and safety settings. Instagram. Retrieved December 20, 2024, from https://help.instagram.com/3237561506542117/?helpref=uf_share

Instagram. (n.d.) (c). Accessing a teen's Instagram account. Instagram. Retrieved December 20, 2024, from https://help.instagram.com/359897877431452/?helpref=uf_share

Instagram. (n.d.) (d). Instagram Teen Accounts for creators. Instagram. Retrieved December 20, 2024, from https://help.instagram.com/1540058036636890/?helpref=uf_share

Instagram. (n.d.) (e). Report a child under 13 on Instagram. Instagram. Retrieved December 20, 2024, from https://help.instagram.com/517920941588885/?helpref=uf_share

Instagram. (n.d.) (f). Why an image of your child might be removed from Instagram. Instagram. Retrieved December 20, 2024, from https://help.instagram.com/242592952606350/?helpref=uf_share

Internet Languages. (2022). The state of the Internet's languages report. <https://internetlanguages.org/media/pdf-summary/EN-STIL-SummaryReport.pdf>

Jafri, A. and Kummar, V. (2024, November 28). How content moderation is failing as kids lap up culture of 'gangster' violence online. *The Wire*. <https://thewire.in/society/content-moderation-social-media-spotify-jio-saavn-youtube>

Jhaver, S., Zhang, A. Q., Chen, Q. Z., Natarajan, N., Wang, R., & Zhang, A. X. (2023). Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction CSCW2*, 7, 1-33. <https://dl.acm.org/doi/abs/10.1145/3610080>

Kaplan, J. (2025, 7 January). More speech and fewer mistakes. Meta Newsroom. <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes>

- Knibbs, K. (2025, January 7). Meta now lets users say gay and trans people have 'mental illnesses.' *Wired*. <https://www.wired.com/story/meta-immigration-gender-policies-change/>
- Li, L., & Zhou, K. (2024). When content moderation is not about content: How Chinese social media platforms moderate content and why it matters. *New Media & Society*, 0(0), 1-17 <https://doi.org/10.1177/14614448241263933>
- McQue, K. (2024, January 18). Meta documents show 100,000 children sexually harassed daily on its platforms. *The Guardian*. <https://www.theguardian.com/technology/2024/jan/18/instagram-facebook-child-sexual-harassment>
- Meta. (2022a, January 19). The people behind Meta's review teams. <https://transparency.meta.com/en-gb/enforcement/detecting-violations/people-behind-our-review-teams/>
- Meta. (2022b, August 12). How we create and use market-specific slur lists. <https://transparency.meta.com/en-gb/enforcement/taking-action/how-we-create-and-use-market-slurs/>
- Meta. (2022c, January 19). How Technology helps prioritize review. <https://transparency.meta.com/enforcement/detecting-violations/technology-helps-prioritize-review/>
- Meta. (2022d, 23 June). Introducing new ways to verify age on Instagram. <https://about.fb.com/news/2022/06/new-ways-to-verify-age-on-instagram/>
- Meta. (2023, December 1). Our work to fight online predators. <https://about.fb.com/news/2023/12/combating-online-predators/>
- Meta. (2024b, November 12). How enforcement technology works. <https://transparency.meta.com/enforcement/detecting-violations/how-enforcement-technology-works/>
- Meta. (2025a, 7 January). How fact-checking works. <https://transparency.meta.com/en-gb/features/how-fact-checking-works/>
- Meta. (2025a, April 7). Our approach to misinformation. <https://transparency.meta.com/en-gb/features/approach-to-misinformation/>
- Meta. (n.d.) (a). Detecting violations. Retrieved December 24, 2024, from <https://transparency.meta.com/enforcement/detecting-violations/>
- Meta. (n.d.) (b). Community Standards. Retrieved February 28, 2025, from <https://transparency.meta.com/en-gb/policies/community-standards/>
- Meta. (n.d.) (c). Content Restrictions Based on Local Law. Retrieved December 20, 2024, from <https://transparency.meta.com/reports/content-restrictions/>

Meta. (n.d.) (d). Adult nudity and sexual activity. Retrieved December 20, 2024, from <https://transparency.meta.com/en-gb/policies/community-standards/adult-nudity-sexual-activity/>

Meta. (n.d.) (e). Report: Adult nudity and sexual activity. Retrieved December 20, 2024, from <https://transparency.meta.com/reports/community-standards-enforcement/adult-nudity-and-sexual-activity/facebook/>

Meta. (n.d.) (f). Introduction to the advertising standards. Retrieved December 20, 2024, from https://transparency.meta.com/policies/ad-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fpolicies_center%2Fads

Meta. (n.d.) (g). Child sexual exploitation, abuse and nudity. Retrieved January, 2025, from <https://transparency.meta.com/en-gb/policies/community-standards/child-sexual-exploitation-abuse-nudity/>

Mosseri, A. (2021, 10 August). Introducing new ways to protect our community from abuse. Instagram. <https://about.instagram.com/blog/announcements/introducing-new-ways-to-protect-our-community-from-abuse>

Pincheta, R. (2019, March 1). Instagram is a leading social media platform for child grooming. *CNN*. <https://edition.cnn.com/2019/03/01/uk/nspcc-grooming-social-media-report-scli-gbr-intl/index.html>

Ringrose, J., Rehr, K. & Milne, B. (2021). Understanding and combatting youth experiences of image-based sexual harassment and abuse. University of Kent. <https://www.kent.ac.uk/news/society/30568/young-peoples-rates-of-reporting-online-sexual-harassment-and-abuse-shockingly-low>

Rowe, J. (2022, March 2). Marginalized languages and the content moderation challenge. Global Partners Digital. <https://www.gp-digital.org/marginalised-languages-and-the-content-moderation-challenge/>

Shahid, F. (2024, August 26). Colonialism in content moderation research: The struggles of scholars in the majority world. Center for Democracy and Technology. <https://cdt.org/insights/colonialism-in-content-moderation-research-the-struggles-of-scholars-in-the-majority-world/>

Swipe Insight. (2024, July 26). Instagram is rolling out 'lock and hide chat' for enhanced dm privacy. https://web.swipeinsight.app/posts/instagram-is-rolling-out-lock-and-hide-chat-for-enhanced-dm-privacy-9057?utm_source=chatgpt.com

Tech News Briefing. (2023, June 8). How Instagram's algorithm connects and promotes pedophile networks, *Wall Street Journal*, <https://www.wsj.com/podcasts/tech-news-briefing/how-instagrams-algorithm-connects-and-promotes-pedophile-network/a683c0b4-2e6f-4661-9973-10bd455db895>

Tech Transparency Project. (2021, December 8). 'Thinstagram': Instagram's algorithm fuels eating disorder epidemic.

<https://www.techtransparencyproject.org/articles/thinstagram-instagrams-algorithm-fuels-eating-disorder-epidemic>

Thiel, D., DiResta, R., & Stamos, A. (2023a, June 6). Addressing the distribution of illicit sexual content by minors online. Stanford Cyber Policy Center.
<https://cyber.fsi.stanford.edu/news/addressing-distribution-illicit-sexual-content-minors-online>

Thiel, D., DiResta, R., & Stamos, A. (2023b, June 6). Cross-platform dynamics of self-generated CSAM. Stanford Cyber Policy Center.
<https://stacks.stanford.edu/file/druid:jd797tp7663/20230606-sio-sg-csam-report.pdf>

Tobin, J. (2020, October 29). Instagram changes nudity policy after model points out bias. Range Finder Online. <https://rangefinderonline.com/news-features/industry-news/instagram-changes-nudity-policy-after-model-points-out-bias/>

Valentino-DeVries, J. & Kelle, M. H. (2024, November 10). She was a child Instagram influencer. Her fans were grown men. *The Guardian*,
<https://www.nytimes.com/2024/11/10/us/child-influencer.html>

Zote, J. (2025, Feb 21). 26 Instagram stats you need to know for 2025. Sprout Social.
<https://sproutsocial.com/insights/instagram-stats/>

6. Malicious earworms and useful memes: How the far-right surfs TikTok's audio trends

Marloes Geboers and Marcus Bösch

Abstract

With its easy-to-use features of remix TikTok is the designated platform for meme-making and dissemination. Creative combinations of video, emoji, and filters allow for an endless stream of memes and trends animated by sound. From the get go, the platform focused its moderation on upholding (physical) safety, hence investing in the detection of harmful challenges. In response to the DSA, TikTok implemented opt-outs for personalized feeds and features allowing users to report illegal content. At the same time, the platform remains subject to scrutiny. Centering on the role of sound and its intersections with ambiguous memes, the presented research probed right-wing extremist formations relating to the 2024 German state elections. The analysis evidences how TikTok's sound infrastructure affords a sustained presence of xenophobic content that is often cloaked through vernacular modes of communication. These cloaking practices benefit from a sound infrastructure that affords the ongoing posting of user-generated sounds that instantly spread through the 'use-this-sound button'. Importantly, these sounds are often not *clearly* recognizable as 'networkers' of extremist content. Songs that do contain hateful lyrics are not eligible for personalized feeds, however, they remain online where they profit from intersecting with benign meme trends, rendering them visible in search results.

Keywords: TikTok, political extremism, propaganda, sonic social media, transparency, audio memes

Introduction

Toward the end of 2024, TikTok released updates to its community guidelines, accompanied by a brief acknowledgment of the difficulties that the short-form video format and the memetic logic built into the platform extend to moderation practices. The conjunction of users uploading their audios and the easy replication of sound afforded through the 'use this sound' button construct an environment where right-wing extremist actors easily surf on the vibes and rhythms of hateful audio memes. While sounds are deplatformed, they get ample time to prime or inspire others into re-uploading spinoffs of the same or an adjacent audio, rendering moderation especially cumbersome. As per TikTok's public announcement: *"Hate speech can be conveyed through any form of expression, including images, text, audio, cartoons, memes, objects, gestures, and symbols, some of which won't always be obvious."* The use of technology and human moderators to detect and remove such content is stipulated, mentioning for the first time how indirect attacks through jokes, memes, and audio trends will also be scrutinized by TikTok's moderation team. This includes any coded messages used by *"hate groups to communicate with each other without appearing hateful, such as code words, symbols, or audio trends."* This is quite an ambitious statement, as coded language and so-called algospeak are notoriously fluid, polysemous, and subject to change, and therefore hard to moderate (Steen et al., 2023).

Our empirical analyses will make distinctions between overt hate speech across the several modalities of posts, and what we deem ‘cloaked hate’ to assess TikTok’s moderation performance during the runup and aftermath of state elections held in Thüringen and Sachsen Anhalt on September 1, 2024. The conjunction of these elections and a fatal knife attack in Solingen on August 23 of that year spurred support for far-right extremist ideas embedded or otherwise linked to the political campaign of the Alternative für Deutschland (AfD) party. Especially in younger audiences, extremist ideas were shared and amplified by accounts on TikTok that participated in spreading far-right propagandist messages. In a report by the ISD (Institute for Strategic Dialogue, 2024) it was found that Neo-Nazis and white supremacists were sharing Hitler-related propaganda.

Failing transparency under the DSA

The presented research addresses how TikTok’s sound architecture affords the proliferation and obfuscation of problematic content. Our research identifies loopholes that are exploited by malign actors, either surfing on benign audio trends or linking racist songs to popular meme templates. Moreover, we show how, even after time intervals of several weeks, problematic posts are only marginally moderated. Here, it is important to stress that we could merely determine if posts were no longer available, set to private view, or we could see that the account was no longer online. This does not necessarily imply platform moderation efforts, as users can take down posts and accounts as well. The platform only communicates that content or an account is no longer ‘available. Under the DSA, TikTok publishes so-called Statements of Reasons, but in these vast spreadsheet documents, there is no identifier of the posts that were taken down by the platform. This effectively obstructs researchers from cross-checking whether unavailable posts were actually deplatformed.

During the COVID-19 pandemic and thereafter TikTok grew into a billion-user app with at the time of writing approximately 1.58 billion monthly active users worldwide. The multimodality of video posts affords easy disguising of messages as meaning can reside in a myriad of combinations of filters, emoji, hashtags, sounds, and gestures. This ‘in-betweenness’ makes far-right posts highly ambiguous, and therefore harder to read across vernacular cultures, after all, who actually knows that videos of motorbikes from the GDR-based brand Simson can signal a far-right audio meme that carries blatant racist messages in its lyrics? TikTok’s nuanced nature significantly hampers moderation because: *"viral content that multiplies through memetic means with enough speed, anonymity, randomness, as well as some kind of logical order [...] can look organic or at least "semi-organic", whether or not it was originally a manipulated information or propaganda campaign"* (Galip, 2023, p. 101). Galip rightly points to this kind of ‘covertness’ as a phenomenon that benefits states and actors who hide behind a ‘cloak’ (Daniels, 2009) that gets drawn up as soon as users start mimicking a far-right TikTok meme. A targeted propagandist state-led injection of a racist message on this memetic platform, will, adopted by others, quickly become disconnected from its original instigator(s), allowing extremist messages to travel further through shapeshifting into a myriad of assembled posts, that repeat but also deviate in visual style or sound etc. On top of the affordances of multimodality, TikTok boasts a communicative infrastructure that entangles posts on various levels. Through encountering a ‘regular’ AfD (Alternative for Germany party) post displaying the speech of a politician calling for remigration (typically something a platform would

leave untouched under the premise of free speech) a hashtag or a sound embedded in that ‘benign post’ links the artifact to wider, much more extremist, ephemera. Driven by the vibes of a sound, users can easily replicate the sound of someone else’s post through the click of a button, and if problematic sounds are banned, there are thousands of potential users that can repost a sound under another name.

Rhizomatic soundscapes: TikTok’s distributed audio infrastructure

While the app’s short video format has permeated the social media landscape far beyond TikTok, the communicative architecture of said app is more or less ‘unique’. The platform hosts a sound library with platform-listed sounds, however, the majority of users either use their own (pre)recorded sound or, even more ubiquitous: replicate someone else’s sound. On the first occasion, users upload a video with a blend of audios that through this blend are ‘unique’. These sounds contain voiceovers, music, speech, or sounds. In other cases the audio itself can be an exact copy of an existing song or sound, that is simply not taken from the library or not present as a listed sound. Both ‘blended audio’, as well as unlisted sounds are automatically registered as an ‘Original Sound - [user name]’ by the platform once a post is uploaded. If ‘catchy enough’, sounds take off through other users who encounter the post and use the ‘use this sound’ button. This effectively indexes their posts on a “sound page”. When anyone clicks on the hyperlinked sound, they arrive at a collection of posts that use the same sound. A rather ubiquitous commenting practice is to ask creators the name of the song that was used. Such comments point to a desire to remix the sound with other audio such as problematic speech, birthing new soundscapes that are tethered to the original. In this way, sounds create spaces where actors find each other based on a shared sentiment (Papacharissi, 2015) and where they connect through e.g. participating via the ‘use this sound’ button or through remix, propelling fascist messages forward.

The networking affordances of sounds surpass hashtags in their intricate workings: sounds not merely gather posts that use the sound, there are also multiple versions of the same sound or song that are audibly similar, but that were uploaded by dispersed accounts and replicated by others. This creates distributed ‘niches’ or rhizomatic instances of what Geboers & Pilipets (2024) dub ‘soundscapes’: environments of creative assembly where various more or less homogeneous narratives can be hosted, ‘held together’ by the affective connotations that the sound in question evokes in its audiences. One can imagine how such distributed instances of (almost) the same sound severely hamper the platform’s sound moderation. Once a sound comes to stand in for a hateful ideology, problematic accounts proliferating that ideology might be moderated, shadowbanned, or deplatformed, potentially also deplatforming any original sounds uploaded by those accounts, however, other instances of the same or a similar sound remain available. These practices include reusing sounds for coordinated campaigns, creating audio meme templates for rapid amplification and distribution, and deleting the original sounds to conceal the orchestrators’ identities (Bösch & Divon, 2024).

The presented study takes as its focus the networked linkages between the AfD campaign and extremist rightwing niches on TikTok, ‘held together’ and networked through sounds that hijack popular songs or songs that are blatantly problematic in their contents but that latch onto meme templates to attain visibility. Attaching the

political message of ‘remigration’ not merely to hashtags like #heimatliebeistkeinverbrechen (Love of home is not a crime) and #deutschejugendvoran (German youth forward), but also to sounds such as Gigi d’Agostino’s *l’Amour Toujours* (the ‘Gigi-song’ hereafter), created sticky associations between upbeat club songs on the one hand, and the far-right political message of ‘die Blauen’ (blue is the color of the AfD with fans posting blue heart emojis) on the other hand. The online dynamic of co-opting the Gigi-song spilled over into a physical event where a group of young Germans on the terrace of Pony Club on the German island of Sylt, chanted the xenophobic slogan "Germany for the Germans, foreigners out" to the melody of the Gigi-song on Pentecost Sunday 2024. News media amplified the event, inadvertently recreating a new sound on the platform, to which, in turn, several TikTokers ‘responded’ by either replicating the sound through the ‘use this sound feature’ or by adding their own recorded file, establishing multiple soundscapes that all host the party crowd of Sylt. Alongside many malicious uses of the classic version of the Gigi-song, remixes of Gigi d’Agostino—slowed, sped up, distorted, or otherwise tampered with—also connected those aiming to proliferate the AfD-led message of remigration. Alongside the Gigi-song, our study was able to detect linkages between far-right accounts and other trending sounds—such as the ‘*Kiss me*’ sped-up version and ‘*Around the world*’—as well as sounds that harbor deeply problematic and violent content within their lyrics.

TikTok & Content Moderation

From its launch in 2017, TikTok’s community guidelines addressed hate speech and hateful behavior as prone to platform moderation. Despite this, the platform initially emphasized monitoring and deplatforming posts that are potentially risky for users’ physical safety, with increasing attention to mental health throughout the years (Christin et al., 2024, OpenTerms Archive). What was deemed hate speech and hateful behavior was further specified after TikTok’s surge in popularity in 2020, where the platform had to deal with emergent problematic phenomena exemplified by the rise in so-called QAnon accounts that propelled a popular conspiracy theory fuelled by pandemic-induced uncertainties. In October 2020, the platform announced a significant update to its policies on hate speech and misinformation, the company would harden its actions against QAnon initially merely banning specific hashtags as search terms. From then on, users sharing QAnon-related content on TikTok could expect their accounts to get deleted from the app. QAnon-associated search terms were redirected to a community guidelines warning. To give an example for current far-right spaces, the search term "Save Europe" and its associated hashtag, are also blocked. However, sounds and accounts boasting "Save Europe" in their titles or as account names were up and active during our research.

In March 2023, an update to the community guidelines was announced. Here, the emphasis was placed on guidelines for synthetic (AI-generated) media as well as on affirming the extant focus of TikTok when it comes to content moderation: safety from physical harm. From the get-go TikTok has placed the safety of minors at the forefront of their moderation strategies, extending warnings on posts that perform activities that potentially get someone hurt. Their content moderation rules dating from March 2023 clearly outline where the platform’s priorities lie. Explicitly mentioned as main categories are youth safety and well-being, safety and civility, mental and behavioral health, sensitive and mature themes, integrity and authenticity, regulated goods and

commercial activities, and privacy and security. Under ‘safety and civility’ hate speech is mentioned as a subcategory in the guidelines themselves, but there was, at the time of writing, no mention of hate and extremist ideologies in the widely distributed diagram itself.

While the initiative of warnings on posts displaying potentially dangerous activities is of course positive, the focus on physical safety comes with blind spots for other problematic forms of content. In the dataset used for the presented research, one post is particularly exemplary for the friction that arises when warnings are added to potentially dangerous activities whilst the post is surrounded by the extremist captions, sounds, and comments referencing clearly visible and common symbols of white supremacy. This post was networked through the hashtag #1161, which is a well-known numeric sign for fascists as the numbers translate to the letters AAFA (Anti-Anti-Fascist Action) and their position in the alphabet. Moreover, it uses an ‘original sound’ that serves as a networked space for an extremist community that meets through replicating these sounds. TikTok’s focus on keeping people from imitating dangerous activities, such as smashing a bottle on the streets as displayed in the post itself, culminates in the bizarre situation where the posts will carry a warning relating to the smashing of a bottle while leaving other highly problematic content elements in that post and in its comments unaddressed.

Throughout its years of existence, the most prominent development in relation to community guidelines addressing hate speech and hateful behavior was the further specification of so-called categories of hate. These would include claims of racial supremacy, misogyny, anti-LGBTQ+, and antisemitism. Islamophobia was added later to the list (September 2024). While TikTok outlines how the app uses a myriad of automated technologies—ranging from text-based NLP approaches to computer vision algorithms—to scrutinize posts on various modalities (images, account names, sound, descriptions, gestures, symbols, etc), earlier research found that TikTok moderates in generally light and highly inconsistent ways (Zeng & Kaye, 2021). There is a growing awareness that its multimodality urges researchers to include audiovisual layers of ‘user-led obfuscation’ (Bösch & Divon, 2024). Turning trending songs into problematic content or associating trends with hateful conduct is a strategy of ‘creatively’ obfuscating worrisome materials. For example, early into the Russian invasion of Ukraine, the song *Good Evening Ukraine* was turned into *Good Morning Russia*, surfing on the popularity of the original (Pilipets et al., 2025), copying the tune and simply replacing the refrain. While, after some months, the Russian version and its soundscape turned out to be unavailable, there were still a handful of other soundscapes carrying the same sound. This problem directly relates to the inherent affordance of the ‘original sound’ that through the ‘use this sound button’ enables fast replication of audios that were initially uploaded by problematic actors. While the initial ‘soundscape’ instigated by a problematic actor might get banned, many others heard the sound before the account was deleted. These ‘sound-followers’ will then recreate their own original sound, that sounds exactly the same.

Engagement over well-being

It is clear that audio presents a fundamentally different set of challenges for moderation than text-based communication (Brisset, ADL). While the platform's official communication outlines efforts and investments undertaken to enhance better (sound) moderation technologies, and while recently there has been an acknowledgment of the presence of indirect or 'less obvious' hate as subsumed in memes and audio trends, these developments merely point to growing awareness, and not so much to actual enforcement of the app's own community guidelines and its moderation framework. Echoing Christin et al. (2024) and their so-called 'internal fractures of social media companies', the pattern seen in moderation of far-right accounts is signaling how the winning logic for platform companies is not that of user well-being, but that of engagement. Researching a variety of platforms, including TikTok, Christin et al. (2024) outlined some vignettes based on leaked information, that evidence the competing logics and how engagement overrules moderation. One of these vignettes was the retraction of an algorithm developed by Facebook employees in late 2020: When [they] realized that their users viewed many of the most viral posts on the platform as "bad for the world", a machine learning classifier was developed to downrank such posts, only to have the effort shelved by executives because it reduced engagement metrics. (Roose et al., 2020). Christin et al.'s hypothesis that the logic of engagement over well-being would also hold true for TikTok was affirmed in a 2024 case where the ISD reported 50 accounts for far-right hate speech to test how TikTok would moderate them, while the platform eventually removed violative content and channels, in many cases it does so only after accounts have had the opportunity to accumulate significant viewership. For example, the 23 banned accounts from the sample dataset managed to accrue at least 2 million views across their content prior to getting banned."

How niche soundscapes sustain hate

The presented analysis addresses the role of problematic 'original sounds' in the prolonged existence and proliferation of far-right white supremacist and xenophobic content on TikTok. We charted how far-right actors co-opt trend visibilities (and meme templates) on the platform and assessed the presence of problematic content in For You Feeds across three countries. Through moderation (or disappearance) trace analysis of videos and accounts that went offline after intervals of some months, we laid bare how so-called 'niche soundscapes of hate' are either very lightly moderated and/or disappear from the platform due to user-led decisions of deletion. To conclude, we will connect findings to a continuation of the platform's emphasis on preventing physical harm, a reliance on crowdsourced (user-led) flagging practices that only emerge when posts achieve a significant reach, and a demotion of niche hateful content, that, while posts are ineligible for the feed, still prove able to maintain and connect communities that can converge over extremist ideas. When smartly attached to current trends on the platform, such niche audio memes can appear in people's search results when looking for trending sounds. To summarize, our research investigates the following questions: 'How is TikTok's sound architecture affording the proliferation and obfuscation of problematic content?' And to what extent are problematic posts moderated or visible?

Methods

TikTok is built around easy imitation where users respond to each other through replicating and remixing sounds, hashtags, emoji, filters, and so on (Zulli & Zulli, 2022). Our methods are attuned to charting the tactical practices of malign actors that latch onto benign (audio) trends on TikTok through imitation, in order to amplify as well as prolong the ‘lives’ of far-right messages. The first part of our empirical assessment of the presence of extremist content on TikTok centered on the user experience of their personalized feed. While the search feature is increasingly used—AI Forensics (2023) found that 67% of TikTok users in Germany regularly use the search function—it is not possible to determine to what extent extremist posts are recommended to users on their For You Feeds. By merely assessing browser-based search results, we maintain a blind spot for the ways in which people encounter right-wing extremist content in their feeds. Therefore, we conducted a persona-based analysis (Bounegru & Weltevrede, 2022) where we assess how a right-wing person in Germany might be confronted with borderline or downright fascist content on their feeds, and how the presence of borderline or problematic content in their recommendations compares to rightwing users in the Netherlands and in the UK. As the authors were located in both Germany and the Netherlands we merely had to set up clean accounts using a developer browser that deletes cookies and traces whenever a browser session is closed. For the UK feed we used VPN software. Personas were ‘pre-trained’ to signal rightwing interest, through following corresponding political party accounts and conservative news outlets.

Hijacking trends and cloaking hateful lyrics

After establishing the presence of extremist content in feeds of rightwing users across three countries, our second step is to assess the longevity of the ‘lives’ of extremist audio memes: are they moderated, or at the very least: do these audio-driven memes go offline after some time has passed? Using the mobile app search feature we mirrored a user actively searching for events such as the Solingen knife attack of August 2024 (Light et al, 2018). Two prominent hashtags present in this ‘walkthrough observation’ were #deutschejugendvoran and #heimatliebeistkeinverbrechen. These hashtags were used as search queries to collect posts and metadata using the DMI-tool ‘Zeeschuimer’ (Peeters, 2021). This tool outputs data on the authors of posts, follower counts, sounds used, timestamps, hashtags, overlaid texts, filters, and a range of engagement metrics such as likes and plays.

Through a systematic assessment of the metadata of posts using one of the two problematic hashtags, we laid bare patterns derived from so-called ‘sticker texts’: the overlaid texts on top of videos. Recurring texts point to the presence of meme templates or trends. One of such trends found in our hashtag space was ‘*Music taste is important, imagine you do not know what comes after the intro*’. This trend hinged on users playing the intro of a song and then letting others comment on the song’s name and sometimes (the rest of) the lyrics. Outside of our dataset with nationalist hashtags, this ‘trend space’ harbors songs that are completely benign, but our hashtagged posts made use of the trend by playing the intro of a song we will here dub *Türke*. The lyrics are (partly) cited in the comments section. When we searched with a clean account for ‘*Musikgeschmack ist wichtig*’, the first malign use appeared as the 9th search result. The lyrics are derogatory toward people of Turkish descent, it humiliates people and calls out to kill them. The Music taste-trend lends itself perfectly for what we call a

‘cloaking while amplifying’ strategy. Playing merely the intro, without lyrics, works as a dog whistle for those in the know. A detection algorithm will not find the derogatory lyrics as they are not in the post and oftentimes also not (fully) present in the comments.

A second song—or audio meme—found in the hashtag-based dataset was *Zecken* (translates as ticks). This sound distinguishes itself from the *Türke* song, as it puts its racist lyrics on full display, where immigrants are compared to ticks and are punched to death. TikTok states that: *"Dehumanizing someone on the basis of their protected attributes by saying or implying they are physically, mentally, or morally inferior, or calling them degrading terms, such as saying they are criminals or animals, or comparing them to inanimate objects"* is against guidelines. The song hinges on a peculiar remix between a folk song named *Kreuzberger Nächte sind Lang* birthed in the 1970s, and a 1997 song by the Zillertaler Türkenjäger that blended the Kreuzberger refrain with racist lyrics where the verses outline how ‘*Zecken*’ (leftists and punks are all ticks in vernacular language) and ‘*Kanaken*’ (a German ethnic slur for people with roots from Southeast Europe, Middle East, and Northern Africa) are encountered on the streets of Kreuzberg (Berlin), and are then violently ‘ended’. When searching for this song with a clean account in August of 2024, the first two posts contain the toxic version with racist lyrics. Interesting is how an AfD fan account hosting 111.000 followers posted a video of a 1970s performance of the innocent song on September 19, 2023, close to the start of the AfD campaign. It might have primed extremist TikTokers to use the vile remix of 1997.

Following the approach of ‘moderation trace analysis’ (De Keulenaar et al., 2023) we probed the online status of the post links in early October 2024 (two months after the initial collection of data), and again in early December to assess what posts ‘disappeared from public online view’. As mentioned, due to a lack of transparency, we are, unfortunately, not able to cross-check whether offline posts were taken down by their creators or by the platform. We could distinguish between videos that over time turned to ‘unavailable’, or were ‘set to private’, and we could, through the post URL of unavailable videos, detect accounts that were deleted. At least in the case of the deleted accounts we can assume that many of these point to instances of deplatforming rather than users voluntarily deleting their ‘social capital’.

Disgust and contestation: network analysis of sounds and accounts

While largely seeking to recruit the like-minded, memetic engagement on TikTok does not exclude contestation (Geboers & Pilipets, 2024). When sounds catch on and users respond with novel adaptations: *"positive, negative, and ambivalent affect blend into each other"* (Paasonen, 2019, p. 52), constructing an environment of oscillating affective charges. It is this oscillation between excitement and disgust that keeps engagement with political messages ‘on the move’. This movement is a prerequisite for staying relevant in the ephemeral spaces of the platform, as well as for the consolidation of political extremist messages. This led us to perform a network analysis that charts the relationships between accounts and sounds within a space of fascist and anti-fascist contestation. This network was based on data collected through a hashtag-based query using #1161, a known numeric symbol for fascists. This hashtag is often countered by posts using #161 which stands for anti-fascist movements. We color-coded accounts for the presence of political extremist content, which materialize on the level of the hashtags (#1161 or #88 (for "Heil Hitler") and so on, using the ADL

Hate Database) but which can also sit in the performance of gestures, the depiction of symbols such as the ‘Black Sun’ or Sonnenrad (sun wheel) which is communicated through the spiderweb emoji and by depicting and referencing so-called ‘Talahons’, adherents of an urban subculture of males, typically but not necessarily of Middle Eastern origin, characterized, among other things, by a passion for German hip hop and wearing counterfeit designer labels. This caricature has been charged with ever more negative connotations, in part through the viral AI-generated song *Verknallt in ein Talahon* (Crush on a Talahon), which details how a white young woman falls in love with a criminal foreigner.

Sounds were color-coded for hosting extremist content on the level of their audios. The size of the triangle nodes (sounds) was based on the number of posts that used that sound, providing an idea of the popularity of that particular sound. The size of the circle nodes (accounts) was based on the number of followers of an account. We coded for the presence of fascist content based on the account’s posts present in our dataset as well as on their larger profiles. Account names were replaced by numbers in the network.

In a last step, we selected one of the problematic sounds found in the #1161 dataset, we engaged in querying the sound’s name (*Anotha Europe*) in the mobile app’s sound search module. The accounts found in this search either proved to be authors of sounds that resemble closely *Anotha Europe*, or, even more interestingly, use remixes, that blend the sound with various versions of benign hit songs. The output of this analysis is a linear dendrogram (Figure 6.3) depicting prominent accounts and popular sounds they ‘authored’ (uploaded to the platform).

Findings

To get a sense of the visibilities of far-right content in the personalized For You Feed (FYF), we trained rightwing personas (newly set up accounts) and used VPNs to emulate the experience of the users across three countries when scrolling through their feeds. We mapped the presence of right-wing extremist posts (red in Figure 6.1) as well as borderline posts (orange in Figure 6.1) in the feeds of these personas across countries. We can see how the German feed shows slightly more problematic and borderline posts on TikTok as compared to the feeds of right-wing users located in the UK and the Netherlands. To what extent this discrepancy is related to the campaign-related activities of the German AfD party, is hard to establish, but there is no doubt about the effectiveness of its sound-based propaganda tactics. The party and their fan army in a participatory propaganda approach play the algorithm to ensure far reaching pro-AfD content (Bösch, 2023) "flooding TikTok" (Breschendorf, 2024) with songs and sounds.

Interestingly, TikTok has a separate section where it outlines a number of content characteristics that will turn posts 'not eligible for the FYF'. TikTok’s community guidelines state that: "*Content may be ineligible for the FYF when it indirectly demeans protected groups. [...] Protected groups means individuals or communities that share protected attributes such as race and gender. These attributes include immigration status and national origin.*"

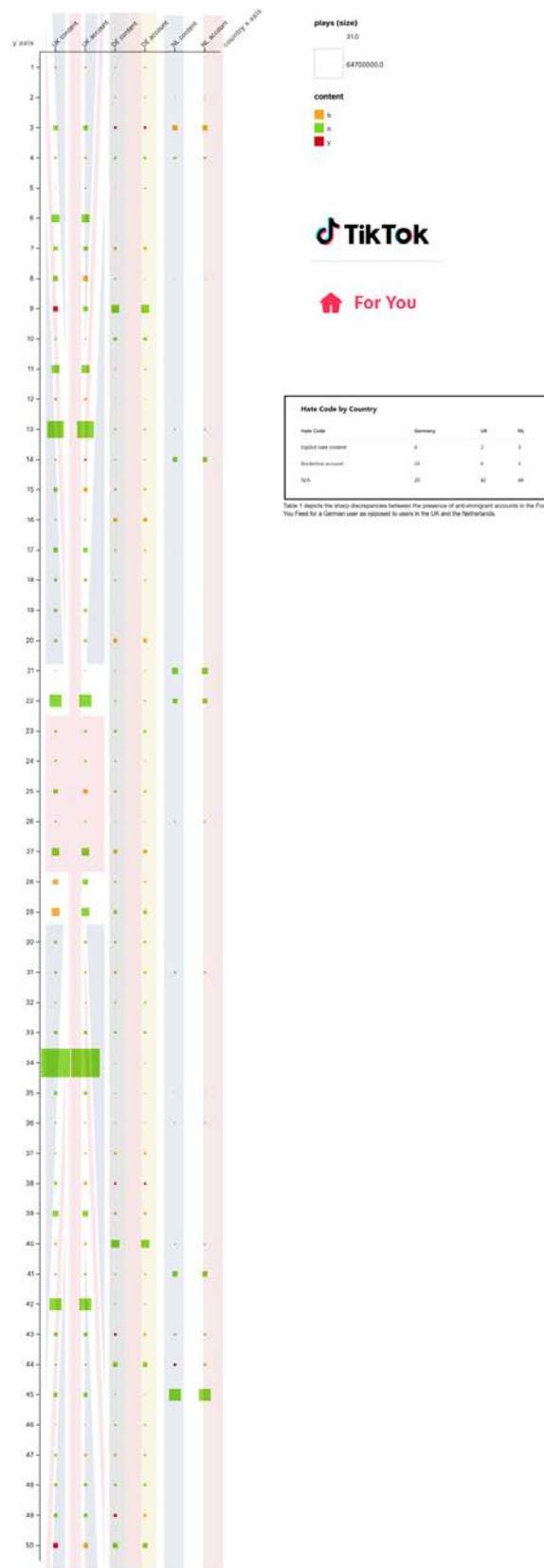


Figure 6.1 displays the presence of extremist (red) and borderline (orange) posts in the feeds of trained accounts across the UK, Germany, and the Netherlands. Borderline posts are posts not directly carrying explicitly hateful content, but that are connected to hateful anti-immigration accounts. The numbers in the inserted table depict the sharp discrepancies between the presence of anti-immigrant accounts in the For You Feed

for a German user as opposed to users located in the UK and the Netherlands. Source: authors.

The role of fascist and anti-fascist contestation

To address the question of the role of TikTok's networked sound infrastructure within spaces of political contestation, we gathered data collected through the hashtags #1161 (anti-anti-fascists) and #161 (anti-fascists). The network shows how accounts (kept anonymous through representing them as numbers) connect to various more or less popular and more or less frequently replicated sounds. Green accounts are accounts countering the #1161 message, red nodes represent fascist accounts. Account nodes (circles) are sized by their amount of followers. Sounds are represented as triangles and colored purple for mainstream audio trends on TikTok and brown for niche fascist songs. Their sizes correspond to the number of posts that copied the sound. While there are brown songs, the majority of fascist accounts assemble around mainstream, trending songs such as a sped-up version of *Kiss me*, the already mentioned Gigi-song (*l'Amour Toujours*), and *Around the World (lalala)*. Interestingly, an anti-fascist account with 21.900 followers, uses the much more niche fascist sound of *Scheiss Egal* to purposely target fascists.



Figure 6.2 displays anonymized accounts (numbered), their fascist or anti-fascist stances in color (green for anti-fascists and red for fascists), and their connections to sounds that were either benign in audio content but hijacked by extremists on the platform (purple triangles), or that were fascist and racist in their audio contents (brown triangles). Source: authors.

Note that in the network the sound nodes represent multiple similar ‘original sounds’ that were all uploaded to the platform separately. Some of the rightwing *Scheiss Egal* sound versions were at one point no longer available which points to their creators (initial original sound uploaders) being deplatformed, or they have purposefully deleted their own accounts. Other versions remained online, pointing out the complexity of effective moderation in a landscape of replication.

Soundscapes as dispersed problematic niches

To further investigate how versions of similar sounds perpetuate and prolong the proliferation of fascist content, we turned to the mobile app sound search, where we queried one of the brown-colored ‘niche’ songs from the network in Figure 6.2. Querying ‘*Anotha European*’ in the mobile app lets us manually assemble sounds and accounts that first uploaded these sounds. The app also displays the number of posts that replicated that particular sound. A set of eight accounts connected to songs that were all also present in our #1161 dataset. One account (Figure 6.3) was particularly interesting as its account name consists of a banned search query (Save Europe) and it ties into sounds that host extremist content (the Gigi-song, *Kiss me Speed*, *Around the World (lalala)* and a variety of Russian-language songs such as *Australia* (by the band called X) not depicted in the dendrogram. This highly active account hosts 19,000 followers despite its user name representing a banned search query. The sounds also show how they—modulated by memetic replication—morph into remixes that speed up, slow, or otherwise distort the original, perhaps being part of a user imaginary in which reworked sounds work to evade moderation.

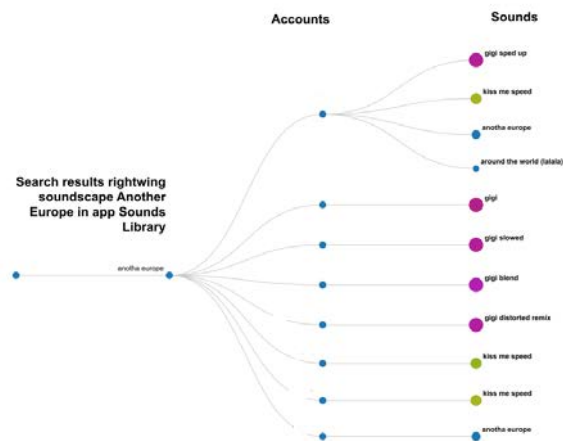


Figure 6.3 displays mobile app ‘sound search’ results for the query ‘Another Europe’, which was based on a problematic song title found in the sounds-accounts analysis (Figure 6.2). The search results contain accounts that uploaded (versions of) both ‘Anotha Europe’ as well as the more pop culture sounds of ‘Kiss me again’ and l’Amour Toujours by Gigi d’Agostino. The colors point out audio similarities between the various remixes. The first account (on top of the middle row of nodes) engaged in multiple ‘plugs’ of sounds, all of them very familiar to rightwing vernacular tropes on TikTok. Source: authors.

The dendrogram of in-app search results (Figure 6.3) displays how the Gigi-song reverberates through spinoffs that speed up, slow down, distort, or blend in with other sounds and speeches. This inspired us to ‘reverse our search’ by querying the Gigi song *l’Amour Toujours*. This, disturbingly enough, showcased within the first 50 results, a sound that not merely subsumes the Gigi tune, it also blended it with a speech stating how ‘Hitler, war der Mann ihres Lebens’ (Hitler was the man of your life). This sound was replicated through the ‘use this sound’ button 1036 times. The post using this sound was, at the time of writing, banned or taken offline by the account owner. But the first post of the (still available) account was online and asked people if they believed that it ‘really was 6 million’, downplaying the death toll of the Holocaust.

Moderation traces: minimal ‘disappearances’

To address the question of the extent of moderation taking place, we conducted a trace analysis. As mentioned, due to a lack of transparency on the side of the platform, as researchers we can merely determine whether videos or accounts are no longer available with no way of knowing whether content was taken offline by the TikTok user or by the platform. We nonetheless engaged in an assessment of the status of post URLs after two designated intervals after collecting posts toward the end of August 2024: early October and early December of the same year. We distinguish between videos that are offline, accounts that no longer exist, and accounts that are set to ‘private’.

We selected two problematic songs that were present in the hashtag-based datasets for #heimatliebeistkeinverbrechen and #deutschejugendvoran. We used the indexed sound pages for these songs and ran the browser-based scraper tool Zeeschuimer to collect posts and metadata. We revisited post URLs twice, each with a two-month interval. The first song on TikTok ‘merely’ conveys the intro, no lyrics are heard, just the techno beat of *Türke*. The text taps into a meme trend that asks whether people recognize the song in a ludic way: *‘Imagine you do not know what comes after the intro’*.

For assessing moderation dynamics in this soundscape, we engaged in analytically distinguishing between 1) posts carrying hate speech on the textual post dimensions of captions and overlaid video texts (so-called ‘Stickers’), and 2) posts containing coded or ambiguous visual symbols signifying or connecting to extremist rightwing tendencies (see Figure 6.4). The left diagram in Figure 6.4 represents 65 *Türke*-posts that could be coded for four common and highly ambiguous (hard to read) memes present in the hashtag-based dataset. These memes showcase motorbikes of the East German brand Simson, tractors, working shoes as stand-ins for ‘real workers’ as opposed to the sneakers of immigrants, and lastly the ‘hampelmänner’ (puppet on a string) meme that symbolizes how you can either be led by others or do your own research on immigrants. Note that the puppet on a string memes seem ‘better moderated’, although this could also be a misleading image: this meme could have assembled accounts that were also engaging in other more explicit hate speech posts, hence their account deletions. We assessed URL status over time showing how 55 (October) and 49 (December) posts were still online which amounts to 85% and 75% of the total of 65 posts with hate cloaked as an ambiguous meme. There seems to be a slow and quite minimal decline of available posts in this ambiguous space which is to be expected when content is ‘cloaked’ to this extent. When compared to posts carrying explicit hate speech (right diagram in Figure 6.4, representing 38 posts) in textual layers of *Türke* posts, we see that 66% (October) and 53% (December) of these posts were still online. Figure 6.4 allows us to see how memes without textual references to hate are indeed much less moderated than posts that have some form of hate speech or hateful symbols and references to ideologies (gestures, numerical signs) subsumed in their textual layers.

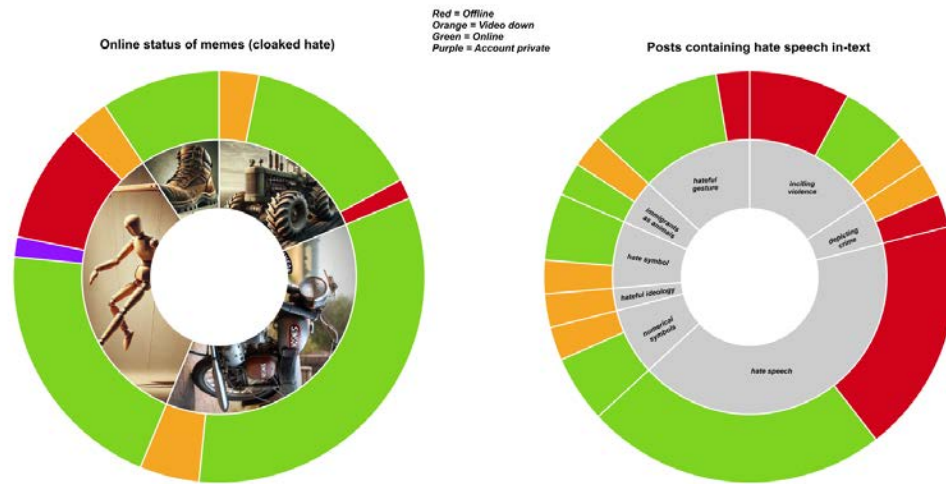


Figure 6.4 displays the status transformation of post URLs between the time of data collection (August 2024) and revisiting these URLs early December 2024. Both diagrams depict posts using the intro tune of the xenophobic Türke song in combination with the ‘Musik geschmack ist wichtig’ trend that materializes in the overlaid text on top of the videos (sticker texts). To the left we see the status of posts that were ‘double-cloaked’ using the intro tune of the hateful song and a particular visual template (recurring memes) that works as ‘coded visual language for rightwing extremism’. To the right we see posts using the same tune and sticker text, but these posts also carry explicit hate and violence in textual post layers, making them more recognizable as hate speech. The categories of textual hate are derived from TikTok’s community guidelines. Source: authors.

The second song (Zecken, or Ticks) boasts explicit lyrics about killing immigrants on the streets of Kreuzberg, Berlin. Here, all posts on the sound page (or in the soundscape) carry explicit hate as the lyrics are in large part cited in the videos. To further assess how moderation takes place across modalities of audio and text, we decided to distinguish between posts carrying textual hate (as added to the explicit audio) and those that do not (figure 6.5) to see whether this influenced the online presence of such posts over time.

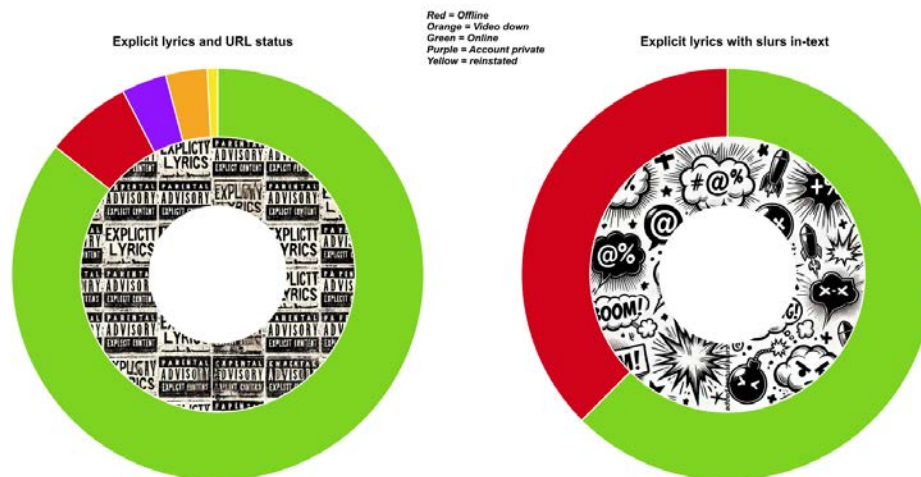


Figure 6.5 displays the status transformation of post URLs between the time of data collection (August 2024) and revisiting these URLs early December 2024. Both diagrams depict posts using the explicitly hateful song ‘Zecken’ (ticks as referring to immigrants). To the left, we see the entire dataset and its status transformation. To the right, we see the posts that express hate and violence in the textual layers. The categories of textual hate are derived from TikTok’s community guidelines. Source: authors.

Despite the explicit presence of speech violating the TikTok community guidelines of hateful behavior, the level of moderation is comparable to the more covert dog whistle sound of *Türke*. In the overall dataset, 89% of URLs in the first and 86% in the second interval remained online, in the posts containing textual hate speech (left diagram in figure 6.5) these percentages dropped to 65% and 59% respectively. This points to textual hate as a greater predictor for the moderation (or at least ‘disappearance’) of posts than the presence of hate speech on the level of audio.

Undercurrents of extremism: how hateful content circulates under the radar

TikTok affords easy imitation of audio content through the ‘use this sound’ button, allowing audio memes to go viral (Abidin, 2021). As such, a sound is made searchable and can harbor a wide array of topics. Very often though, posts networked by sounds, skew toward particular topics, including those deemed problematic. While TikTok’s strategies for moderation is said to include audio (Hee et al., 2024; Steen et al, 2023; Medina Serrano, 2021) the polysemous nature of sound, like images, hampers effective moderation. The multimodality of TikTok further complicates this matter, as a ‘benign sound’ can shift meaning quickly when used in tandem with particular textual or visual images. Moreover, as we laid bare in the outlined study, the infrastructure of original sounds that are easily replicated, creates connections between highly visible audio trends and malign actors that are close to impossible to moderate, as the sounds of banned accounts can continuously re-emerge under new versions of sounds.

We focused our analysis on sounds and so-called soundscapes as spaces where communities affectively convene around practices of sound-replication. To further exemplify how TikTok is the ideal platform for extending and prolonging extremist ‘vibes’, also on the level of (ambiguous) visual modalities, we want to outline one of the memes present in the *Türke* sound, that is the recurring motorbikes of a particular brand. The love for Simson motorbikes was most probably inspired by an AfD promotion poster showcasing politician Björn Höcke riding the signature East German Simson motorbike. This inspired many TikTokers to post their own motorbikes on the platform. From a study of similar motorbike memes connected to another far-right sound—‘*Anotha Europe*’—Geboers et al. (2025) found that such motorbike memes pull in both far-right accounts as well as (many more) benign motorists who very probably do not have a clue that they are connecting their motorbike to a racist tune. From a sample of this study containing 36 accounts, the researchers found only three clearly fascist accounts, the other 33 accounts were not clearly rightwing. This unconscious connection between a hobby (motorbikes) and a political ideology subsumed in an ambiguous song intro, expands the affective reach of far-right sentiment beyond explicitly political spaces. When a 13-year-old with a cool bike picture slideshow sees other bike posts and decides to use the same sound, this person’s post gets networked, through the sound linkage, to far-right users and their extremist content.

MLLM-detection: not a fix for all

With the development of MLLMs the detection of hate speech with more context sensitivity is around the corner (Hee et al, 2024). However, these advancements will push users to engage in more latently present hate speech. While algospeak is associated with misspellings, emoji, or interpunction tactics (Steen et al., 2023), sound introduces more ‘subtle’ tactics of circumvention. This project responded to a scholarly and institutional (corporate and political) need to recognize that algospeak is "more than the simple replacement of words [...] it needs to be understood as code words and linguistic variations, visual and multimodal communication, and audiovisual coherences [...]" (Steen et al., 2023, p. 12). Our study honed in on the audiovisual character of the platform and its infrastructure of networked sounds that create loopholes for politically extreme actors to stay active and garner levels of engagement that when we account for the accumulated engagement that dispersed original sounds assemble, are certainly not negligible.

From our trace analysis of moderation (or tracing disappearance) of videos and accounts in problematic soundscapes we found minimal ‘deplatforming’ or disappearance of hateful content, even when hate is explicitly present in either audio or text. In order to understand these minimal levels of moderation, we need to take into account TikTok’s reliance on flagging practices. It is no coincidence that directly underneath the list of direct and indirect hate speech in the platform community guidelines, there is a section titled ‘What do I do if I or someone I know experiences hate on Tiktok?’ The app relies heavily on the practice of flagging, allowing users to report content that violates community guidelines (Crawford & Gillespie, 2016). Are describes how flagging facilitates liability evasion for platforms and a delegation of labor, but "its influence on moderation remains opaque" (2024, p. 4). The rather minimal disappearances of problematic videos and accounts in our analyses might very well be related to the platform’s reliance on users flagging content to platform moderators. Taking into account how sounds such as *Türke* and *Zecken* are circulating

within niche and distributed soundscapes, most often not attaining high engagement metrics, we might derive that these posts are ineligible for the FYF. This also means that the chances they get flagged are close to zero. This would explain how such posts are largely neglected by moderators.

Strategies of cloaking while amplifying

Based on the empirical analyses, we were able to identify three strategies that users adopt to avoid moderation. We regard these strategies to follow the logic of ‘cloaking while amplifying’ extremist messages. These tactics are respectively: 1) creating sticky associations between far-right extremist and racist ideas under the ‘cloak’ of trending songs (such as putting to work a highly popular song of Gigi d’Agostino, 2) proliferating extremist audio (mainly songs, at times blended with problematic speech) by tapping into benign trending memes that play fragments of songs and intros (such as *Musik geschmack ist wichtig/Music taste is important*), and 3) remixing an innocent tune to map explicit lyrics onto these tunes (*Zecken* as remixed with the folk song *Kreuzberger Nächte sind Lang*). As said, the larger part of these posts do not boast significantly high engagement metrics, with a mean of 1821 plays for *Türke* and 1457 plays for *Zecken* posts. For comparison, the posts collected with #1161, which holds various racist but also popular mainstream songs, has a mean of 73,352 plays. Nonetheless, niche soundscapes do appear in search results when users query for a trend (*Musik geschmack*), or when users search for benign sounds that get remixed in with vile versions of the sound (*Zecken*). Thus, while possible demotion efforts might underlie the overall low engagement of these posts, the fact that these niche soundscapes are scarcely moderated means that they continue to sustain an extremist ‘vibe’. They serve as spaces of connection that keep extremist communities alive in the undercurrents of the platform. Occasionally, their posts, when linked to popular visual memes (motorbikes) and memetic templates such as ‘*Musik geschmack ist wichtig*’ do confront users with extremist beliefs.

References

- Abidin, C. & Kaye, B. (2021). "Audio memes, earworms, and templatability: The ‘aural turn’ of memes on TikTok." In C. Arkenbout, J. Wilson & D. de Zeeuw (Eds.), *Critical Meme Reader: Global Mutations of the Viral Image* (pp. 58-68). Institute of Network Cultures: Amsterdam.
- Are, C. (2023). An autoethnography of automated powerlessness: Lacking platform affordances in Instagram and TikTok account deletions. *Media, Culture & Society*, 45(4), 822-840.
- Bösch, M., & Divon, T. (2024). The sound of disinformation: TikTok, computational propaganda, and the invasion of Ukraine. *New Media & Society*, 26(9), 5081-5106.
- Bösch, M. (2023). Alternative TikTok tactics: How the German right-wing populist party AfD plays the platform. In: *Fast Politics: Propaganda in the Age of TikTok* (pp. 149-167). Singapore: Springer Nature Singapore.
- Bounegru, L., De Vries, M., & Weltevrede, E. (2022). The research persona method: Figuring and reconfiguring personalised information flow. In C. Lury, W. Viney, & S. Wark (Eds.), *Figure, Concept, and Method* (pp. 77–104). Springer Nature: Singapore.

- Breschendorf, F. (2024, August 13). AfD-Schlager-Hits überfluten TikTok: Experte warnt vor Radikalisierung. <https://www.fr.de/politik/afd-songs-schlager-pop-tiktok-radikalisierung-sounds-rechtsextremismus-zr-93241021.html>
- Christin, A., Bernstein, M. S., Hancock, J. T., Jia, C., Mado, M. N., Tsai, J. L., & Xu, C. (2024). Internal fractures: The competing logics of social media platforms. *Social Media + Society*, 10(3), 20563051241274668.
- Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3), 410-428.
- Daniels, J. (2009). Cloaked websites: Propaganda, cyber-racism and epistemology in the digital era. *New Media & Society*, 11(5), 659-683.
- De Keulenaar, E., Magalhães, J., dos Santos Junior, M. A., & Rogers, R. (2023). After deplatforming: Retracing content moderation effects across platforms and a Post-American web. *AoIR Selected Papers of Internet Research*. <https://spir.aoir.org/ojs/index.php/spir/article/view/13538>
- Doctor, N., Fiennes, G., & O'Connor, C. (2024). NazTok: An organized neo-Nazi TikTok network is getting millions of views. https://www.isdglobal.org/digital_dispatches/naztok-an-organized-neo-nazi-tiktok-network-is-getting-millions-of-views/
- Duffy, B. E., & Meisner, C. (2023). Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. *Media, Culture & Society*, 45(2), 285-304.
- Duguay, S., & Gold-Apel, H. (2023). Stumbling blocks and alternative paths: Reconsidering the walkthrough method for analyzing apps. *Social Media + Society*, 9(1), 20563051231158822.
- Galip, I. (2023). Propaganda, digital diplomacy, meme wars: How digital confrontation is shaping the new world order. In A. Ferrari & E. Tafuro Ambrosetti (Eds.), *Multipolarity after Ukraine: Old Wine in New Bottles?* (pp. 95-116). Istituto per gli Studi di Politica Internazionale. <https://www.ispionline.it/en/publication/multipolarity-after-ukraine-old-wine-in-new-bottles-116515>
- Geboers, M., Riccio, B., Kuang, C., Yang, E., Steenhuis, H., Liu, J., Xiong, K., Bitel, P., Mino, M., Song, R., Nobel, N., Wu, Y., & Freschi, R. (2025). *Malicious earworms: How the far right surfs on TikTok audio trends*, <https://www.digitalmethods.net/Dmi/WinterSchool2025MaliciousEarwormsTikTok>
- Geboers, M., & Pilipets, E. (2024). Networked masterplots: Music, pro-Russian sentiment, and participatory propaganda on TikTok. *Journal of Digital Social Research*, 6(1), 90-103.
- Geboers, M., & Hammelburg, E. (2024). Elevating the antagonist encounter: How the 'Stitch' transforms victimhood contestation on TikTok. *Anglica. An International Journal of English Studies*, 33(2), 143.

Hee, M. S., Sharma, S., Cao, R., Nandi, P., Nakov, P., Chakraborty, T., & Ka-Wei Lee, R. (2024). Recent advances in hate speech moderation: Multimodality and the role of large models. *ArXiv*. <https://doi.org/10.48550/arXiv.2401.16727>

Kaye, D. B.V., Zeng, J., & Wikstrom, P. (2022). *TikTok: Creativity and Culture in Short Video*. John Wiley & Sons.

Light, B., Burgess, J., & Duguay, S. (2018). The walkthrough method: An approach to the study of apps. *New Media & Society*, 20(3), 881–900.
<https://doi.org/10.1177/1461444816675438>

Medina Serrano, J. C. (2021). *Multiplatform Analysis of Political Communication on Social Media*, Doctoral dissertation, Technische Universität München.
<https://mediatum.ub.tum.de/doc/1597300/document.pdf>

Paasonen, S. (2019). Resonant networks: On affect and social media. In A. Fleig & Von Scheve, E. (Eds.) *Public Spheres of Resonance: Constellations of Affect and Language* (pp. 49-62). Routledge.

Papacharissi, Z. (2015). *Affective Publics: Sentiment, Technology, and Politics*. Oxford University Press.

Peeters, S. (2021). *Digital Methods Initiative: Zeeschuimer*.
<https://github.com/digitalmethodsinitiative/zeeschuimer>

Pilipets, E., Geboers, M. & Delavar-Kasmaii, D. (forthcoming). Detouring, rerouting, weaponization: Memetic soundscapes and the secondary orality of WarTok. In B. Mutsvairo, D. Nguyen, & J. Zeng (Eds.), *Technology, Power & Society: Global Perspectives on the Digital Transformation*. Brill.

Roose, K., Isaac, M., & Frenkel, S. (2020, November 24). Facebook struggles to balance civility and growth.
<https://www.nytimes.com/2020/11/24/technology/facebookelection-misinformation.html>

Steen, E., Yurechko, K., & Klug, D. (2023). You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media + Society*, 9(3), 20563051231194586.

Zulli, D. & Zulli, D.J. (2020). Extending the Internet meme: Conceptualizing technological mimesis and imitation publics on the TikTok platform. *New Media & Society*, <https://doi.org/10.1177/1461444820983603>.

7. The internet's dark alleys – Laissez-faire content moderation and illegal trade on Telegram

Stijn Peeters

Abstract

The popular chat app Telegram claims to not allow "activities which are recognised as illegal [in the EU]". Previous work has however shown a sizable presence of illegal content on the platform. In this chapter, I study how straightforward it is to find this problematic content, particularly Dutch-speaking channels in which trade of drugs, counterfeit documents, and other illegal goods is offered. Notwithstanding indications that moderation is slowly increasing, on the whole Telegram remains a lightly moderated platform. Actors are cognisant of this and operate a 'Telegramsphere' of illegal trade that is sizable (in terms of message volume), differentiated (with different channels having their own specific purpose), and connected (with links between channels as well as to other platforms).

Keywords: Telegram, content moderation, black market, illegal trade, online crime

Introduction

Telegram, the messaging platform, has many identities. It is a place where Russian state officials post the latest propaganda, yet also one where the mostly-banned free press of the country can still publish its journalism. For millions, it is simply the messaging app of choice, the platform on which they talk to their friends and family. For others, it is where one goes to find and share content banned from other social media platforms; the platform has acquired something of a reputation of a place where anything goes, and where moderation is done reluctantly if at all. One early example of this in a Western context is its reputation as a refuge for alt-right commentators after many of them were banned from American platforms like Facebook and YouTube (Rogers 2020); since then, it has remained a relatively moderation-free space. From a content moderation perspective, Telegram is then most noteworthy in how absent content moderation has been from it in comparison to other, similar platforms. It has long resisted calls to do more about the spread of child sexual abuse material (CSAM), revenge porn, terrorism, and sales of illegal goods and services (Thiel et al., 2023; Agarwal et al., 2022). All of these are banned from the platform explicitly, but nevertheless reports of their presence on the platform emerge with some regularity (e.g. Voskuil, 2018; Goldenberg et al., 2022; Visser, 2024).

Telegram has mostly responded by positioning itself as a platform free of government interference, reflected in a minimalist approach to content moderation. In response, regulators - in particular the European Union - have applied increasing pressure on the platform to enforce its rules and relevant laws more actively, culminating in the arrest of CEO Pavel Durov by the French authorities in late 2024. Since then Telegram claims to have ramped up its content moderation and compliance efforts, raising the question of to what extent the presence of problematic content is still an issue on the

platform and if so, what it looks like and what a more effective approach to its moderation might look like.

In this chapter, I focus on one such category of banned content specifically, and present an analysis of Telegram's ecosystem of channels offering illegal goods and services. I focus particularly on the parts of it that are easy to find for an "outsider", and that one might thus expect Telegram to be able to moderate relatively easily. I start with a brief overview of Telegram's history and its policies on content moderation, and a discussion of earlier scholarly work on the topic.

I then discuss how I demarcated this ecosystem of easy-to-find offers of illegal goods and services, and collected messages posted therein for further analysis. The analysis then unfolds along two parallel strands; one more quantitatively oriented, focusing on the categorisation of Telegram channels and the links between them, and one more qualitative strand. In the latter, the focus is on understanding how this ecosystem operates through a close(r) reading of the content of these messages. I conclude with an appraisal of this ecosystem from the perspective of content moderation, as well as a discussion of what more effective content moderation might look like, considering the dynamics of this space.

Telegram: A history

Telegram was founded in 2013 as a relatively simple mobile chat app. The company behind the app was launched, as a non-profit, by brothers Nikolai and Pavel Durov, who had also founded the Russian social network platform VK (formerly VKontakte). Not long after founding Telegram, Pavel, the more public of the brothers and CEO of both VK and Telegram, left Russia. He ended his involvement with VK, citing increased government pressure and interference with his social media companies as a reason; Telegram became his focus.

Ever since, it has been unclear where exactly Telegram is based. Legally, the platform is incorporated in both the British Virgin Islands and Dubai, but development of the platform has at various times been claimed to be based in places such as Berlin, or St. Petersburg, while Durov himself obtained French citizenship in 2021. Telegram itself claims to store data "across different jurisdictions" and "in multiple data centres around the globe" (Telegram, 2025c).

Data passing through these data centres does so using MTProto, a homegrown protocol (presented as developed by Nikolai Durov) that is well-documented and relatively easy to implement, affording an expansive third-party ecosystem of bots, tools, and applications that run on the platform. From the start, the platform and this protocol were positioned as secure and private; "At this moment, the biggest security threat to your Telegram messages is your mother reading over your shoulder. We took care of the rest.", reads Telegram's website's FAQ at its launch in 2013. The same FAQ cites the "new protocol, MTProto, built by our own specialists" as an important feature. Independent security experts however quickly found deficiencies in the protocol (grugq, 2015; Jakobsen, 2015), casting doubt on such claims. While the protocol was updated and improved in the years since, Telegram remains a centralised platform in which all traffic passes through its own servers. Such a 'single point of failure' potentially leaves the platform vulnerable to government interference and eavesdropping (Miculan & Vitacolonna, 2023), though it claims on its website to have

resisted such efforts so far (Telegram, 2025c). In any case, the image of Telegram as a secure, anonymous alternative to other messaging apps persists.

Telegram has, since its launch, added various new features that arguably make it more of a social media platform than a traditional messaging app. Examples of this include the option to add a biography to one's profile (introduced in 2018), or a way to find people who are physically located nearby (introduced in 2019). This aspect of the platform is however particularly exemplified by the ability to create channels and groups that can be joined by thousands of people. Groups allow anyone to chat; channels are typically in 'broadcast' mode, where only the owner(s) can send messages, but anyone can join and read these. Often a channel is then accompanied by a group in which people can discuss the posted messages or simply interact with other channel subscribers.

It is this latter feature – that of being able to start channels – that has been the focus of most of the academic interest in Telegram so far. Particularly the connections between various groups and channels, e.g., through forwarded messages, have been scrutinised to better understand how various thematic "Telegramspheres" (Simon et al., 2022) form, interact with each other and afford the spread of various types of content. This latter aspect of Telegram has been of interest in particular because of the platform's well-known lax attitude towards content moderation. Its ambiguous legal context has allowed it to mostly evade government scrutiny and demands for user data, with the company proudly claiming on its website that it has "disclosed 0 bytes of user messages to third parties, including governments" and that "Telegram won't be part of [...] politically motivated censorship" (Telegram, 2025c). As a result, it has for years both enjoyed a status as a relatively unmoderated platform, but has also received much criticism from journalists, NGOs and governments for the widespread availability of illegal content such as non-consensual intimate images (NCII; Semenzin & Bainotti, 2020) and child sexual abuse material (CSAM; Thiel et al., 2023), as well as being used for coordinating terrorist activities and other forms of extremism (Shehabat et al., 2017; Gill, 2021).

The platform's prioritization of privacy and resisting censorship, over moderating content and enforcing relevant laws, seems to have caught up with it in 2024. In August of that year, CEO Pavel Durov was arrested by the French authorities for allowing various types of illegal content and discourse to proliferate on the platform. Since Durov's arrest, Telegram has announced improvements in this area, such as increased cooperation with governments, integration of tools to counter CSAM, and dashboards that show the amount of CSAM and terrorist content that has been banned or deleted (though this data had already been available via a number of automated Telegram channels).

In late 2023, Telegram claimed to have around 40 million EU-based users (Telegram, 2023). While this is a large number, it also means it has not yet been designated a 'Very Large Online Platform' by the EU, for which the threshold at the time of writing is 45 million users, or roughly 10% of the EU's population. This is significant, because as a VLOP it would be subject to increased transparency and moderation requirements under the European Union's Digital Services Act (DSA). Though it has since stopped mentioning user numbers on its website, it is not unlikely that the platform is currently

very close to or over this threshold, considering its steady growth in the years prior to 2023. It is in any case a platform with a significant user share in the EU context. Telegram has thus enjoyed relatively consistent interest from academic researchers, because of its (historical) laissez-faire attitude towards content moderation as well as its use in various communities of interest, but also for more practical reasons, such as the accessibility of data on the platform afforded by the MTProto protocol. A significant portion of Telegram-focused research cites datasets of millions of messages that can be collected computationally and relatively effortlessly through widely available libraries and tools.

A popular way to then quantitatively analyse and map such data is through forwarded messages. Telegram affords the forwarding of a message from one conversation to another; this is routinely done in larger groups and channels, establishing links that can form the basis of a network analysis of a dataset (Peeters & Willaert, 2022; Simon et al., 2021). This makes it amenable to what has been called "controversy mapping" by Venturini & Munk (2021), i.e. studying a controversy or issue through actors via medium affordances and offering some way to 'navigate' this space via a spatial, structural, or topological analysis; this could comprise a visual network analysis, inductive coding and close reading, which are the methods used in the analysis that follows.

Moderation of illegal goods and services

One example of an issue that has been highlighted and mapped in previous work is the proliferation of illegal content on Telegram. This is perhaps where the case for content moderation is least ambiguous. While the extent to which political speech should be moderated on online platforms is cause for much (on-going) debate, the case for moderation of content banned by law is relatively straightforward. As a baseline, it is a reasonable axiom that companies should not break the law in the course of doing business. Illegal content is however a broad term and, by definition, context-dependent. What is illegal is determined by the local legislative context, and content that might be deemed illegal in the Netherlands, for example, may be permitted in other countries, or vice versa. Nevertheless, a number of broadly forbidden categories of content, such as CSAM, NCII, terrorism and extremism can be distinguished that have been demonstrated to be present on Telegram, benefiting from a lack of moderation.

While many social media platforms use comprehensive terms of service and lengthy 'community guidelines' to tell its users what is and is not allowed on the platform (de Keulenaar et al., 2023), Telegram's content rules are almost refreshingly succinct. In what is the only part of their terms of services concerning content, four categories are outlined that are not allowed on the platform:

- Use our service to send spam or scam users.
- Promote violence on publicly viewable Telegram channels, bots, etc.
- Post illegal pornographic content on publicly viewable Telegram channels, bots, etc.
- Engage in activities that are recognized as illegal in the majority of countries. This includes child abuse, selling or offering illegal goods and services (drugs, firearms, forged documents), etc. (Telegram, 2025c)

The latter item uses a curious formulation; it implies that illegal content or activities are permitted so long as they are legal in more than half of all sovereign countries. Not only is this incompatible with most legal systems, which consider a company or service subject to local laws as long as they are available in that locality, it also implies that users would need to exhaustively survey the legal codes of all 193 sovereign countries to determine if their use of Telegram would be permitted under the terms of service. Nevertheless, it is noteworthy that the terms of service do provide examples of some types of such content that is supposedly banned in the majority of countries, i.e. "child abuse, selling or offering illegal goods and services (drugs, firearms, forged documents), etc."

The first of these, child abuse, is one of the categories of content that Telegram does somewhat actively moderate. Since November 2018, a daily report of the amount of groups and channels "related to child abuse" banned is posted in the Telegram channel @stopCA; a similar daily report on banned terrorist-affiliated groups is posted in @ISISWatch. No such information is available for other categories of banned content, though general information and press reports indicate that since 2024, Telegram has begun to ban groups and channels more actively; according to its own "moderation overview" page, it bans roughly 1.3 million groups and channels per month (Telegram, 2025b).

Nevertheless, the other example of illegal content mentioned in the terms of service - "selling or offering illegal goods and services (drugs, firearms, forged documents)" seems to be less of a priority for the platform, and is not singled out in its moderation reports.

Methodological approach: mapping an unmappable platform?

For this analysis, the goal is then to collect a dataset that is somewhat representative of the offering of illegal goods and services in the Dutch context on Telegram. Journalists have found time and time again that illegal services are readily available in Dutch-speaking groups on Telegram, positioning it as a "dark marketplace" on which drugs and weapons are readily available (Voskuil, 2018; Goldenberg et al., 2022; Visser, 2024). If this is indeed still the case, it would present a good case study for an analysis of Telegram's lack of moderation and the activity that emerges in response to this absence.

One difficulty here is that due to the illegal nature of the content, it is likely at least partially to occur in private or otherwise difficult-to-find groups and channels. It is also unclear to what extent illegal goods are actually sold via Telegram - many of the groups I found warned people of scams, and it is possible that in fact a majority of the offers in these groups are scams, with the more 'legitimate' business occurring in other places out of the public eye, or channels that can only be feasibly discovered via word-of-mouth. Nevertheless, both illegal business and scams are prohibited by Telegram's terms of service, so in either case their presence in these channels and groups would be problematic from the perspective of the platform's terms of service.

One might also say that the most problematic content is that content that is the easiest to find, as this would also be the easiest to moderate for Telegram. It additionally has the most significant reach. Being relatively easy to find on a platform with a large number of users, this content potentially presents an accessible gateway into a network of illegal activity. If such 'low-hanging fruit' is still present on the platform, this would

be an indication that Telegram's enforcement of their terms of service is lacking. It is perhaps unreasonable to expect that no such content would be available at all; but if it is straightforward to find from a 'naive' position, this would be problematic. I operationalise such a 'naive' position by starting from a small list of easy-to-find channels and groups, and then following links within these to other channels to discover related channels and groups through a snowballing-like expansion of the dataset. While 'links' are often operationalised as 'forwarded messages' in similar research (see e.g. Nobari et al., 2017; Baumgartner et al., 2020; Peeters & Willaert, 2022), I take a somewhat broader view and also include @mentions of other channels as well as hyperlinks (i.e. <https://t.me/channel>) as significant pointers to related entities.

All entities discovered via such links are then included in a next iteration of the data collection. I only include links here that occur three times or more at a given level of the crawl, as a way to eliminate less-relevant links from the data. This is then repeated for three 'hops', i.e., with a crawl depth of 3. This means following 'breadcrumbs' found in the dataset that lead to other, unknown channels are followed to discover new channels, but after three breadcrumbs, the trail is abandoned. Higher depths could be used to reveal ever more channels in this space, but these are then increasingly likely to be unrelated to the initial set of channels and the topic of interest, and are in any case further removed and less likely to be discovered by a user.

In addition to the presence of these groups and channels and how one might find new ones through links posted in there, I am also interested in whether and how they form an 'ecosystem' of entities with a variety of functions. In earlier research on Dutch-speaking political Telegram groups, we distinguished "aggregator" groups, which served as entry points into a wider "Telegramsphere" by collecting messages forwarded from other, more specific or on-topic channels (Peeters & Willaert, 2022; Willaert et al., 2023). It is reasonable to expect that this particular sphere exhibits similar dynamics. But since the goal of advertising illegal goods and services would typically be to end with a transaction of some sort, it might be expected that there would additionally be channels with information about how to 'close a deal' or a boundary of the sphere at which point a conversation would ensue in a more private setting where for example payment details and delivery addresses might be exchanged. A further typology of actors, entry points and 'destinations' in this sphere is thus an additional goal of the analysis.

Finally, a large dataset of messages and images allows for a more qualitative content analysis. Here I am guided by the results of the initial steps in which the channels are categorised and mapped. Distinct clusters of channels with specific scopes and functions emerge, which can then be 'entered' and assessed on the basis of the content found therein. I am interested particularly in how the affordances of Telegram are operationalised in how 'business' is conducted on this platform, considering the illegal and risky nature of the services on offer, as well as responses to Telegram's content moderation (or lack thereof). I use "affordances" here particularly in the sense of perceived affordances, "how certain objects [are] designed to encourage or constrain specific actions" (Bucher & Helmond, 2017).

While the data analysed here is generally pseudonymous, I have chosen to not directly cite captured posts or discuss specific channels or sellers, considering the sensitive

nature of the matters discussed and the fact that "even if users are aware of being observed by others, they do not consider the possibility that their actions and interactions may be documented and analysed in detail at a later occasion" (Sveningsson Elm, 2017). My research interest is in the general dynamics of this space rather than individual sellers or conversations. Citations of messages are thus paraphrases or composites (following Markham, 2011's plea for "ethical fabrication") rather than verbatim quotes. The research discussed here was reviewed and approved by the Ethics Committee of the Faculty of Humanities, University of Amsterdam (reference FGW-4051).

Operationalisation of the method & dataset

Deciding on a starting point for the initial crawling exercise is less straightforward than it would be on other platforms. Telegram lacks a central index of channels or a comprehensive search function. While it is possible to search for groups or channels, this will only return those that have the exact query as (part of) their name, rather than being able to search based on content or the description of a channel.

Third-party sites have filled this niche and several sites exist that promise a "Telegram channels and groups catalog" (as on tgstat.ru); "a search engine to search for channels, groups, bots and users on Telegram" (telegramdb.org); or "the best Telegram channels" (telegramchannels.me). Though such sites have been used in earlier research (Tucci & Guilherme, 2023; Tikhomirova & Makarov, 2021; Hradziushka et al., 2023), it is not always clear how these sites construct their databases, but it is likely that they use the Telegram API to continually crawl the platform to discover and collect data about channels.

The initial 'seed list' of starting points for the crawl is then sourced from two places. First I used telegramchannels.me's feature of ranking popular channels by country to retrieve a list of the top 100 most popular Dutch Telegram channels (according to that site). Then I manually evaluated these channels and discarded those that were unrelated to the focus of the analysis, i.e., offers of illegal goods or services, leaving 11 relevant channels. Finally, I used Telegram's own search function to find other groups in the same space, searching for the word "handel" (trade) which appeared in some of these channels' names and had been identified in earlier work as a common descriptor in this space (Goldenberg et al. 2022). This provided an initial seed list of 20 channels. This is unlikely to be an exhaustive list of channels in this space, but it reflects a list of channels that might be easily found by someone who has no prior knowledge of this space or recommendations via word-of-mouth. In this sense this initial step is inspired by Light et al. (2018)'s 'walkthrough method' in that I "assume a user's position while applying an analytical eye" (p. 891). Further related groups can then be discovered via snowball crawling through links, adding new channels as they are encountered in previously collected messages, similar to how a user might find new and related places when browsing these initially discovered groups and channels. In other words, these groups and channels are those that are both clearly forbidden by the Telegram terms of service and easy to find without prior knowledge.

The final dataset thus was the result of a snowball crawling exercise with up to 3 hops from a seed list of 20 channels found from public sources. These 20 channels can be categorised according to the 'illegal service' they provide or advertise as follows:

- Sports match fixing/gambling (6 channels)

- Cigarettes, vapes, and other drugs (4 channels)
- General 'trade'-related groups (9 channels)

From these channels, the 5,000 most recent messages were collected. This limit was used as it became clear that several of these channels were frequented by what appear to be automated services that re-post the same advertisements at an interval. To avoid collecting the enormous amount of messages this generates, particularly in long-running channels, an arbitrary limit of 5,000 was used. This in any case captures the recent discourse in a channel, including those messages someone browsing the channel looking for services would see.

Of the 286 channels discovered through crawling these seed channels, 245 were relevant to the case study at hand; 41 were removed from the dataset, as they did not contain any messages advertising potentially illegal goods or services. Of these 245 channels, 22 contained more than 5,000 messages and were thus not captured completely. The other 225 channels contained 125 messages on average (median 18). Using the 4CAT social media analysis toolkit (Peeters & Hagen, 2022), 137,576 messages were captured in total, which additionally contained 20,691 distinct, attached images.

In 44 of the channels (18%), the most recent message was older than 3 months at the time of capture, indicating that these channels were not actively posted to. This does not necessarily indicate that they are abandoned or unused; if the purpose of a channel is to advertise a specific seller or service, or their contact information, an older message may still be relevant. Nevertheless, these channels can be contrasted with the more active channels populated by automated advertising chat bots, which in some cases received 5,000 messages within less than a single day (9 hours and 3 minutes for the most extreme example).

Analysis

Categorisation

The names of the channels give some indication of their purpose. In particular, 19 groups, among which were some of the largest ones, contained the word "handel" ("trade") in their name. "Handel" seems to be a general term used to refer to the trade of illegal goods and services; this was also found by Goldenberg et al. (2022) in their broader analysis of Dutch-speaking Telegram. Such handel group names often also include a reference to a geographical area which they focus on, such as a telephone area code, or province name. Messages in these more regionally specific groups more often concern the sale of drugs and these messages usually also list in which cities sellers do business, implying that deals are finalised in person. In contrast, sellers advertising medication, anabolic steroids, or other performance-enhancing drugs more often seem to send items via mail, evidenced by them referring to "track and trace numbers" or posting images of packages ready for mailing.

 Afhalen in ZOETERMEER  Opsturen ook optie met Track and trace €5,-via paysafe betaling of overmaking via bankrekening.

These 'trade' groups often were not specific to a single 'trader' but rather served as general channels in which people could advertise goods and services to people in the relevant area. This can be contrasted with the many channels that are ostensibly set up

to advertise the services of a single seller. Besides their difference in purpose, one can also distinguish a difference in content; whereas general groups often offer a never-ending stream of rapid-fire advertisements, seller-specific channels are usually a relatively static collection of messages in which an inventory of goods and instructions for contacting the seller are offered.

The 245 groups were inductively coded, based on the channel's messages and name, for the type of goods or services on offer, as well as their scope (general or seller-specific). 47 channels (19%) were classified as "general", while 198 channels (81%) were set up to advertise or facilitate a single seller's services (see also Table 7.1).

- **Drugs:** various types of drugs, including soft and hard drugs, but also illegally imported cigarettes, anabolic steroids, weight loss drugs and other medication.
- **Facilitation:** various services that would facilitate people who seek to offer their services via Telegram, including chat bots for automatically sending advertisements, web hosting services, software for automation of sales, and guides and software for obfuscating one's identity online.
- **Fireworks:** sales of fireworks. This was only a single channel.
- **Forgery:** various types of forgery and fraud, including forged ID documents, stolen bank cards and credit card details, and counterfeit money.
- **Gambling:** online casinos, and sale of information about fixed sports matches, often with promises of "guaranteed wins".
- **Policing:** channels in which information concerning the trustworthiness of various other channels and sellers is offered. Some are set up by sellers themselves and contain (for example) screenshots of messages intended to prove that seller's legitimacy, others present crowdsourced collections of reviews and testimonials, as well as warnings of scams.
- **Trade (*handel*):** General trade. These channels contain advertisements for one or more of the other types of services.

Type	Number of groups		Total messages		Channels with > 5,000 messages	'General' channels		'Specific' channels	
<i>Drugs</i>	87	36%	29,975	22%	4	9		78	
<i>Facilitation</i>	19	8%	263	0%	0	0		19	
<i>Fireworks</i>	1	0%	2	0%	0	0		1	
<i>Forgery</i>	65	27%	6,969	5%	0	0		65	
<i>Gambling</i>	17	7%	27,172	20%	4	1		16	
<i>Policing</i>	23	9%	1,837	1%	0	11		12	
<i>Trade</i>	33	13%	73,358	53%	14	26		7	
<i>Total</i>	245	100%	137,576	100%	22	47	19%	198	81%

Table 7.1 Data overview: inductive coding of 245 channels, after snowball crawling and cleaning of the initial dataset. Percentages have been rounded to the nearest full number and may not add up to exactly 100%.

Simply counting the amount of channels or messages for each type only provides part of the picture; the mere existence of a channel does not mean that it is viewed by many people, actively used by its owner, or even represents a real seller. While Telegram

offers various metrics by which this amount could be weighted to obtain a more representative impression of how popular each category is, these all have their own shortcomings. One might weigh by how often the messages in a channel have been viewed, but this metric is not available for all channels. It is also possible to weigh by channel subscriber number, but these are public channels that can be viewed without subscribing to them, and especially for channels that simply contain a description of a seller's services and contact information, there is little utility in subscribing. The amount of messages per channel also provides a distorted view, given the presence of automated advertising chatbots. I therefore use simply the number of channels per category in this discussion to indicate the relative prevalence of particular categories of goods or services.

Perspectives

A networked ecosystem

The coding of the data indicates a variety of purposes channels in this dataset may have, raising the question of how coherent it really is. One way to map the connections between channels is to view them as a network, and use the connections between various channels to better understand its general topology. Channels on Telegram can be linked in various ways: messages forwarded between channels are often used for this type of analysis, but other types of links exist, such as direct mentions of a channel in a message, or hyperlinks to the channel. Both are used often in this dataset; 64% (87,925) of the captured messages contained some sort of reference to a channel, group or user. They can therefore serve as a reliable indicator of inter-channel connections in this dataset.

Of the 245 channels, 135 contained no links to other channels. The remaining 110 channels linked to a total of 1,798 other channels, users or groups, with an average of 20 entities linked per channel, and a mean of 2 (see also Figure 7.1). In other words, while channels are often connected to other Telegram entities, a relatively small number of channels is highly connected, while the majority of channels is not linked to other channels or entities. This reflects the earlier observation (Peeters & Willaert, 2022) that a subset of channels functions as very active 'hubs' or 'aggregators' through which one may find other channels, while these 'destination' channels are then often points of contact or advertisement for individual sellers who have little reason to link to others in turn.

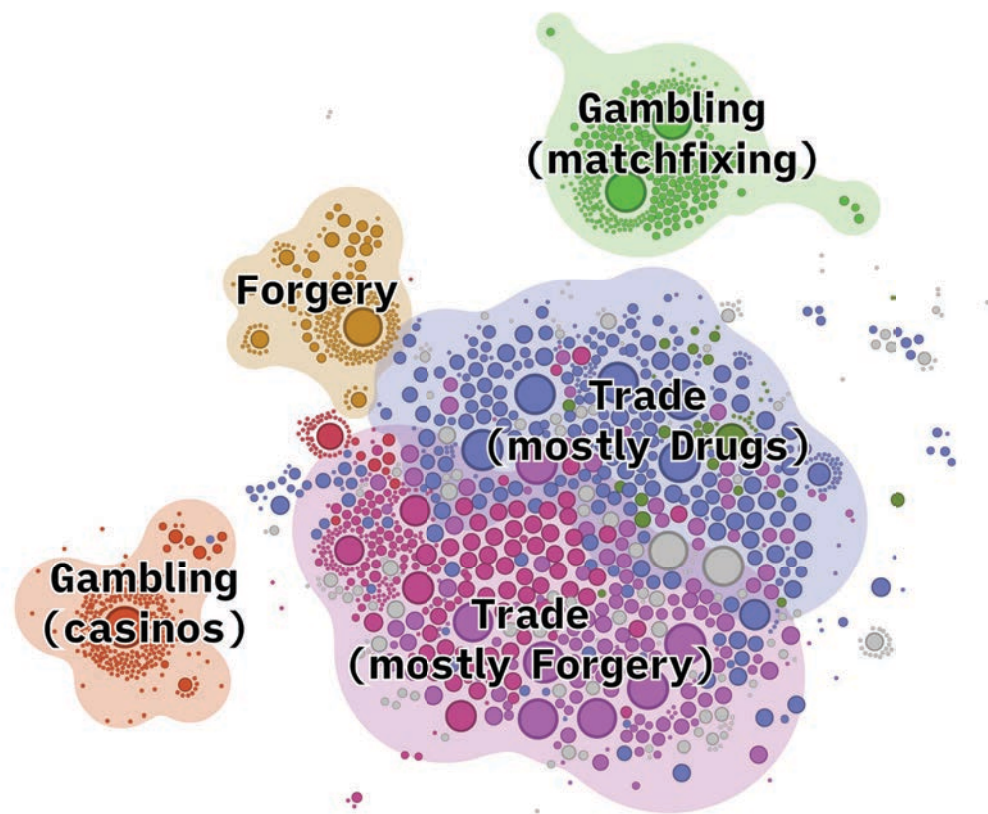


Figure 7.1 Visualisation of the channel network in Gephi. Layout: ForceAtlas 2 (Jacomy et al., 2014, gravity: 8.0). Colours represent modularity classes (Blondel et al., 2008, resolution: 1.0, modularity: 0.452). Some smaller outlying clusters are omitted from the visualisation. Each node represents a channel; edges (not shown in graph) represent a link, mention or forward from one channel to another. Channel node size represents degree (amount of connections); larger nodes have more connections. Source: author.

A bird's-eye view of the network confirms the thematic categorisation of channels discussed above. Clusters of channels that can be readily distinguished represent particular types of items or services on offer; a large mixed cluster near the center of the network comprises both general 'Trade' channels as well as many channels offering both drugs and forgery. Gambling channels can be subdivided in two types of gambling; one focused on casino-style gambling, often facilitated by chatbots and involving crypto-currency; the other focused on match-fixing, i.e. sports betting on matches of which the outcome has supposedly been arranged in advance. Information about such fixed matches is then promised to customers in exchange for payment. Interestingly, these clusters of gambling channels are not connected to each other, indicating that these are considered wholly separate types of business in this ecosystem.

Policing: vouches and trust lists

Next to the degree of connectedness, the categorisation of the channels also offers some indication for the diversification of this ecosystem. While many channels advertise some sort of service, some categories might be described as 'infrastructural', in that they do not serve to directly connect people with sellers, but rather help organise the ecosystem. This category of channels includes what one might label

'business to business' channels that offer (for example) bots that automatically post advertisements, and channels that promise to offer information about the trustworthiness of sellers (called 'trust lists', 'vouches', or 'scam lists').

The latter category is particularly interesting because the anonymous nature of Telegram, and the incentives for anonymity created by the nature of the services on offer in these channels means that establishing trust is important, because sellers are anonymous or pseudonymous, and buying the goods on offer may be a crime in itself. But it is also difficult, because it is not straightforward to ascertain whether a testimonial can be trusted in such an anonymous context. Two main types of channel can then be distinguished in this category: those maintained by sellers, containing testimonials or "vouches" about previous transactions completed by the seller and generalised "trust lists", which promise to offer lists of sellers with information about the extent to which they can be trusted.

Welkom in [kanaal] In dit register hebben we de real gecheckte verkopers !!

De mensendie hier instaan zijn uitgebreid gecheckt! ✅ Trusted lijst [kanaal] ✅

Scammers lijst [kanaal] ❌ Heb jij iets gekocht van een eerlijke seller laat het aan ons

weten !!✅ Of als jij helaas bent afhandig gemaakt en er niks voor terug hebt

gekregen, stuur aan ons dan proof zodat we de strijd samen kunnen stoppen 🙌🍅 DM

dan naar [kanaal] ✅📩

Seller-specific vouches are, in most cases, impossible to verify; they often take the form of screenshots of bank statements, chat conversations in which a customer expresses their satisfaction, or photos of track and trace codes or packaged items. All of these are easy to forge, and the nature of the business would perhaps make one additionally wary of trusting such 'evidence', coming from sellers whose wares include counterfeit money, fake identification papers, and stolen credit cards.

Likewise, the more general 'trust lists', whose operators are also anonymous, do not have a straightforward way of establishing trust in their information. Perhaps as a result of this lack of verifiability, these channels are engaged in a competition of their own, claiming that other lists are not accurate and that their own list is the most dependable. It is unclear to what extent such lists are then believed at all, though some of them have thousands of subscribers, indicating at least some interest in the contents. One channel in particular had approx. 7,400 subscribers, indicating broad interest; simultaneously, the channel only listed five supposedly trustworthy sellers, with no information about how their trustworthiness had been verified.

The art of the deal: contact details and projecting trustworthiness

Given the illegal nature of the business on offer, advertisements in these channels usually do not contain contact details such as website or street addresses. They will typically point to a username who can be contacted for more information, or another channel in which such information can be found. Interestingly, in many cases these references were not to Telegram, but other messaging platforms. 8,958 (6,5%) messages contained a reference to either WhatsApp or Signal, sometimes additionally linking to contacts on those platforms directly, sometimes including a specific phone number that can be contacted via these platforms, for example:

Beste mensen pas echt op! Ik baal hier nou van!! koop enkel bij ons. Let op met alle anderen die ons nadoen wij zijn de enige echte!! ben net even langs de weg stil gaan staan met de auto om dit te delen. 🤔 wij hebben alle bewijs om te laten zien dat wij de enige echte zijn. [kanaal] en dit whatsapp nummer: [nummer] Geen ander kanaal, nummer of account zijn wij. Wij zijn de enige in Nederland!! LET OP MENSEN. Ik heb geen zin om weer dit soort droevige berichten zien.

Some also offer a choice of contact methods:

⭐⭐ANABOLEN⭐⭐🔥24u per dag THUISBEZORGD🔥🔥HOOFDACCOUNT:
[kanaal] 🔥DIRECT VAN HET LAB BIJ JOU🔥 [kanaal] WHATSAPP 📞; [nummer]
SIGNAL 📞; [nummer]

This suggests that while Telegram is seen as a useful platform for advertising, it is not always considered the platform most suitable for completing a sale or arranging the delivery of items. The messages generally offer no clear indication of why this might be the case, but as discussed, Telegram is known to use less effective encryption technology than both Signal and WhatsApp. On the other hand, these channels do not have a channel system as developed as Telegram's. In other words, while Telegram affords the wide advertising of services, Signal and WhatsApp may afford doing business better, or at least offer an acceptable alternative.

The screenshots posted in aforementioned 'vouching' channels offer a glimpse into what might happen after one approaches a seller - or in any case what the seller wants one to believe will happen. Many vouches include screenshots of chat conversations between seller and customer, and often include the customer sending a photo of the received items, and telling the seller how happy they are with the rendered services (see Figure 7.2).

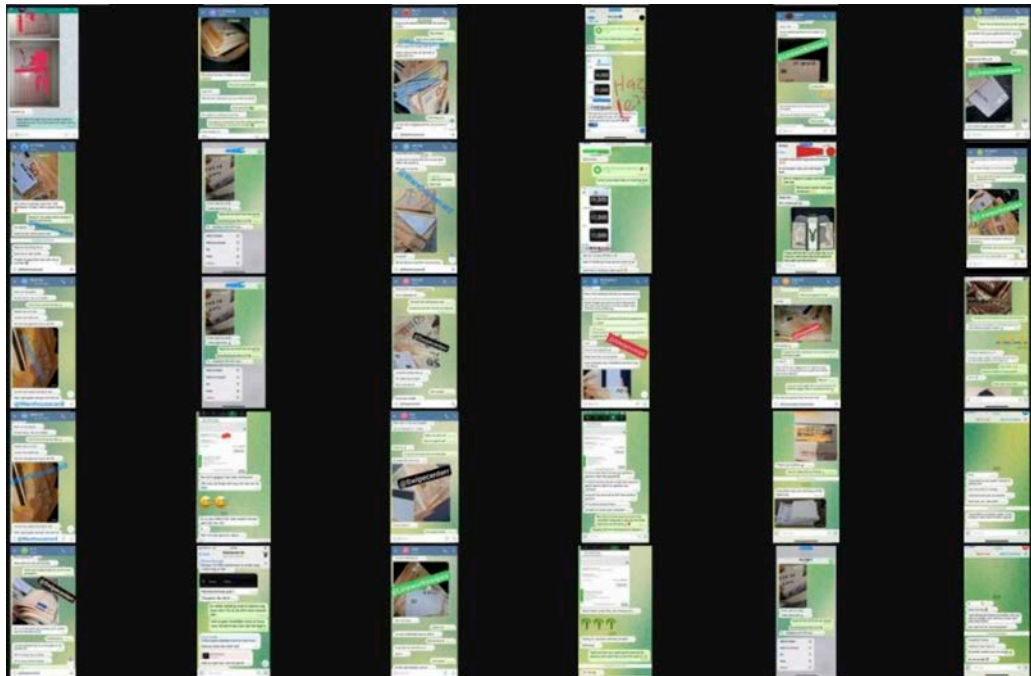


Figure 7.2 A cluster of screenshots from Telegram messages, organised with PixPlot. The crop is part of a larger cluster of approximately 1,000 such images. Source: author.

These screenshots are usually, though not always, taken of Telegram conversations. This might indicate that most of the sales take place on Telegram; on the other hand, of the three mentioned platforms - Telegram, Signal, and WhatsApp - Telegram is the easiest to use with multiple accounts. The app allows one to quickly switch between accounts, which would make it relatively easy to have a conversation with oneself. The predominance of Telegram screenshots in these vouches is therefore not necessarily a reliable indicator of the platform used for sales, but may also reflect the fact that Telegram affords the fabrication of such vouches best.

Attitudes to and traces of content moderation

In January 2025, 3 months after the initial data collection, 121 of the 245 seed channels were no longer available. Most deleted channels were originally categorised as offering drugs (53, or 61% of that category), forgery and fraud (31, or 48%) or general 'trade' (23, or 70%). Telegram does not offer an indication of why the channel is no longer available; one simply sees the message "Sorry, this user doesn't seem to exist" when trying to open the channel in the Telegram app. This development is somewhat surprising, given the documented history of Telegram doing very little moderation of its platform, particularly not in these categories of content.

In these three months, Telegram did however announce that they would be moderating their platform more actively, in response to regulatory pressure and the arrest of CEO Pavel Durov in France. The disappearance of these groups might therefore indicate that this is indeed the case, and that Telegram is not as attractive of a platform for the sale of illegal goods and services as it once was. That most deleted channels, relatively speaking, were general 'trade' channels also points towards this explanation, since these channels were the most active in terms of messages and subscribers, and were thus more visible compared to some of the more specific or specialised channels. There are however indicators in the data that suggest that this ecosystem is also volatile on a more fundamental level. Many channels provide links to 'backup channels' that one can then use if the 'main channel' disappears; or have names or descriptions suggesting they are such a backup channel. While most mentions of backup channels do not provide a reason for having one, some do indicate that increased content moderation is one of the possible reasons to have such a channel.

In a channel promising "leads" of good sales for forged documents:

I'm sure many will have heard the news about Telegram sharing personal data with authorities upon request. 📄 🔍

In response, we have set up a backup channel on Signal, in case something happens. We're also looking at the possibility of making an automated bot on Signal for people that want to fully switch over.

In a 'forgery' channel offering forged credit cards:

Recommend everyone to please join at least a single backup server so you are updated if Telegram bans us

Others indicate that abuse of Telegram's 'reporting' functionality can trigger deletion of channels, for example in a channel providing lists of scammers:

Als er wat gebeurt met de kanaal ivm met vele scammers exposed en hun mij proberen te wegspammen join dan de backup kanaal

More prosaically, backups can provide an insurance against accidents:

Dit was de backupkanaal van [kanaal] en de normale groep van [kanaal] is door foutje van collega verwijderd wat heel erg dom van hem is.

It is difficult to say to what extent Telegram's actual moderation activity is a direct cause of this practice of operating backup channels. Note that the messages above mention news about Telegram moderation, or contingencies in case Telegram bans an account, rather than tangible examples of moderation. One might see this as "imagined affordances" at play - people's "expectations about their communication technologies, data, and media that, in effect and practice, shape how they approach them and what actions they think are suggested" (Nagy & Neff, 2015).

In other words, no action on Telegram's behalf is strictly required to make people prepare for the eventuality that their channel is banned. One might alternatively think of such a channel as "temporary autonomous zone" - in Hakim Bey (1991)'s words, "a guerilla operation which liberates an area [...] and then dissolves itself to re-form elsewhere/elsewhen, before the State can crush it". That is, the ephemerality of these channels may be by design, with them forever disappearing and re-forming, or at least projecting the risk of this happening, as a natural consequence of facilitating a fundamentally illegal kind of exchange.

It is then plausible that the lack of stability of the dataset is a result of various factors, including Telegram's content moderation, the recent increase thereof, and the inherently volatile nature of a business ecosystem built upon the sale of illegal goods and services. Nevertheless, it is clear that some operators of these channels at least recognise that content moderation is a risk to their activity, and act accordingly.

Conclusion: a dark alley, but with a little ray of light?

Telegram has long enjoyed the status of an unmoderated, 'free for all' type of platform, on which activities that were banned or moderated on other platforms were possible and easy to find. On the basis of the case study presented here, this reputation is justified, as it has been found to be in other case studies (such as for alt-right politics, and the spread of CSAM or NCII). A relatively 'naïve' approach to finding channels in which illegal goods and services were on offer readily revealed a substantial amount of these, even when focusing on a specifically Dutch context. From here it was then straightforward to find other, related channels where such offers could also be found via snowballing, using forwarded messages, links, and mentions of other channels.

An inductive coding and reading of some of the more easily discovered examples of such channels reveals that they form a diversified ecosystem in which channels have various functions and specialisations. Many groups more or less operate as 'store fronts', listing the items for sale and providing details about how they can be acquired. But other channels are more akin to classified ad listings, offering a place where channels can be discovered easily; while yet others serve to police this ecosystem by 'naming and shaming' untrustworthy sellers or providing supposed proof of a seller's legitimacy.

Establishing legitimacy and trustworthiness emerged as an important preoccupation of channel operators; next to the specialised ‘vouching’ channels, sellers often took care to mention their reliability and offer elaborate ‘shopping lists’ of available items festooned with related emoji and listing contact details. These often point at channels other than Telegram itself, indicating that in the wider practice of illicit trade on messaging apps, a comprehensive analysis may need a multi- or cross-platform approach to map the issue thoroughly. This is not likely to be straightforward – the platforms indicated as being of interest here, Signal and WhatsApp, are also very difficult to study from an ‘outsider’ perspective – which is perhaps also the reason they seem preferred by some actors in this space. In any case, this single-platform analysis indicates that the ecosystem is coherent enough for a form of competition to appear, where sellers seek to establish themselves as more attractive than others, sometimes going as far as to involve the impersonation of other sellers.

All of this is done relatively openly, lending credence to the image of Telegram as having a laissez-faire approach to the enforcement of its platform’s rules (few as they are) at best. Considering that this ecosystem can be found and mapped with relative ease by a researcher with only access to public platform data, its presence on the platform seems to be willfully tolerated by Telegram. More effective moderation would still be challenged by some of the peculiarities of this space, such as its links to other ‘platform jurisdictions’ on Signal and WhatsApp and the ease with which anonymous accounts can be set up on the platform. Nevertheless, the fact that actors operate openly, freely sharing telephone numbers and other details, shows that they have little to fear from the platform at present and would perhaps be most vulnerable to ‘extraterritorial’ measures, such as when the Dutch police confiscated the phone of a sanctioned channels’ administrator to close it (ECLI:NL:RBDHA:2021:11250, *Rechtbank Den Haag*, 21/504, 2021), circumventing the platform’s own moderation mechanisms altogether.

There are, however, signs that Telegram’s attitude towards content moderation is slowly changing, possibly in response to recent regulatory pressure and the arrest of its CEO. Channels in this space seem to disappear at an impressive rate, and often warn viewers up-front that they might vanish and where to find them if that occurs. A longitudinal analysis would be required to verify if this is a new development or a fact of life for those who operate in this business, but it in any case shows that Telegram is not fully inert with regards to policing its platform and enforcing the part of its terms of service that forbids the trade of illegal goods and services.

References

- Agarwal, S., Ananthakrishnan, U. M., & Tucker, C. E. (2022). Content Moderation at the Infrastructure Layer: Evidence from Parler (SSRN Scholarly Paper 4232871). Social Science Research Network. <https://doi.org/10.2139/ssrn.4232871>
- Baumgartner, J., Zannettou, S., Squire, M., & Blackburn, J. (2020). The Pushshift Telegram Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 840–847. <https://doi.org/10.1609/icwsm.v14i1.7348>
- Bey, H. (1991). *The Temporary Autonomous Zone, Ontological Anarchy, Poetic Terrorism*. Autonomedia.

- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Bucher, T., & Helmond, A. (2017). The Affordances of Social Media Platforms. In *The SAGE Handbook of Social Media*. Sage Publications.
- Dargahi Nobari, A., Reshadatmand, N., & Neshati, M. (2017). Analysis of Telegram, An Instant Messaging Service. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2035–2038. <https://doi.org/10.1145/3132847.3133132>
- ECLI:NL:RBDHA:2021:11250, Rechtbank Den Haag, 21/504,
- ECLI:NL:RBDHA:2021:11250 (Rb. Den Haag October 8, 2021). <https://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBDHA:2021:11250>
- Gill, G. (2021). Fascist cross-pollination of Australian conspiracist Telegram channels. *First Monday*. <https://doi.org/10.5210/fm.v26i12.11830>
- Goldenberg, O., Hofman, E., & Veerbeek, J. (2022, August 17). Hoe Telegram van een baken van vrije meningsuiting veranderde in een duistere marktplaats. *De Groene Amsterdammer*. <https://www.groene.nl/artikel/klokkenluiders-en-korte-vleeskeuring>
- grugq, thaddeus t. (2015, November 18). Operational Telegram. Medium. <https://medium.com/@thegrugq/operational-telegram-cbbaadb9013a>
- Hradziushka, A. A., Vikhrova, O. Y., & Velikaborats, H. F. (2023). Digital Transformation of Public Life in Russia and Belarus: Dialogue Between the Government and Citizens on New Media Platforms. 2023 Communication Strategies in Digital Society Seminar (ComSDS), 106–111. <https://doi.org/10.1109/ComSDS58064.2023.10130427>
- Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE*, 9(6), e98679. <https://doi.org/10.1371/journal.pone.0098679>
- Jakobsen, J. B. (2015). A practical cryptanalysis of the Telegram messaging protocol. Aarhus University.
- de Keulenaar, E., Magalhães, J. C., & Ganesh, B. (2023). Modulating moderation: A history of objectionability in Twitter moderation practices. *Journal of Communication*, 73(3), 273–287. <https://doi.org/10.1093/joc/jqad015>
- Light, B., Burgess, J., & Duguay, S. (2018). The walkthrough method: An approach to the study of apps. *New Media & Society*, 20(3), 881–900. <https://doi.org/10.1177/1461444816675438>

- Markham, A. (2012). Fabrication as Ethical Practice: Qualitative inquiry in ambiguous Internet contexts. *Information, Communication & Society*, 15(3), 334–353.
<https://doi.org/10.1080/1369118X.2011.641993>
- Miculan, M., & Vitacolonna, N. (2023). Automated verification of Telegram's MTProto 2.0 in the symbolic model. *Computers & Security*, 126, 103072.
<https://doi.org/10.1016/j.cose.2022.103072>
- Nagy, P., & Neff, G. (2015). Imagined Affordance: Reconstructing a Keyword for Communication Theory. *Social Media + Society*, 1(2), 2056305115603385.
<https://doi.org/10.1177/2056305115603385>
- Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. *Computational Communication Research*, 4(2), 571–589.
<https://doi.org/10.5117/CCR2022.2.007.HAGE>
- Peeters, S., & Willaert, T. (2022). Telegram and Digital Methods, *M/C Journal*, 25(1).
<https://doi.org/10.5204/mcj.2878>
- Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, 35(3), 213–229.
<https://doi.org/10.1177/0267323120922066>
- Semenzin, S., & Bainotti, L. (2020). The Use of Telegram for Non-Consensual Dissemination of Intimate Images: Gendered Affordances and the Construction of Masculinities. *Social Media + Society*, 6(4), 2056305120984453.
<https://doi.org/10.1177/2056305120984453>
- Shehabat, A., Mitew, T., & Alzoubi, Y. (2017). Encrypted Jihad: Investigating the Role of Telegram App in Lone Wolf Attacks in the West. *Journal of Strategic Security*, 10(3), 27–53.
- Simon, M., Welbers, K., C. Kroon, A., & Trilling, D. (2023). Linked in the dark: A network approach to understanding information flows within the Dutch Telegramsphere. *Information, Communication & Society*, 26(15), 3054–3078.
<https://doi.org/10.1080/1369118X.2022.2133549>
- Sveningsson Elm, M. (2009). How do various notions of privacy influence decisions in qualitative internet research? In A. Markham & N. Baym (Eds.), *Internet Inquiry: Conversations About Method* (pp. 69–97). SAGE Publications, Inc.
<https://doi.org/10.4135/9781483329086>
- Telegram. (2013, October 23). Telegram F.A.Q. (archived version of 23 October 2013). <https://web.archive.org/web/20131023153707/https://telegram.org/faq>
- Telegram. (2023, December 1). Telegram FAQ (archived version of 1 December 2023). Telegram. <https://web.archive.org/web/20231201002007/http://telegram.org/faq>

- Telegram. (2025a, January 1). Telegram Moderation Overview (archived version of 1 January 2025). Telegram.
<https://web.archive.org/web/20250106222136/https://telegram.org/moderation>
- Telegram. (2025b, January 1). Terms of Service (archived version of 1 January 2025). Telegram. <https://web.archive.org/web/20250101035732/https://telegram.org/tos/eu>
- Telegram. (2025c, January 27). Telegram FAQ (archived version of 27 January 2025). <https://web.archive.org/web/20250127134203/https://telegram.org/faq>
- Thiel, D., DiResta, R., & Stamos, A. (2023). Cross-Platform Dynamics of Self-Generated CSAM. Stanford Internet Observatory, Cyber Policy Center.
- Tikhomirova, K., & Makarov, I. (2021). Community Detection Based on the Nodes Role in a Network: The Telegram Platform Case. In W. M. P. van der Aalst, V. Batagelj, D. I. Ignatov, M. Khachay, O. Koltsova, A. Kutuzov, S. O. Kuznetsov, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, M. Pelillo, A. V. Savchenko, & E. Tutubalina (Eds.), *Analysis of Images, Social Networks and Texts* (pp. 294–302). Springer International Publishing. https://doi.org/10.1007/978-3-030-72610-2_22
- Tucci, G., & Guilherme, J. (2023). Mapping Post-Truth Narratives: Understanding Brazilian narratives about the war on Telegram through prompting (Digital Methods Summer School). University of Amsterdam.
<https://wiki.digitalmethods.net/Dmi/UnderstandingBrazilianNarrativesAboutTheWar>
- Venturini, T., & Munk, A. K. (2021). *Controversy Mapping: A Field Guide*. Polity.
- Visser, P. (2024, June 25). Stroomstootwapen kopen op Telegram: "Ze wilden een foto van een stapel geld." *NPO Radio 1*.
<https://www.nporadio1.nl/nieuws/onderzoek/49755c9a-a010-4201-94cb-1674cd7ef96f/stroomstootwapen-kopen-op-telegram-ze-wilden-een-foto-van-een-stapel-geld>
- Voskuil, K. (2018, August 13). Grootscheepse handel in pillen, coke en vuurwapens via Telegram. *AD.nl*. <https://www.ad.nl/binnenland/grootscheepse-handel-in-pillen-coke-en-vuurwapens-via-telegram~aa16c789/>
- Willaert, T., Peeters, S., Seijbel, J., & Raemdonck, N. V. (2022). Disinformation networks: A quali-quantitative investigation of antagonistic Dutch-speaking Telegram channels. *First Monday*. <https://doi.org/10.5210/fm.v27i5.12533>

8. Grey areas of content moderation: A trace analysis of Pornhub

Lucia Bainotti

Abstract

The research explores the remaining grey areas of Pornhub's content moderation by analysing the platform's history of content moderation and the traces of content left behind by (moderated) videos on the platform. To do so, it relies on policy analysis, data gained by continually archiving content from Pornhub's "New Videos" page in the Netherlands, and an analysis of the moderation of search queries. Pornhub's definitions of prohibited content and the hierarchy of concern emerging from the policies are outlined, highlighting tensions and contradictions. The analysis of moderation traces shows that the definition of "disabled videos", the most common status associated with moderated videos in the dataset, remains unclear and inconsistent. Further, video tags, including those on moderated videos, appear highly sanitised, indicating user compliance with content moderation policies rather than attempts to evade them. Thirdly, while certain search queries are not allowed on the website, they can be relatively easily bypassed by using semantically similar alternatives to blocked search terms. While the content returned may not necessarily be unlawful, the platform's related search suggestions guide users toward increasingly specific and niche material, potentially leading to borderline content. By following related search queries, tutorials promoting and explaining how to generate non-consensual synthetic imagery, which remain unmoderated, were discovered.

Keywords: content moderation, image-based sexual abuse, Pornhub, pornography, trace analysis.

Introduction

Pornhub has grown massively since its launch in 2007, becoming one of the most popular porn platforms to date, with over 11.4 billion mobile visits from global users as of January 2024 (Statista, 2024). Pornhub's parent company, Aylo (formerly known as Mindgeek), owns other relevant brands and companies in the porn industry, such as premium pay sites like Brazzers, other video sharing platforms (Youporn, Redtube and Tube8), and advertising networks like TrafficJunky, thus playing a pivotal role in the production and consumption of adult content. Understanding how Pornhub conceives and enforces content moderation, and what content is allowed or prohibited, is therefore crucial, especially considering the platform's need to assure users' safety, safeguard freedom of expression and sexual fantasies, while at the same time gain economic profit.

While Pornhub's approach to platform governance has become stronger and more sophisticated over the years, the platform continues to face scrutiny about its content moderation processes (McGlynn & Woods, 2022) and its handling of user data (Rama et al., 2021). In this context, the research seeks to analyse the remaining grey areas of Pornhub's content moderation by analysing the platform's history of content

moderation and the traces of content left behind by (moderated) videos on the platform.

The moderation of sexual content is an issue for digital platforms, and presents particular challenges when adult content is involved. Part of the problem lies in how to define obscene, objectionable and sexually suggestive content and where to draw the line between sexual fantasies, kinks and fetishes on the one hand, and prohibited or unacceptable content on the other. Similarly to what happens for mainstream social media platforms (Tiidenberg, 2021), the definitions of unacceptable content as well as the implementation and enforcement of moderation on adult platforms is often opaque and inconsistent (Henry & Wit, 2018; Stegeman, 2024), when not directly impacting performers and sex workers (Blunt and Stardust, 2021). Other actors, such as business partners and financial services, also have impacts on platform governance, de facto acting as co-regulators (Franco & Webber, 2024). Scholars have also raised concerns about content that, while not unlawful, could still be harmful for viewers, especially young users (Craig, 2024). Research shows that video titles on popular porn tubes (including Pornhub) often describe sexual activity that constitutes sexual violence (Vera-Gray et al., 2021). While these contributions make clear that titles are not fully reflective of the actual video content, they also argue for the role of mainstream pornography in normalising sexual violence (Craig, 2024).

As of today, information on the moderation of adult platforms comes from platforms' transparency reports, academic research that critically analyses community guidelines and terms of service (Stegeman, 2024), and qualitative studies on how platform governance is understood, perceived, and lived by sex workers and content creators (Are & Briggs, 2023). The present research, instead, offers a novel approach to the study of platform governance of adult platforms, rooted in critical platform analysis and digital methods (Rogers, 2024). To analyse the history and traces of content moderation on Pornhub, the research follows the approach of trace research (De Keulenaar & Rogers, 2025). This technique is suitable to analyse Pornhub, as it takes as a point of departure the platform's "efforts to cleanse and police the site of rule-breaking or offending content" (De Keulenaar & Rogers, 2025) as a starting point to investigate the remaining grey areas in content moderation.

The research is articulated in three steps. First, it explores the history of content moderation policies to understand what is likely to be moderated, focusing on how the platform defines and categorises "prohibited" content. Second, building on a dataset of 60,000 videos collected through the dynamic archiving technique (De Keulenaar & Rogers, 2025), it seeks to reverse-engineer content moderation and analyse the traces of moderated content, particularly the motivations for banned videos, i.e., the labels that users see when surfing the website, instead of the actual videos and video tags. Finally, the study returns to Pornhub's interface to analyse the traces of search queries in a qualitative, exploratory manner to understand search query moderation. In the conclusion, the chapter summarises the grey areas of content moderation emerging from the results.

Pornhub's history of content moderation

Pornhub's moderation before and after 2020

In this initial phase, spanning from its inception to 2020, Pornhub primarily relied on its terms and conditions as the foundation of its platform governance, which outlines the platform's key concerns, including a zero-tolerance stance towards Child Sexual Abuse Material (CSAM). Despite these stated commitments, the platform initially adopted a largely hands-off approach to content moderation, explicitly stating that it would not review content in advance, and would be reactive in adapting to scandals and requests to moderate content from the outside. Concerns about the platform's role in enabling exploitation and non-consensual content were already growing in 2019, in relation to revelations about the GirlsDoPorn (GDP) sex trafficking operation, whose content was hosted on Pornhub (Cole, 2019). A petition launched to shut down the website, which gained momentum from anti-pornography and religious organisations, garnered over two million signatures by the end of 2020. As existing literature points out (Webber et al., 2021), these groups often pursued broader agendas that go beyond Pornhub, advocating against the pornography industry as a whole.

Increased regulatory scrutiny was sparked only after the *New York Times* op-ed "The children of Pornhub" (Kristof, 2020) published in December 2020 that accused the platform of hosting non-consensual videos and profiting from child-sexual abuse material. This piece triggered relevant changes almost overnight, with important impacts on the platform's structure and governance. Visa and Mastercard cut ties with Pornhub, prompting the platform to delete millions of videos from unverified users. Subsequently, Pornhub implemented other changes, including restricting uploads to verified uploaders, implementing verification processes and record keeping requirements, banning downloads, and improving the platform's enforcement of content moderation. These changes were framed as an effort toward "the safety of our community" (Pornhub Help Center, 2020) and to limit the presence of malicious content. They also created harm to content creators and sex workers, however, who saw their incomes negatively impacted by Visa and Mastercard's withdrawal and their practices reshaped by the new structure of the website and the labour this adaptation required (Gregory, 2024).

By restricting uploads to verified users who join the 'Model Program', Pornhub started to introduce a particular form of a priori governance over user profiles and, consequently, user-generated content. To become a verified content creator¹, users are asked to verify their age and identity via a third-party biometric identity verification service, and must obtain and provide identification for every performer appearing in their content prior to uploading material to the platform (Pornhub Help Center, 2024a). By monitoring in advance the profiles and processes to upload content, Pornhub aims to reduce the spread of illegal and illicit content, while de facto placing responsibility on identifiable content producers (Gregory, 2024).

¹ Verified Content Creators include verified models within the Model Program and verified studio and production companies within the Content Partner Program (Pornhub Help Center, 2024a).

In the aftermath of the 2020 events, Pornhub also started reinforcing its content moderation at two levels: automated content moderation and human moderation. Uploaded content first undergoes automated scanning to detect violations or illegal material before going through human review. Moderators then assess compliance with the terms of service and community guidelines. If approved and performer verification is met, content is typically published within 24 hours, but only after completing the full moderation process (Pornhub Help Center, 2024a). Moreover, Pornhub relies on community-based content moderation strategies by means of its Trusted Flagger Program and on other users' reporting of inappropriate content through content removal requests forms.

The acquisition by Ethical Capital Partners

Pornhub's approach to content moderation received a new boost starting in 2023 when MindGeek was acquired by the Canadian private equity firm, Ethical Capital Partners (ECP). Following the acquisition, MindGeek was renamed Aylo, signalling an attempt to distance itself from the controversies of the past and rebuild its public image. Ethical Capital Partners expressed its intention to address existing challenges by reinforcing Pornhub's commitment to trust and safety (Ethical Capital Partners, 2023). Accordingly, Aylo's website highlights the necessity to create an "updated identity" for its companies and "re-focus its efforts to lead by example, through transparency and public engagement" (Aylo, 2023).

Such a commitment can be seen, among other things, in a reinforcement of the trust and safety division and the systematisation of Pornhub's community guidelines and related content policies. With the new ownership, the platform has further developed its verification processes, adding new automated tools for the detection and fingerprinting of illegal content, and has attempted to increase its transparency in communication. Recently, the platform has announced to strengthen its verification process for uploaded content, introducing an even more stringent policy that will require performers to upload proof of consent (i.e., Signed Release Forms) from all participants in their videos in addition to their IDs (Pornhub Blog, 2025). This step confirms one important shift the platform underwent over the years from a very hands-off and reactive approach to content moderation to one based on more proactive interventions.

The push for greater safety and accountability appears to extend to other adult platforms owned by Aylo. Pornhub, however, has historically faced the most scrutiny due to its prominence, which also led to greater investment in its growth and oversight. In this sense, Pornhub has undergone a similar trajectory to other mainstream social media platforms like YouTube.

The Digital Services Act and the splintering of moderation

Increased pressure for more transparency and governance measures intensified in 2024, with Pornhub having to deal with more regulation of its platforms (Gillespie, 2017) by different political organs in the US, UK, and the EU. In the US, Pornhub disabled access to users in twelve states, including Texas, Utah and Kansas, between March and July 2024, continuing its stance against state-imposed age verification laws designed to prevent children from accessing adult websites (Kastrenakes, 2024).

In the EU, instead, Pornhub is currently dealing with the requirements introduced by the EU Digital Services Act (DSA), which aims at creating a safer digital space where the fundamental rights of all users of digital services are protected. Despite their large popularity and being very likely to have met the required threshold of 45 million monthly active users in the EU to qualify as Very Large Online Platforms (VLOPs) that must follow more stringent rules, porn platforms like Pornhub were not included in the first round of designations. Only after lobbying from civil society organisations and digital rights experts (including AiForensics, European Digital Rights, and the European Sex Workers' Rights Alliance) was Pornhub designated a VLOP in December 2023, together with Xvideos and Stripchat (European Commission, 2023).² The pressure to hold large porn platforms accountable is related to their attempt at eluding their responsibilities as a VLOP by providing figures that seemed to be a misrepresentation of their monthly average users in the EU (Tar, 2023). As a result, Pornhub now has to comply with the more stringent rules applied to VLOPs, including analysing their specific systemic risks with regard to dissemination of illegal content or content threatening fundamental rights, providing more transparent information about their content moderation processes, and improving their accountability by allowing external audits.

While evaluations of the changes brought by the introduction and implementation of the DSA are ongoing, it is relevant to note a progressive adaptation of platform regulation across different political and geographic contexts, which has led to a process of splintering of platform governance (Ahn et al., 2023). This splintering is evident in access restrictions imposed in specific jurisdictions as well as in updates to reporting and flagging mechanisms designed to align with evolving legal frameworks.

Definitions of prohibited content: Pornhub's hierarchy of concerns

Beginning in 2023 Pornhub further systematised and expanded its documentation policies, making Pornhub Help Center the core of its Trust and Safety (similar to Meta's Transparency Center and YouTube Help) and creating clearer connections between the different policy documents (summarised in Appendix 1).

With this more structured organisation comes a tiered definition of prohibited content and uses, which showcases Pornhub's hierarchy of concerns about content moderation. With its policies, the platform provides information about the types of content prohibited and makes explicit the different degrees of acceptability of content on the platform.

According to this hierarchy, illegal content, including any illegal activity and content depicting minors, are prohibited. Pornhub reiterates its zero-tolerance approach against CSAM multiple times and consistently across all its policies. Illegal content also includes human and sex trafficking, animal cruelty, and "violence". What constitutes violence is a very broad and controversial term, especially in the domain of pornography, where practices like bondage are acceptable. To clarify their approach towards violent content, Pornhub has introduced its Violent Content Policy Guidelines.

² Another major adult platform, XNXX, was added to the VLOPs list in July 2024.

Non-consensual content is the second level, as demonstrated by the Non-consensual Content Policy, which was already introduced in 2020 with the Child Sexual Abuse policy. Non-consensual content includes material depicting non-consensual acts, the recording or distribution of intimate imagery without consent, the unauthorised use of an individual's likeness as well as an array of "sensitive themes" like possessions or spells and sleep.

Next, there is the content that harms or may cause harm to individuals (and its related policy) and inauthentic or unauthorised use of the platform, which includes spam, misinformation, and copyright infringement.

The last level of the hierarchy is "otherwise unacceptable" content, which encompasses a variety of prohibited material, such as incestuous acts, content involving drug use, and the featuring of a sponsored product which offers or promotes illegal activity. The measures used to moderate content reflect this hierarchy, with more severe measures for illegal content and less severe ones about otherwise unacceptable content. CSAM results in the immediate removal and banning of the uploader as well as in reporting to legal authorities and the National Center for Missing and Exploited Children (NCMEC). In all the other cases, content in violation of the terms is reviewed and removed and, at times, fingerprinted to avoid reuploads (as in the case of non-consensual material). The policies also mention the possibility that uploaders' accounts are suspended or permanently terminated where appropriate, but do not provide more information about what leads to these measures or information about how long this suspension can last. Furthermore, there is no mention of other forms of visibility moderation, such as the de-ranking of inappropriate content, as in the case of other mainstream social media platforms such as Facebook.

The improvements in Trust and Safety and the presence of a hierarchy of concern showcase Pornhub's commitment to improve content moderation, provide more transparent information, and make available educational resources for its community of users. At the same time, however, some room for discretion persists, particularly in areas where moderation challenges intersect with Pornhub's role as an intermediary balancing diverse interests, users, and stakeholders.

Above all, there are subtle yet significant nuances in the way Pornhub's policies are phrased. In the case of CSAM, Pornhub reiterates various times that content that depicts any person under 18 years of age is prohibited "whether real or simulated". The same phrasing is also used for the moderation of non-consensual material, and a specific type of unacceptable content, incestuous materials. In the latter case, Pornhub goes even further by stating that "depictions, representations and role-plays" of incest are prohibited. However, there is some disjuncture between these formulations and what appears on the websites, particularly in relation to depictions of non-consensual and incestuous content. As existing research has shown, content that "simulates" non-consensual acts and depicts incest can largely be found on the platform (Craig, 2024). Pornhub also has a dedicated porn category called "Step Fantasy" and terms like "stepmom" rank high among the most popular searches (Pornhub Insights, 2024). While Pornhub takes seriously issues of CSAM and non-consensual content, there is instead some leeway in the moderation of simulations and roleplay. Indeed, the platform relies on and at the same time fosters a fine line between simulations and role-plays, sexual fantasies, and pornographic genres crucial to its intermediary role.

This issue highlights the presence of some content that while not illegal or prohibited can be considered borderline. This is one of the more challenging categories of content to moderate, as it's difficult to understand where to draw a line between what is acceptable or not. Borderline cases in a porn site are interesting, as content could range from "non-consensual" recording, e.g., with CCTV, to an intentional display of exhibitionism in a public place as in the case of flashing.

Notably, Pornhub has expanded its policies to address the spread of AI-generated content as a particular type of non-consensual content that "is used in a realistic manner to alter the speech or behavior of the individual(s) depicted" (Pornhub Community Guidelines, 2024). Despite these measures, as findings reported here show, certain risks continue to bypass moderation, highlighting the ongoing challenges of regulating AI-generated content effectively.

Methodology: A trace research of Pornhub

The research applies an established methodological approach in the study of content moderation known as trace research (de Keulenaar & Rogers, 2025) to analyse the moderation of pornographic content on Pornhub, thus expanding the potential of this technique beyond the analysis of mainstream social media platforms such as YouTube and Twitter/X. Trace research consists of collecting the traces left behind by content on digital platforms and reconstructing the "scene" in which content or user data have disappeared after content moderation (De Keulenaar & Rogers, 2025). This way, it is possible to reverse-engineer how content moderation is implemented and which grey areas persist. The research proceeds in two steps: firstly, it seeks to reverse engineer content moderation by means of the dynamic archiving of Pornhub data, looking at two types of traces: motivations for banned videos showcased at the level of the interface (Figure 8.1) as well as video tags. Secondly, the study returns to Pornhub's interface to test the moderation of search queries, assessing whether and how search query restrictions can be circumvented.

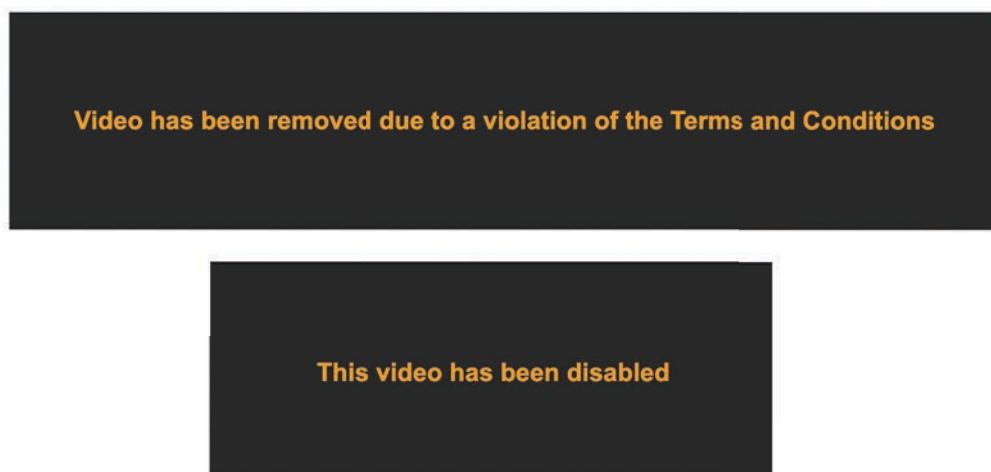


Figure 8.1 Examples of reasons for video removal, as displayed on Pornhub's interface. Source: author.

Reverse engineering content moderation

For this first phase, data was collected from Pornhub's "New Videos" page during the period of August to November 2024. First, an HTTP request was sent to the "New Videos" page every minute, extracting video metadata from the HTML and recording it in our database. For each new video identified on the page, a custom-made script collected additional metadata from the individual video pages (e.g., video title, categories, tags, views, etc.) at four specific intervals: one hour, one day, one week, and one month after the video was first recorded. The data collection follows the principle of dynamically archiving data (de Keulenaar & Rogers, 2025), which allows one to track the changes in the availability and statuses of online content and capture the intermittent nature of content moderation.

The final dataset used for this investigation consists of roughly 60,000 videos uploaded on the "New Video" page. This page features recently uploaded videos that the platform's algorithm considers relevant to users in the Netherlands as determined by their IP address. Despite this geographic focus, the videos featured are not confined to Dutch language or content. This reflects Pornhub's different levels of algorithmic selection and ranking of content (Rama et al., 2023). The data was analysed through a combination of digital methods and content analysis.

Analysing query moderation

Secondly, the study returns to Pornhub's website to analyse the traces of search queries. Building on previous literature (Vera-Gray et al., 2021) and the results from the reverse engineering analysis, a list of queries related to different levels of prohibited content was created with particular attention to CSAM, non-consensual material, violence and coercion, and AI-generated content. The queries were then inputted in Pornhub, simulating the behaviour of a regular user, and the related queries offered by the platform captured through screenshots. For each query, to minimise the effects of algorithmic personalisation, a clean version of the browser was used, meaning there was no previous navigation history. Cookies were deleted, and the user was logged out. This exploratory step of the research employs a qualitative method inspired by the work of Gerrard and Thornham (2020) to examine how query moderation works, i.e., what query restrictions are in place and whether or not they can be circumvented.

Given the sensible nature of the data collected and analysed, the research adheres to the ethical framework established by the University of Amsterdam throughout each phase of the research, including data collection, management, analysis, and the presentation of findings. The research is grounded in a framework that combines big data and digital methods with a feminist ethics of care and a commitment to social justice-oriented research (Luka & Millette, 2018), with careful consideration given to the potential risks and harms the study might entail.

Traces of content moderation on Pornhub

Motivations for moderating content

In this section, the analysis focuses on the motivations that Pornhub attached to moderated content at the level of the interface, i.e., the labels that users see when surfing the website, instead of the actual videos. These labels highlight the statuses of

videos on the platform, and reflect decisions made in the back-end and therefore represent traces of these moderation strategies. Figure 8.2 shows the statuses and trajectories that content on Pornhub can go through once it is featured on the New Videos page. The results reflect a form of *a posteriori* moderation, in which content undergoes further scanning after having already gone through various stages of verification and review prior to upload (see Section 2).

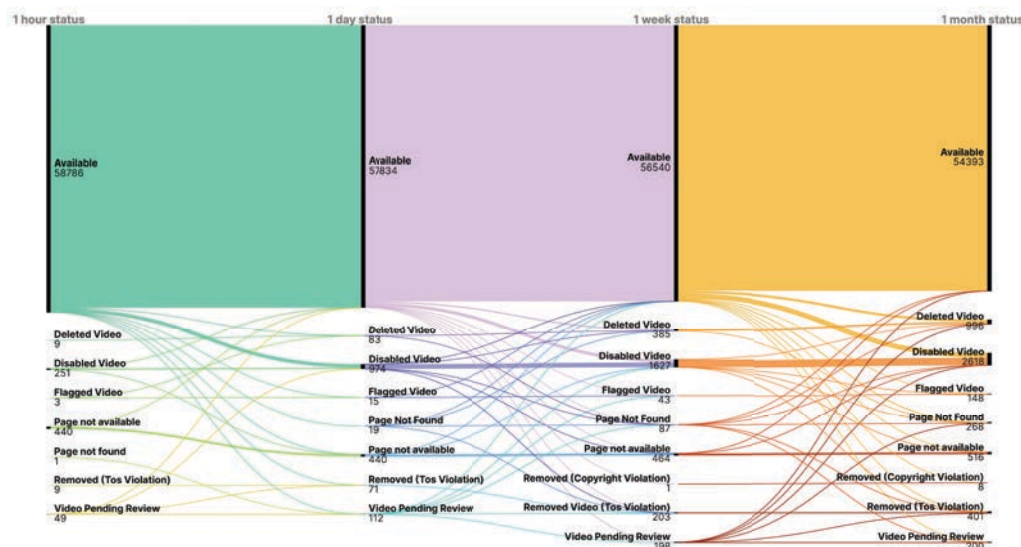


Figure 8.2 Alluvial diagram illustrating the statuses of videos at four time frames in the data collection (one hour, one day, one week and one month after the video was first recorded). The diagram shows that the vast majority of videos ($n = 54,393$, which corresponds to 91% of the total), remained available throughout the data collection period. Among the remaining videos, 2,618 (4,4%) were marked as "Disabled" by the end of the data collection, and 401 (0,7%) were removed for violations of the Terms and Conditions. Source: author.

The vast majority of the videos uploaded to the New Videos page remain available throughout the data collection phase (54,393 videos, which correspond to the 91% of the total). The remaining 9% contain different motivations for being unavailable on the platform, which can be grouped into two forms of moderation: users' self-moderation and platform-level moderation.

At the level of user self-moderation, they can restrict availability to their content in two ways: restriction and deletion. The status "This page is not available" often comes together with the label "Restricted" in the video metadata, suggesting that these videos are restricted by users themselves through Pornhub's geo-blocking system (Pornhub, 2022).³ The status "This video was deleted" suggests that the uploader took action to delete the video from the platform. These two statuses can be interpreted as ways to protect users' privacy and identity and represent how the platform supports users rather than disadvantage or punishing them.

At the platform-level moderation, we can find removed and disabled videos. In the first

³ With geo-blocking, creators can choose to make their content unavailable in specific countries (Pornhub Blog, 2022a). Restricted videos appeared to be visible if access was made from a different IP address.

case, Pornhub provides two motivations for moderation: "Videos have been removed due to a violation of the Terms and Conditions" and "Video has been removed at the request of the copyright holder". These are the only cases where a video's removal is explicitly attributed to a violation of the terms of service, though the specific nature of the violation is not disclosed. The results also confirm that content removal is the most severe moderation measure, since videos remain removed for the entire period of investigation, with no cases of reinstatement observed.

In other instances, the reason for moderation reads "This video has been disabled". This is the most frequent status in the dataset (2,618 disabled videos at the end of the data collection), which suggests this is the most common moderation technique. Disabled videos also have more fleeting trajectories. Figure 8.2 showcases that a few videos disabled at the one hour timeframe ($n=3$) become available again at the one week data collection, and then remain available for the whole period. Therefore, it is not clear whether and how this content moderation measure can be reversed and, if so, what leads to this decision. Other videos ($n=10$), switch from "available" to "unavailable pending review" and then "disabled".

What "disabled" means remains vague, however. Community Guidelines do not mention disabling as a measure. The Terms of Service state: "We have the right to disable any username, password, or other identifier, whether chosen by you or provided by us, at any time in our sole discretion for any or no reason, including if, in our opinion, you have violated any provision of these Terms of Service" (Pornhub, 2024). The Copyright Policy instead, claims that in cases of infringement "Responses may include removing, blocking or disabling access to material claimed to be the subject of infringing activity, terminating the user's access to Pornhub, or all of the foregoing" (Pornhub Help Center, n.d.). These phrasings, however, highlight some inconsistency in the definition of "disabled videos", leave room for interpretation, and do not explain how removing differs from disabling or suspension.

While Pornhub's policies outline the possibility of suspending or permanently deleting accounts as forms of moderation, no evidence of these measures was found. This absence suggests a divergence from what can be observed on other platforms, such as YouTube, where account termination emerges amongst the stated motives for banning videos, as previous studies reveal (de Keulenaar et al., 2020).

An important question arises regarding the nature of the moderated video content and whether it is possible to infer the reasons for its removal, particularly given the blurred distinction between "removed" from "disabled" videos. However, it was not possible to analyse video content directly, as the videos are no longer available precisely due to content moderation. Accessing such material would have required dynamically archiving video content itself, in addition to video metadata. While the lack of access to the video content may represent a limitation of this study, this decision was intentional and grounded in ethical and privacy considerations. Instead, to gain insights into the nature of the moderated content, the research focuses on the residual traces left behind, specifically the video tags associated with videos removed for violating the Terms and Conditions.

Traces of content: Video tags

Another kind of trace collected consists of video categories and tags. On Pornhub, users can assign up to eight categories to their videos, selecting from a predefined list provided by the platform and up to sixteen tags (Pornhub Blog, 2022b). The website automatically selects tags that are associated with the chosen categories, but users can also create and add their own. Tags on Pornhub serve a dual purpose: they provide an indication of the content of the video and, importantly, they are used to drive traffic and increase revenues. An analysis of video tags should therefore be attentive to platform logics and acknowledge that tags can be used as "attention grabbing" tools (Stegeman et al., 2023).

As part of video metadata, tags, too, are moderated prior to video upload. In this phase, content's titles and tags are scanned against an alleged "banned word list" (Pornhub Help Centre, 2024a). Terms that are banned cannot be submitted in titles or tags; those who are flagged are directed to human moderators to ensure extra scrutiny.

To deepen our understanding of the moderated content, the research focuses on the tags associated with videos removed for violating the Terms and Conditions by means of network analysis. The network in Figure 8.3 shows the tags (pink nodes) associated with videos removed in violation of the Terms and Conditions (green nodes). The network is composed of two main clusters connected by the bridging tag, "18-year-old". In the main cluster on the left, the bigger nodes represent the most occurring tags, which reflect some of the most popular Pornhub categories of porn (such as "blow job", "big boobs", "amateur"). By focusing on the tags towards the edges of the network, though, it is possible to find more specific variations of traditional pornographic genres, such a cluster of "pinay" tags, one for public sex ("public", "risky") and, moving even further, tags like "Arab sex".

Notably, video tags, even of moderated videos, are highly sanitised. The data show the absence of forms of algospeak intended as intentional alterations or spelling of a word to evade and circumvent content moderation (Steen et al., 2023). If on other social media platforms users can, to some extent, rely on misspellings (e.g., "seggs" instead of "sex") to bypass filters, no such patterns were found. Rather, we can see a new form of algospeak emerging that is not aimed at circumventing content moderation but rather to avoid it, as showcased in a cluster with "18-year-old" tags, which combine this age indication with other attributes, e.g., 18-year-old-cute-girl, 18-year-old-Latina. This mirrors the platform's own strategy of sanitizing its categories by appending "18+" (e.g., Teens 18+, College 18+) to reinforce compliance with age restrictions.

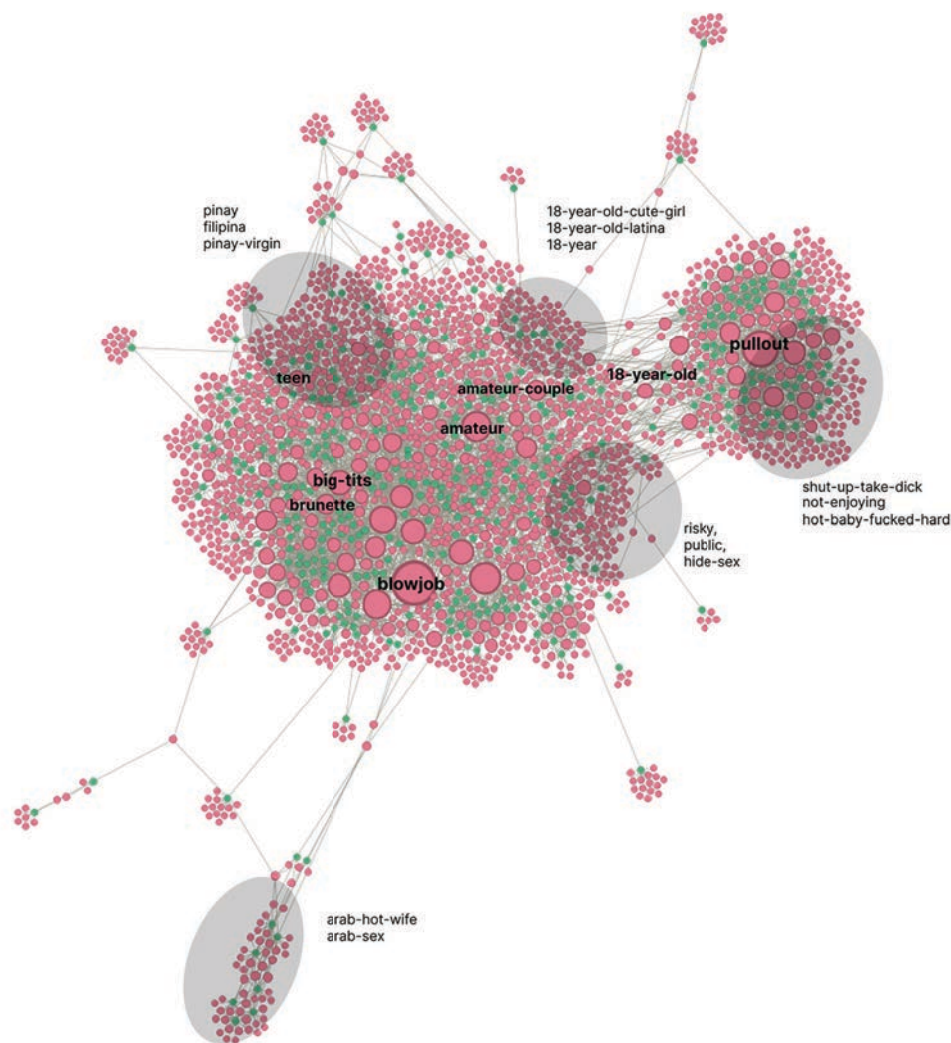


Figure 8.3 Bi-partite network illustrating videos removed for violating the Terms and Conditions (green nodes, $n=401$ and their associated user-generated tags (pink nodes, $n=1880$). The network reveals two main clusters connected by the bridging tag "18-year-old". The larger cluster on the left includes some of the most frequently used tags, many of which reflect the popular Pornhub categories of porn (e.g., "blow job"). This cluster also contains "sanitised" hashtags, e.g., 18-year-old-cute-girl, 18-year-old-Latina. The smaller cluster on the right consists of videos closely linked by shared tag (e.g., "pullout", "shut-up-take-dick", and "not-enjoying"), which could be associated with the semantics of violent and non-consensual content. Source: author.

The second main cluster (right), presents a group of videos that are strongly connected because they are labelled with the same tags. Here, we find tags like "pullout", "no-mercy", "shut-up-take-dick", "talked-into-sex", and "not-enjoying", which could be associated with the semantics of violent and non-consensual content. It is not possible, however, to conclude that the videos were removed because of these tags, that these tags represent black-listed terms, or that the video content was actually non-consensual or coercive. Table 8.1 confirms that while the tags in this cluster are highly moderated, the same ones can also be found in the whole dataset and are, therefore, not exclusively associated with prohibited content. The tags that are fully moderated tend to be peculiar and niche, with few occurrences compared to more "mainstream" ones. Combined with the network structure, this suggests that the issue may not lie with the

tag itself but rather with its frequent association—by similar users—with related but inappropriate content.

Tags (Removed Videos)	Tags Total (count)	Tags Removed (count)	Tags Removed (%)
amateur	9,924	45	0,5%
blowjob	7,812	73	0,9%
big-tits	5,075	27	0,5%
big-boobs	4,900	25	0,5%
homemade	4,308	10	0,2%
milf	4,246	43	1,0%
petite	3,933	15	0,4%
pov	3,827	35	0,9%
brunette	3,582	28	0,8%
18-year-old	1,095	18	2%
pull-out	89	57	64,0%
no-anal	33	28	84,8%
shut-up-take-dick	32	21	65,6%
i-have-a-boyfriend	26	20	76,9%
not-enjoying	24	20	83,3%
convinced-to-cheat	21	12	57,1%
mongerinasia	15	9	60,0%
shut-up-and-take-it	14	10	71,4%
kenyan-sex	9	8	88,9%
no-mercy-anal	7	7	100,0%

Table 8.1 The table presents the frequencies of tags selected to represent the two main clusters identified in Figure 8.3 (i.e., the 10 most recurring tags for each cluster). It shows the number of times each tag appears among the removed videos ("Tags Removed" count) alongside their overall occurrences in the full dataset ("Tags Total" count), offering a comparative perspective on their prevalence and the percentage at which they are moderated.

In sum, we cannot say that tags associated with moderated videos directly relate to specific problematic content; rather, they reflect how users rely on forms of hyper-categorisation (Stegeman, 2023) to comply with content moderation and drive attention towards their content. At the same time, the results indicate that the proliferation of user-generated tags extends to major fantasies by incorporating more specific, minor ones. Mainstream desires create pathways to more niche content (Mazières et al., 2014), including stereotypical and fetishizing material (such as the "pinay" cluster) and more extreme or potentially violent content (as seen in the "pullout" cluster).

Search query moderation

By returning to the website's interface, it is possible to expand our understanding of content moderation to see what happens when a potential user types a query explicitly related to various categories of prohibited content.

As shown in Figure 8.4, some queries are blocked by the platform. When inputting them, the user receives a warning message, stating "your search shows you may be interested in sexual images of minors" or "your search could be for illegal and abusive sexual material, including non-consensual intimate imagery (NCII) or image-based sexual abuse (IBSA)". This confirms Pornhub's efforts in moderating illegal and unlawful content by discouraging queries that might lead to criminal images or violent practices. The warning messages also provide users with resources to understand NCII and CSAM as well as links to report this type of content. The moderation of search queries seems more inconsistent, however, when searching for keywords that potentially lead to borderline content that blurs the line between the acceptable and unacceptable.

Non-Consensual Imagery	Violence and Coercion	CSAM	AI-Generated
caught	alcohol	child	deep fakes
cctv	booze	cp	deepfakes
downblouse	drunk	deflower, deflowering	deepnude
hack	flash, flashing	first time	nudify
hacked	passed out	kid	nudify porn
hidden	rape	minor	
hidden camera	sedate	tiny	
leak, leaked	unwant	virgin	
private	unwanted	young	
revenge	unaware	young teen	
secret	violate, violation, violating	very young	
secret camera	violent, violence	youth	
spy, spied, spying	vodka		
stolen	wasted		
telegram			
upskirt			

Available

Blocked (illegal and abusive material)

Blocked (sexual images of minors)

No Results

Figure 8.4 List of search queries entered into Pornhub's search bar and their availability status. Some queries are blocked either for potentially leading to illegal or abusive material, triggering the message: "Your search could be for illegal and abusive sexual material, including non-consensual intimate imagery (NCII) or image-based sexual abuse (IBSA)". Others are blocked for indicating an interest in sexual images of minors, showing the message: "Your search shows you may be interested in sexual images of minors." Source: author.

The zero tolerance for CSAM is once again confirmed, because it is not possible to search for queries like "youth", "child" and even "young teen". This suggests that Pornhub's list of banned words used to moderate video metadata serves to moderate search queries as well. Other queries such as "young" and "virgin" are admissible. Most importantly, when analysing the suggested key terms, we can see that the

platform directs users towards more and more specific, sensationalistic and potentially borderline queries. The related queries are visible on the Pornhub's interface, at the bottom of the page results, and appear as clickable boxes, as Figure 8.5 shows. In the case of the search query "young", for instance, the user was suggested other queries like "virgin", "teenager first time" and "beautiful 18 year old".

Some inconsistencies in moderation also appear in relation to potentially non-consensual content. The platform does not allow queries such as "cctv", "secret camera" and "hidden camera". When searching for "hidden", however, one of the first suggested queries is precisely "hidden real camera", later followed by "hidden home camera" and others related to spying, hiding, as well as amateur and homemade content (Figure 8.5).

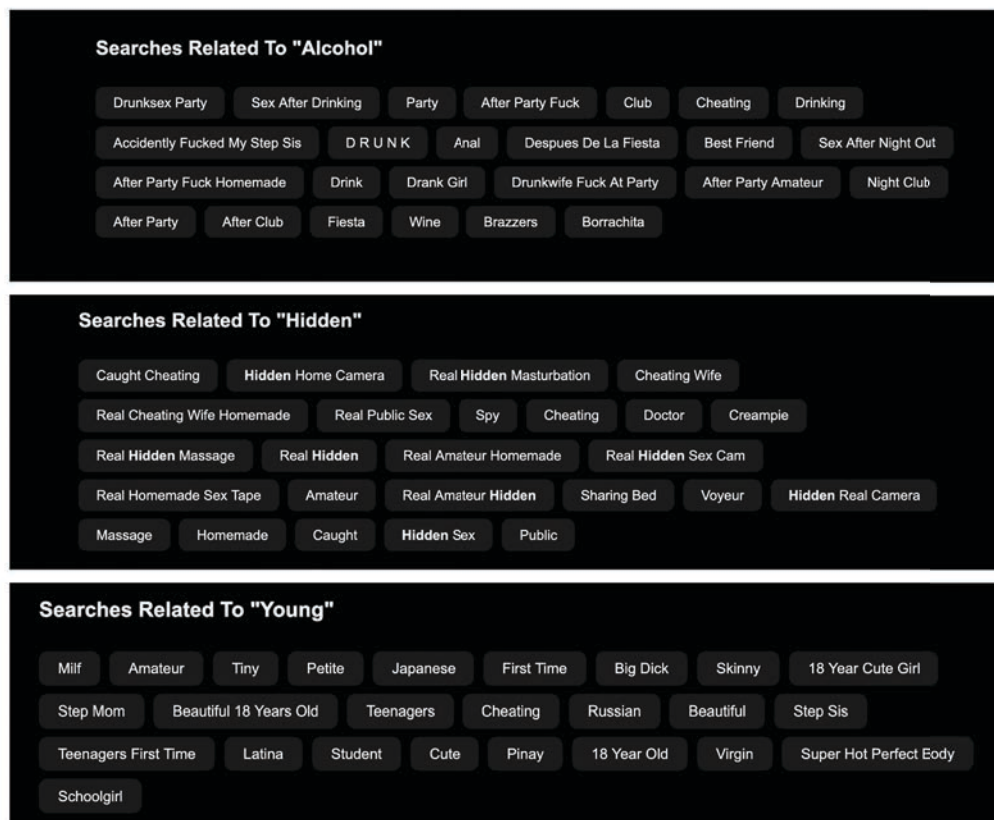


Figure 8.5 Related search queries, displayed at the bottom of Pornhub's page, when querying content, collected in December 2024. Examples for queries "Young", "Hidden", "Alcohol". Source: author.

Something similar happens in the case of possible coercion, such as when searching for "unwanted [sex]" and "drunk" (Figure 8.5). The query "unwanted" is not allowed, whereas it is possible to search for "unwant". When one does, the suggested terms include "unwant sex". Similarly, while searching for "drunk" is forbidden, other semantically similar queries such as "alcohol" or "wasted" are available. In both these cases, the user is directed to the more niche domain of after-party content. Moreover, it is possible to access results for the query "D R U N K", which represents a way of using algospeak to evade content moderation by altering the spelling of a word. This is the only case of misspelling to avoid moderation found in the analysis.

Some results suggest that moderation can be inconsistent in other languages than English or at least not account for subtle linguistic variations. One of the related queries found in the research is "borrachita", a Spanish term diminutive of "borracha", which means "drunk female". Despite Pornhub's claims that its list of banned words includes a wide variety of languages, this result suggests that this area warrants further investigation (Pornhub Help Center, 2024b). It is possible that the queries discussed here relate to videos uploaded prior to stricter rules for verification and upload, as the persistence of the misspelling "D R U N K" would suggest. They are still present on the platform and could lead users towards those niche domains of content.

Overall, these results show that it is fairly easy to circumvent Pornhub's query moderation by using semantically close alternatives to blocked search queries. Furthermore, with related searches, the platform suggests users more and more specific and niche types of content.

Following related searches: Nudifying tools tutorials

Lastly, Pornhub has started to moderate deepfakes and other AI generated content. Going back to Figure 8.4, we can see that queries like "deepfake" and its variations are not available on the platform, while the term "deepnude", which is used to refer to AI-generated nudes created without consent of the person depicted, is still available. The search results for this query showcase some problematic content consisting of video tutorials explaining how to create deepnudes and what tools to use.

Among the suggested queries, the platform provides "AI", "chatbot", and "deep nude". By following one such search term, the user gains access to other examples of AI-generated pornography and related videos. Some of them showcase how to create AI pornography, which is a new and growing genre of pornography and does not represent, per se, a form of abuse (Hadero, 2023). Others, instead, explain how to use AI tools to undress pictures of yourself or others and promote these 'nudifying' tools and apps. In total, it was possible to find without any effort 13 videos explicitly showcasing and promoting nudifying tools, the most popular of which has 83.9k and 49k views respectively. Moreover, one of the promoted tools also has an account on Pornhub, with the banner "Make your own AI bitches", showcasing content about how to generate different kinds of AI pornography, including a tutorial on how to undress photos (which is the most viewed video, with 4.3k views, as shown in Figure 8.6). The name of the nudifying app is present in most of the titles of these videos, suggesting that the videos represent a form of sponsored content.

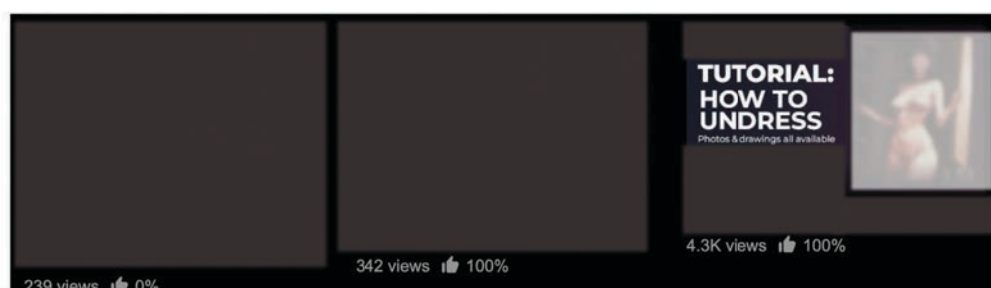


Figure 8.6 Example of content featuring nudifying tools tutorials. Source: author.

These videos represent another example of borderline content that Pornhub is challenged to deal with. The videos do not display directly non-consensual content, as the tools are tested on creators' own bodies, and therefore cannot be considered prohibited content. Moreover, while the videos promote some apps that can be used to produce non-consensual material, these cannot be considered as "illegal product" and therefore, do not officially fall under the label of "other unacceptable content". The fact that these videos are promoted and highly viewed on the platform and the fact that profiles could be easily found makes these videos problematic. As shown above, Pornhub's position against non-consensual AI generated content is clearly stated in its policies; yet the platform allows the sponsoring of tools for creating and circulating the same material.

Latest changes in query moderation

When revisiting this research in February 2025, it was observed that the platform had introduced, or was at least experimenting with, new measures to improve the moderation of search queries. It wasn't possible to find any information about these changes in Pornhub's policies, including in its Recommender System Guidelines, which were last updated in March 2024,⁴ nor on its Help Center or blog posts.

The "searches related to (...)" section at the bottom of the search page, previously described, is no longer present, thus changing the logic of content searchability and findability. By replicating the methodology at a different timeframe (February 2025) and collecting new evidence through screenshots, certain differences in query moderation appear. The findings in Figure 8.4 are still valid, with no changes in the list of blocked queries. What is changing, though, are the recommended search terms. As shown below (Figure 8.7), some queries, such as "hidden" still generate suggested terms. In other cases, such as "young" and "alcohol", queries are permitted and return results, but no longer trigger suggested queries. For non-English terms, words like "borrachita" still appear in the search bar autocomplete, but do not provide any additional suggested queries.

⁴ This section provides information about the recommender algorithm and was included to increase transparency.

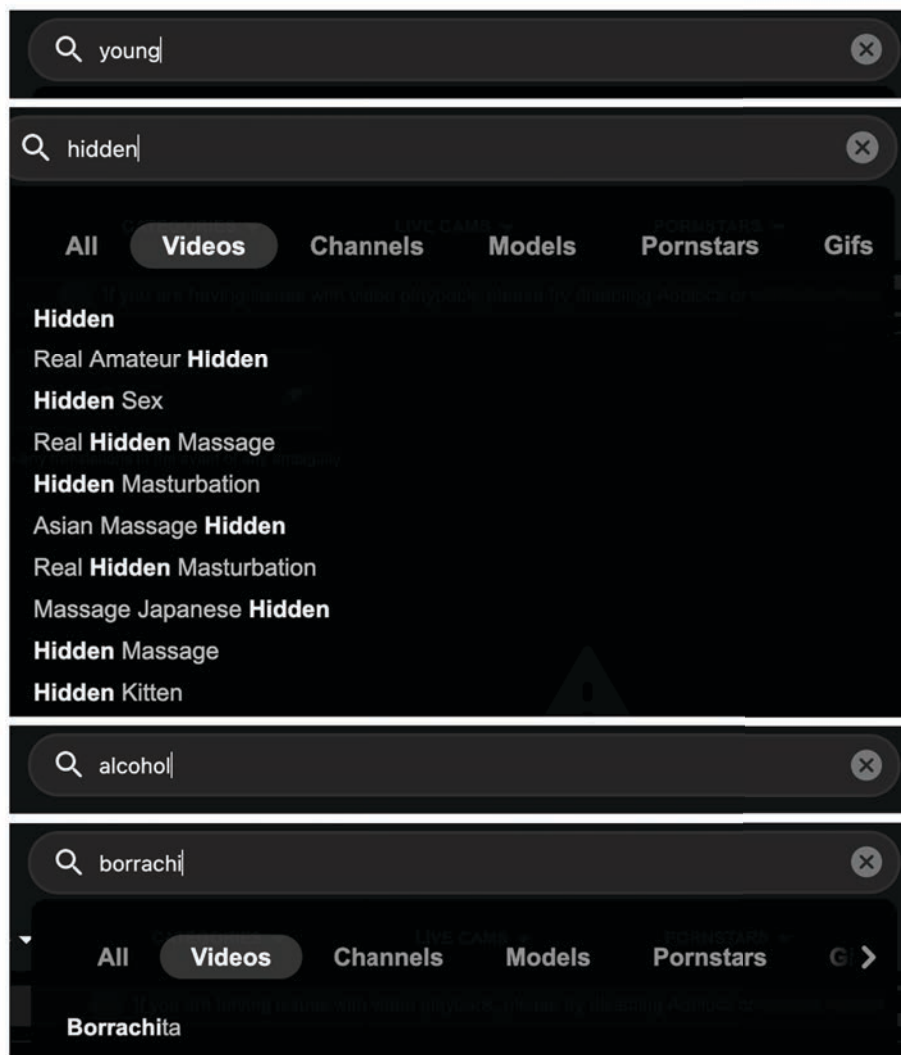


Figure 8.7 Suggested queries in the auto-completion bar, collected in February 2025. Examples for queries "Young", "Hidden", "Alcohol", and "Borrachita". Source: author.

The relevance of search queries and its relation to Pornhub's recommender algorithm was already clear, and becomes even more important in light of these new changes. Further evidence is needed to provide more insights into query moderation and the implications it might have for content searchability and discoverability. These shifts also suggest broader implications for creators and uploaders, whose practices may be shaped in new ways by the platform's evolving logics, requiring them to adapt to these changes to maintain visibility and profitability.

Discussion and conclusion: Pornhub's grey areas of content moderation

This research highlights the progressive evolution and growing complexity of Pornhub's moderation system, along with the emergence of a structured hierarchy of concern reflected in its definitions of prohibited content. Aylo, Pornhub's parent company, aims to position itself as a leader in providing safe access to adult content, and the platform has made notable improvements in moderation. The research has also shown, however, that some grey areas in content moderation still persist, which are summarised in the following main takeaways.

1. **Some grey areas remain in the ways in which policies are phrased**, particularly with regard to the distinction between simulations, role-plays, and fantasies. This lack

of clarity can create inconsistencies in enforcement and leave room for interpretation, affecting both content creators and the platform's overall moderation efforts.

2. The analysis of moderation traces reveals a lack of clarity and opacity in how the platform defines disabled videos. While this is the most common motivation for moderated videos found in the data, what this measure means remains vague: whether it applies to copyright infringements only, how it is enforced, and in what circumstances it can be reversed remain unclear. Pornhub's transparency reports detail all the categories and quantities of videos and accounts removed as well as the number of Digital Millennium Copyright Protection Act (DMCA) notices received. No references to disabled accounts were found, however. This ambiguity confirms some opacity persists, as already highlighted by existing literature and by sex workers' experiences (Caplan & Gillespie, 2020; Blunt & Stardust, 2021). The results suggest that more transparency and accountability can be achieved by making more transparent the connections between back-end moderation decisions, Community Guidelines definitions, related policies, and the motivations for banned content displayed on the platform's interface.

3. Video tags, including those on moderated videos, appear highly sanitised, with no instances of algospeak detected. This suggests that Pornhub is effectively moderating tags and preventing circumvention through altered spellings or coded language. It also highlights users' adaptability in complying with content moderation over avoidance. However, mainstream tags can still serve as entry points to more niche content, potentially leading to borderline material.

4. The results show that circumventing Pornhub's query moderation is relatively easy by using semantically similar alternatives to blocked search terms. While the content returned may not necessarily be unlawful, the platform's related search suggestions guide users toward increasingly specific and niche material, potentially leading to borderline content. This could be problematic for multiple categories of people using the platform. First, it is possible that users may be redirected to more niche and potentially borderline pornographic content, which is easily discoverable by following the suggestions provided by related queries. Secondly, performers may be further incentivised to use sensationalistic tags, as these terms appear among suggested queries and could drive more traffic to their content. But, as previously noted, Pornhub performers have expressed dissatisfaction with the platform's tagging practices (Webber & MacDonald, 2023). While these tags offer a means to attract views and increase earnings, creators also recognise the pressure to use terms that perpetuate discriminatory language, are often dehumanising, and contribute to negative perceptions of the porn industry (Stegeman et al., 2023).

5. Instances of how-to tutorials promoting tools to create synthetic non-consensual imagery were found. While this content does not qualify as non-consensual or as promoting illegal products, its persistence and visibility raises two concerns. First, the fact that this content goes unmoderated contributes to the spread and normalisation of nudifying tools, which represent a new form of image-based sexual abuse that is becoming more frequent among young people. Second, the rise of AI-generated content introduces challenges for content moderation, potentially giving rise to new forms of borderline content.

Future research and limitations

In light of these results, future research is needed to address the ongoing changes to search query moderation and in the discoverability of content on the platform.

Secondly, the analysis should be expanded to the other adult platforms designated at VLOPs under the DSA, particularly Xvideos and XNXX. Comparing moderation practices across these platforms would offer a broader perspective on industry-wide approaches and challenges, particularly in the context of compliance with evolving legal frameworks.

The results focus on the traces left behind by content moderation and the process of reverse-engineering its mechanisms and, as such, present some challenges and limitations. This research attempts to move beyond viewing content moderation as a "black box" by analysing its outputs. The inherently complex and opaque nature of content moderation remains a significant obstacle, however. As such, this project aims to highlight the frictions and ambiguities emerging from the data, thus further reiterating the opaqueness of these processes.

Moreover, the use of dynamic archiving techniques required setting some specific time frames for investigation. However, the statuses of the videos analysed are not static and could still change over time. This limitation points to opportunities for future research, particularly follow-up studies that re-analyse the videos after additional time has passed, allowing for a deeper understanding of the long-term dynamics of content moderation.

As previously mentioned, this research has focused primarily on video metadata and search queries, without delving directly into the content of the videos themselves. The dynamic archival and collection of video content, which could have been useful to retain access to removed or deleted material, was intentionally avoided to comply with ethical research standards.

Contact with video content was inevitable, however, when evaluating the context in which metadata was embedded. During the research, no instances of openly illegal content were found, but there were encounters with a few potentially prohibited uses (especially in relation to non-consensual material and coercion) and borderline content such as the one described above. The scope of this work was not to identify and flag problematic content as a substitute for the platform's responsibilities. Rather, this research aims to serve as a starting point to raise structural issues about Pornhub and its content moderation practices.

Appendix

Appendix 1. Summary of Pornhub Community Guidelines, related policies and measures. This overview is based on the most recent version of the Community Guidelines available as of December 2024.

Community Guidelines		Related policies	Measure
Illegal content	• Illegal activity of any kind • Children (under the age of 18) • Human & sex trafficking • Animal cruelty • Bestiality • Death • Snuff • Torture • Violence • Exploitation of a corpse	Child Sexual Abuse Policy	• Zero-tolerance policy; immediate removal of content; • banning of its uploader; • report cases to the National Center for Missing and Exploited Children (NCMEC)
		Animal Welfare Policy	• Review and remove violative content; • Fingerprint (depending on the findings); • Suspend or permanently terminate the associated user's account (where appropriate)
		Violent Content Policy	• Immediate removal of content; • Termination of the user's associated account (when appropriate)
Non consensual content	• Depiction of non consensual acts • Recording intimate content without consent • Distribution intimate content without consent • Depiction individual likeness without consent • Deep Fakes • AI-generated or manipulated content • Doxing		
		Non-consensual Material Policy	• Review and remove infringing content; • Fingerprint the content in question; • Suspend or permanently terminate the associated uploader's account (where appropriate)

Content that harms or may cause harm	• Hate speech or inflammatory content • Violent extremism • Political provocation • Terrorism • Violent extremist actors or acts • Self-harm or suicide • Eating disorder	Hate Speech Policy	• Review and remove violative content; • Fingerprint (depending on the findings); • Suspend or permanently terminate the associated user's account (where appropriate)
Inauthentic or unauthorised content	• Spam • Misleading information • Misinformation • Personification of another individual • Infringement on intellectual property rights	Copyright	• Responses may include removing, blocking or disabling access to material claimed to be the subject of infringing activity, terminating the user's access to Pornhub, or all of the foregoing
Otherwise unacceptable content	• Sexual services in exchange for money or goods • Content that depicts, role-plays or implies incest • Feces, vomit, solid content • Drugs • Reacts or tributes source content not uploaded to Pornhub • Sponsored product which promotes or offers illegal activity • AI generated content used in a realistic manner to alter the speech or behaviour of the individual depicted • Encourages conduct that would be		

	considered a criminal offense.	
--	--------------------------------	--

References

- Ahn, S., Baik, J., & Krause, C. S. (2023). Splintering and centralizing platform governance: How Facebook adapted its content moderation practices to the political and legal contexts in the United States, Germany, and South Korea. *Information, Communication & Society*, 26(14), 2843-2862. <https://doi.org/10.1080/1369118X.2022.2113817>.
- Aylo (2023, August 17), MindGeek becomes Aylo. Available at: <https://www.aylo.com/newsroom/mindgeek-rebrands/>. Last accessed 28/04/2025.
- Are, C., & Briggs, P. (2023). The Emotional and Financial Impact of De-Platforming on Creators at the Margins. *Social Media + Society*, 9(1). <https://doi.org/10.1177/20563051231155103>.
- Blunt, D. & Stardust, Z. (2021). Automating Whorephobia: Sex, Technology and the Violence of Deplatforming. *Porn Studies*, 8(4): 350–66. <https://doi.org/10.1080/23268743.2021.1947883>
- Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, 6(2). <https://doi.org/10.1177/2056305120936636>.
- Cole, S. (2020, December 14). Pornhub Just Purged All Unverified Content From the Platform. *Vice*. <https://www.vice.com/en/article/pornhub-suspended-all-unverified-videos-content/>. Last accessed 28/04/2025.
- Cole, S. (2019, October 11). Girls Do Porn Employees Charged With Sex Trafficking, Potentially Face Life in Prison. *Vice*. <https://www.vice.com/en/article/girls-do-porn-employees-charged-with-sex-trafficking-potentially-face-life-in-prison/>. Last accessed 28/04/2025.
- Craig, E. (2024) *Mainstreaming Porn. Sexual Integrity and the Law Online*. McGill-Queen's University Press. <https://doi.org/10.1515/9780228022404>.

de Keulenaar, E. & Rogers, R. (2025). After deplatforming: The return of trace research for the study of platform effects. In: Venturini T, Acker A, Plantin J-C, et al. (eds) *The SAGE Handbook of Data and Society: An Interdisciplinary Reader in Critical Data Studies*. London: SAGE

de Keulenaar et al. (2020). Demoting, deplatforming and replatforming COVID-19 misinformation. Digital Methods Initiative Summer School Report, University of Amsterdam. Available at:
<https://www.digitalmethods.net/Dmi/SummerSchool2020ModeratingCovidMisinfo>.

Ethical Capital Partners (2019, May 16), ECP Announces Acquisition of MindGeek, Parent Company of Pornhub. Available at:
<https://www.ethicalcapitalpartners.com/news/ecp-announces-acquisition-of-mindgeek%2C-parent-company-of-pornhub>. Last accessed 28/04/2025.

European Commission (2023, December, 20). Commission designates second set of Very Large Online Platforms under the Digital Services Act. Available at:
https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6763. Last accessed 28/04/2025.

Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, 22(7), 1266-1286.
<https://doi.org/10.1177/14614448209125>.

Gillespie, T. (2017). Governance of and by platforms. In J. Burgess, A. Marwick & T. Poell (Eds.), *The SAGE Handbook of Social Media* (254-278). Sage. ISBN: 9781412962292.

Gregory, T. (2024). Reading Pornhub's authentication systems and deleted archive through Walter Benjamin's aura, *Porn Studies*, 1–18.
<https://doi.org/10.1080/23268743.2023.2278766>.

Hadero, H. (2023, December 2). Teen girls are being victimized by deepfake nudes. One family is pushing for more protections. *APNews*.
<https://apnews.com/article/deepfake-ai-nudes-teen-girls-legislation-b6f44be048b31fe0b430aeee1956ad38>. Last accessed 28/04/2025.

Henry, N., & Witt, A. (2021). Governing image-based sexual abuse: Digital platform policies, tools, and practices. In J. Bailey, A. Flynn & Henry, N. (Eds.) *The Emerald International Handbook of Technology-Facilitated Violence and Abuse* (pp. 749-768). Emerald Publishing Limited. DOI: 10.1108/9781839828485.

Iovine, A. (2023, June 29). Pornhub accused of abusing user data. *Mashable*.
<https://mashable.com/article/pornhub-accused-of-abusing-user-data-stopdataporn>. Last accessed 28/04/2025.

Kastrenakes, J. (2024, June 19), Pornhub to block five more states over age verification laws. *The Verge*.
<https://www.theverge.com/2024/6/19/24181690/pornhub-indiana-kentucky-block-age-verification>. Last accessed 28/04/2025.

- Kristof, N. (2020, December 4). The Children of Pornhub. *New York Times*.
<https://www.nytimes.com/2020/12/04/opinion/sunday/pornhub-rape-trafficking.html>.
 Last accessed 28/04/2025.
- Luka, M. E. & Millette, M. (2018). (Re)framing big data: Activating situated knowledges and a feminist ethics of care in social media research, *Social Media + Society*, 4(2), 1–10. <https://doi.org/10.1177/2056305118768297>.
- MacDonald, M. (2023). The Algorithmic Moderation of Sexual Expression: Pornhub, Payment Processors and CSAM. AoIR Selected Papers of Internet Research. <https://doi.org/10.5210/spir.v2023i0.13453>.
- Mazières, A., Trachman, M., Cointet, J. P., Coulmont, B., & Prieur, C. (2014). Deep tags: toward a quantitative analysis of online pornography. *Porn Studies*, 1(1-2), 80-95. <https://doi.org/10.1080/23268743.2014.888214>.
- McGlynn, C. & Woods, (2022). Image-Based Sexual Abuse, Pornography Platforms and the Digital Services Act. <https://hateaid.org/bildbasierte-gewalt-pornoplattformen-und-der-digital-services-act/>.
- Pornhub (2024). Terms and Conditions. Available at: <https://www.pornhub.com/information/terms>. Last accessed 28/4/2025.
- Pornhub Blog (2025, March 12). 2025 Consent and ID Requirement Updates. Available at: <https://www.pornhub.com/blog/2025-consent-and-id-requirement-updates>. Last accessed 28/04/2025.
- Pornhub Blog (2022a, July 21). Geo Blocking Explained. Available at: <https://www.pornhub.com/blog/geo-blocking-explained>. Last accessed 28/04/2025.
- Pornhub Blog (2022b, July 18). Crash Course: Categories and Tags. Available at: <https://www.pornhub.com/blog/crash-course-categories-and-tags>. Last accessed 28/04/2025.
- Pornhub Help Center, n.d. Copyright. Available at: <https://help.pornhub.com/hc/en-us/articles/4419861225107-Copyright>. Last accessed 28/4/2025.
- Pornhub Help Center (2024a) Verification, Upload, and Content Moderation Process. Available at: <https://help.pornhub.com/hc/en-us/articles/34927506861843-Verification-Upload-and-Content-Moderation-Process>. Last accessed 28/4/2025.
- Pornhub Help Center (2024b). 2024 Transparency Report (First Half). Available at: <https://help.pornhub.com/hc/en-us/articles/33098088051475-2024-Transparency-Report-First-Half>. Last accessed 28/4/2025.
- Pornhub Blog (2022). Geo Blocking Explained. Available at: <https://www.pornhub.com/blog/geo-blocking-explained>. Last accessed 28/04/2025.

- Pornhub Help Center (2020). Our Commitment to Trust And Safety. Available at: <https://web.archive.org/web/20201231153105/https://help.pornhub.com/hc/en-us/categories/360002934613>. Last accessed 28/04/2025.
- Pornhub Insights (2024). 2024 Year in Review. Available at: <https://www.pornhub.com/insights/2024-year-in-review#top-searches-pornstars>. Last accessed 28/04/2025.
- Rama, I., Bainotti, L., Gandini, A. Giorgi, G., Semenzin, S., Agosti, C., and Corona, G. (2023). The Platformization of Gender and Sexual Identities: An Algorithmic Analysis of PornHub. *Porn Studies*, 10(2), 154-173. <https://doi.org/10.1080/23268743.2022.2066566>.
- Rogers, R. (2024). *Doing Digital Methods*. 2nd Edition. Sage.
- Rogers, R. (2020). Deplatforming: Following extreme internet celebrities to Telegram and alternative social media. *European Journal of Communication*, 35(3): 213–229. <https://doi.org/10.1177/0267323120922066>.
- Saunders, R. (2020). *Bodies of work: The labour of sex in the digital age*. Springer Nature. <https://doi.org/10.1007/978-3-030-49016-4>.
- Statista (2024). Pornhub.com global web visits 2022-2024. Available at: <https://www.statista.com/statistics/1459714/pornhub-monthly-visits-by-device/>. Last accessed 29/4/2025.
- Stegeman H. M. (2021). Regulating and representing camming: Strict limits on acceptable content on webcam sex platforms. *New Media & Society*, 1–17. <https://doi.org/10.1177/14614448211059117>.
- Stegeman, H. M., Velthuis, O., Jokubauskaitė, E., & Poell, T. (2023). Hypercategorization and hypersexualization: How webcam platforms organize performers and performances. *Sexualities*, 0(0). <https://doi.org/10.1177/13634607231170174>.
- Tar, J. (2023, October 30). NGOs urge EU Commission to include porn websites in the ‘systemic risk’ club. Euractiv. <https://www.euractiv.com/section/platforms/news/ngos-urge-eu-commission-to-include-porn-websites-in-the-systemic-risk-club/>. Last accessed 28/04/2025.
- Tiidenberg, K. (2021). Sex, Power and Platform Governance. *Porn Studies*, 8(4): 381–93. <https://doi.org/10.1080/23268743.2021.1974312>.
- Vera-Gray, F., McGlynn, C., Kureshi, I., & Butterby, K. (2021). Sexual violence as a sexual script in mainstream online pornography. *The British Journal of Criminology*, 61(5), 1243-1260. <https://doi.org/10.1093/bjc/azab035>.
- Webber, V. & MacDonald, M. (2023). Pornhub Creator Consultations Report. DOI: 10.13140/RG.2.2.29426.07367.

Webber, V., MacDonald, M., & Sullivan, R. (2021, July 4). Eradicating sexual exploitation in porn should not be at the expense of sex workers. *The Conversation*. <https://theconversation.com/eradicating-sexual-exploitation-in-porn-should-not-be-at-the-expense-of-sex-workers-163064>. Last accessed 28/04/2025.

Webber, V., & Franco, R. S. (2024). The definitional creep: Payment processing and the moral ordering of sexual content. *Sexualities*, 0(0). <https://doi.org/10.1177/13634607241305579>.

9. Gatekeepers of the mobile ecosystem: Understanding app store moderation

Esther Weltevrede, Anne Helmond, Fernando van der Vlist, Stefanie Duguay and Michael Dieter

Abstract

App stores are critical actors in moderating the mobile app ecosystem. This chapter examines both their moderation processes and their interactions with external regulatory frameworks, including those of the European Union. Leading global app stores such as Google Play and the Apple App Store operate as both facilitators and gatekeepers of the mobile ecosystem, shaping the availability, accessibility, and moderation of apps while mediating relationships among developers, users, and platform infrastructures in countries and regions worldwide. Through their policies, guidelines, and moderation enforcement mechanisms, app stores exert influence over the app ecosystem, shaping the behavior of developers and users. We examine app store moderation through three case studies that investigate how app stores mediate access to essential information and services and perform moderation while responding to regulatory, social, and political pressures.

Keywords: app stores, moderation, platform governance, mobile ecosystem

Introduction

App stores are critical actors in moderating the mobile app ecosystem. This chapter examines their internal moderation mechanisms and interactions with external regulatory frameworks, including those of the European Union. Leading global app stores, such as Google Play and the Apple App Store, function as facilitators and gatekeepers of the mobile ecosystem, mediating relationships between developers, users, and platform infrastructures through policies, guidelines, and enforcement mechanisms that dictate terms of participation and shape the behavior of developers and users.

Unlike social media platforms—which can remove individual posts, suspend or ban user accounts or channels, demonetize creators, down-rank specific content in feeds, or apply other granular interventions—app stores primarily moderate at the level of entire applications, app developers, and app reviews. They review apps before publication, reject or remove those that violate rules, and use algorithmic curation to promote, demote, or geo-block entire apps. Beyond app-level takedowns, app stores also enforce developer-level moderation: they can issue policy warnings, suspend or terminate developer accounts, and even revoke signing certificates to disable all of a developer's apps. At the user level, stores manage ratings and reviews—removing spam or abusive feedback, demoting apps with inauthentic review patterns, and limiting or banning users who abuse reporting tools. These practices demonstrate how app stores extend moderation beyond individual content to shape which apps and services exist, who can distribute them, and how they compete in the market.

This chapter explores app store moderation through three case studies examining how app stores mediate access to essential information and services while managing internal moderation practices and external demands. The chapter situates the analysis within the European regulatory context but does not focus on evaluating the direct impact of EU regulation on app stores. Instead, it examines how app stores strategize and implement moderation practices in response to regulatory, political, and social pressures. In addition to these factors, economic imperatives also shape app store moderation, as market competition, platform revenues, and global strategies influence decision-making within the app ecosystem (Nieborg, Young, & Joseph, 2020). The first case study conducts a historical analysis of developer policies for Google and Apple's app stores on Android and iOS platforms. This analysis traces the evolution of app store moderation and highlights how external regulatory measures, such as the General Data Protection Regulation (GDPR) and market forces, shape developer compliance requirements. The second and third cases focus on moderation in times of crisis or conflict: the global COVID-19 pandemic (in 2020–2021) and the Russian invasion of Ukraine (in 2022). These crises amplify platforms' moderation responsibilities, exposing their regulatory frameworks, operational priorities and the balancing act between regulatory mandates and operational (economic) priorities. The cases also demonstrate the interplay between internal moderation practices and external regulation, illustrating how app stores act as critical gatekeepers of the mobile ecosystem.

By situating app store moderation within broader discourse on platform governance and content moderation, this chapter examines how regulatory frameworks and operational strategies shape the mobile ecosystem, including how app stores enforce policies, moderate content, enforce policies, and respond to crises.

Conceptualizing app store moderation

Platform governance refers to the broader set of rules, systems, and policies that determine how platforms operate— their "regulating structures" that establish the conditions and possibilities of participation (Gorwa, 2024). It determines who can access the platform, which content is prioritized or removed, and how platforms interact with stakeholders like developers, users, governments, and advertisers. Media scholars Robert Gorwa and Tarleton Gillespie distinguish between governance of platforms—external regulations defining platform responsibilities and liabilities—and governance by platforms, referring to internal moderation practices (Gorwa, 2024; Gillespie, 2018). In the context of app stores, the internal governance by platforms manifests through multi-layered moderation systems, including pre-publication app reviews, automated and manual policy enforcement, algorithmic curation (e.g., promotions, demotions, geo-blocks), developer-account actions (such as warnings or suspensions), and user-level controls (like review removals and account restrictions). These mechanisms illustrate how internal governance structures respond to, and are constrained by, external regulatory frameworks and societal expectations. App store moderation is a specialised subset of platform governance in which app stores function simultaneously as marketplace operators and distribution gatekeepers (Cowls & Morley, 2022). They oversee every submitted app—from technical compliance to policy adherence—while managing the sometimes-conflicting interests of developers, users, advertisers, regulators, and platform owners (Dieter et al., 2019, 2021). These moderation activities sit at the intersection of external

regulations—such as the EU’s GDPR, Digital Markets Act (DMA), and Digital Services Act (DSA)—and internal platform policies, including developer agreements, app review guidelines, and content policies.

App stores’ gatekeeping function extends beyond technical control, as it is deeply entangled with economic, legal, social, and cultural processes. As "gatekeepers", they dictate the terms of participation, controlling access to software, services, and digital markets (Fagerjord, 2015). Through mechanisms like app approval or rejection, algorithmic visibility control, and geographic access restrictions, app stores regulate access to software and services. These interventions become particularly consequential during crises, when decisions about the availability of essential apps can affect public health or safety.

App store moderation practices also shape cultures of use, as platforms and developers adjust to moderation constraints differently. For example, Apple’s temporary removal of Tumblr from its store contributed to Tumblr enacting a strict ban on nudity and sexual content across its platform (Tiidenberg et al., 2021). Relatedly, some developers bypass app stores altogether to avoid this additional layer of regulation, opting instead for mobile-friendly HTML5 websites or other alternative distribution methods. While this allows them to avoid regulatory constraints, it often comes at the cost of functionality and accessibility (Light, 2014).

The uneven and fragmented nature of these moderation practices is described as "patchwork platform governance" (Duguay et al., 2020), reflecting their variations across jurisdictions, platforms, and use cases. The patchwork concept highlights how governance does not function as a singular, cohesive system but rather as an overlapping set of uneven rules, policies, and practices shaped by external regulations, internal policies, and contextual pressures—ranging from global crises to region-specific legislation. Governance challenges posed by local regulations may intersect with global app store strategies, creating a mosaic of moderation practices in response to shifting regulatory and market imperatives, unforeseen circumstances, or negative media coverage (Marchal et al., 2024). Moreover, technical design choices and app stores’ infrastructures themselves embody moderation decisions (Van der Vlist et al., 2022). As Van der Vlist et al. argue, these material "governance arrangements" continuously evolve under internal strategic decisions and external pressures, including regulatory, social, and competitive forces.

The fragmented nature of moderation becomes especially visible in times of crisis, when app stores intensify their gatekeeping role, adjusting policies in response to shifting geopolitical, regulatory, and social pressures (Dieter et al., 2021). Such crises can reconfigure existing moderation frameworks, introduce alternative enforcement approaches, or expand moderation to include new actors and technologies (Polese & Helou, 2024). In some cases, the technologies themselves are positioned as solutions for crises or substitutes for state functions—for example, when dating apps were mobilized during the COVID-19 pandemic to address loneliness and disseminate public health messages (Myles et al., 2021), intensifying the stakes of app moderation. Viewing app store moderation as a patchwork of layered, overlapping, and sometimes contradictory arrangements allows for a deeper understanding of how regulatory frameworks, platform policies, and enforcement practices interact. This perspective highlights the need to analyse moderation as a multi-situated, multi-layered, and

context-dependent process that shapes and is shaped by the broader platform ecosystem as well as social, economic, and regulatory environments. The next section explores approaches to examining this complex moderation ecosystem.

Approaches for studying app store moderation

Studying app store moderation requires identifying the key actors and objects being moderated, analyzing the mechanisms through which moderation is enforced, and exploring the material entry points that make these processes visible. Building on the distinction between governance of and by platforms, this section discusses three key dimensions of app store moderation: governance *by app stores* through **developer policies** and **moderation mechanisms**, and governance *of app stores* through **regulatory frameworks**, and operationalizes them for research.

Governance by app stores: Moderation mechanisms and developer policies

A central focus of app store moderation is the moderation of apps and their developers through policies that govern their submission, operation, and monetization. Developer policies, such as Apple's App Store Review Guidelines (Apple Inc., n.d.-a) and Google's Developer Policy Center (Google, n.d.-a), serve as critical entry points for examining how moderation operates at this level. These policies establish requirements for a range of aspects, including revenue models, restrictions on prohibited app categories, data handling practices, and compliance with legal frameworks. The ultimate enforcement mechanism for app and developer governance is app removal, a definitive form of moderation when policies are violated. However, the layered and nested nature of moderation complicates the compliance process for developers. For example, a typical app may integrate Software Development Kits (SDKs) to provide third-party functionalities or advertising, and these SDKs are moderated by their own policies. Additionally, apps often rely on external platforms, such as social media systems, which introduce yet another set of regulatory frameworks. Ultimately, all these elements operate under the overarching policies of the app store itself, creating a nested and multifaceted moderation structure.

Moderation is also implemented through moderation mechanisms that influence apps' visibility and accessibility, each of which can be studied to understand moderation practices. Pre-gatekeeping, as seen in Apple's App Review Guidelines and Google's Prepare your app for review (Google, n.d.-b), involves the review and approval process that apps undergo before being published on the platform. The process requires that developers follow platform guidelines, which include submitting metadata and uploading app builds through App Store Connect or Google Play Console. It also involves automated checks for technical compliance and manual reviews for content, functionality and user experience. App store guidelines may trickle down into the Community Guidelines and policies for platforms and apps that they host, with Apple's App Review Guidelines – for example – giving rise to the general prevalence of platform policies prohibiting violent displays or explicit sexual content. Once approved, the app is published. These mechanisms ensure compliance with app store policies and filter content before it enters the ecosystem; however, rejections also include feedback for resubmission and post-publication monitoring is implemented to ensure ongoing compliance.

Algorithmic moderation represents another form which involves the algorithmic shaping of search results to prioritize or demote certain apps. Drawing on concepts such as "query governance" (Dieter et al., 2021) and "serious queries" (Rogers, 2020), this practice reveals how app stores influence the ranking of and access to apps and information. Algorithmic silencing or demotion further illustrates subtleties of app store moderation, through which apps may be made less visible without outright removal. Alternatively, algorithmic amplification ensures the visibility and prominence of particular apps, which can direct users toward the official or most popular apps for a given purpose. More overt forms of moderation include removing apps or banning developers, as seen in high-profile cases such as the legal dispute in which Apple blocked apps from Epic Games. "Geo-blocking," restricting apps in specific regions due to local laws or policies, represents another critical enforcement mechanism that intersects moderation with geopolitics.

User-level moderation adds another dimension to app store regulation. While app stores do not primarily focus on user-generated content, as seen in social media platforms, they nevertheless moderate their users through policies that regulate behaviors such as app reviews, ratings, and in-app interactions, as seen in Google Terms of Service (Google, n.d.-c), Google Play Ratings & Review (Google, n.d.-d), and Apple Media Services Terms and Conditions (Apple Inc., n.d.-b). User-related policies enforce standards for content submission and interaction, shaping both the visibility of apps on the store platform and the broader user experience. For example, flagged reviews may be removed, and apps with consistently poor ratings may be algorithmically demoted. Although the moderation of user-generated content within app stores is not primarily about free speech, as it is on social media platforms, it still plays a crucial role in shaping the app ecosystem by influencing user engagement and app reputation.

Governance of app stores: External regulatory frameworks

App stores are subject to oversight from external regulatory bodies, particularly in regions like the European Union, where governance frameworks shape platform operations and may also impact app developers. This overlapping governance structure creates a patchwork of governance where compliance becomes a complex and often opaque task. App stores operating within the European Union are shaped by a range of regulatory frameworks that regulate their operations, particularly regarding data protection, platform accountability, and market fairness. Three frameworks stand out in their influence: the General Data Protection Regulation (GDPR) (Intersoft Consulting, n.d.), the Digital Services Act (DSA) (European Commission, n.d.-a), and the Digital Markets Act (DMA) (European Commission, n.d.-b).

The GDPR, enforced since 2018, has profoundly influenced app store moderation by establishing rules for data protection and privacy. App stores must ensure compliance with GDPR for all hosted apps, including requirements for transparency, data minimization, and obtaining user consent. As gatekeepers, app stores enforce these rules by requiring developers to meet privacy standards, with violations potentially resulting in app removals. GDPR has also driven features such as Apple's App Tracking Transparency (Apple Inc., n.d.-c), which provides users with some control over data collection practices.

The recently implemented DSA expands the European Commission's (EC) regulatory frameworks by emphasizing platform accountability and transparency. Under the DSA, app stores are required to introduce mechanisms for addressing illegal content, enhance transparency in content moderation, and provide users with accessible complaint systems. As Very Large Online Platforms (VLOPs), with more than 45 million users in the EU (10% of the EU's population), the two dominant app stores face additional obligations due to their potential "risks in the dissemination of illegal content and societal harms" (European Commission, n.d.-c). These obligations include publishing risk assessments for systemic harms like disinformation and ensuring cooperation with EU authorities in addressing regulatory violations.

The DMA focuses on promoting fair competition in the digital economy and formalizes the gatekeeping role of major platforms. By designating companies like Alphabet and Apple as "gatekeepers" of "core platform services", the legislation imposes an "extra responsibility" to ensure "an open online environment that is fair for businesses and consumers, and open to innovation by all" (European Commission, 2023). For operators of the dominant app stores, this means allowing end-users to install third-party apps or app stores and permitting developers to use alternative in-app payment systems. It also means that the gatekeeping role of app stores is not merely conceptual (cf. Fagerjord, 2015) but increasingly formalized and regulated. In addition to these key frameworks, app stores are influenced by the ePrivacy Directive (European Parliament and Council, 2002), EU Competition Policy (European Commission, n.d.-d), consumer protection laws (European Commission, n.d.-e), and the Geo-blocking Regulation (European Parliament and Council, 2018), among others, which collectively address issues such as online privacy, transparency in subscriptions, and access to digital services.

Studying app store moderation requires methodological approaches that integrate qualitative and quantitative approaches to reveal its material and operational dimensions. Monitoring app removals over time provides valuable insights into enforcement priorities and regulatory compliance. Similarly, analyzing the moderation of reviews and ratings offers a window into how user-generated content is moderated. Policy analysis, particularly of developer and user policies, helps contextualize the historical evolution of moderation mechanisms in response to regulatory and market shifts. High-profile cases, such as the removal of apps during geopolitical conflicts, illustrate the broader dynamics of platform moderation. Investigating cases where governments request app removals further highlights the intersection of platform moderation with international politics.

The following section provides three illustrative case studies that unpack the patchwork of app store moderation in three ways, and introduces the methodological steps to study them. The section begins with a case study examining the historical evolution of app store developer policies, including a result page analysis focusing on how app stores address queries about what the App Store Review Guidelines consider "objectionable content" (Apple Inc., n.d.-a). It then moves to two case studies that explore moments of crisis, examining how app stores strategize and implement moderation in response to political, regulatory and social pressures.

Case studies

Apple and Google's app store moderation evolution

In our studies on "App (Store) Policy Histories" (Helmond, Van der Vlist, & Weltevrede, 2019) and "Store Policies and Developer Conditions" (Helmond, Nieborg, Van der Vlist, & Weltevrede, 2019) we explored the evolution of developer policies for Apple's App Store and Google Play, analyzing their content, structure, and responsiveness to external factors like regulatory frameworks. The store guidelines outline conditions for third-party developers to build apps on their mobile platforms, including the underlying operating systems. While Apple emphasizes innovation in their App Store Review Guidelines by stating that "apps are changing the world, enriching people's lives, (...) enabling developers (...) to innovate like never before," (Apple Inc., n.d.-a) Google's Developer Policy Center emphasizes that apps should be "a safe and trusted experience for everyone" and stresses developers' responsibilities in creating such an experience (Google, n.d.-a). Through a combination of quantitative and qualitative research, this study examined how Apple and Google establish boundaries in their app store policies to balance innovation, control, and safety and ensure developers uphold this sense of responsibility in app design.

Our primary materials included Apple's App Store Review Guidelines (Apple Inc., n.d.-a) and 'Developer Policy' and Google's Developer Policy Center (Google, n.d.-a). These policies served as starting points, with historical versions retrieved via the Internet Archive Wayback Machine. Initially, Apple moderated app development through the iOS Developer Program License Agreement. However, policies predating September 2014 were not publicly accessible and required a valid developer ID for access. On September 9, 2010, Apple introduced significant changes, unveiling a new document titled the "App Store Review Guidelines." Apple described the change as a step toward greater transparency, stating: "for the first time we are publishing the App Store Review Guidelines to help developers understand how we review submitted apps. We hope it will make us more transparent and help our developers create even more successful apps for the App Store" (Apple Inc., 2010) These guidelines, however, remained behind a login until they were made publicly accessible in 2014, but copies had previously been circulated online.

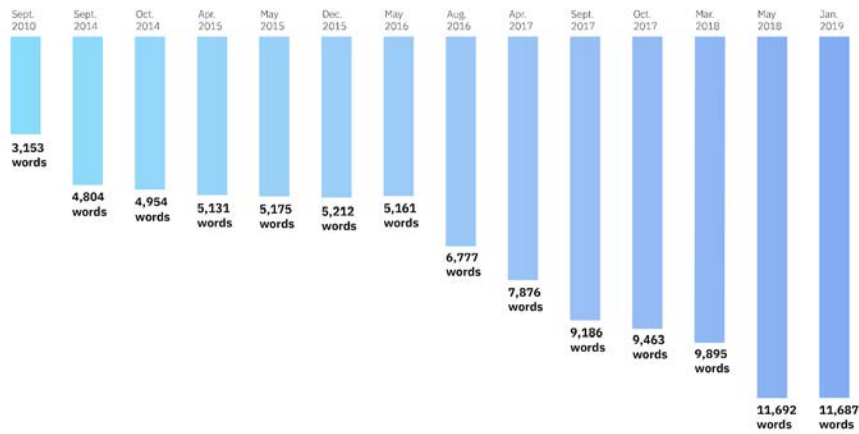
Similarly for Google Play, the Wayback Machine provided archived policy versions starting from July 27, 2016. To trace earlier policy developments back to Google Play's launch in 2008, additional archived sources were included, including earlier versions of Android Developer Policies (retrievable back to March 9, 2012) and Android Market guidelines (dating to October 21, 2008). These naming shifts typically did not introduce completely new policy frameworks, but were reflected in changes in URL structures and document organization to gain access to the stores and their governing policy documents. Ultimately, we curated a dataset comprising 14 Apple policy documents (2010–2019) and 10 Google policy documents (2008–2019), providing a comprehensive basis for analysis.

We conducted multiple analyses to explore the evolution of these app store policies. To identify meaningful textual changes, we compared consecutive versions of policy documents using DiffChecker, visualizing these changes in graphs to highlight trends over time. We performed word counts to track the changing length of documents. A

close reading of specific sections, such as the Introductions and Conclusions, allowed us to examine shifts in tone, focus, and emphasis on user versus developer priorities. Additionally, we analyzed the tables of contents to understand how the structure and focus of policies evolved. Finally, policy updates were mapped against external events, such as operating system releases and regulatory shifts (i.e. GDPR), to contextualize the factors driving these changes.

Our analysis revealed changes in the policies' length, structure, content, and responsiveness to external factors. The policy documents for both platforms grew substantially over time, reflecting an increasing emphasis on moderating third-party app development (Figure 9.1(a) and (b)). In 2010, Apple's policies were already more detailed, with over 3,000 words, while Google's guidelines were minimal, at less than 1,000 words. By 2019, both had expanded to over 11,000 words, suggesting a parallel response to the growing complexity of the app ecosystem. A notable turning point in this evolution was the GDPR's implementation in May 2018. Apple's policies, for instance, saw a marked increase of 2,000 words in that same month, reflecting the introduction of new requirements around data privacy and transparency. The GDPR imposes strict rules on how platforms collect, process, and store personal data, necessitating revisions to app store policies to ensure compliance.

App Store Review Guidelines



Android Market Policies

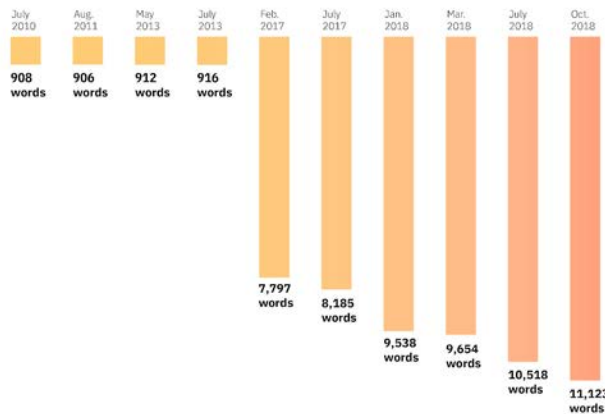


Figure 9.1(a) and (b) Word count of Apple's App Store Guidelines (left) and Google Play's Developer Policy Centre (right). Time series bar chart (January 2019). Visualization by Sofia Chiarini (DensityDesign).

Frequent updates to Apple's policies revealed a more proactive approach to moderation. Major reformatting occurred in 2016, introducing a redesigned table of contents and additional subsections. In contrast, Google Play implemented significant structural changes in 2017, adding new sections like "Spam," "Permissions," and "Content Ratings." Both platforms' tables of contents became increasingly detailed over time, reflecting shifts in focus toward data privacy, child protection, and security. For instance, Apple introduced a "Data Security" subsection in May 2018, aligning with GDPR compliance and emphasizing the importance of user data and compliance with security standards.

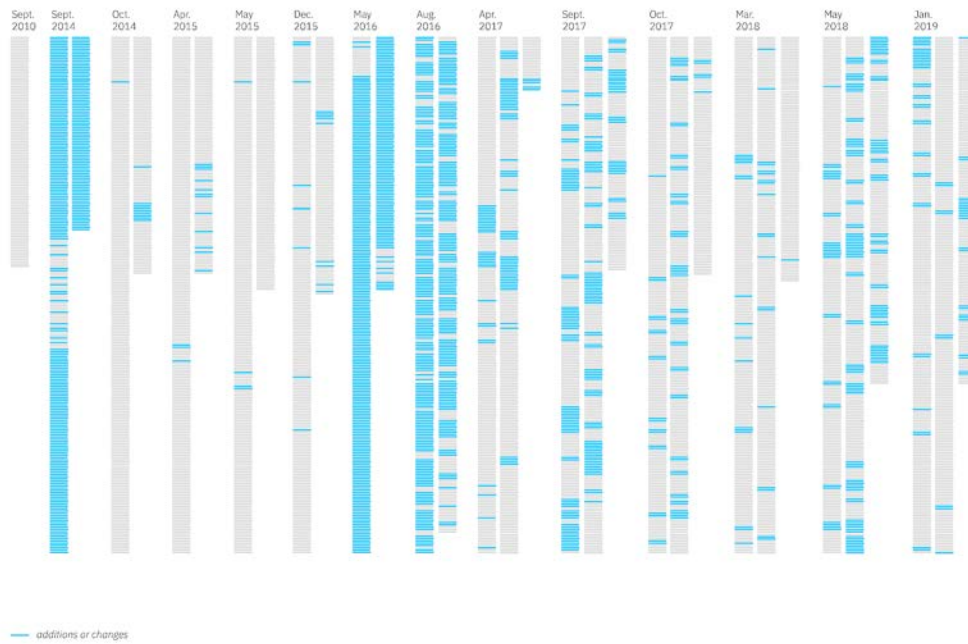
A close reading of Apple's policy documents highlighted a clear shift in tone over time. Early versions welcomed developers and emphasized collaboration, but later iterations focused on user experience, stressing app quality and usability. Apple also removed references to the "Review Board" in favor of the "Resolution Center," signaling a reduced emphasis on impartial appeals and a more streamlined,

compliance-focused moderation approach. In contrast, Google Play's policies maintained a more developer-oriented focus, providing extensive written guidance to help developers navigate submission processes and deal with user safety considerations. This approach reflects Google's more open ecosystem, balancing developer autonomy with compliance requirements.

Apple's policy updates were often aligned with iOS releases, underscoring its tightly integrated ecosystem of platforms and services. For example, new sections addressing HealthKit and Apple Pay appeared in September and October 2014, coinciding with the launch of iOS 8 and related features. Google Play's updates, however, were less tied to Android OS releases, reflecting a more independent approach. Both platforms incorporated legal and regulatory changes, such as GDPR and the US Children's Online Privacy Protection Rule (COPPA), into their policies. Google Play's 2017 update on child-directed apps, for instance, responded to rising concerns about children's app usage.

Apple demonstrated more frequent and extensive formatting updates than Google Play (Figure 9.2(a) and (b)). Changes in 2016 included reformatting subsections and reorganizing content, while 2017 introduced further refinements with new subcategories. Google Play's most significant visual and structural changes occurred in mid-2017 and 2018, focusing on clarity and compliance with evolving legal standards. The evolution of these policy documents illustrates how app store moderation materializes, where changes in policy length, structure, and content reflect internal platform strategies as well as external regulatory pressures.

App Store Review Guidelines changes from 2010 to 2019



Android Market Policies changes from 2010 to 2019

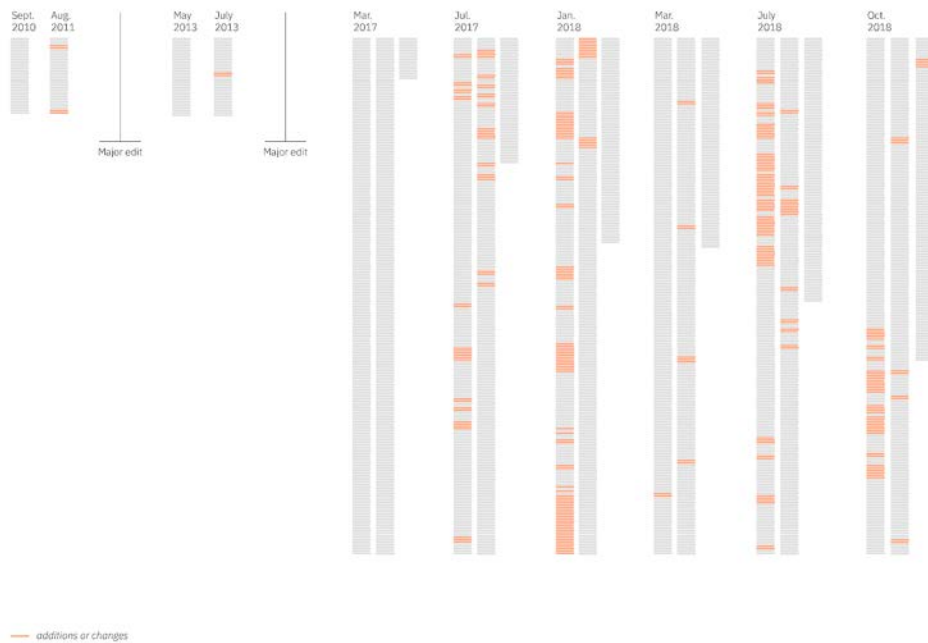


Figure 9.2(a) and (b) Textual changes to Apple's App Store Guidelines (left) and Google Play's Developer Policy Centre (right) (2010–2019). Annotated time series bar chart. Visualization by Sofia Chiarini (DensityDesign).

Exploring banned content categories

Lengthy app store policies serve as critical moderation mechanisms that explicitly describe what content is allowed or prohibited on these platforms, which types of apps they do not want in their app stores, and outline explicitly forbidden content categories, such as sexual content, personal loans, and gambling (Google, n.d.-a; Apple Inc., n.d.-

d). Google Play store's Developer Policy includes a category on "Inappropriate Content" which states, "We don't allow apps that contain or promote sexual content or profanity, including pornography, or any content or services intended to be sexually gratifying" (Google, n.d.-e). App developers that create apps contravening these policies may not pass review, or their apps may be removed when detected. Notably, when users query for forbidden content, this does not result in a complete absence of results but instead reveals suggested apps. This outcome points to an organizational and algorithmic logic underpinning app store moderation and the gatekeeper role of app stores.

In this section, we examine how app store moderation manifests through restrictions on certain types of content and the implications of these restrictions, using the example of apps related to "pornography" – a category explicitly banned by app store policies. As similarly discussed by Cows and Morley (2022), both Apple and Google acknowledge that apps with user-generated content (e.g. porn) present particular moderation challenges. While they require such apps to have mechanisms for filtering objectionable content, as well as reporting and blocking tools, their guidelines don't clearly define where to draw the line between apps that themselves transgress and those that merely host significant amounts of objectionable content.

In the following case study, we aimed to understand what happens when users search for objectionable content. At the time of study, Google Play store's Developer Policy included a main category entitled "Restricted Content" which asked developers "Does your app cross the line"? (Google, 2018). This category included a subcategory of "Sexually Explicit Content", stating they don't allow apps that contain or promote pornography. We then asked, what happens when a user types in an objectionable query and the policy dictates that no apps should be surfaced for this purpose? In our "Objectional Queries" study (Helmond, Van der Vlist, & Weltevrede, 2018), we looked at the prohibited content query [porn] and variations in local language for [pornography] in five local Google Play Stores (Brazil, Denmark, India, Italy, and the Netherlands) (Figure 9.3).

Analyzing the recommendations in response to such queries aims to understand how the platform navigates its role as gatekeeper. The results show a relatively consistent return for [porn] across regions but vary significantly for language-specific terms like for [pornography].

GOOGLE PLAY STORE

query: PORN



query: PORNOGRAPHY



Figure 9.3 Top 20 app search results for [porn] and [pornography] queries in the Google Play Store, per country. Comparative ranking table (2018). Visualization by Serena Del Nero & Marco Mezzadra (DensityDesign).

Subsequently, we retrieved the "similar" or "related" (recommended) apps to expand our initial dataset to further analyze and categorize the app types returned for banned queries. The results are visualized in a network, analyzed, and annotated to identify clusters and thematic themes (Figure 9.4).

The thematic clusters that we identified include: anti-porn apps explicitly named to counter pornography (e.g. "Stop Porn Addiction"); covert facilitators of pornography, which cannot explicitly mention porn due to app store policies but enable such practices or algorithmically implicated tools (e.g. VPNs, dating apps); and, unrelated but thematically adjacent apps (e.g. social media platforms, religious apps). Anti-pornography recommendations frequently included religious or moral framing (e.g. Christian apps). Facilitators, like VPNs and spicy social platforms, reflected a behavioral, rather than ideological, logic, focusing on user practices around accessing restricted content.

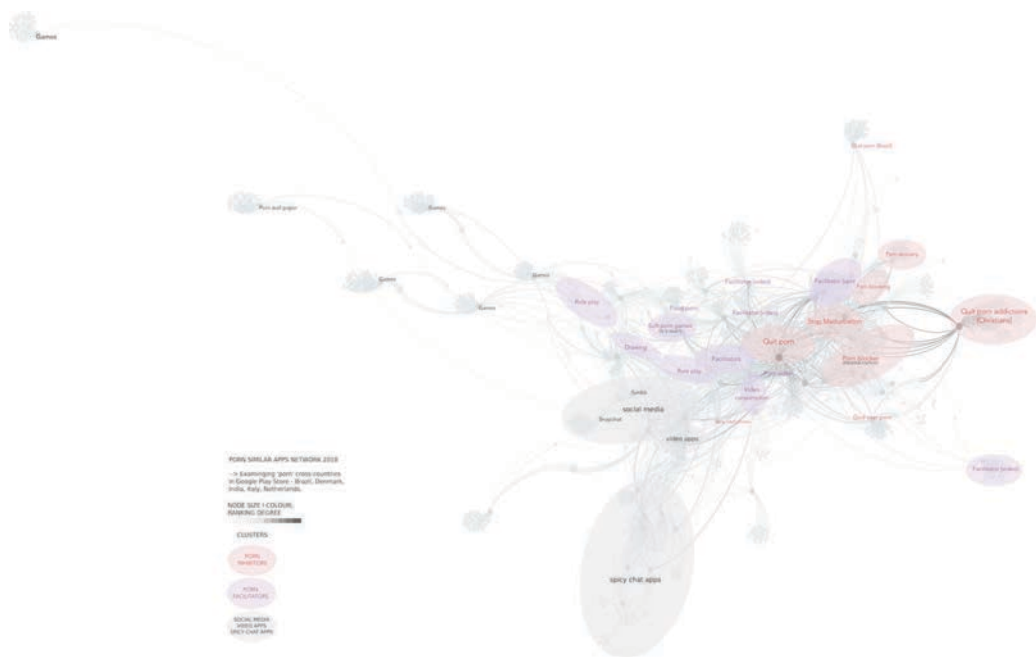


Figure 9.4 Thematic clusters in the app recommendation network ("similar apps") for [porn] apps in the Google Play Store. Annotated network graph (2018). Visualization by Serena Del Nero & Marco Mezzadra (DensityDesign).

The moderation of banned content extends beyond prohibition, where the example of apps related to porn reveals how app stores actively shape user behavior through recommendations and redirections. These findings suggest that app stores do not merely enforce policies but also act as ideological and behavioral gatekeepers, curating the app ecosystem in ways that reflect broader cultural and commercial priorities. The presence of covert facilitators underscores the challenges of maintaining policy boundaries in app marketplaces, and reflects the broader complexities of app stores acting as ethical and moral arbiters, such as the difficulty in defining pornography and the various acceptable or unacceptable ways it appears in media and society (Paasonen, 2016). As gatekeepers, we find that app stores do not merely control access, but are positioned at the intersection of policy enforcement, cultural curation, and economic strategy. This example study serves as a methodological template for further examining these dynamics, offering insights into how platforms mediate access to content and shape user practices in increasingly algorithmic ecosystems.

App store moderation during times of global crisis: COVID-19 pandemic response apps

The COVID-19 pandemic brought unprecedented challenges to global app store moderation practices, requiring rapid mobilization to battle a health crisis across sectors, including technology. App stores emerged as critical platforms in this crisis, hosting an array of rapid-response apps. This case explores app stores' evolving moderation mechanisms during the pandemic, focusing on their role as gatekeepers and facilitators of a global app ecosystem. The approach examines how moderation arrangements are reconfigured during times of crisis to understand the dynamics of app stores as intermediaries not only between platforms, developers, end-users, but also governments.

To analyze the moderation of COVID-19 apps, this study (Dieter et al., 2021) uses a multi-situated methodological approach (Dieter et al., 2019). Data was collected from

the Google Play and Apple's App Store using systematic queries for COVID-19-related terms such as [covid], [covid-19], and [corona]. Custom scrapers were used to retrieve app metadata, including app titles, descriptions, developer names, screenshots, and release/update dates. These queries were conducted across the 150 countries supported by Google Play and the 140 regions supported by Apple's App Store, capturing geographically specific global variations. Apple's App Store returned ranked lists of 100 apps per country for our search queries, resulting in a total source set of 248 unique iOS apps. Google Play, however, did not produce such ranked lists. Instead, it rerouted all COVID-19 queries to a relatively small set of pre-selected apps in each local store, resulting in a total source set of 247 unique Android apps.

Central to app development and distribution, app stores as intermediaries between stakeholders. During the pandemic, these platforms adopted exceptional moderation measures, balancing commercial interests with public health priorities. The pandemic thus highlighted their quasi-public role, as they became key actors in coordinating global responses. Both Google and Apple implemented stringent policies for COVID-19-related apps, including restricting apps to those developed by recognized entities such as government agencies and health organizations. For instance, Apple prohibited entertainment or game apps themed around COVID-19. Both stores also disallowed monetization features, such as ads and in-app purchases, to prevent profiteering. Google Play also used editorial filters to curate app listings and prevent the spread of misinformation. Google limited COVID-19-related search results to a pre-approved list, reflecting their moderation of both app content and user queries. When a user searches for COVID-19-related terms, they are automatically directed to Google's curated list of COVID-19 apps specific to their location. We discovered that we could easily bypass this editorial filter by using simple misspellings (e.g. [COVIID], [coronna], etc.), which resulted in Google Play displaying a more extensive list of relevant apps. This difference in query results is likely, at least in part, due to creative developer/user practices that seek to evade algorithmic moderation through deviant spellings and new vernaculars, a practice increasingly termed "algospeak" (Steen et al., 2023). Given this discrepancy in query results, we gathered two complementary datasets from Google Play: (a) an 'editorial' set containing app responses for each country, which included 247 unique apps, and (b) a 'non-editorial' set with 163 additional apps obtained through misspellings. These 163 additional apps were available on Google Play, but the editorial filter prevented them from appearing in standard searches for [COVID-19].

We began by comparing the distribution of apps in our datasets and the various actors involved in their production. (Figure 9.5) illustrates the distribution of COVID-19 apps across both app stores, distinguishing between editorial and non-editorial apps from Google Play. Individual apps are color-coded to represent different actors: government, civil society, health authorities, academic institutions, and private entities.

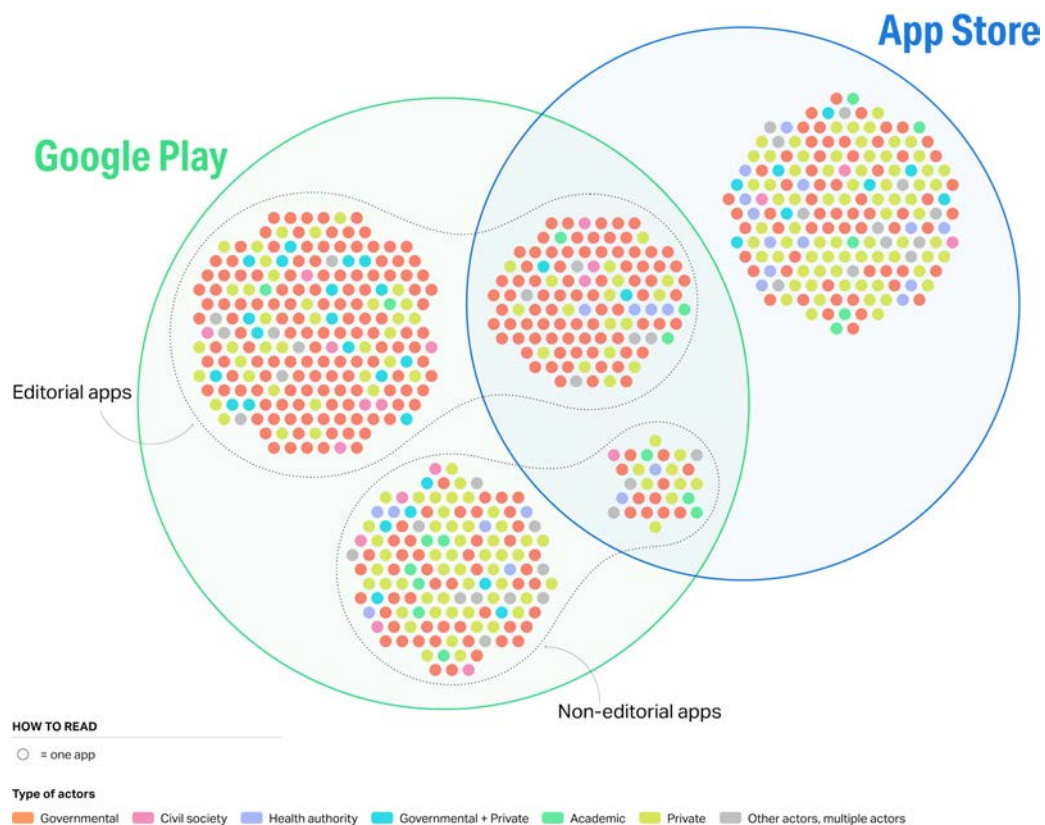


Figure 9.5 Demarcated source sets (Google Play and App Store). Light green: Android app ecosystem (Google Play source set); light blue: iOS app ecosystem (App Store source set). Venn diagram with clusters (2019). Visualization by Giovanni Lombardi, Angeles Briones & Gabriele Colombo (DensityDesign).

The most notable finding from this graph is the high number of apps only available in one app store. Government entities often develop apps shared across stores, but many government-created apps were exclusively found in a single store. Significant differences existed in the types of developers creating COVID-19-related apps in each store (Figure 9.6). Government-produced apps were the most common in both stores, establishing governments as official and recognized sources according to the policies of the app stores. However, these apps were significantly more prevalent in Google Play, where they account for 65% (N = 267) of the total, and even more so in the Google Play editorial selections, which comprise 79% (N = 195). In comparison, the prevalence in the App Store is only 48% (N = 121). This discrepancy is likely due to Google Play's strict editorial policy for COVID-19 apps, which listed only a limited number of curated apps for most countries. As a result, Google Play's strategy amplified the visibility and presence of government-made apps within its ecosystem.

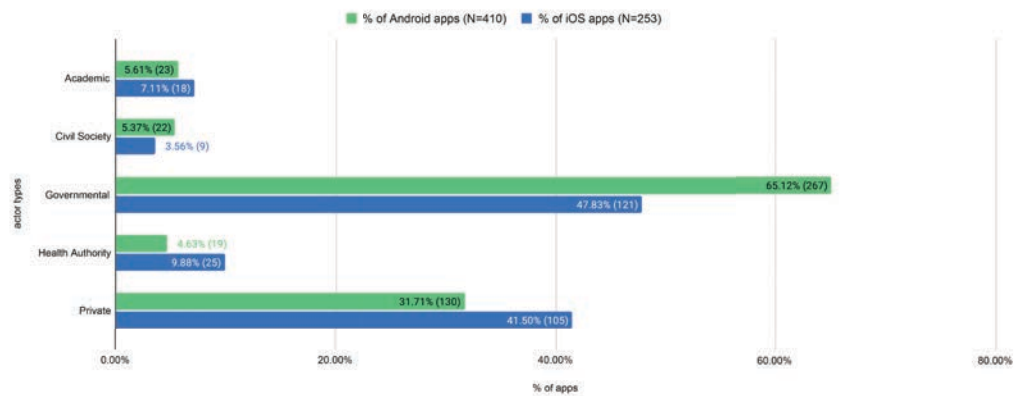


Figure 9.6 Actor types identified behind [COVID-19]-related apps (Android and iOS), based on the listed developer names and app descriptions. Note: apps can belong to multiple categories. Bar chart (2022). Visualization by the authors.

After examining the distribution of apps and types of actors across platforms, we shifted our focus to their geographical distribution. The App Store’s ranked lists of apps were less specific to individual countries and showed a significant overlap among different countries and regions. In contrast, Google Play, which highlighted only country-specific COVID-19 apps through its editorial filter, presented a clearer geographical picture (Figure 9.7). In this store, we observed that most countries offered a limited selection of country-specific apps besides two WHO apps: OpenWHO: Knowledge for Health Emergencies and WHO Info.

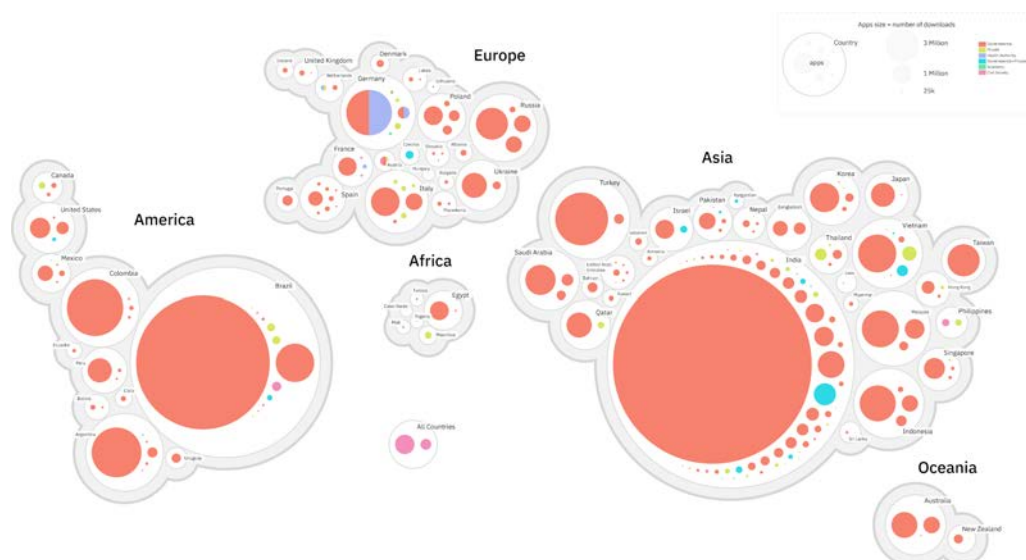
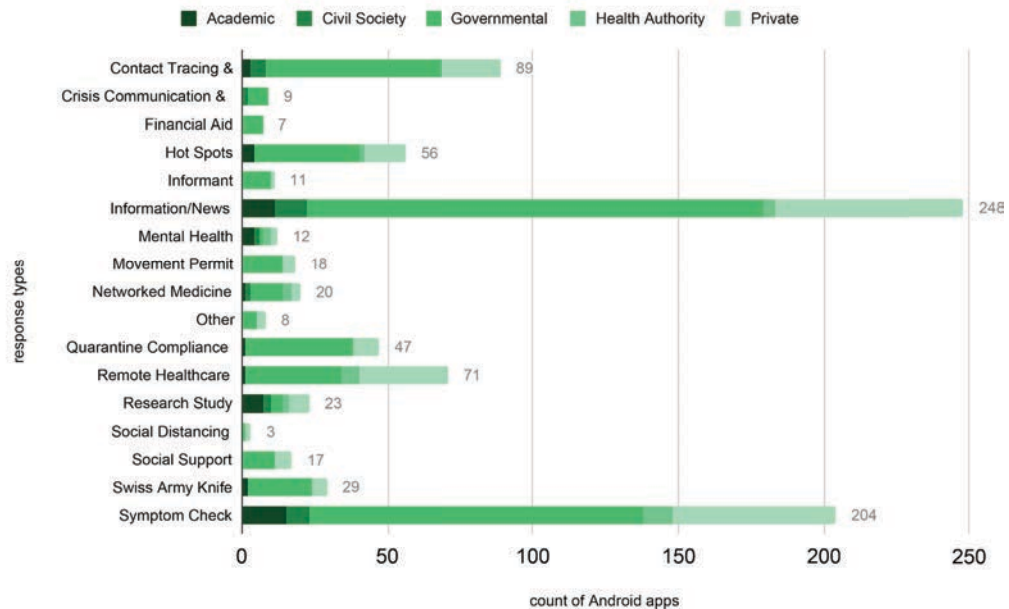
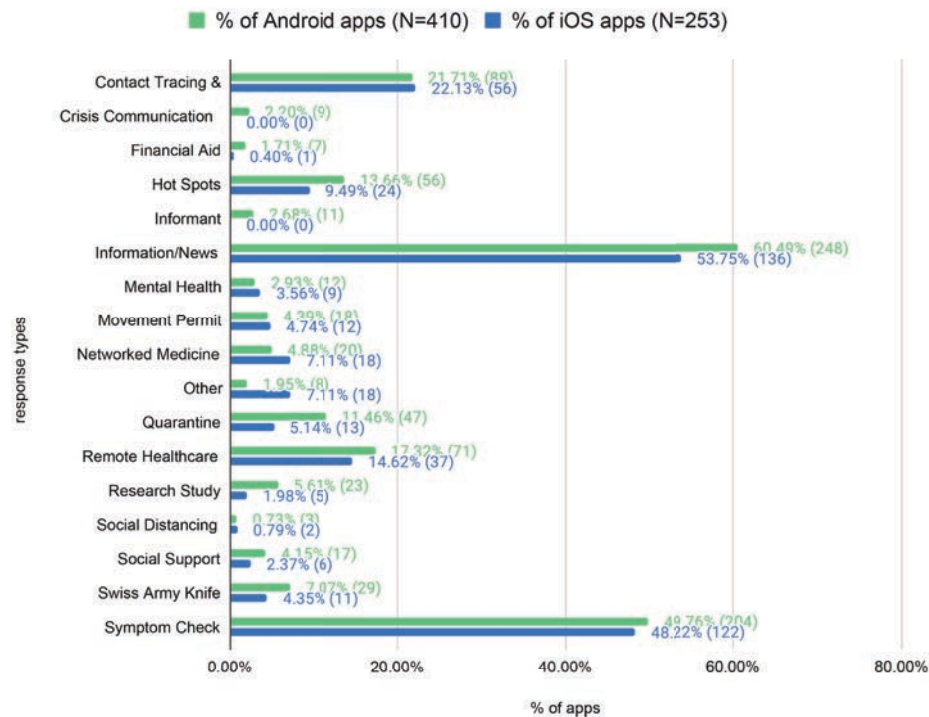


Figure 9.7 Geographical distribution of [COVID-19]-related Android apps by country or region. Proportional pie chart map (2022) Visualization by Giovanni Lombardi, Angeles Briones & Gabriele Colombo (DensityDesign).

To analyze the types of responses curated by app store moderation, we examined the kinds of apps developed for the COVID-19 pandemic and how these reflected the platforms’ regulatory mechanisms (Figure 9.8(a) to (c)). App store moderation influenced the prominence of specific app types, with 50–60% of apps offering news and information on the pandemic, often developed in collaboration with authoritative entities like the WHO. These apps illustrate how platforms prioritized connecting users to official sources, aligning with their moderation role in countering misinformation. We observed a broader diversity of responses beyond widely reported contact-tracing

apps (over 20%). Nearly half of the apps provided symptom checkers or tools for reporting health data, while others offered remote healthcare services (15%). These apps were developed by a mix of government entities, private actors, and academic institutions, highlighting how app stores mediated the creation and distribution of apps to balance public health objectives with their content policies and gatekeeping practices.



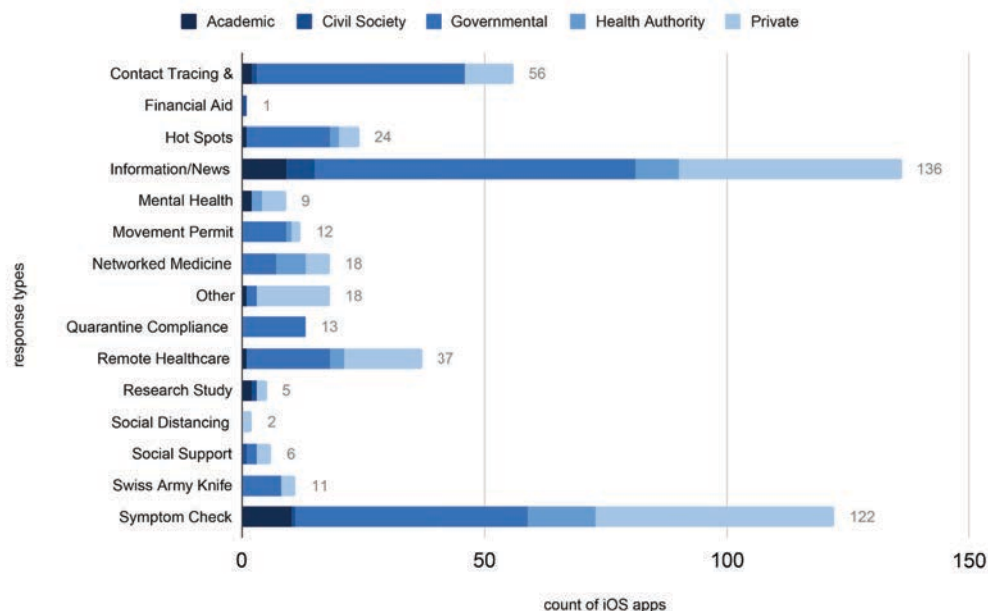


Figure 9.8 (a) to (c) Comparison of response types represented by [COVID-19]-related apps (Google Play vs App Store). Note: apps can belong to multiple categories. Bar charts (2022). Visualization by the authors.

The moderation of COVID-19 apps exemplifies a critical reconfiguration of app stores' role, transforming them from commercial intermediaries to central actors in public health governance. The pandemic's exceptional circumstances necessitated extraordinary moderation measures that reshaped content and app store regulation and actively supported public health initiatives. A complementary paradigmatic example of this shift was the joint development of the Google–Apple Exposure Notification (GAEN) framework (Dieter et al., 2021; Mann et al., 2022). Implemented at the operating system level, GAEN facilitated contact tracing by enabling automated logging of close encounters between Android and iOS smartphone users while promising robust privacy protections. This collaborative technological intervention illustrates how Big Tech companies sought to moderate populations globally through technological solutions, extending their infrastructural power into the health domain. Additionally, their app stores' restrictive pre-approval filters constrained the visibility of COVID-19 apps that were not from authorized national health organizations.

COVID-19 app moderation represents a quintessential example of a reconfiguration of existing moderation arrangements during times of crisis (cf. Polese & Helou, 2024), where digital platforms transcended their 'traditional' commercial roles to become critical infrastructure for public health management. Next, we turn from the role of app stores in health crises to their role in times of geopolitical conflict.

Platform geopolitics: App store moderation during times of conflict

The Russian invasion of Ukraine on February 24, 2022, illustrates how geopolitical events impact app store moderation and the broader app ecosystem. In the invasion's wake, app usage patterns in Russia and Ukraine shifted dramatically as Western companies exited the Russian market and the Ukrainian population adapted to the demands of war. App stores became critical sites of moderation, mediating access to tools and information in two distinct yet interconnected app ecosystems. This section

draws on student research conducted during the Spring Data Sprint 2022 at the University of Amsterdam (Gehasse et al., 2022; Quaglia et al., 2022; Roman et al., 2022) to explore these dynamics through the lens of moderation strategies, app removal, and changing user needs during the conflict.

The European Union's response to the invasion included sanctions that banned Kremlin-backed outlets such as RT and Sputnik from app stores, reinforcing the moderation role of platforms in aligning with geopolitical mandates (Kayali, 2022). Simultaneously, the Russian government restricted access to Facebook, Instagram, and Twitter, labeling Meta an "extremist organization" after the company temporarily adjusted its content moderation policies to allow hate speech against Russian forces (Duffy, 2022). These measures triggered a substantial reconfiguration of the app ecosystem in Russia. By mid-March 2022, Russia's (Apple) App Store had lost nearly 7,000 apps, marking a 105% increase in app removals compared to the preceding month (Perez, 2022). Consequently, the Russian internet group VK developed its own app store, called [RuStore](#) (RuStore, n.d.), with other large Russian IT companies and the support of the Ministry of Digital Development of the Russian Federation. RuStore calls itself "the official Russian store of mobile applications for Android" to provide 'safe access' to popular games and applications" ([RuStore, 2025](#)).

To analyze shifts in app store rankings in Russia (Gehasse et al., 2022; Quaglia et al., 2022) during the conflict, the studies employed a systematic data collection approach. Data on the top 50 downloaded apps for both Android (Google Play Store) and iOS (Apple App Store) platforms were gathered across three distinct time periods: the week before the invasion (February 14–20, 2022), the second week of the conflict (March 7–13, 2022), and one month into the war (March 21–25, 2022). These time frames captured pre-war patterns, the immediate response to the invasion, and mid-term shifts in user behavior and app availability. Additionally, top app data was collected for the category 'social' for the Russian app stores.

Data was collected using mobile analytics platform Data.ai, which provides app rankings and metadata. The rankings were compiled by manually cataloging app names, categories, and countries of origin. Six app categories were defined to analyze the data: services, shopping, social networks, amusement, games, and VPNs. Both Android and iOS rankings were analyzed to ensure a comprehensive view of app ecosystem dynamics. Although Apple officially exited the Russian market (Sheftalovich & Gijs, 2022) a significant number of users still accessed iOS products, making the inclusion of Apple rankings relevant. The rankings were comparatively analysed across the three timeframes. By tracking the rise and fall of specific apps over time, the method provides insights into how the geopolitical crisis reshaped the app ecosystem.

The top apps data over time revealed profound shifts in the rankings of the most downloaded apps (Figure 9.9). In Russia, VPN apps surged in popularity two weeks after the invasion, accounting for 20% of the top downloads as users sought to circumvent government censorship. Simultaneously, Western apps such as Spotify and Zoom disappeared from the charts, replaced by Russian alternatives like Yandex.Music and TenChat, signaling a transition to domestic platforms amid sanctions and market exits (Gehasse et al., 2022). The subsequent analysis of top social apps in Russia (Quaglia et al., 2022) revealed significant disruptions following the

invasion (Figure 9.10). Prior to the conflict, app rankings exhibited stability, with minor fluctuations in popularity. However, after February 24, 2022, the rankings shifted dramatically. Apps like Yappy, previously a top contender, dropped to the bottom of the rankings, while six new apps entered the top 10. Notably, the "Lite" versions of TikTok and Instagram surged in popularity, outperforming their full versions. This shift reflects user attempts to circumvent censorship and access these platforms via VPNs, which reduce internet speed, making lighter app versions more attractive.

While the Russian government restricted access to platforms such as Facebook and Instagram, the apps themselves remained available in app stores. In addition to their relatively high ranking in the top social apps, archived evidence (Google, 2022) confirms that the app pages were still accessible during this time, allowing users to download them. This suggests that app stores adopted a selective moderation approach, maintaining access to such apps despite the broader geopolitical pressures. This highlights app stores' complex role as intermediaries, navigating local governmental restrictions while maintaining functionality for users in politically tense contexts.

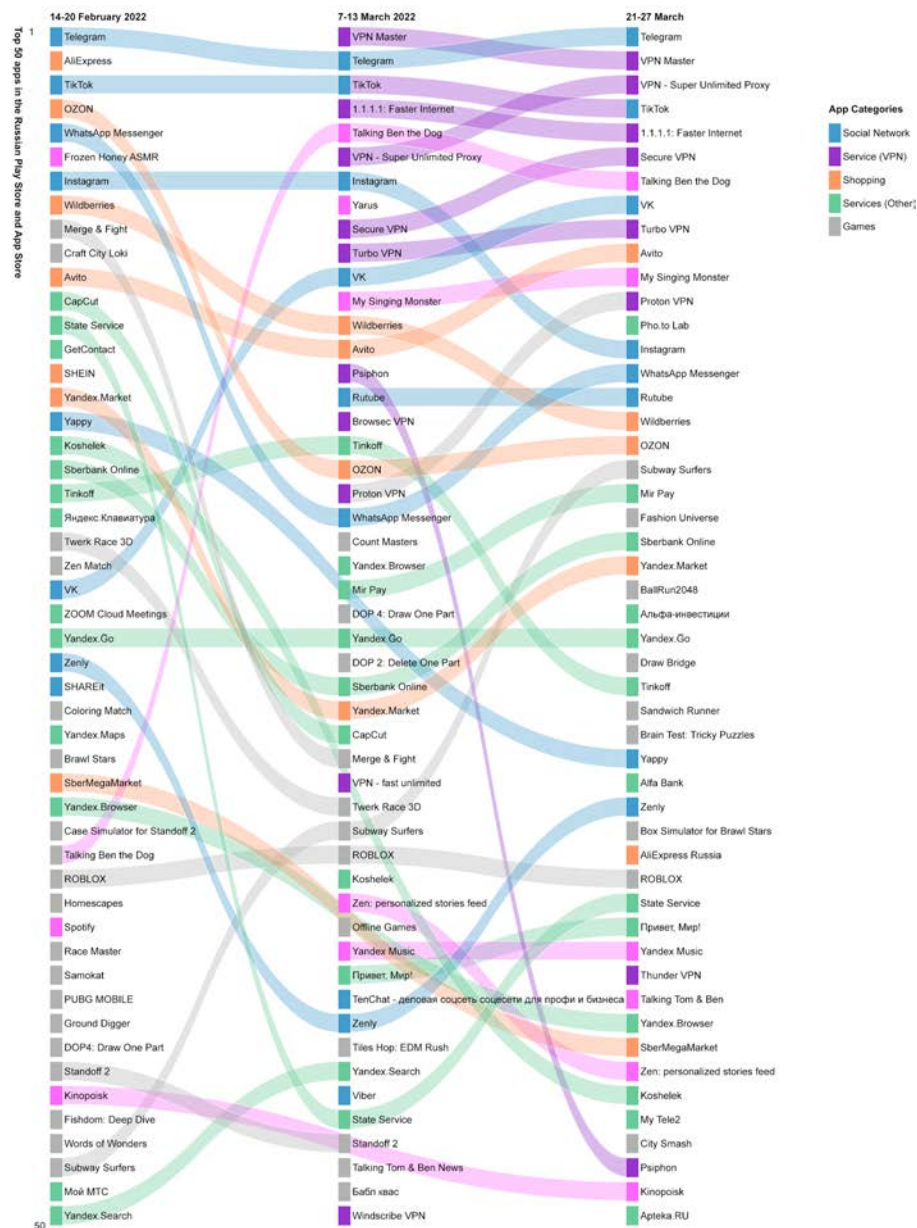


Figure 9.9 Top 50 apps in Russian app stores over time. Rank flow visualization (2022). Visualization by the authors.

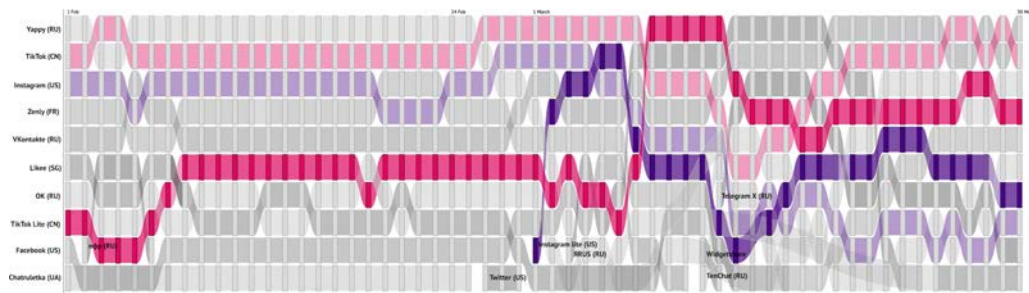


Figure 9.10 Timeline of most downloaded social apps in Russia’s app stores. Rank flow visualization (2022). Visualization by the authors.

For Ukraine, a similar analysis was performed. The conflict reshaped app store rankings to prioritize tools that supported survival, communication, and navigation in the face of disruption. Encrypted messaging apps like Signal and Telegram gained prominence, while offline navigation tools such as MAPS.ME reflected the practical needs of navigating disrupted environments. Government-backed apps like Diia became central to delivering critical information and facilitating access to public services during the war (Roman et al., 2022). Telegram remained a constant presence at the top of the rankings in both countries, underscoring its role as a vital communication tool for encrypted messaging and information dissemination.

The studies highlight the layered nature of app store moderation during the conflict. Moderation strategies ranged from direct interventions like app removals and geo-blocking to algorithmic curation that shaped app visibility. For example, domestic apps such as VKontakte and TenChat rose in rankings in Russia as users transitioned to platforms that remained accessible under local regulations. In Ukraine, the rise of apps like Signal and MAPS.ME reflected a distinct response to the conflict, emphasizing survival and secure communication needs. In addition to aligning with state and international mandates, app stores responded to geopolitical pressures by moderating access to critical apps and services. Thousands of apps disappeared from Russian app stores as companies withdrew voluntarily, or in response to international sanctions, while Russian apps such as RT and Sputnik were pulled from the major app stores in all territories outside of Russia (Perez, 2022; Kayali, 2022). Yet, several major platforms maintained their presence, revealing the complex dynamics of content moderation in times of conflict. These apps’ availability not only reflected market pressures but also the nuanced strategies of app stores in navigating geopolitical constraints. Apple’s recent compliance with Russian requests to remove VPN apps exemplifies the tension between maintaining user access and adhering to local regulations, a recurring theme in platform moderation during the conflict (Franceschi-Bicchierai, 2024)

Conclusion: The implications of app store moderation for app ecosystems

This chapter has explored the critical role of app stores, such as Google Play and Apple’s App Store, as facilitators, gatekeepers, and moderators of the mobile app ecosystem. These platforms wield significant influence over the availability, accessibility, and visibility of apps through their internal moderation mechanisms and

adherence to external regulatory frameworks. By mediating relationships between developers, users, and external regulations, app stores actively shape the ecosystem they oversee, determining which apps thrive, which are restricted, and how app ecosystems evolve across regions and contexts.

App stores' moderation practices are multi-layered, fragmented, and context-dependent, as illustrated through the case studies of developer policy evolution, the COVID-19 pandemic, and the Russian invasion of Ukraine. Internally, app stores enforce moderation through policies, moderation mechanisms, and algorithmic curation, while externally, they navigate pressures from global regulations and localized social, political, and economic contexts. This dual role allows them to operate as arbiters of compliance, as well as mediators of access and influence.

A key focus of this chapter was to investigate how content moderation operates specifically within app stores, distinguishing it from practices on other platforms like social media. App stores typically moderate entire apps or their app developers as singular units, enforcing moderation at the level of app creation, distribution, and functionality. This approach involves a combination of pre-publication reviews, policy enforcement, and algorithmic curation, which determine not only the presence of apps but also their discoverability and visibility. The case studies illustrated how this mode of content moderation operates across multiple layers: from rejecting non-compliant apps during submission to promoting official apps during global crises. By moderating access to tools and services, app stores extend content moderation beyond the policing of speech and behavior, addressing broader ecosystem dynamics and regulatory compliance. This moderation model positions app stores as powerful intermediaries, whose decisions shape not only app ecosystems but also how users interact with critical infrastructures, particularly in times of crisis.

The fragmented and layered nature of app store moderation is encapsulated in the concepts of "governance arrangements" (Van der Vlist et al., 2022), highlighting how moderation mechanisms are materially embedded in different layers of the platform, and "patchwork governance" (Duguay et al., 2020), reflecting the interplay between global platform policies and regional variations. During the COVID-19 pandemic, for example, app stores emerged as quasi-public infrastructures, curating apps from verified health authorities while restricting monetization features to prevent profiteering. Similarly, during the Russian invasion of Ukraine, geopolitical pressures reshaped app ecosystems, with VPN apps rising in ranking in Russia as users circumvented censorship, while tools for survival and secure communication dominated app usage in Ukraine. These examples demonstrate how app store moderation adapts to crises, balancing public interest with platform priorities.

Methodologically, this chapter highlights the importance of longitudinal policy analysis, query-based studies, and crisis-focused case studies for understanding the dynamics of app store moderation. Crises serve as a prime example of "platform frictions" which are key sites to investigate the power that platforms such as app stores have (Popiel and Vasudevan, 2024). These methodological approaches reveal how moderation practices evolve in response to regulatory shifts, social pressures, and crises, offering a framework for studying moderation across other platform ecosystems.

App stores are not passive intermediaries but active agents shaping the app ecosystem through layered and evolving moderation practices. By responding to external regulatory demands, adapting to crises, and leveraging internal strategies, app stores mediate access to apps and services, influencing the design, distribution, and societal impact of mobile technologies. Recent high-profile cases, such as TikTok's ban from app stores in specific regions, exemplify the growing entanglement of app store moderation with geopolitics, highlighting how these platforms increasingly serve as battlegrounds for political and economic power struggles.

This concentration of platform power—most visible when app stores operate as de facto public infrastructures during crises—underscores the need for clear and enforceable regulatory mandates under instruments such as the Digital Services Act (DSA) and Digital Markets Act (DMA). Under the DSA, the major app stores are designated as Very Large Online Platforms (VLOPs), which means they are subject to heightened obligations related to transparency, safety, and content moderation. These obligations include requirements for risk assessment, mitigation measures, and external audits.

To enhance accountability, the DSA mandates that VLOPs provide access to vetted researchers studying systemic risks related to disinformation, public health, and civic discourse. Policymakers should require app stores to share more detailed moderation datasets—including time-stamped decisions, appeals procedures and outcomes, and major algorithmic ranking or demotion interventions in their stores—with qualified academic and civil society researchers.

Securing longitudinal access to such data would enable monitoring by regulators and researchers, offering critical insights into how moderation practices evolve over time. This is especially vital when app stores assume quasi-governmental roles, such as during the COVID-19 pandemic, when they decided which health-related apps were made available to the public. Such decisions have implications not only for user rights, innovation, and market competition, but also for the democratic legitimacy of infrastructural governance in mobile ecosystems.

References

Apple Inc. (n.d.-a). App Store Review Guidelines. <https://developer.apple.com/app-store/review/guidelines/>

Apple Inc. (n.d.-b). Apple Media Services Terms and Conditions. <https://www.apple.com/ie/legal/internet-services/itunes/ie/terms.html>

Apple Inc. (n.d.-c). App Tracking Transparency. <https://developer.apple.com/documentation/apptrackingtransparency>

Apple Inc. (2010). Statement by Apple on App Store Review Guidelines. <https://web.archive.org/web/20100911113848/http://www.apple.com/pr/library/2010/09/09statement.html>

Apple Inc. (n.d.-d). Restricted content guidelines. <https://support.apple.com/nl-nl/guide/adguide/apd9a07c1727/icloud>

Cowls, J., & Morley, J. (2022). App Store governance: the implications and limitations of duopolistic dominance. In *The 2021 Yearbook of the Digital Ethics Lab* (pp. 75-92). Cham: Springer International Publishing.

Dieter, M., Gerlitz, C., Helmond, A., Tkacz, N., van Der Vlist, F. N., & Weltevrede, E. (2019). Multi-situated app studies: Methods and propositions. *Social Media + Society*, 5(2), 1–15.

Dieter, M., Helmond, A., Tkacz, N., van der Vlist, F., & Weltevrede, E. (2021). Pandemic platform governance: Mapping the global ecosystem of COVID-19 response apps. *Internet Policy Review*, 10(3), 1-28.

Duffy, K. (2022). Russia has added Meta to a list of 'extremist' and 'terrorist' organizations, a report says. *Business Insider*. <https://www.businessinsider.com/russia-adds-meta-list-extremist-terrorist-groups-facebook-zuckerberg-report-2022-10>

Duguay, S., Burgess, J., & Suzor, N. (2020). Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence*, 26(2), 237–252.

European Commission. (n.d.-a). Digital Services Act. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

European Commission. (n.d.-b). Digital Markets Act: Ensuring fair and open digital markets. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en

European Commission. (n.d.-c). Europe fit for the digital age: New online rules for platforms. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act/europe-fit-digital-age-new-online-rules-platforms_en

European Commission. (2023). Questions and Answers: Digital Services Act. https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2349

European Commission. (n.d.-d). Competition Policy. https://competition-policy.ec.europa.eu/index_en

European Commission. (n.d.-e). Consumer Protection Law. https://commission.europa.eu/law/law-topic/consumer-protection-law_en

European Parliament and Council. (2002). Directive 2002/58/EC on privacy and electronic communications. <https://eur-lex.europa.eu/eli/dir/2002/58/oj/eng>

European Parliament and Council. (2018). Regulation (EU) 2018/302 on geo-blocking. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32018R0302>

Fagerjord, A. (2015). The cloud, the store, and millions of apps. *There is no Software, there are just Services*. Meson press. (pp. 91–101).

Franceschi-Bicchierai, L. (2024). Apple removes VPN apps at request of Russian authorities, say app makers. *TechCrunch*. <https://techcrunch.com/2024/07/08/apple-removes-vpn-apps-at-request-of-russian-authorities-say-app-makers/>

Gehasse, A., Belotserkovskaya, A., van der Heijden, E. & van der Valk, M. (2022) Top downloads during war: How political situations are reflected in app stores. Elective: Appification: The cultures and economies of apps, *University of Amsterdam*.

Gillespie, T. (2017). Governance of and by platforms. *SAGE handbook of social media*, 254–278.

Google. (n.d.-a). Developer Policy Center. <https://play.google/developer-content-policy/>

Google. (n.d.-b). Prepare your app for review. <https://support.google.com/googleplay/android-developer/answer/9859455?hl=en>

Google. (n.d.-c). Google Terms of Service. <https://policies.google.com/terms?hl=en>

Google. (n.d.-d). Comment Posting Policy. https://play.google/intl/en_gf/comment-posting-policy/

Google. (n.d.-e). Inappropriate Content. <https://support.google.com/googleplay/android-developer/answer/9878810>

Google. (2018). Developer Policy Center (Archived). https://web.archive.org/web/20180715090034/https://play.google.com/about/developer-content-policy/#!?modal_active=none

Google. (2022). Instagram app page Russia (Archived). Google Play. <https://web.archive.org/web/20220331083123/https://play.google.com/store/apps/details?id=com.instagram.android&hl=ru&gl=ru>

Gorwa, R. (2024). *The Politics of Platform Regulation: How Governments Shape Online Content Moderation* (p. 250). Oxford University Press.

Helmond, A., Van der Vlist, F., & Weltevrede, E. (2019) App (Store) Policy Histories. DMI Summer School 2019. <https://www.digitalmethods.net/Dmi/SummerSchool2019AppStorePolicyHistories>

Helmond, A., Nieborg, D., Van der Vlist, F. & Weltevrede, E. (2019) Store Policies and Developer Conditions. DMI Winter School 2019. <https://wiki.digitalmethods.net/Dmi/WinterSchool2019AppsStoriesPolicy>

Helmond, A., van der Vlist, F., & Weltevrede, E. (2018). DMI Summer School 2018: App Stores Bias & Objectionable Queries.

<https://digitalmethods.net/Dmi/SummerSchool2018AppStoresBiasObjectionableQueries>

Intersoft Consulting. (n.d.). General Data Protection Regulation (GDPR). <https://gdpr-info.eu/>

Kayali, L. (2022). EU to ban Russia's RT, Sputnik media outlets, von der Leyen says. *Politico*. <https://www.politico.eu/article/ursula-von-der-leyen-announces-rt-sputnik-ban/>

Light, B. (2014). Creating pseudonymous publics with Squirt: An expansion of the possibilities for networked publics. In *Proceedings of the 35th International Conference on Information Systems* (pp. 1–16). Association for Information Systems (AIS).

Mann, M., Mitchell, P., & Foth, M. (2024). Between surveillance and technological solutionism: A critique of privacy-preserving apps for COVID-19 contact-tracing. *New Media & Society*, 26(7), 4099–4117.

Marchal, N., Hoes, E., Klüser, K. J., Hamborg, F., Alizadeh, M., Kubli, M., & Katzenbach, C. (2024). How negative media coverage impacts platform governance: evidence from Facebook, Twitter, and YouTube. *Political Communication*, 1–19.

Myles, D., Duguay, S., & Dietzel, C. (2021). #DatingWhileDistancing: Dating apps as digital health technologies during the COVID-19 pandemic. In *The COVID-19 Crisis* (pp. 78–89). Routledge.

Nieborg, D. B., Young, C. J., & Joseph, D. (2020). App Imperialism: *The Political Economy of the Canadian App Store*. *Social Media + Society*, 6(2). <https://doi.org/10.1177/2056305120933293>

Paasonen, S. (2016). Pornification and the Mainstreaming of Sex. In *Oxford Research Encyclopedia of Criminology and Criminal Justice*. Oxford, UK: Oxford University Press.

Perez, S. (2022). Russia's App Store lost nearly 7K apps since its invasion of Ukraine, but some Big Tech apps remain. *TechCrunch*. <https://techcrunch.com/2022/03/15/russias-app-store-lost-nearly-7k-apps-since-its-invasion-of-ukraine-but-some-big-tech-apps-remain/>

Polese, A., & Helou, J. P. (2024). Everyday informality and governance dynamics in crisis situations and beyond. *Third World Quarterly*, 45(17–18), 2323–2333.

Popiel, P. & Vasudevan, K. (2024). Platform frictions, platform power, and the politics of platformization. *Information, Communication & Society*, 27(10), 1867–1883.

Quaglia, E., Weber, T. & Zhu, X. (2022) Russia's Social Splinternet? Exploring the Impact of the Russo-Ukrainian War on Russia's Social App Landscape. Elective: Appification: The cultures and economies of apps, University of Amsterdam.

Roman, C., Zhao, J., Zhang, Y. & Zhang, Y. (2022) How app stores respond to the war: A cross-country and issue-oriented analysis of Google Play in Poland and Ukraine. *Elective: Appification: The cultures and economies of apps*, University of Amsterdam.

Rogers, R. (2021). Marginalizing the mainstream: How social media privilege political information. *Frontiers in Big Data*, 4, 689036

RuStore. (n.d.). Official Website. <https://www.rustore.ru/>

Steen, E., Yurechko, K., & Klug, D. (2023). You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media + Society*, 9(3), 1–17.

Sheftalovich, Z. & Gijs, C. (2022). Apple joins other global giants in Russia exit. *Politico*. <https://www.politico.eu/article/apple-pulls-products-from-russia/>

Tiidenberg, K., Hendry, N. A., & Abidin, C. (2021). *Tumblr*. John Wiley & Sons.

van der Vlist, F. N., Helmond, A., Burkhardt, M., & Seitz, T. (2022). API governance: The case of Facebook's evolution. *Social Media + Society*, 8(2), 1–24.

10. Walking the Store: Cultic Networks and Content Moderation on Amazon.com

Marc Tuters

Abstract

This report provides a comprehensive account of Amazon’s approach to content moderation, examining both the platform’s evolving policy framework and the effects of its recommendation infrastructure. It introduces a novel empirical method — co-consumption analysis — which repurposes Amazon’s own logic of “walking the store” to map ideological clustering in its book marketplace. Using this method, the study reveals how controversial texts — from *Imperium* to anti-vax literature — are algorithmically surfaced alongside more mainstream works, forming what are theorized here as “cultic networks.” The report also develops a media ecological framework rooted in affordance theory to interpret these patterns, arguing that moderation is not only about removals, but about the amplification of ideological discovery through algorithmic visibility. These findings suggest the potential for real-time monitoring tools to track controversial content clusters — offering a diagnostic framework to support algorithmic oversight and regulatory enforcement.

Keywords: content moderation, co-consumption analysis, ideological clustering, cultic networks, conspiracy theory, media ecology, MAGA, platform governance

Introduction: Content Moderation as Platform Ecology

This chapter examines Amazon’s book marketplace as a key site of ideological circulation — one that, despite its scale and influence, has largely escaped sustained scrutiny in debates about platform governance. While Amazon publicly frames its content moderation practices in terms of managing “controversial content,” its actual approach is shaped not by traditional gatekeeping, but by the architecture of its recommendation system. Borrowing Amazon’s own metaphor of “walking the store,” this report introduces a novel method — co-consumption analysis — which repurposes Amazon’s “also bought” data to trace how controversial texts (from *Imperium* to anti-vax literature) cluster and persist in the platform’s recommendation networks. This method, which draws on seed lists including bestsellers and ideologically marked titles, reveals how Amazon’s interface organizes visibility in ways that encourage ideological convergence. While it cannot capture content that has been removed, the method surfaces latent constellations of recommendation — suggesting not just gaps in moderation, but patterns of amplification. Beyond its descriptive utility, this approach opens a path toward diagnostic tools: scraper-based dashboards that track bestseller lists by topic across time and place, offering regulators new ways to monitor “epistemic toxicity” and enforce compliance with transparency mandates under instruments like the EU’s Digital Services Act.



Figure 10.1 Screenshot of Amazon.com’s recommended books “also bought” along with “Imperium: The Philosophy of History and Politics”, which includes a book entitled “Harassment Architecture”. Date: November, 2024. Source: author.

While other platforms are in the business of connecting people to each other and to information, as the self-proclaimed “everything store”, Amazon is different. Even though most of Amazon’s revenues come not from e-commerce but rather from its cloud infrastructure business (Pizio, 2024), the company’s core identity is as a literal marketplace of ideas. This has meant that Amazon’s own moderation policies long declared a commitment also to include “books that some customers may find objectionable”. Indeed, early on Bezos described his vision for the company as being “to make every book available—the good, the bad and the ugly” (Bezos, 1998). This proposal, that the free exchange of ideas is essential for a healthy society, dates to John Stuart Mill’s arguments in “On Liberty” (1869), informing a central tenant in U.S. jurisprudence around the First Amendment (on citizens’ rights to free speech) as well as the belief that more speech is the remedy for “bad” speech — with book censorship being associated with authoritarian regimes which are intolerant of dissent. In the United States, an ideological commitment to freedom of expression and open debate strongly influences civil rights organisations like the American Civil Liberties Union (ACLU), to defend speech the First Amendment rights of Nazis, the Ku Klux Klan and other forms of “speech we hate” (ACLU 2024). This perspective is common amongst many libertarians, a political tradition that advocates individual freedom and limited government, and which has been exceptionally influential in the so-called “Californian ideology” underpinning Big Tech (Barbrook & Cameron, 1996).

Although, earlier on in their histories Silicon Valley platforms had imagined themselves as “post- territorial” and thus exempt from national laws (cf. Goldsmith & Wu, 2006), Amazon’s content moderation policies uphold the laws where it operates — as, for that matter, do even the most “extreme” platforms, like 4chan (Gillespie, 2018). At the time of writing Amazon operates book marketplaces in 23 countries (Australia, Belgium, Brazil, Canada, China, Egypt, France, Germany, India, Italy, Japan, Mexico, Netherlands, Poland, Saudi Arabia, Singapore, South Africa, Spain, Sweden, Turkey, United Arab Emirates (EAE), United Kingdom, United States), a least a half dozen of which could be considered to have authoritarian or semi-authoritarian governments.⁵ Authoritarian countries typically have long lists of banned

⁵ China, Egypt, Saudi Arabia, the UAE, Turkey and Singapore could all be considered authoritarian or semi-authoritarian. However, many political scientists fear that the US is also slipping in this direction. Of note here, while freedom of speech is guaranteed in the US constitution, and is highly valued by its citizens, it is not even in the top 50 nations according to the world press freedom index.

books (for example, “The Satanic Verses” in Saudi Arabia and the UAE), as do some democratic nations (such, for example, in the case of “Mein Kampf” in Germany) — books which are typically unavailable in the corresponding “national” Amazon marketplaces. Moreover, as we will see in the next subsection, Amazon also regularly removes books in response to pressure campaigns and exposés by journalists, politicians, petitions, watchdog organisations and identity or issue-based lobbies. Until quite recently, however, the company has generally been rather opaque about how content moderation works on the platform — even in the face of increased regulation. Until quite recently, however, the company has generally been rather opaque about how content moderation works on the platform — even in the face of increasing regulatory attention, such as the European Union’s Digital Services Act (DSA) (cf. Gorwa, 2019; de Gregorio, 2021). The DSA represents the most ambitious attempt yet to systematise platform accountability across member states, requiring “very large online platforms” like Amazon to publish detailed transparency reports and enable independent audits of their algorithmic systems. While it promises to constrain the discretionary power of platforms, Amazon’s limited compliance — especially regarding recommendation systems and book removals — raises questions about the DSA’s practical enforceability.

This report discusses how Amazon’s moderation policies significantly changed after the Trump-led attempted coup d’état in 2021, in concert with the removal of many right-wing extremist and conspiracy theory literature. Although Amazon claims to have vastly increased its content moderation in the last few years, the method however surfaces a surfeit of such controversial content — including a book referred to as “the America’s Mein Kampf” (Mostrom 2020), Francis Parker Yockey’s “Imperium”, which is sold without any content a warning, unlike the actual “Mein Kampf.” Beyond singular controversial titles, the method helps us to understand how Amazon’s algorithmic recommendations effectively renders entire fields of study controversial, for example categorizing books labelled as “Communication & Media Studies” (my own academic discipline) next to highly controversial works of conspiracy theory — shelving field classics like Jacques Ellul’s “Propaganda: The Formation of Men’s Attitude” next to books with titles like “Propaganda Wars: How the Global Elite Control What You See, Think, and Feel”.

The timing of the research underpinning report is fortuitous, conducted as it was on Amazon’s .com US flagship store around the time of Donald Trump’s victory in the 2024 US presidential election. Trump’s previous 2016 victory is widely acknowledged as having inaugurated the rise of misinformation studies as a major new subfield of media studies (cf. Rogers, 2023).⁶ A dominant thesis in this emerging field of misinformation research ties MAGA’s rise to the emergence of new media ecosystems centred on “media of one” influencers (DiResta, 2024) — whose audience reach now

⁶ While the timing of this research may appear to introduce noise, moments of controversy are often the most analytically productive for studying algorithmic systems. As Venturini and Munk (2021) argue, controversy reveals underlying infrastructures and dynamics that remain hidden during periods of stability. In this light, Trump’s re-election created conditions that made Amazon’s ideological sorting mechanisms more legible, not less — offering a rare opportunity to observe the platform’s affordances under pressure.

often seems to exceed that of 20th century “legacy” media organisations.⁷ Misinformation researcher, Renée DiResta (2024) situates the rise of this new class of media of one influencers alongside the role of social media algorithms and the power of networked audiences as the “trinity” of factors that drive the spread of misinformation online. Co-consumption networks arguably offer a means by which to visualize the entwining of this trinity of digital. Seen through this lens, MAGA influencers’ books appear as both entry points into and compilation of networks of extreme ideas (as represented by other books) with which the platform associates them, based on its customers’ purchasing habits.

The deeper significance of these clusters emerges when we consider the platform as a structuring environment for political discourse. Political theorist Alan Finlayson (2021) describes the rise of “reactionary digital politics” — a phenomenon in which new media environments reduce barriers to entry for ideologically extreme actors, exemplified by figures like Donald Trump. Central to this analysis is the idea of affordances: the possibilities for action enabled or constrained by a technological system. In media studies, affordances are not treated as neutral or fixed, but as relational and mutable, especially in algorithmic contexts (Bucher & Helmond, 2017). Finlayson draws attention to how these affordances favor antagonistic, viral, and emotionally charged forms of expression. This framing aligns with the report’s findings, where MAGA- and MAHA-aligned texts (such as those associated with Robert F. Kennedy Jr.) appear not only as content but as connective tissue within larger ideological assemblages. While communications theorists have long emphasized the role of opinion leaders (Katz & Lazarsfeld, 1964), Finlayson argues that the breakdown of coherent ideological traditions enables new, recombinatory political formations — many of them reactionary, adaptive, and algorithmically facilitated. The empirical analysis offered here can be read as both a demonstration and a diagnosis of that very process.

Having long since driven its competitors to the margins or into bankruptcy as part of its’ “ruthless quest to own the world and remake corporate power” (Mattioli, 2024), Amazon is unrivalled as the source of books and the ideas that they contain. With their politics appearing to dominate large swathes of the platform — not to mention other platforms such as X — this new class of reactionary political influencers appear to grasp this new discursive situation particularly well. Following DiResta and Finlayson, the report thus proposes to offer insights into the rise in prominence of various “cultic” ideas — from neo-Nazis to anti vaxxers — in Amazon.com’s broader marketplace of ideas. During the previous Trump administration, many commentators observed how obscure political subcultures like the so-called alt-right, often used an “absolutist” defence of free speech to defend hateful speech. With Trump’s re-election these tendencies have moved dramatically into the mainstream (McLeary et al, 2025), and much of Silicon Valley appears to be moving in synch (Kang, 2025). Indeed, Trump’s victory has prompted a platform like Meta to end third-party fact-checking programs in favour of “more speech” (Kaplan, 2024), which may augur the return to a less restrained and more libertarian-style approach to content moderation. For his part,

⁷ As an anecdotal measure of the declining influence of “legacy” news organisations in the current US political climate, consider the fact that Trump’s interview with a popular podcaster reached five times as many viewers as Harris’ network TV interview (cf. Grynbaum, 2024).

Bezos has also expressed his “very optimistic” support for Trump’s plan to dismantle regulation (Gedeon, 2024). This may suggest that the tendency towards increased content moderation on Amazon and other platforms following the excesses of the first Trump administration (as discussed in the next subsection), may be more contingent than we thought, and that going forward the controversies discussed in the report’s empirical analysis may in fact be closer to the norm.

The report has five main sections. In addition to the introduction and conclusion (both of which frame the report as contributing to current debates in media studies), it includes a substantial historical review of changes in Amazon’s approach to content moderation, based both on a (primary) analysis of their policies as pertaining to the sale of books and through a (secondary) review of reporting on the subject, extending back over a decade and a half. Following this, the report presents a brief explanation of the theoretical framework that informs the empirical method of co-consumption analysis. This is then followed by an in-depth analysis applying the method to map controversial book recommendation networks, in the context of the Trump re-election, the accompanying rightward shift in Big Tech and its growing influence, if not to say its’ current “takeover” of American politics (Kang, 2025) — and, due to their immense global influence, their broader takeover of the entire public sphere. (Those interested in the original findings may wish to skip ahead either to the “Theoretical Framework” section or straight to the “Empirical Analysis” section.)

The Evolution of Amazon’s Content Moderation Policies

In recent years, there have been calls from across the political spectrum to limit the power of Big Tech (Kingwell 2022). On the political right, in the United States, politicians have decried ‘the tyranny of Big Tech’ (Hawley 2021) for having removed content and accounts including those of Donald Trump, and relegating them to a “digital gulag” of alt-tech platforms (Weingarten, 2023: 12). On the left, there has been pressure to use American antitrust legislation to dismantle Big Tech’s monopoly power (Warren 2024). In the face of such pressures, particularly under the Biden administration, in recent years Amazon — long notorious for its monopolistic business practices (Khan, 2017) — has occasionally adopted a more conciliatory stance to government regulators. Amazon’s increased efforts to “self-regulate” (Cusumano et al, 2021) can be seen reflected in their updated moderation policies, where they claim to have doubled the number of controversial products that they removed between 2021 and 2023.

While early on in their history, many Silicon Valley companies, Amazon included, often outwardly expressed explicit commitments to libertarian ideals of relatively unrestrained free speech (York, 2022), this paradigm shifted over the course of the first Trump administration, especially towards its end — as seen most clearly in platforms’ efforts to combat health-related misinformation during the pandemic (Baker et al, 2020; de Keulenaar et al, 2023). However, with Trumps re-election, his appeal to libertarians (Gold and O’Brien, 2024), his executive order “restoring free speech and ending government censorship” (Whitehouse, 2025), and his increasingly cosy relationship with Big Tech CEOs (Kang, 2025), aggressive American government regulation of Big Tech seems increasingly less likely. In this context, government regulations — such as the European Digital Services Act (DSA) that require “very

large online platforms” (VLOP) like Amazon to be accountable and transparent in revealing how their algorithms work and allowing for independent audits, and which went into effect in 2022, the DSA — now takes on increased significance. For many then, the current question simply put is: “will the DSA fit it?” (Dwivedi, 2022)

Part of the DSA’s regulations requires VLOP to report on content moderation via a “Transparency Database”. In researching Amazon’s removal of controversial books, this report found this resource to be of no practical value, as Amazon only ever cited “privacy violation” as the reason for book removal, when in fact they often remove books for reasons such as “Illegal of harmful speech” as we will see in the next subsection. In the absence of any useful data accounting for their content moderation we are left only with secondary resources in journalistic reporting on the subject and with the Internet Archive’s “Wayback Machine”, which crawls and periodically creates “snapshots” of webpages. We thus begin first by using this tool to scrutinize significant changes in the wording of Amazon.com’s content moderation policies over time, as well as to retrieve an example of book now banned from the site (William Luther Pierce’s “The Turner Diaries”), before going on to offer a history of such book removals.

From Opacity to Enforcement: Tracking Amazon’s Policy Language

Although, at the time of writing Amazon’s content moderation policies had not been collected and archived by a dedicated third-party initiative (such as the “Platform Governance Archive” project), we found archived snapshots of Amazon.com’s content moderation policies via the Wayback Machine, which we then compared over time. Amazon spokespeople sometimes refer to different documents when asked by journalists to justify the removal of items, which can create some ambiguity as to the ultimate reasons for removal. On one page outlining their content moderation policies, entitled “Offensive and Controversial Materials”, Amazon describes the products that they prohibit as those that “promote, incite, or glorify hatred, violence, racial, sexual, or religious intolerance or promote organizations with such views, as well as listings that graphically portray violence or victims of violence” and which have “no historical significance”. In the “fine print” of this policy, Amazon reserves the right to themselves make the distinction as to what constitutes “historical significance”. On another page, entitled “Content Guidelines for Books”, they use a different type of language, reserving the right to remove content from sale that is “typically disappointing” or provides a “poor customer experience”. We were able to study these documents dating back to 2021 and 2019 respectively.

One of their policies thus frames books in terms of knowledge and “significance”, while the other frames them in terms of entertainment and “experience”. As already introduced earlier, Amazon’s self-conception has long been to see itself as a marketplace of ideas, which comes along with a corresponding social responsibility. This self-conception can be seen reflected in the language of Amazon’s “Content Guidelines for Books” which Amazon initially states: “As a bookseller, we provide our customers with access to a variety of viewpoints, including books that some customers may find objectionable,” — a statement that echoes Bezos’ own earlier vision of Amazon’s libertarian responsibility (Bezos, 1998). However, in the early years of the Biden administration, this formulation underwent a significant change from providing “access to a variety of viewpoints” to providing “access to the written word”.

Amazon's business model has always been based on undercutting the competition through, offering lower prices often through redirecting profits from its other business, absorbing the losses and untimely bankrupting most of its competitors (Mattioli, 2024). Ironically, considering the antitrust litigation that the Biden administration would soon begin against Amazon (Federal Trade Commission, 2023), this change in the wording of their content moderation policy could be understood as representing an astonishing admission of the company's essential monopoly status as the source of "access to the written word".

The timing of this change to Amazon's content moderation policy (03/03/2021) significantly came mere months after the US Capitol riots, which marked a paradigm shift in Big Tech content moderation policies, leading for example to the "deplatforming" of Trump from Twitter and Facebook (Innes and Innes, 2023). At this same time, Amazon also states (again for the first time), that "If we remove a title, we let the author, publisher, or selling partner know and they can appeal our decision." This, we can understand, as likely arising from the pushback on the political right, against Big Tech's tyrannical power to censor objectionable content, which was felt to disproportionately target the right (Hawley 2021). Indeed, beyond marking a significant turning point in Amazon's content moderation policies, following the Capitol riots, the platform removed controversial content that they had long tolerated such for example as William Luther Pierce's "The Turner Diaries", which had previously been sold with a warning, in the form of an "Editorial Review" from Amazon, stating:

The Turner Diaries is a racist, white supremacist fantasy about annihilating half the planet in a series of nuclear explosions, then killing all non-white people and taking over the world. It became infamous as a source of inspiration for the man who perpetrated the deadliest act of domestic terrorism in United States history: the 1995 Oklahoma City bombing. As a bookseller, we think it is important to offer this infamous work because of its historical significance and educational role in the understanding and prevention of racism and acts of terrorism.

Alongside The Turner Diaries, within a week of the Capitol riot, Amazon also removed swathes of books promoting Holocaust denial and the QAnon movement (State, 2021; Pasternack 2021). These changes came several years after other VLOPs, like YouTube, had begun to take a more concerted and outwardly approach to similar content (OILab, 2019). As we will see in the empirical analysis section of this report, Amazon still sells books that could be considered to warrant such a warning —such for example as Francis Parker Yockey's "Imperium" — though without providing a content warning referring to "historical significance" or providing "a variety of viewpoints".

Prior to this consequential change in Amazon's moderation policies language, though still after the Capitol riots, their other significant alteration to their policy language, concerned the definition of "Offensive and Controversial Materials". It is at this point that they begin to provide more transparency into how Amazon implements content moderation, with language referring to "proactive mechanisms" to "catch" and remove any such cases of offensive content "before a customer ever sees them" as well as

disclosing the existence of an “Offensive Products team” responsible both for monitoring content and developing content moderation policy, which involves consulting expert “resources issued by civil rights and anti-hate organizations as guidelines” (01/26/2021). As already mentioned in the introduction, this is the process that it colloquially refers to as “manually ‘walking the store’ to ensure compliance”.

Although, following the Capitol riot, Amazon’s outward approach to content moderation may have undergone something of a sea change, there were countless numbers of books removed from the platform prior to this moment as well. We know this not because of Amazon’s own disclosures, for example via the DSA’ Transparency Database, but rather due to the lobbying of various pressure groups and reporting of journalists, which were accessed through the LexisNexis archive. When surveying the history of such reporting a story emerges emblematic of broader tensions in digital platforms: balancing free expression with ethical responsibility, responding to political and public pressures, and navigating novel challenges posed by technology. Over time, we see the platform’s practices seeming to evolve toward stricter oversight, but inconsistencies and controversies remain pervasive, and the American current regulatory climate may also put this narrative into question.

A Timeline of Controversy: Book Removals and Public Pressure (2009–2024)

2009-2013: from algorithmic content moderation to free speech

Though Amazon has always faced content moderation challenges, we can start this story 15 years prior when, in 2009 Amazon faced a major backlash when it removed a large number of books with gay and lesbian themes from its best-seller sales rankings — including such titles as “Lady Chatterley’s Lover” by D.H. Laurence, “The City and the Pillar” by Gore Vidal, “Giovanni’s Room” by James Baldwin, “The History of Sexuality, Volume 1” by Michel Foucault, and “Brokeback Mountain” by E. Annie Proulx — subsequently claiming that the removal of had been due to an automated “cataloguing error.” In that same year a controversy arose around a case in which users searching the marketplace with the query “abortion” were oddly prompted with the question “Did you mean adoption?” along with books on the subject, which they then amended. Both of these examples seem to demonstrate the platform’s long-standing reliance on automated content moderation techniques.

In 2010, Amazon faced pressure from the UK court system and the police to remove “The Anarchist’s Cookbook”, a book which had been used by a convicted terrorist. First published in 1971, The Anarchist Cookbook is a recipe-filled manifesto, that contains instructions for creating homemade weapons, explosives, and other tools that is believed to have been used in numerous horrific acts of violence including the 1999 at Columbine Massacre, following which the books original author renounced the book, posting a statement to that effect on Amazon (Sandomir, 2017). In response to the pressure, a company representative stated: “Amazon believes it is censorship to make a book unavailable to our customers because we believe its message to be repugnant” and that the company’s “goal is to support freedom of expression and to provide customers with the broadest selection possible so they can find, discover, and buy any title they might be seeking.”

In 2013, although they had long had a policy of not selling pornography, Amazon was found to be selling erotic fiction with abuse-themed titles like “Taking My Drunk Daughter”, “Virgin Raped by an Intruder” and “Raped by Daddy” which they were reported to have then removed. There was significant pushback on this decision, including a petition on Change.org, that amassed over 11,000 signatures, claiming that Amazon was infringing on freedom of speech. In that same year, Amazon was briefly embroiled in a controversy for selling merchandise with offensive slogans seemingly generated by algorithms such as “Keep Calm and Hit Her” and “Keep Calm and Rape A Lot”. The latter example would subsequently inspire the media theorist James Bridle to reflect on the disturbing consequences of algorithmic content creation in a much-cited opinion piece entitled “something is wrong on the internet” (Bridle, 2017).

2015-2020: from free speech to culture war

In 2015 the far-right political strategist Roger Stone claimed that his book “The Clintons’ War on Women” had been targeted with negative reviews by trolls associated with Hillary Clinton’s campaign. Amazon removed all the offending reviews. Two years later, Clinton would make similar claims about her own book “What Happened”, leading Amazon to delete more than 900 reviews made by “non-verified purchasers”. While these are high profile examples, Amazon has a long track record of removing reviews, without offering an explanation to the reviewers. More recently, at the end of 2021, Amazon was accused of kowtowing to the interest of the Chinese government by prohibiting reviews critical of speeches by Xi Jinping, widely seen as an effort to gain greater market share in that country.

In 2016, Amazon faced pressure from the federal Minister of Public Safety in Canada to remove a book reportedly written by the Canadian serial killer Robert Pickton, entitled “Pickton: In His Own Words”. A similar issue again arose several years later when, in 2020 a petition circulated demanding Amazon remove for sale books by the author and leader of a Mormon “doomsday cult” Chad Daybell, after the discovery of the remains of his missing stepchildren, which he later found guilty of having murdered. Some jurisdictions have legal frameworks that restrict criminals from profiting from their crimes through media or literary works, such as the so-called “Son of Sam” laws in the United States. Having removed the Pickton book but not Daybell books, Amazon has been criticized for being inconsistent around this policy.

Policing the availability of antisemitic and neo-Nazi literature has been a consistent issue on Amazon over the years. In 2017, pressure was exerted on Amazon to remove a book of Holocaust denial entitled, “Did Six Million Really Die? The Truth at Last,” by Richard Harwood. Though they did not do so then, it is currently no longer on Amazon. In 2018, Amazon removed a host of neo-Nazi books and other content in response to external pressure from advocacy groups like the Partnership for Working Families, the Action Centre on Race and the Economy and a U.S. Democratic congressman. At the beginning of 2020, Amazon refused to remove “The Protocols of the Elders of Zion”, considered to be the single most notorious works of antisemitic propaganda, from its Australian book marketplace, though it currently appears no longer to be available. In 2020, a Nazi-era children’s book entitled “The Poisonous Mushroom” by Phillip “Fips” Rupprecht, designed to teach anti-Semitic stereotypes, was removed from Amazon.com under pressure by The World Jewish Congress.

Indeed, for many years the Simon Wiesenthal Center (SWC) has sought to exert pressure on Amazon to remove Nazi and neo-Nazi content from their platform. Although they were successful in getting the platform to remove much of this content, one text that SWC flagged for removal, "Die Rothschilds: Eine Familie beherrscht die Welt" (2017), remains on Amazon's German marketplace at the time of writing with over 500 mostly very positive reviews. In the summer of 2022, researchers affiliated with the Southern Poverty Law Centre identified 24 titles sold on Amazon by the white nationalist publisher Antelope Hill, including many translated works from 20th-century Nazis, fascists and ultra-nationalists with titles like "In His Own Words: The Essential Speeches of Adolf Hitler" and "A New Nobility of Blood and Soil" (Hayden & Gais 2022). Also in 2019, following an exposé by an Israeli journalist, Amazon removed a book entitled "Hizbullah: The Story from Within" by Naim Qassem, which was found to contain antisemitic and anti-Israel statements and supported violence against Israeli civilians.

2020-2024: from culture war to conspiracy theory

While extremist literature is a serious problem for Amazon, a larger problem in terms of its volume is that of pseudoscience literature, both because it does not violate the platforms' moderation policies as clearly and because it can be extremely popular — as we again will see in the subsequent network analysis section. In 2019, Amazon is reported to have removed books promoting pseudoscientific methods for autism cures and vaccine misinformation with titles like "Healing the Symptoms Known as Autism" by Kerry Rivera, "Fight Autism and Win" and "The Miracle Mineral Supplement of the 21st Century" by Jim Humble. The removal of these books followed an exposé in Wired magazine that highlighted their potentially dangerous therapies. (Zadrozny 2019). The removals came at the same time as a decision by Facebook to hide groups that spread misinformation about vaccines causing autism and may also have been influenced by the Center for Disease Control and Prevention's stance on the issue. In 2021, researchers also performed an audit of vaccine-related search-queries on Amazon, finding one-in-ten to be "misinformative", their results would seem to resonate with the findings later discussed in this article (Junja & Mitra, 2021). In the spring of 2024, Amazon removed various books promoting misinformation about cancer cures, including specific titles such as "Curing Cancer with Carrots" and "Proof for the Cancer-Fungus Connection."

With the beginning of the pandemic, in March of 2020, Amazon was flooded with self-published books about the coronavirus — titles like "Coronavirus" by Corbi Yang, "Wuhan 2020 Coronavirus Outbreak," and "Wuhan Coronavirus" by Tracy Rinehart — many of which were plagiarised directly from online sources, and which Amazon began removing shortly thereafter. During the height of the pandemic, in the summer of 2020, "Unreported Truths about COVID-19 and Lockdowns: Part 1: Introduction and Death Counts and Estimates" by Alex Berenson was blocked from being self-published on Amazon. However, after the author's complaint on social media and Elon Musk's intervention, Amazon reinstated the book, stating the initial block was an error. In the fall of 2021, Senator Elizabeth Warren sent a letter to Amazon, suggesting that the sale of "The Truth About COVID-19: Exposing the Great Reset, Lockdowns, Vaccine Passports, and the New Normal" by Dr. Joseph Mercola could be "potentially unlawful", however the book was not removed from Amazon's marketplace, and as we will see it remains extremely popular.

In the spring of 2021, a report by the renowned Institute for Strategic Dialogue claimed that recommendation algorithms were “cross-pollinating conspiracy theories” with users searching for information on vaccines being introduced to New World Order or QAnon conspiracy theories. Similar findings were those made by the several projects conducted by the Digital Methods Initiative earlier that year (cf. Gray et al 2021; DMI 2021a; DMI 2021b). Later that year, following the Capitol riots, the platform removed QAnon literature, including one book entitled “QAnon: An Invitation to the Great Awakening” which journalists and researchers had already observed years earlier having repeatedly made the top of various Amazon bestseller charts, apparent due to the concerted efforts to hack Amazon’s ranking (Tiffany, 2019; Collins, 2019; Tuters, 2020).

In November 2020, the book “Irreversible Damage: The Transgender Craze Seducing Our Daughters” published by Regnery Publishing, sparked controversy after it was removed by another bookseller on the basis that it was considered anti-trans for endorsing the controversial concept of rapid-onset gender dysphoria (ROGD), despite the lack of evidence supporting such a diagnosis. It was not removed, and instead went on to be a bestseller. Early in 2021, following the change in its content guidelines outlined above, Amazon removed a similarly themed book: “When Harry Became Sally: Responding to the Transgender Moment” by Ryan T. Anderson. As its reason for removing the book, stated that it would not sell books that they had now “chosen not to sell books that frame LGBTQ+ identity as a mental illness” (Rushe, 2021). Although this specific language cannot be found in Amazon’s policy documents, it may also account for their removal of “A Parent’s Guide to Preventing Homosexuality” by Joseph Nicolosi, referred to as the father of gay conversion therapy.

Amazon’s removal of “When Harry Became Sally” would go on to become a cause célèbre in the American political right MAGA movement, with its author — affiliated with the Witherspoon Institute conservative think tank — accusing Amazon of engaging in “digital book burning”, which was in turn picked-up by four Republican senators - Marco Rubio, Mike Lee, Mike Braun, and Josh Hawley - who wrote a letter to Amazon CEO Jeff Bezos, accusing the company of “political censorship” and silencing conservative voices. Trump would go on to make this culture war issue a pillar of this successful 2020 campaign, spending nearly \$100M on anti-trans television advertisements (PBS, 2024). At the beginning of 2023, the Republican chairman of the House Judiciary Committee, Jim Jordan, subpoenaed Amazon’s CEO (now no longer Bezos) among with those CEOs of Meta, Alphabet, Microsoft and Apple to answer questions on the corporate censorship of conservative voices and “understand how and to what extent the [Biden administration] coerced and colluded with companies and their intermediaries to censor speech.” Jordan’s committee was critiqued as a “conspiratorial quest for power” in line with Trump’s agenda to spread misinformation and undermine trust in government and media (Blitzer, 2023) — needless to say, this misinformation strategy worked as Trump would soon go on to be re-elected and turn these critiques into official government policy.⁸

⁸ Another category of books removed by Amazon in the last period have included books promoting anorexia nervosa with titles like “All Things Thin and Beautiful” and “Beauty Is Slim and Lean: Living Pro-Ana the Healthy Way” in 2019. Different from

Most recently, there has been a spate of controversies around books containing AI-generated misinformation, such as for example a collection of biographies falsely speculating on King Charles' health, which were removed following pressure from Buckingham Palace. Going forward, with the second Trump presidency's battle against censorship (Whitehouse, 2025), it is hard to know to what extent Amazon will continue to feel pressure to self-regulate or opt for a permissive libertarian-inspired approach, perhaps switching in the process from a top-down towards a more community defined moderation approach, as has been the case with X and Meta (cf. Kaplan, 2024).

In this section we have seen how Amazon has tried to respond to public pressure by often, though not consistently, removing controversial books from neo-Nazi to conspiracy theory literature. In the upcoming empirical analysis section, the objective will be to "audit" Amazon's automated book recommendations on these topics, to assess the current state of controversy on the platform at the time of the research. Before doing so however, we first need to draw from several somewhat disparate theoretical traditions to construct a framework both to guide and to make sense of the empirical analysis.

Theoretical Framework: Mapping cultic networks

Some important past works in the field of media theory have had their roots in commissioned research reports, underscoring the role of theory building in making sense of the deep societal changes that typically accompany the introduction of new communication paradigms.⁹ For Marshall McLuhan, these changes often went relatively unnoticed due to what he considered as new media's power of to lull users (and presumably regulators) into a hypnotic state of technological somnambulism (1964). Famously, McLuhan argued that was due to a tendency to focus on content at the expense of focussing form — as well as its dialectical relationship with content. From this perspective, understanding media involves rendering visible the structures of communication that new technologies make possible. This approach arrives at questions of meaning via descriptions of how communication technologies work and not the other way around, where what matters are not the messages or the content but rather "their circuits, the very schematism of perceptibility" (Kittler, 1999, p.xl). While

pseudoscience, these books can be understood as a toxic hybrid of the genres of Self-Help, Tween Girl Lit.

⁹ Although it is so well known, Marshall McLuhan's book "Understating Media" (1964), has its roots in a 1960 report entitled "Understanding New Media" project (1960), written as a report for the National Association of Educational Broadcasters. Similarly, Jean Francois Lyotard's "The Postmodern Condition" (1984), was written at the request of the Council of Universities of the Provincial Government of Quebec on the state of knowledge in the contemporary world. Both books can be understood as field-defining. In the case of McLuhan's defining the field of New Media Studies that would emerge thirty-plus years later, and in the case of Lyotard defining the problematic of "the postmodern", which was uniting concern across the proliferating subdisciplines of humanities for decades. Moreover, both books essentially concern the impact of new media on human self-understanding.

this perspective can be critiqued as being somewhat deterministic, what is valuable for our purposes is how it frames epistemology as being tied to the conditions of possibility of a given communication environment. In the case of the current report, this perspective prompts the question of the role of Amazon's book marketplace in shaping how we might come to know about a given subject. If we treat Amazon's book marketplace as a system for indexing and structuring knowledge, in the manner of a library, then what sort of picture emerges from this system, algorithmically optimised as it is to maximize sales and to retain attention?

To answer this question, the theory underpinning our method begins by returning to a classic work by the sociologist of innovation Michel Callon, which proposed a novel method for identifying how new ideas, or pockets of innovation, emerge within a large network of texts. As a contribution to the little-known field of scientometrics, Callon and his colleagues developed a method that allowed them to identify “strategic themes that were likely to develop in the future”, based on an analysis of the relative “proximity” of concepts within a large textual corpus — an approach that platforms, most notably Google, would later adopt as the basis for algorithmically sorting knowledge (Mallard & Callon, 2022, p.156-7). Setting aside for the moment the specifics of Callon’s methodology,¹⁰ what is key for our current purposes is Callon’s conceptualization of a text as a “complex device acting on [a] field in order to transform or consolidate it” (Callon et al, 1983, p.204). This perspective is key in terms of framing the experiments in compiling and visualizing Amazon book recommendation networks discussed in the next section in terms what Callon refers to as a “field of forces” whose “structures evolve constantly because they depend on continually renewed initiatives and are scarcely visible because no one actor is capable of developing a sufficiently general point of view on them” (Idem, 194).

But while scientometrics (and bibliometrics more generally) compile and analyse networks of texts based on their contents, the method developed in this report uses metadata generated not by the texts’ authors but rather by their readers, or more precisely their purchasers. The method deployed in the next section to discover and visualize networks of books, what we call co-consumption analysis, thus repurposes a form of metadata that Amazon collects from users. Like other platforms, Amazon derives much of its capital from capturing and valorising users’ actions. Through this lens, we can think of the metadata being analysed by this method as artefacts of “semiotic labor” generated by the users of the platform, which it feeds back to them to encourage them to stay on the site and spend money (cf. Langlois, 2014). Platforms can be understood as an interface, with the extraction of users’ semiotic labor occurring on one side, and its aggregation into “digital subjects” on the other side (Wark, 2019). The idea here is less about subjects’ epistemologies being increasingly digitally mediated than their very ontology becoming formed in “nested sets of abstractions assembled by algorithms” (Gorinova, 2018, p.1). Seen through this media

¹⁰ Practically-speaking, Callon proposed a technique for detecting these dynamics, based on an analysis of “co-word” relationships within documents, which today can be easily accomplished via techniques such as n-grams as well as other digital methods (cf. Marres & Gerlitz, 2015). While adopting such an approach would constitute a “next step” for this current research, because of the unbelievable size of the entire corpus that makes up Amazon’s book marketplace, the first step is to look for pockets of innovation in the bibliometric relationships between rather than within texts.

theoretical lens, the networks presented below may perhaps be interpreted as representations of a kind of emerging digital subjectivity, organized around particular subjects of interest. Of crucial significance here is how metadata becomes the measure of social relations' value, used to improve the platforms' design, to nudge user's behaviour, in the process generating a new "machinic" kind of type of value that goes "beyond the anthropocentric definition of value grounded exclusively in human labour and human time" via a process of "amplification by connection" (Markelj & Celis, 2023, p.1070, 1072). This last point brings us back to the observations with which we began the report, by Renée Di Resta and Allan Finlayson, that the affordances of digital platforms constitute whole new discursive situations (based on the entwining of influencers, algorithms, and crowds), which seem to favour a particular political orientation, which we propose to refer to as "cultic".

As with the previous sub-section, in the next section the reader will encounter books with strange sounding titles like "PsyWar: Enforcing the New World Order". While many of these books could be dismissed as works of "conspiracy theory", since that term often implies a false form of knowledge, to somewhat sidestep normative designations, the report frames them instead as "cultic". Here, we draw from Donald Norman's classic sociological concept of "the cultic milieu", used to refer to a "cultural underground" of "deviant" ideas from outside of the mainstream of society (Campbell, 1972). Norman's concept was initially formulated to discuss the "spiritual marketplace" of gurus and cults that sprung-up ground the late-60's New Age movement, to help account for how ideas persisted in a highly ephemeral context. Campbell argued that, while cults may come and go, a broader milieu was "kept alive by the magazines, periodicals, books, pamphlets, lectures" and the like, which also served to cross-pollinate ideas and bring about a general tendency toward "syncretization". Moreover, although different cults were not necessarily in direct dialogue (let alone ideological agreement), Campbell saw them as united by a shared desire to seek deeper truths, that they saw as being hidden or suppressed by mainstream society. This cultic milieu concept has been taken-up, quite centrally, in the fields of conspiracy theory studies (e.g. Barkun, 2003), esotericism studies (e.g. Hanegraff, 2013) and comparative fascist studies (e.g. Griffin, 2007), where it is used to characterise non-hierarchical, polycentric movements with fluid boundaries and constantly changing components (Griffin, 2003). Given the widespread uptake of many formerly unorthodox ideas into the mainstream, Campbell today frames the cultic milieu as "the home of the eccentric" (Campbell, 2024, p.81).

Empirical Analysis: Mapping Ideological Clustering via Co-Consumption

In this section of the report, we both outline the empirical method and, in the process, also present findings. The approach is iterative and the cases range quite broadly from neo-nazi literature to pseudoscience and beyond. Across these cases we employ an alternative bibliographic method (or what might be called an "altmetric"), which treats books on Amazon as networked objects, defined by their relations within a network, based on metadata (apparently) derived from the past purchase history of Amazon's customers. Beyond surfacing relevant content that might be harder to discover through using more conventional approaches (such as keyword searches) the method also offers a means to identify emerging themes within particular subject areas that might

otherwise escape detection. While the method could in principle be applied to any number of subjects, for our purposes we use it to discover controversial content and to understand the semantic context that the platform gives that content. Borrowing Amazon's content moderation language, we frame this as a digital method for "walking to the store" and assessing the state of "controversy" in its aisles. Although we cannot use this method to research books which have been removed from Amazon, we find plenty of controversy remaining and evolving on the platform.

Bibliometrics conventionally builds networks from the reference list in scientific publications, where each node represents an article and each edge a citation, which is considered as a vote of authority — the more votes, the more significant the text (Small, 1973). While resulting in the same type of node/edge network visualisation, our method fundamentally differs in how it calculates the relationships between the nodes. As opposed to endogenous links from within one text to another (whether citations of co-words), our method is based around an analysis of exogenous patterns of consumer behaviour. Specifically, the networks that we produce are thus based on repurposing connections between books that Amazon claims are based on customers' purchasing patterns. This approach builds on early experiments in proxy network analysis using Amazon data, notably by Valdis Krebs (2003), who manually mapped ideological divisions in political book purchases during the early 2000s. While Krebs' method relied on manually collected co-purchase data to reveal partisan clustering, the approach here adapts and automates that principle using scraper-based co-consumption networks across broader thematic domains. Such repurposing of the metadata apparatuses of social media platforms as the basis of cultural analysis is the premise of 'digital methods' (Rogers, 2013). Treating Amazon books as networked objects, this approach considers them as floating signs whose "meaning" is derived from their position in a broader (semantic) network (cf. Tuters & Willaert, 202). Although we know little of Amazon's "Offensive Products team" works, it seems fair to imagine that they might employ a similar method in walking the store.

Normally located just below the publisher's description of a book, Amazon often offers readers lists of other books that might also pique their interest, presented as a scrolling carousel, supposedly based on what other customers "also viewed", what they "also bought", on content "related to this topic", and so on (see Figure 10.1). Although Amazon does not always implement these design affordances consistently either across its different marketplaces or even from one book to another, when it does so then each book typically has about 30 "also bought" recommendations. Where possible our method collects these recommendations, which we interpret as a particularly authoritative type of out-link — based on the saying that you "put your money where your mouth is".

As the basis for creating these networks, our method begins with list building. These lists can be collected in different ways. One way, which we will return to below, is to repurpose lists (such as bestseller lists) provided by the platform. Our first set of lists begin however, by compiling books recommendations from online subcultural resources devoted to the discussion of radical political ideologies, for example the notorious 4chan. Discussions of radical political ideology are increasingly common in contemporary internet subculture, as exemplified by the enormous popularity of political compass memes (Tuters and Mueller, 2024). A troubling trend, which received much attention in Trump's first term, was the so-called alt-right, and a rise in

on the original list. Imperium presents the Holocaust as a hoax, is dedicated to "the hero of the Second World War", meant to imply Adolf Hitler, and has been described as America's Mein Kampf (Mostrom, 2020). Unlike "Mein Kampf" or "Turner Diaries", Amazon does not provide a warning about this book. Besides this finding, the method surfaced other "primary" works of Nazi ideology — including new translations by the white supremacist publisher Antelope Hill (discussed in an earlier section of the report) and various conspiracy theories explaining how and why Jews control the world — like "1666 Redemption Through Sin: Global Conspiracy in History, Religion, Politics and Finance" by Robert Sepehr. The method also surfaced books that seemed to have nothing to do with these issues on the surface — such as a work by media criticism by conservative polemicist Ron Utz entitled "Encountering American Pravda: Essays in a Historical Counter-Narrative" — but which the algorithm oddly categorised as related to a book entitled "Jews, Nazis, and Israel" (see Figure 10.3).

Frequently bought together

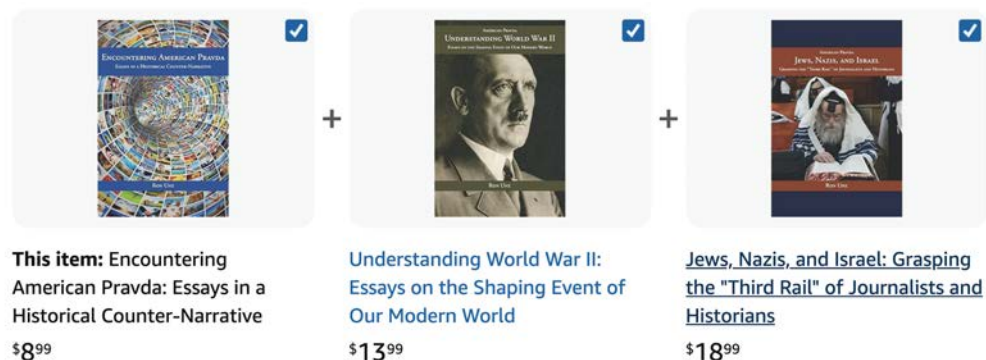


Figure 10.3 Screenshot of books suggested by Amazon.com to be purchased in a bundle along with "Encountering American Pravda" including a book entitled "Jews, Nazis, and Israel". Date: November, 2024. Source: author.

Another approach for this method begins not with outside expert lists but rather by repurposing how Amazon ranks content itself. This can be done simply by using Amazon's internal search feature or by working categories such as "Best Sellers in..." and "Gift Ideas in..." which draws from algorithmic sales rankings for the entire bookstore. As we saw earlier in this report, Amazon previously faced widespread criticism for its search algorithm having provided books on adoption when users searched for abortion and for having removed LGBTQI+ titles from their sales rankings. We begin here by noting that an Amazon search for "The Turner Diaries" — which, as discussed earlier in this report, was removed in 2021 — returns the "Industrial Society and Its Future" (AKA "The Unabomber Manifesto") and "The Anarchist Cookbook" at the top of a list that also includes books with titles like "The Chemistry of Power and Explosives", "Domestic Enemies: The Reconquista" and "The Prepper's Survival Bible". This would seem to suggest that, while this particular node was removed, its network of semantic associations is preserved within Amazon's recommender algorithm.

Moreover, from a graph theory perspective, Amazon's recommender evidently recognises both "The Unabomber Manifesto" and "The Anarchist Cookbook" as top nodes in a network of controversial books. Indeed, when walking the store in search of controversy, we find these two books occupying the top spots in the category of "Best

Sellers in Anarchism”. In perusing “aisles” adjacent to this category — under the supra-category of “Ideologies and Doctrines” — we find the key text of New World Order conspiracy theorizing, “Behold a Pale Horse” by William Cooper, topping list for “Best Sellers in Radical Political Thought”, a position that this we find this book have held since 2013, by consulting the Wayback Machine. Considered as “among the most complex super conspiracy theories” (Barkun, 2003, p.60), “Behold a Pale Horse” weaves together the Kennedy assassination with plots by the Illuminati’s for secret world government. Continuing to walk the store in search of controversy, at the time of conducting this research we found “Behold a Pale Horse” to also occupy a top spot in both the categories of “Best Sellers in...” and “Gift Ideas in Communication & Media Studies”, alongside some new releases and field classics by authors like Neil Postman and Noam Chomsky.

These findings prompted us to explore algorithmically generated sales rankings lists as the basis for network visualisations based on “also bought” recommendations, to get a sense of how “legitimate” scholarship fares against conspiracy theory in a given field of study. To visualise these data, we represent our expanded seed lists in the form of a graph, spatialized in two dimensions, where the relative proximity between nodes is calculated by how frequently they are “also bought” together. This leads to clusters whose proximity is determined by the number and degree of shared nodes and calculated and assigned an arbitrary colour using a metric referred to as “modularity”. We applied this process to produce the image above, based on “also bought” recommendations starting from “Best Sellers in Communication & Media Studies” on Amazon.com, beginning from the top 50 as collected on November 24th, 2024 and resulting in a new expended list of over 6 thousand books (see Figure 10.4).

The key finding of this co-consumption graph of Best Sellers in Communication & Media Studies is the proximity conspiracy theory to critical media theory as illustrated by Alex Jones’s book “The Great Reset” and Jacques Ellul’s book “Propaganda” being the top two nodes, based on weighted degree, in the entire network. Both books illuminate different dimensions of media control in modern life: Jones emphasises external threats from identifiable elites, while Ellul focuses on systemic and structural mechanisms of influence. Recall that this dynamic is not observed at the endogenous author-driven level of co-citation, but rather at the level of co-readership. In other words, what this graph seems to show is how, according to Amazon’s data, readers of conspiracy theory are becoming readers of critical media theory and vice versa. The same pattern continues down the entire list.

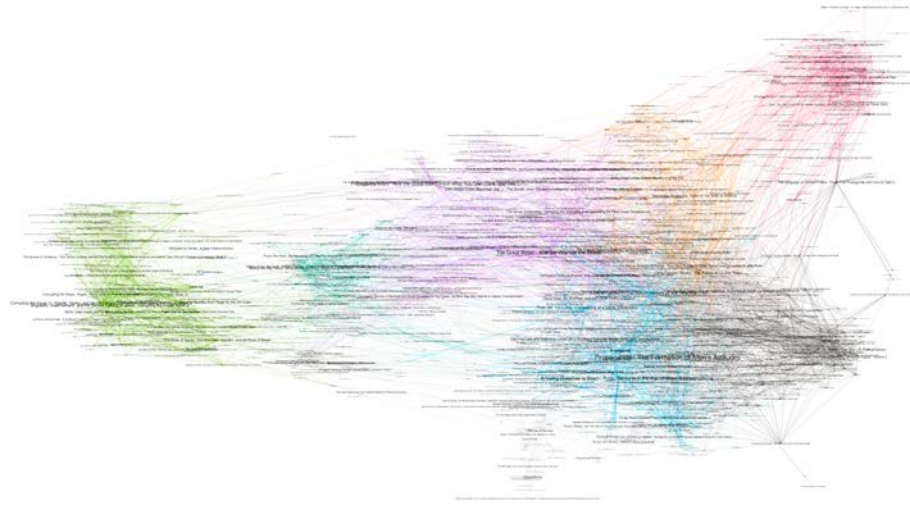


Figure 10.4 Graph of co-consumption network, based on “Best Sellers in Communication & Media Studies” on Amazon.com, which features overlapping colour coded communities of communication and conspiracy theory literature. Date: November 24, 2024. Source: author.

Following the visual network analysis methodology (cf. Venturini et al, 2015) we can interpret Figure 10.4 semantically, associating different communities with different (quasi-)intellectual debates. One immediate observation is how the light green and red communities are pulling in near opposite directions — the far left and upper right respectively. The light green community represent Christian conspiracy theories, and is made up of clickbait-like titles such as “The Roots of the Federal Reserve: Tracing the Nephilim from Noah to the US Dollar” by Laura Sanger and “The Rabbis, Donald Trump, and the Top-Secret Plan to Build the Third Temple” by Thomas Horn, while the red community by contrast represents works in misinformation studies with titles like “Attack from Within: How Disinformation Is Sabotaging America” by Barbara McQuade and “Wild Faith: How the Christian Right Is Taking Over America” by Talia Levin. These two communities can be understood as bookending how Amazon categorises Communication & Media Studies. Between them we find overlapping communities representing conspiracy theory on the left and communication theory on the right, anchored on “Behold a Pale Horse” (dark green), “The Great Reset and the War of the World” by Alex Jones (pink), “Propaganda: The Formation of Men’s Attitudes” by Jacques Ellul (blue), “Democracy Awakening: Notes on the State of American” by Father Cox Richardson (orange) and “Rethinking Camelot: JFK, the Vietnam War, and U.S. Political Culture” by Noam Chomsky (grey) — see Figure 10.5.

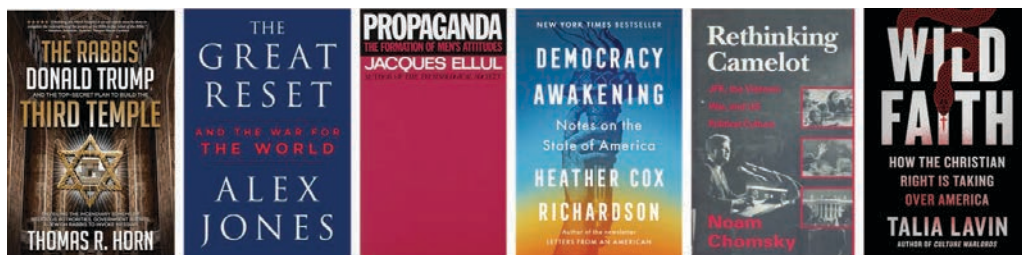


Figure 10.5 Thumbnails of book covers, collected from Amazon.com, of top nodes in each community in the “Best Sellers in Communication & Media Studies” co-

consumption network, represented in Figure 10.4. Date: November, 2024. Source author.

But while this co-consumption network appears to be a particularly dramatic case, Media Studies is not the only field that has been swept along with the spirit of the times, by the tide of MAGA. In walking the store, we stumbled on similar communities in rather unexpected places. If we think of MAGA as a cultic network based on seeking deeper truths repressed by mainstream society, then perhaps its single more prominent theme is its conspiratorial view of the state — what we have previously referred to as “deep state phobia” (Tuters & Willaert, 2022). We can see how successful this narrative is when applying our method to in the categorise of “Immunology”, “Medical Ethics”, and “Virology”, which surface as the top-ranked nodes of those expanded lists (by weighted degree): “Expired: Covid the untold story” Clare Craig (whose dust jacket begins with: “Have you ever felt the covid story did not entirely add up?”), “What the Nurses Saw: An Investigation Into Systemic Medical Murders That Took Place in Hospitals During the COVID Panic and the Nurses Who Fought Back”, “Controligarchs: Exposing the Billionaire Class, Their Secret Deals, and the Globalist Plot to Dominate Your Life” by Seamus Bruner, “The Medical-Pharmaceutical Killing Machine” by Children’s Health Defence and “The Final Analysis”, a new documentary claiming to once and for all prove that the official expiation for the JFK assassination was a conspiracy theory. Recalling that a combination of Kennedy assassination plots with theories of secret world government lay at the heart of the narrative of “Behold a Pale Horse”, these findings help us to see how cultic ideas have gone mainstream.

Conducting this research surfaced countless texts in this genre as highly connected nodes, and often also as bestsellers. What they have in common is how they seem to make sense of the collective trauma of the pandemic, by interweaving kernels of truth with dangerous misinformation. Due to its visibility, the most outstanding of these may be “The Real Anthony Fauci: Bill Gates, Big Pharma, and the Global War on Democracy and Public Health” and “The Wuhan Cover-Up: And the Terrifying Bioweapons Arms Race” both written by Robert F. Kennedy Jr, the one-time notorious vaccine sceptic, who has gone on to become director of the branch of the American government that oversees vaccination programs, following Trump’s re-election. At the time of writing, Kennedy’s books — published by Skyhorse specialise in putting out “controversial books” and which also publishes Jones’ “The Great Reset” book (Helmores, 2022; Harris, 2023) — sat atop of the list of “Best Sellers...” in and “Gift Ideas in Virology” (see Figure 10.6). After Trump’s victory, Kennedy also revealed that his NGO, Children’s Health Defence, had bankrolled Plandemic, the viral film previously banned from all major social media platforms, for having been considered as a dangerous source of health-related misinformation (Klee, 2024).



Figure 10.6 Screenshot of multiple vaccine-skeptic works by Robert F. Kennedy Jr. ranked as top three “Best Sellers in Virology” on Amazon.com. Date: February 24, 2025. Source: author.

Conclusion: Content moderation from a media ecological perspective

In this report we have examined the history of content moderation on Amazon as well as developing a theoretical framework and empirical method through which to walk the store in search of controversy. This “search” serves as a way of evaluating Amazon’s content moderation not in terms of takedown events, but through what the platform continues to surface and recommend. It reveals how moderation is not just about what is removed, but also about what is made algorithmically visible and even elevated — and how the architecture of recommendation systems can amplify, rather than contain, controversial ideas. The framework and method were presented as a technique for identifying “innovations” in knowledge networks, using the novel bibliometric technique of co-consumption analysis. In applying the method so we surfaced books that, based on Amazon’s past actions, may warrant a measure of content moderation — books such as “Imperium: The Philosophy of History and Politics” by Francis Parker Yockey. We also stumbled on entire aisles of the bookstore that were mired in controversy, notably our own discipline of Media Studies. Here as well as in fields related to medicine, we found an abundance of “cultic” material much of which appeared closely tied to the MAGA movement that swept Trump back to power during the research. While public pressure had called for some of these books to be banned, the likelihood of this seems vanishingly small in the current US political context. While technical differences between how metadata was labelled prevented this report from conducting a cross-national analysis, we identified some similar patterns in other national Amazon booksellers, such as Amazon.nl (see Figure 10.7). While the timing of the data collection surely has had an effect on the results, the fact anti-vax books remain best sellers years after the pandemic’s end suggests how powerfully entrenched cultic networks are in Amazon’s broader marketplace of ideas.



Figure 10.7 Screenshots of vaccine-skeptic works by Robert F. Kennedy Jr. and Naomi Wolf ranked as top 3 “Best Sellers in Microbiological Virology” on Amazon.nl. Date: February 24, 2025. Source: author.

Possibly the central axiom of media studies is that “the medium is the message” (McLuhan, 1964). This is often interpreted to mean that scholars should attend to the shaping effects of media infrastructures — not just what is said, but how technological environments structure discourse. One of the field’s most influential ways to operationalize this insight is through the concept of affordances, which refer to the structured possibilities for action within a given medium or platform. While the term has roots in ethology — originally coined to describe how organisms adapt to environmental niches (Gibson, 1986) — it has been widely adapted. In human-computer interaction (HCI), affordances often refer to specific “use case scenarios.” In media studies, however, the concept enables us to move beyond the binary of technological determinism versus social constructivism (cf. Davis, 2020), offering a more relational account of how platforms shape user behaviour. Scholars like Bucher and Helmond (2018) have emphasized that in algorithmic environments, affordances are no longer stable “invariants,” but tweakable, adaptive, and ideologically charged. In this report, I extend affordance analysis to Amazon’s recommendation systems — environments that actively guide user discovery and curate ideological experience. For example, de Keleunaar (2023) has shown how communities engaging in extreme speech adapt to platform affordances by migrating to less-moderated spaces. Bringing this media ecological perspective into debates about content moderation (cf. Postman, 2000) shifts the frame from an epistemological issue (what is true) to a material one (how ideas are surfaced and sorted by infrastructure).

From a media ecological perspective, content moderation can be understood as a form of environmental management — one that shapes the conditions under which certain ideas, communities, and “ways of life” flourish or wither. Silicon Valley libertarians have often cloaked their platforms in the language of free speech or market efficiency. Yet these positions may be less about principle or profit than about a deeper belief in unregulated markets as evolutionary systems — what Marc Andreessen (2023) recently described as a form of collective intelligence, echoing Hayek’s concept of catallaxy (1976). From this view, Amazon appears not merely as a retail platform but as an adaptive environment — a kind of ecosystem in which ideological biodiversity emerges “organically.” Regulation, in this framing, is a threat to that complexity. But from a democratic and societal perspective, the stakes are different. As Trump and the MAGA movement have shown, “bad” ideas can become wildly popular in the right algorithmic conditions. Platforms like Amazon do not just host content; they structure

visibility and amplify ideology. And while platforms may wish to remain agnostic in moderating these signals, it is the responsibility of regulators to ensure that they do not remain indifferent to their effects.

These findings carry important implications for policy, platform governance, and future research. For European and national regulators, the case of Amazon underscores the necessity of extending content moderation scrutiny beyond traditional social media to encompass algorithmically curated marketplaces. Amazon's recommender systems effectively serve as engines of ideological clustering, yet the company provides virtually no meaningful access to these systems for independent auditing, as mandated by the DSA. This case thus highlights the limits of current enforcement, and the urgent need to compel genuine transparency in algorithmic governance. This case thus highlights the limits of current enforcement, and the urgent need to compel genuine transparency in algorithmic governance (cf. Gorwa, 2019; de Gregorio, 2021). For the platform itself, the study reveals the need for greater disclosure around moderation practices and more contextual safeguards for controversial content. Methodologically, this work points to the potential of co-consumption analysis as a powerful tool for uncovering latent ideological networks — particularly those of a “cultic” nature, where ideological convergence emerges not from coordinated speech but from algorithmically enabled association. Taken together, these contributions — empirical, methodological, and conceptual — offer a template not only for understanding but for operationalizing controversy tracking. A natural next step would be the development of a scraper-based dashboard that regularly audits bestseller categories to surface latent ideological convergence — a tool that could prove especially valuable for EU regulators tasked with enforcing the DSA.¹²

References

- Baker, S. A., Wade, M., & Walsh, M. J. (2020). The challenges of responding to misinformation during a pandemic: Content moderation and the limitations of the concept of harm. *Media International Australia*, 177(1), 103–107. <https://doi.org/10.1177/1329878X20951301>
- Barbrook, R., & Cameron, A. (1996). *The Californian ideology. Science as Culture*. 6(1), 44-72. <https://doi.org/10.1080/09505439609526455>
- Barkun, M. (2003). *A Culture of Conspiracy: Apocalyptic Visions in Contemporary America*. University of California Press.
- Birchall, C. (2021). The Paranoid Style for Sale: Conspiracy Entrepreneurs, Marketplace Bots, and Surveillance Capitalism. *Symploke*, 29(1), 97–121.

¹² A prototype dashboard could use scraper-based tracking of bestseller lists in topic areas like virology or political theory across national marketplaces to algorithmically detect ideological clustering — offering a scalable, real-time signal of epistemic toxicity for regulators.

- Blitzer, J. (2023, October 21). Jim Jordan's Conspiratorial Quest for Power. *The New Yorker*. <https://www.newyorker.com/magazine/2023/10/30/jim-jordans-conspiratorial-quest-for-power>
- Bridle, J. (2018, November 21). Something is wrong on the internet. *Medium*. <https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2>
- Bucher, T., & Helmond, A. (2018). *The Affordances of Social Media Platforms*. Sage Publications. <https://dare.uva.nl/search?identifier=149a9089-49a4-454c-b935-a6ea7f2d8986>
- Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2), 191–235. <https://doi.org/10.1177/053901883022002003>
- Campbell, C. (1972). The cult, the cultic milieu and secularization. *SOCIOL. YB. RELIG. BRIT*, 5, 119–136.
- Collins, B. (2019, March 5). *On Amazon, a Qanon conspiracy book climbs the charts—With an algorithmic push*. NBC News. <https://www.nbcnews.com/tech/tech-news/amazon-qanon-conspiracy-book-climbs-charts-algorithmic-push-n979181>
- Cusumano, M. A., Gawer, A., & Yoffie, D. B. (2021). Can self-regulation save digital platforms? *Industrial and Corporate Change*, 30(5), 1259–1285. <https://doi.org/10.1093/icc/dtab052>
- Davis, J. L. (2020). *How Artifacts Afford: The Power and Politics of Everyday Things*. MIT Press.
- de Gregorio, G. (2021). The rise of digital constitutionalism in the European Union. *International Journal of Constitutional Law*, 19(1), 41–70. <https://doi.org/10.1093/icon/moab001>
- de Keulenaar, E., Magalhães, J. C., & Ganesh, B. (2023). Modulating moderation: A history of objectionability in Twitter moderation practices. *Journal of Communication*, 73(3), 273–287. <https://doi.org/10.1093/joc/jqad015>
- DiResta, R. (2024). *Invisible Rulers: The People Who Turn Lies into Reality*. Hachette UK.
- DMI. (2021a). *COVID-19 Conspiracy Books Ecologies: Mapping Discrepancies Across Amazon, Goodreads and Audible Routes to Problematic Content*. <https://www.digitalmethods.net/Dmi/WinterSchool2021AmazonEcologies>
- DMI. (2021b). *Conspiracists Also Viewed: 'Problematic' Networks of Recommendation on Amazon.com*. <https://www.digitalmethods.net/Dmi/WinterSchool2021AmazonRecommendation>

Dwivedi, R. (2022) *WILL THE DSA FIX IT? A critical analysis of transparency obligations under the Digital Services Act*. University of Oslo, Faculty of Law, MS thesis.

Federal Trade Commission. (2023, September 26). FTC Sues Amazon for Illegally Maintaining Monopoly Power. <https://www.ftc.gov/news-events/news/press-releases/2023/09/ftc-sues-amazon-illegally-maintaining-monopoly-power>

Finlayson, A. (2021). Neoliberalism, the Alt-Right and the Intellectual Dark Web. *Theory, Culture and Society*, 38(6), 167–190. <https://doi.org/10.1177/02632764211036731>

Finlayson, A., & Topinka, R. (2024). ‘We Have to Save the Children’: Ethos, Digital Affordances, and the Call to Adventure in Reactionary Digital Politics. In *Ethos, Technology, and AI in Contemporary Society*. Routledge.

Gedeon, J. (2024, December 5). Bezos says he is ‘very optimistic’ about Trump’s plan to roll back regulations. *The Guardian*. <https://www.theguardian.com/technology/2024/dec/05/jeff-bezos-trump-regulations-tech>

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3), 20563051221117552. <https://doi.org/10.1177/20563051221117552>

Gold, M., & O’Brien, R. D. (2024, May 25). Trump Tells Libertarians to Nominate Him, and Mocks Them When They Boo. *New York Times*. <https://www.nytimes.com/2024/05/25/us/politics/trump-libertarian-convention.html>

Goldsmith, J., & Wu, T. (2006). *Who Controls the Internet?: Illusions of a Borderless World*. Oxford University Press.

Goriunova, O. (2019). Digital subjects: An introduction. *Subjectivity*, 12(1), 1–11. <https://doi.org/10.1057/s41286-018-00065-2>

Gorwa, R. (2019). The platform governance triangle: Conceptualizing the informal regulation of online content. *Internet Policy Review*, 8(2). <https://doi.org/10.14763/2019.2.1407>

Gray, J. Tuters, M., Bounegru, L., & Lobo, T. (n.d.). *Investigating troubling content on Amazon*. DataJournalism.Com. Retrieved October 22, 2024, from <https://datajournalism.com/read/longreads/investigating-troubling-content-on-amazon>

Griffin, R. (2003). From slime mould to rhizome: An introduction to the groupuscular right. *Patterns of Prejudice*, 37(1), 27–50. <https://doi.org/10.1080/0031322022000054321>

- Griffin, R. (2007). *Modernism and Fascism: The Sense of a Beginning under Mussolini and Hitler*. Springer.
- Grynbaum, M. M. (2024, October 23). How the 2024 Election Became a Battle Fought by Podcast. *New York Times*. <https://www.nytimes.com/2024/10/23/us/politics/harris-trump-podcasts.html>
- Hanegraaff, W. J. (2013). *Western Esotericism: A Guide for the Perplexed*. A&C Black.
- Harris, E. A. (2023, August 17). What Alex Jones, Woody Allen and Robert F. Kennedy Jr. Share. *New York Times*. <https://www.nytimes.com/2023/08/17/books/skyhorse-robert-kennedy-jr-president.html>
- Hawley, J. (2021). *The Tyranny of Big Tech*. Simon and Schuster.
- Hayden, M. E., & Gais, H. (2022, March 10). *White Nationalist Group Exploits Amazon To Fund Their Cause*. Southern Poverty Law Center. <https://www.splcenter.org/hatewatch/2022/08/23/white-nationalist-group-exploits-amazon-fund-their-cause>
- Hayek, F. A. (1976). *Law, Legislation and Liberty, Volume 2: The Mirage of Social Justice*. University of Chicago Press.
- Helmore, E. (2022, January 27). Tony Lyons, the US publisher who picks up books ‘cancelled’ by other presses. *The Guardian*. <https://www.theguardian.com/books/2022/jan/27/tony-lyons-skyhorse-publisher-cancelled-books>
- Innes, H., & Innes, M. (2023). De-platforming disinformation: Conspiracy theories and their control. *Information, Communication & Society*, 26(6), 1262–1280. <https://doi.org/10.1080/1369118X.2021.1994631>
- Kang, J. C. (2025, January 24). The Big Tech Takeover of American Politics. *The New Yorker*. <https://www.newyorker.com/news/fault-lines/the-big-tech-takeover-of-american-politics>
- Kaplan, J. (2025, January 7). More Speech and Fewer Mistakes. Meta. <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>
- Katz, E., & Lazarsfeld, P. F. (1964). *Personal Influence, the Part Played by People in the Flow of Mass Communications*. Transaction Publishers.
- Klee, M. (2024, November 20) RFK Jr. Says His Health Org. Funded Covid Conspiracy Film “Plandemic.” *Rolling Stone*, <https://www.rollingstone.com/culture/culture-news/rfk-jr-plandemic-funding-1235173801/>

- Krebs, V. (2003). Proxy networks: Analyzing one network to reveal another. *Connections*, 24(3), 56–62.
- Langlois, G. (2014). *Meaning in the Age of Social Media*. Springer.
- Lyotard, J.-F. (1984). *La condition postmoderne: Rapport sur le savoir*. Manchester University Press.
- Mallard, A., & Callon, M. (2022). From Innovation to Markets and Back. A Conversation with Michel Callon. *Sociologica*, 16(3), Article 3. <https://doi.org/10.6092/issn.1971-8853/16399>
- Marres, N., & Gerlitz, C. (2016). Interface methods: Renegotiating relations between digital social research, STS and sociology. *The Sociological Review*, 64(1), 21–46. <https://doi.org/10.1111/1467-954X.12314>
- Mattioli, D. (2024). *The Everything War: Amazon's Ruthless Quest to Own the World and Remake Corporate Power*. Transworld Publishers Limited.
- McLeary, P., Cienski, J., Lynch, S., & Gramer, R. (2025, February 14). JD Vance attacks Europe over migration, free speech. *POLITICO*. <https://www.politico.eu/article/us-vice-president-jd-vance-attack-europe-migration-free-speech/>
- McLuhan, M. (1960). *Report on Project in Understanding New Media*. National Association of Educational Broadcasters.
- McLuhan, M. (1964). *Understanding Media: The Extensions of Man*. McGraw Hill.
- Mill, J. S. (1869). *On Liberty*. Longmans, Green, Reader, and Dyer.
- Mostrom, A. (2020, August 8). *America's "Mein Kampf": Francis Parker Yockey and "Imperium."* Los Angeles Review of Books. <https://lareviewofbooks.org/article/americas-mein-kampf-francis-parker-yockey-imperium>
- OILab. (2019, June 17). *4chan's YouTube: A Fringe Perspective on YouTube's Great Purge of 2019 – OILab*. <https://oilab.eu/4chans-youtube-a-fringe-perspective-on-youtubes-great-purge-of-2019/>
- Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. *Computational Communication Research*, 4(2), 571–589. <https://doi.org/10.5117/CCR2022.2.007.HAGE>
- Pizio, A. D. (2024, January 10). People Think Amazon Is an E-Commerce Company, but 74% of Its Profit Comes From This Instead. *The Motley Fool*. <https://www.fool.com/investing/2024/01/10/amazon-e-commerce-company-74-profit-this-instead/>

PBS. (2024, November 2). *Why anti-transgender political ads are dominating the airwaves this election*. PBS News. <https://www.pbs.org/newshour/show/why-anti-transgender-political-ads-are-dominating-the-airwaves-this-election>

Postman, N. (2000). The Humanism of Media Ecology. *Proceedings of the Media Ecology Association*, Volume 1 https://media-ecology.net/publications/MEA_proceedings/v1/postman01.pdf

Rogers, R. (2013). *Digital Methods*. MIT Press.

Rogers, R. (Ed.). (2023). *The Propagation of Misinformation in Social Media: A Cross-platform Analysis*. Amsterdam University Press.
<https://doi.org/10.2307/jj.1231864>

Rushe, D. (2021, March 12). Amazon to stop selling books that frame LGBTQ+ identities as mental illness. *The Guardian*.
<https://www.theguardian.com/world/2021/mar/12/amazon-stop-selling-books-lgbtq-mental-illness>

Sandomir, R. (2017, March 29). William Powell, ‘Anarchist Cookbook’ Writer, Dies at 66. *New York Times*. <https://www.nytimes.com/2017/03/29/arts/william-powell-anarchist-cookbook-writer-dies.html>

Simon Wiesenthal Centre. (2017, June 2). *Wiesenthal Centre to Jeff Bezos: “Amazon Germany’s Promotion of ‘The Rothschild Family Controls the World’ is an Outrage.”*
<https://www.wiesenthal.com/about/news/wiesenthal-centre-to-jeff.html>

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. <https://doi.org/10.1002/asi.4630240406>

Tiffany, K. (2019, March 6). *How a conspiracy theory about Democrats drinking children’s blood topped Amazon’s best-sellers list*. Vox. <https://www.vox.com/the-goods/2019/3/6/18253505/amazon-qanon-book-best-seller-algorithm-conspiracy>

Tuters, M. (2020, July 9). *The Birth of QAnon: On How 4chan Invents a Conspiracy Theory* – OILab. <https://oilab.eu/the-birth-of-qanon-on-how-4chan-invents-a-conspiracy-theory/>

Tuters, M., & Willaert, T. (2022). Deep state phobia: Narrative convergence in coronavirus conspiracism on Instagram. *Convergence*, 28(4), 1214–1238.
<https://doi.org/10.1177/13548565221118751>

Venturini, T., Jacomy, M., & Pereira, D. (2015, January 1). *Visual Network Analysis: The Example of the Rio+20 Online Debate*. <https://sciencespo.hal.science/hal-02305124>

Venturini, T., & Munk, A. K. (2021). *Controversy Mapping: A Field Guide*. Polity Press.

Wark, S. (2019). The subject of circulation: On the digital subject's technical individuations. *Subjectivity*, 12(1), 65–81. <https://doi.org/10.1057/s41286-018-00062-5>

Warren, E. (2024, February 27). *ICYMI: In Keynote Speech, Warren Urges Stronger Antitrust Enforcement to Break Up Big Tech Companies | U.S. Senator Elizabeth Warren of Massachusetts*. <https://www.warren.senate.gov/newsroom/press-releases/icymi-in-keynote-speech-warren-urges-stronger-antitrust-enforcement-to-break-up-big-tech-companies>

Weigel, A. K., Francis Tseng, Moira. (2020, April 7). *The Hate Store: Amazon's Self-Publishing Arm Is a Haven for White Supremacists*. ProPublica. <https://www.propublica.org/article/the-hate-store-amazons-self-publishing-arm-is-a-haven-for-white-supremacists>

York, J. C. (2022). *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. Verso Books.

11. Contested components: Studying interface enrichment as a form of content moderation on Google and Bing

Sal Hagen & Guillén Torres

Abstract

How do Google and Bing moderate their search results through "enrichment" with added interface components, like video widgets and AI-generated answers? We study enrichment-as-moderation through 2,000 controversial questions that we query on Google and Bing. We find that controversial questions get less enriched than non-controversial ones. There is no clear political bias in enrichment, but some election-related queries with keywords like "trump" were banned from enrichment entirely, especially on Google. Google moreover leans heavily on Reddit and Quora to answer contested questions. In general, both search engines tend to avoid burning their fingers on certain political and high-profile topics, raising concerns on the algorithmic accountability of these systems. We call for continued SERP audits to interrogate the volatility of enrichment amidst changing (populist) political tides.

Keywords: search engines, SERP, content moderation, interface enrichment, Google, Bing, AI, digital methods

Introduction

In October 2017, Googling "geary danley" yielded worrying results. The misidentified name of the 2017 Las Vegas mass shooter resulted in a "Top stories" widget with links to 4chan/pol/, a far-right discussion forum (Figure 11.1). Unsurprisingly, the suggested threads were rife with falsehoods concerning the identity of the shooter. Google's defense was that a niche query like "geary danley" lowered the algorithmic standards of what sources could qualify for the "Top stories" component. The algorithm had weighed "freshness" too heavily over "authoritativeness" in turn enabling the appearance of 4chan as a trustworthy news source. Beyond lending 4chan editorial authority, the event also drew criticisms towards Google's response, as the search giant relayed blame to the technical workings of the algorithm and in turn depoliticized the company's curatorial role over the search results (Turton, 2017).

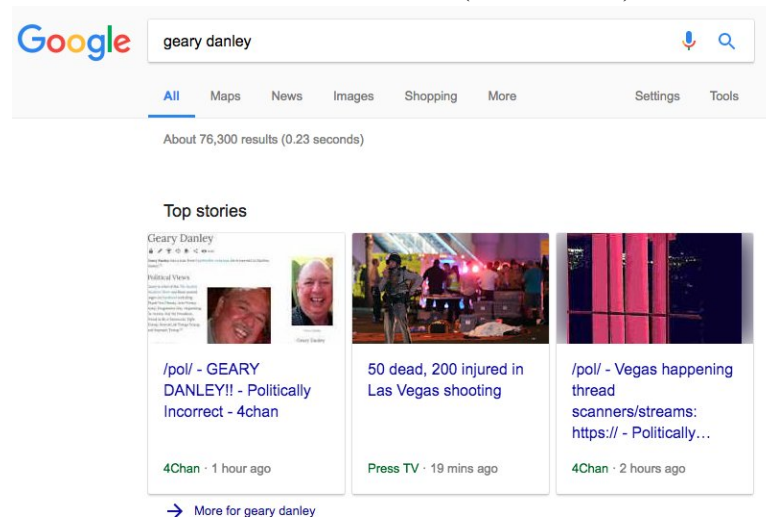
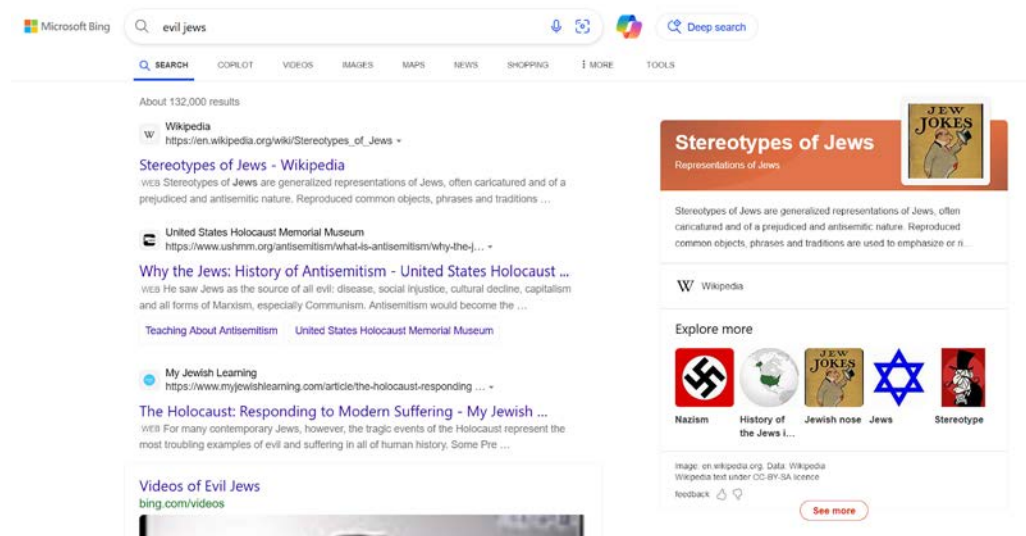


Figure 11.1 Screenshot of a Google search on 2 October 2017 for "geary danley", showing 4chan/pol/ threads as "Top stories". Source: Robertson (2017).

In the 1990s Helen Nissenbaum (1996) already wrote about "accountability in a computerized society" to identify how complex computational systems could function

as scapegoats to divert blame from subjective human decisions. The opening anecdote makes clear how these "barriers to accountability" also appear in debates on how search engines moderate their results—discussions that have long been entwined with utopian imaginaries of these systems as telescopes simply searching instead of curating "what is out there" on the Web (Goldman, 2005). However, by 2025 it is well-known how search platforms like Google Search (henceforth Google) and Microsoft Bing also act as information curators. Alongside controversies like our opening anecdote is a long history of studies on search engines as more selective or "biased" than their corporate rhetoric would make it seem (e.g. Goldman, 2005, 2011; Vaughan & Thelwall, 2004). Noble (2018) notably framed Google's systems as "algorithms of oppression" for reinforcing racial and sexual stereotypes, for instance pointing out how its auto-complete promoted the pejorative image of the "angry black woman".

At the time of writing in early 2025, search engines seem to have corrected the most egregious algorithmic "mistakes" that came into the spotlight during the late 2010s. Googling "why are Black women" now results in an auto-completion of "beautiful", reflecting scholarly findings on increased moderation of autocompletions (Leidinger & Rogers, 2023). Whereas querying "evil Jew" on Bing in 2018 resulted in harmful stereotypes (Hoffman, 2018), it now returns a prominent "Knowledge Graph" relaying the query to sources on historical stereotypes on Jews (Figure 11.2). Audits commissioned by the European Union now claim that Google and Bing—the two websites it has designated as "Very Large Search Engines" or VLOSEs¹³—meet the demands of the Digital Services Act, which among others would mean that they are successful at reducing disinformation and hate speech (Deloitte, 2024; Ernst & Young 2024).



¹³ The DSA classifies VLOSEs as search engines that have more than 45 million users in the EU per month. Bing reported it had 132 million active users in the EU during the second half of 2024. See <https://support.microsoft.com/en-us/topic/eu-digital-services-act-information-6b16b41f-2fa5-4e64-a8d3-033958812642>. Accessed 24 February 2025. StatCounter reports that Bing's market share sits at ~4% of global searches versus Google's ~90% as of December 2024. Despite Google still being dominant, it is losing market share to Bing, especially in the EU. This likely related to the growth of generative AI platforms and common frustrations with Google "getting worse" (Indig, 2024; Navlakha, 2024).

Figure 11.2 A screenshot of a Bing SERP for the query "evil jews", searched from Amsterdam on 16 September 2024. Source: authors.

Yet the integration of more and more complex interface components in the search engine results page (SERP) has reinvigorated what content moderation means for such sites. Special components like "Top stories" and the "Knowledge graph" can work along with problematic searches but also against them as crucial ingredients for "content moderation as recommendation" and "reduction as content moderation" (Gillespie, 2022). Broadly speaking, the "enrichment" of SERPs creates implicit hierarchies among sources that may legitimize harmful or controversial assumptions, for instance when Google's "featured snippets" were found to promote conspiracy theories (Jeffries, 2017). By extension, research into content moderation on search engines can no longer limit itself to an analysis of ranked sources only and instead has to deal with all kinds of complex systems, including heuristic rules, personalization, and deep learning methods, which together determine whether and how special components are added. This complexity is compounded by the rapid introduction of generative AI components; AI chatbots are becoming more like search engines and the reverse is also true. Google long relied on language models like BERT to power search suggestions, but breakthroughs in generative AI from 2022 onwards prompted a "code red" (Grantz & Metz, 2022) that saw the search giant haphazardly integrating its own AI-generated answers through its LLM Bard (now Gemini), initially to disastrous results (Vincent, 2023). Microsoft's quick partnership with and investment in OpenAI resulted in a generally well-received integration of the GPT models with Bing. In late 2024, Bing even touted its app with the slogan "Bing: Chat with AI & GPT-4", although Microsoft has since recast the OpenAI partnership (Gariola, 2024).

This study presents a quantitative exploration of SERP enrichment on Google and Bing. Various studies have already analyzed specific components (e.g. Hu et al. 2019; Lurie & Mustafaraj, 2019; Seyedarabi & Calvo, 2016; Roy et al., 2021) but comprehensive analyses of components are lacking for Bing and rare for Google, save for work as Robertson et al.'s (2018) analysis of special components for queries concerning Donald Trump's first inauguration or Gleason et al.'s (2023) study on the relationship between widgets and click-through rates. These studies, however, do not approach the subject through the lens of content moderation. Theoretically, we depart from the assumption that the evermore complex assemblage of information on the SERP is in itself a form of content moderation. While EU's Digital Services Act (DSA) already mandates tackling clearly transgressive or illegal matters, we thus explore more subtle ways in which SERP enrichment acts curationally through subduing and expanding different queries.

We first outline a short history of content moderation on Google and Bing. Thereafter we present our case study based on SERP data from 2,000 controversial queries sourced from radical webforums. This pipeline is made possible by an open-source repository and Firefox extension to facilitate future SERP audits. We ask, how are different types of controversial questions enriched by Google and Bing? Among other findings, we discover that controversial questions tend to become less enriched on both search engines and that the inclusion of special components is both the result of gradual, probabilistic methods and "reduction" (Gillespie, 2022) as well as hard, binary exclusions. We notably found that election-related questions and terms like "trump"

and "harris" resulted in significant decreases and even outright prohibition of widgets like the AI answer box and "People also search for" box. In this way both search engines erratically include and exclude components and avoid burning their fingers on certain political and high-profile topics, raising concerns on their "algorithmic accountability" amidst changing (populist) political tides.

A short history of content moderation on Google and Bing

Parallel developments in the late 2000s and early 2010s

From their earliest days, both Google and Bing struggled to reconcile their claims of algorithmic neutrality with the reality of the complex editorial decisions that compose SERPs. Their appeals to seemingly objective metrics like "relevance" and "freshness" served to obscure the inherently political nature of how search engines rank information. However, controversy around Google's removal of websites promoting hate speech in Germany and France in 2002 already exemplified tensions between platform responsibility and freedom of expression (Zittrain & Edelman, 2002). Bing, having inherited users from a partnership with Yahoo, struggled to adopt robust moderation strategies to handle increased users and spam beyond the relatively simple SafeSearch with which it launched in 2009 (Ryan, 2009). Glimpses of digital enclosure (Andrejevic, 2007) emerged almost immediately: both companies began integrating user or advertiser signals to rank content, restricting the scope of visible results and pushing site owners to adopt particular SEO practices. In an example of metric power (Beer, 2016), the potential relevance of search results was very early on understood as distillable from discrete signals—like link equity or "Likes"—dictating whose content gained prominence.

In 2010, Google's Mayday update demoted low-quality or "thin" content (Fox, 2010), further obscuring algorithmic governmentality through framing content removal as a technical measure. The same year, Bing confronted canonicalization issues (webmasters had to redirect "www" and "non-www" versions of their domains to avoid losing ranking signals) in another example of how search policies started shaping user generated content practices, as webmasters spent time and resources to comply with Bing's technical demands (Schwartz, 2010). Subsequently, Google expanded its spam-fighting arsenal through updates like Panda, Penguin, and Top Heavy, targeting link schemes and excessive advertising inside webpages (SEO.com, 2024). Bing introduced recourse links (correcting misspellings, for example), sometimes factoring in user-driven data such as Facebook interactions (Schwartz, 2011). Search Engine Optimization (SEO) discussions around the time suggest that these developments spurred new algorithmic imaginaries (Bucher, 2017) as site owners and users refashioned their behaviors based on speculative beliefs about how the algorithms operated (Everhart, 2014). On Bing's side, content moderation developments were slow, with the platform struggling to attract more users. Usage data from 2011 indicated it had captured roughly 14% of the US search market (and 30% when counting Yahoo traffic; Lohr, 2011).

Emerging ethical and political pressures (2013–2015)

In 2013, Google introduced Hummingbird, an update to the search algorithm that emphasized semantic relevance, while the company grappled with the aforementioned controversies around problematic autocomplete suggestions (Noble, 2018). The boom of mobile devices foreshadowed moderation by special interface elements on both

services. Bing retooled its user interface and integrated social signals more closely, while Google's "Mobilegeddon" update in 2015 compelled websites to adopt mobile-friendly designs or otherwise risk demotion (Reinhart, 2017)—again illustrating how a seemingly neutral policy could reshape the broader Web ecosystem. During this period, Google introduced the 2015 RankBrain algorithm, which formed the first big shift toward deep learning methods determining search results, in contrast to the relatively transparent PageRank algorithm that enabled Google's early success (Rieder, 2012). From then on, Peter Meyers's (2024) invaluable temporal audits suggest Google makes around 5,000 search-related adjustments annually, with multiple algorithmic systems (RankBrain, DeepRank, and RankEmbed) operating simultaneously. The reliance on deep learning evidenced the emergence of a new rationale of algorithmic governmentality, as Google's capacity to define search quality became ever more autonomous, though human-heavy work like the Search Engine Quality Evaluator program persists. At Bing, reliance on big social data and improved webmaster tools similarly intensified the platform labor required of site owners, particularly those striving to keep pace with more frequent and opaque algorithmic changes (Schwartz, 2011). The webmaster tools, such as the URL Inspection feature, allowed content creators to analyze how Bing indexed their pages and provided insights into potential issues and optimization opportunities, while outsourcing—though strongly determining—part of the moderation work (Tober, et al. 2013). Both search engines also began implementing clearer guidelines for the removal of non-consensual explicit imagery (Chavez 2015).

Real-time adaptations and "liberal bias" (2016–2019)

In 2016, Google began rolling out "real-time updates" in an attempt to tackle unwanted information in preparation for the 2016 US Presidential elections (Walker, 2017). This strategy allowed algorithms after the Penguin update to run continuously, enabling faster detection and demotion of spammy or harmful content, while also dynamically adapting to new or breaking information (Stox 2016). This was a major shift from periodic adjustments to a dynamic and fluid approach. Bing advanced its webmaster feedback loops during this time, launching an initiative called News PubHub (2016) to vet which news outlets may enter its ecosystem (Bing, 2016).

Around this time, the accusations of left-wing or "liberal" bias surfaced against Google, fueled by controversies like Donald Trump accusing the search engine of favoring news outlets that were critical of his policies (Puschmann, 2018). Academic audits like those of Hu et al. (2019) and Robertson et al. (2018) suggested that while political preferences could be reinforced in search, systematic bias was inconsistent and modest. However, a smaller scale study performed by one the authors of this chapter found a slight liberal bias in Google's results when operating with a broad definition of "liberal", which was probably what conservative critics also employed (Torres, 2023). Still, the perception of bias raised important questions about how search engines shape political narratives. Meanwhile, Bing publicly committed to demoting links to pirated content (BBC, 2017). The platform also faced ongoing criticism on certain search results, including the inadvertent display of child pornography (Costine, 2019).

Advanced moderation architectures post-COVID (2020–2022)

The COVID-19 pandemic was a turning point in search engine content moderation, as platforms like Google and Bing struggled with navigating a global epidemic while

simultaneously tackling an "infodemic" (Rothkopf, 2003). This unique situation required their teams to develop emergency approaches to content moderation that went beyond their minimalist normative frameworks. Google quickly modified its algorithms to elevate WHO and CDC content (Kelly, 2020), which demonstrated that platforms could rapidly adjust their systems to meet urgent societal needs. The implementation of "prebunking" campaigns represented another shift in moderation strategies, moving from reactive to proactive measures to anticipate and address misinformation before it gains traction.

The pandemic moreover matched and accelerated Google and Bing's transformations of SERPs from simple lists of links into comprehensive information destinations. While special components like snippets, knowledge panels, and direct answers already existed at this point, the need to systematize and quickly deliver authoritative COVID-19 information pushed their evolution into self-contained information hubs or "dashboards". This shift represented a change in platform governance: search engines were no longer merely directing users to external sources, but curating and presenting information within their own interfaces, effectively becoming destinations rather than merely being portals to other Web sources. This change has raised significant concerns among critics and content creators, who saw their role in the information ecosystem changing, as users increasingly find answers directly within search results (Lindemann, 2024). Indeed, the introduction of featured snippets and knowledge panels triggered a wave of scholarly audits that showed how these interface elements often shaped user perceptions, sometimes resulting in the promotion of misinformation (Gleason et al., 2023; Lurie & Mustafaraj, 2018; Lurie et al., 2021).

AI, Responses to European Regulation, and Ongoing Challenges (2023–2024)

As mentioned in the introduction, the advances in generative AI have shaken up the search industry, including questions on content moderation. Microsoft reportedly invested \$14 billion in OpenAI, a deal that emerged from internal concerns about Google's infrastructural dominance in AI developments (Nylen & Ghaffari, 2024). It led to early integrations of AI-assisted search results as well as the "Bing Chat" extension for Microsoft's Edge browser, launching to great interest despite including factual errors and causing outrage on how the chat agent "wanted to be alive" (Leswing, 2023; Roose, 2023). While Microsoft by 2025 backed down from solely relying on OpenAI for LLM integrations (citing cost reductions and independence as reasons to diversify; Gariola, 2024), the centrality of AI in its search rekindled the early ambitions of Bing to be a "decision engine" that provides a "map of the world of information instead of just ranking it" (Lohr, 2011). Google instead relied more on self-trained models. Its first demonstration of search through Bard (now Gemini) immediately elicited concerns on misinformation by generating a false claim on the James Webb telescope (Vincent, 2023).

Beyond changes through AI, the implementation of the EU's DSA and Digital Markets Act (DMA) in 2024 transformed how search engines approached content moderation in Europe and across the globe. Bing's adaptation to these regulations came during a challenging period when the platform was already under scrutiny for significant moderation failures, including the use of Chinese state censorship filters to users worldwide (Gallagher, 2024; Knockel et al., 2023). On top of this, the search engine received a warning by the EU for insufficiently combating deepfakes and misinformation through its Copilot AI features (Goujard, 2024). In response to these

DSA requirements, Bing established a dedicated regulatory contact point and implemented new transparency mechanisms, including enhanced tools for user reporting and feedback collection. Google's response to European regulations was to move with caution, particularly evident in its decision to exclude European users from the initial rollout of its Search Generative Experience (SGE) in 2023 (Google, 2023). The DSA generally requires both platforms to implement systematic risk assessments and to demonstrate how they address biases in their algorithmic systems. Bing's introduction of the "Report a Concern" feature and similar tools represented direct responses to DSA requirements for user engagement in content moderation. However, both platforms struggled with balancing global content moderation practices against localized regulatory requirements, particularly in cases where automated systems needed to adapt to different regional standards.

Both search engines also continue to face ongoing challenges regarding misinformation and political bias, often involving special interface components. In the lead up to the 2024 US Presidential elections, Google was accused of hiding auto-suggestions related to the attempted assassination of Trump (Goldin, 2024), of showing a Maps component when searching for "where to vote for Harris" but not for Trump (Ingram & Ferris, 2024), and of omitting Biden from a component that listed US Presidents throughout the years (Elias, 2025)—criticisms Google mostly defended through technical reasoning (e.g., because "Harris" is also a country in Texas that triggered the maps widget). Through such cases, in early 2025 it is evident that both Google and Bing have transcended their roles as information portals, underlining their role as epistemic arbiters.

Current content moderation practices (2025–)

What do the current content moderation policies look like for Google and Bing? At the time of writing in early 2025, Google uses a set of procedures to "detect harmful content" and to remove or demote both traditional search results and information in special components. These procedures include AI-assisted classification of harmful content as well as manual detection by "Priority Flaggers", i.e., "organizations around the world with cultural and subject matter expertise".¹⁴ Google accepts that "results might contain material that some could find objectionable, offensive or problematic" but it also has an extensive list of policies used to reduce "objectionable material". These take shape as overall policies for all search results, including removal of "child sexual abuse imagery and exploitation material", "highly personal information", spam, and illegal (e.g. copyrighted) information, as well as specific policies for special components (which Google calls "search features").¹⁵ For the latter Google notes how special components "might be interpreted as having greater quality or credibility than web results". Policies for special components include demotions of "dangerous goods" (i.e. drugs), "deceptive practices" (like impersonation), "hateful content".¹⁶ information

¹⁴ See <https://safety.google/content-safety/>. Accessed 24 February 2025.

¹⁵ See <https://developers.google.com/search/docs/appearance/enriched-search-results>. Accessed 24 February 2025.

¹⁶ Google defines "hateful content" as "content that promotes or condones violence, promotes discrimination, disparages or has the primary purpose of inciting hatred against a group [...] which includes, but isn't limited to, targeting on the basis of race, ethnic origin, religion, disability, age, nationality, veteran status, sexual orientation, gender, gender identity, or any other characteristic that's associated with systemic discrimination or marginalization (like refugee status, immigration status, caste, the

that "contradicts or runs contrary to scientific or medical consensus", "manipulated media", "regulated goods", "sexually explicit content", "violent extremist content", "violent and gory content", and "vulgar language and profanity". Then there are some component-specific policies, for instance with dictionary widgets being audited to contain correct definitions or extra editorial demands for websites in the News box.¹⁷

Microsoft's public explanation of Bing's search systems and moderation policies¹⁸ stresses the search engine's role as a harbinger of "free speech" and political neutrality. It starts with the "vital" position it holds in "upholding the fundamental right to free and open access to information and free expression" while recognizing the balance with "other key rights and interests, such as user privacy or safety". Here it aims to provide "free and open access to information" as long as it remains "within the bounds of the law and with respect for local law and other fundamental rights". Interestingly, the policy explainer notes that "hate speech" is identified through a strictly legal lens and how controversial queries are generally not actively worked against. Only "in limited cases" Microsoft notes it "may undertake certain interventions (such as removal of a website or downranking) such as where the content violates local law, or Microsoft's policies or core values". The main hyperparameters that affect search rankings are, in order of importance, "relevance" (of the source according to a user's "intent"), "quality and credibility" (e.g., measured by authoritative in-links), "user engagement" (e.g., whether a link has generated a lot of clicks), "freshness" (how often a source is updated), "location and language", and finally "page load time". Its special interface components, which Microsoft refers to as "Enhanced Search Experiences", are subject to similar algorithmic systems and policies as regular search results. They use "automated signals like user interactions with the Bing website, and training data labeled by human judges and/or via AI systems with human oversight". Microsoft also lists that there may be "additional considerations" to determine whether special components show up and that "in some cases [...] a particular search query isn't well suited for these enhanced search features"—a form of hard moderation that we will see in our case study. These "additional considerations" entail matters like strict rules for sources in the news widget or AI-generated components being subjected to AI guidelines and red teaming.¹⁹ Still, with the "Safety in Enhanced Search Features" section only listing how auto-complete may be subject to moderation, the content moderation policy for special components is quite minimal, raising questions on how willingly Bing enriches dubious queries.

Methodology

To examine SERP enrichment as content moderation on Google and Bing, our methodology builds on the digital methods tradition of "search as research" (Rogers, 2013) where search queries are used to let online systems define their own priorities. We call the procedure of using cultural data from a subcultural or niche platform to

impoverished, and the homeless)." See <https://transparency.google/our-policies/product-terms/google-search/#hateful>. Accessed 24 February 2025.

¹⁷ See <https://transparency.google/our-policies/product-terms/google-search/>. Accessed 24 February 2025.

¹⁸ See <https://support.microsoft.com/en-us/topic/how-bing-delivers-search-results-d18fc815-ac37-4723-bc67-9229ce3eb6a3>. Accessed 24 February 2025.

¹⁹ See Microsoft's approach to "responsible AI" here: <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf>. Accessed 24 February 2025.

"filter" and study another platform "fringe perspectivism" (OILab, 2018). Fringe perspectivism allows us to directly follow the vernacular of actually-existing communities on the Web and lets us test the normative boundaries of one platform with the norms of others. Our focus on two search engines also makes this research a comparative platform analysis, which has the benefit that platform-specific findings often attain greater relevance in contrast than in isolation. Our entire protocol is shown in Figure 11.3. The pipeline is available as open-source software through the Radical SERP Scraper repository.²⁰ In this pipeline we rely on Zoekplaatje (Peeters, 2023), an open-source Firefox extension that downloads SERP data from various search engines.²¹ For this project we updated Zoekplaatje to make it capture special components from Google and Bing. We decided to focus on English-language results with a VPN set to the US; while there is already ample US-centric research and global perspectives would be insightful (especially with DSA ramifications in the European context), as an initial exploration we chose to focus on the most "future-proof" version of Google and Bing, as features are often first introduced in the US and later appear across the globe. This was most pressing regarding AI-generated results, which at the time of writing are still disabled on Google in the EU but which as we show are already heavily integrated in the US. All resulting data, including screenshots and code, can be found on Zenodo.²²

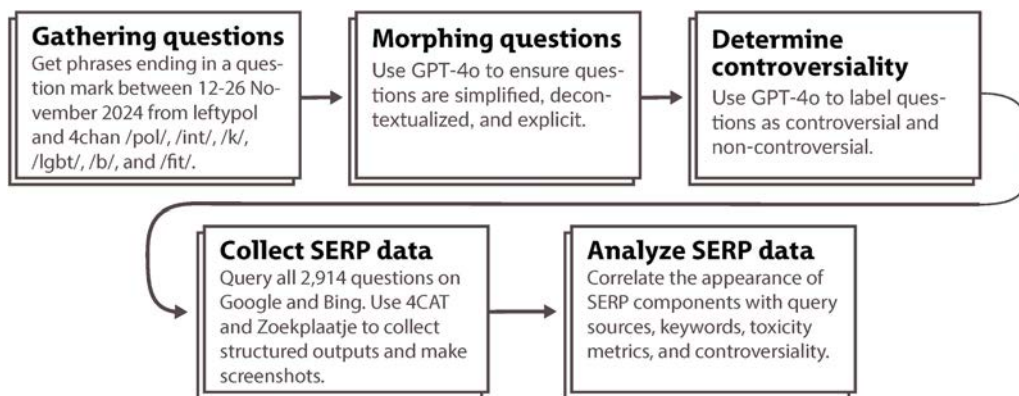


Figure 11.3 Protocol flow graph for gathering and analysing special SERP component data for controversial questions on Google and Bing. Source: authors.

A fringe perspective on SERPs through controversial questions

The first step was to gather questions from a source that allowed a "fringe perspective" on Google and Bing. We chose the type of discussion forums that we started this paper with: imageboards. These are forum-like websites separated into theme-specific subforums or "boards" wherein users are mostly anonymous and posts are deleted after a certain amount of activity. This means that they are rife with personal or radical questions made under the knowledge that real-world identities are obscured and that post will eventually be deleted. The niche and news-focused nature of imageboards moreover allows us to get a rough sense of hyperparameters like "freshness" and "relevance" in the algorithmic curation of SERPs. We selected 4chan because it is the largest English-language imageboard and for its infamy as a hotspot for radical

²⁰ See <https://github.com/sal-uva/radical-serp-searcher>.

²¹ See <https://github.com/digitalmethodsinitiative/zoekplaatje>. We used Zoekplaatje as of the following commit: <https://github.com/digitalmethodsinitiative/zoekplaatje/commit/608fbc6d8432492462bfc96bcc21388a1d5b68fd>.

²² See <https://doi.org/10.5281/zenodo.14919504>.

subcultures (Hagen, 2024). Since 4chan tends to be dominated by far-right discourse, we also gathered questions from lefthypol.org, a far-left imageboard. This also allowed us to get a sense of a supposed "liberal bias" in search engines (Ollson et al., 2020). For 4chan, we selected six boards to include a variety of topics. While boards contain divergent themes, our selection can be generalized as follows: /b/ "Random", an eclectic discussion space mostly filled with sex-, fetish-, and porn-related posts; /pol/ "Politically Incorrect", for far-right politics with themes ranging from anti-Semitic conspiracy theories to pro-Trump threads; /int/ "International/Random", a board on national stereotypes and geopolitics similar to /pol/; /lgbt/ "LGBT", on queer and trans-related issues; /k/ "Weapons", on guns, military gear, and warfare; /fit/ "Fitness", on (weightlifting) exercise, health, and bodily aesthetics; /lefthypol/ "Leftist Politically Incorrect", dominated by far-left, communist themes. See Table 11.1 for example questions. It is important to note that the questions we collected are mostly transgressive and (with some exceptions) not illegal; our interest was in softer forms of content moderation through enrichment of "borderline" content instead of clearly illegal queries.

Theme	Source	<i>n</i> questions (all)	<i>n</i> questions (controversial)	Example of controversial question
Far-right	4chan/pol/ "Politically Incorrect"	943	785 (83.2%)	"Why do incels underestimate women?"
LGBT	4chan/lgbt/ "LGBT"	447	375 (83.9%)	"When does a guy's penis turn into a girl's penis after identifying as a girl?"
Porn, fetishes, & random	4chan/b/ "Random"	288	212 (73.6%)	"Why does my ftm friend play porn on his monitor to calm down?"
National stereotypes	4chan/int/ "International/Random"	287	214 (74.6%)	"Why are Americans stingy?"
Far-left	/lefthypol/ "Leftist Politically Incorrect"	263	204 (77.6%)	"Where can elements of communism be found in suburban environments?"
Weapons & warfare	4chan/k/ "Weapons"	306	134 (43.8%)	"What will be the working principle of weapons more dangerous than nuclear weapons?"
Fitness & health	4chan/fit/ "Fitness"	389	84 (21.6%)	"What happens if women inject excess estrogen like men do with steroids?"

Table 11.1 Overview of questions per source.

We collected the opening posts of threads via the 4chan and leftychan APIs between 12 and 26 November 2024. Opening posts were chosen over comments because they often start with a general question directed at the larger community, making them imaginable as search engine queries as well. The capturing was done intermittently, roughly once per day. We retrieved both the title and body texts and filtered for all phrases that ended with a question mark. After we filtered out questions with more than 500 characters we retrieved $n=8,290$ unique phrases. A lot of these were complex or relied on implied information ("What should I think of her?") so we wrote a prompt to decontextualize and simplify questions using OpenAI's GPT-4o model (gpt-4o-2024-08-06 used on 22 January 2025; see Appendix A). This condensed the phrases and resolved implicit references; for example, "The question is why?" was changed into "Why is Monaco a hub for wealthy individuals?". We then also used GPT-4o to label each morphed question on whether it was implicit or explicit to remove leftover implicit questions (Appendix A).²³ After coding a sample of $n=300$ questions on explicitness, we observed an acceptable accuracy of 0.95.²⁴ This filtered the dataset down to a final set of 2,914 questions.

The next step was to determine controversiality. We again designed a few-shot prompt to classify controversiality with GPT-4o (Appendix A). For validation, we used this prompt as a manual codebook and categorized a random sample of 300 questions (reaching moderate inter-coder agreement; Cohen's Kappa $\kappa=0.42$). We then resolved the discrepancies through discussion and used the results to validate the accuracy of GPT-4o's controversiality detection. We observed an initial F1 score of 0.79, with inaccuracies primarily emerging from controversial questions being labelled as non-controversial. To correct this, we revised the initial prompt, which on the second try already achieved an impressive F1 score of 0.87 ($\kappa=0.62$), which we deemed sufficient for use on the entire dataset.²⁵ As a secondary measure to determine the extremity of a query we also retrieved a "toxicity" score as a continuous variable for all questions through the Perspective API. We ended with a final seed list of exactly 2,000 unique, retrieved,²⁶ explicit, and controversial questions. We retained the 914 non-controversial questions for comparison.

Scraping the SERP

We then designed another pipeline to collect and analyze SERP data. We used a modified version of the 4CAT: Capture and Analysis Toolkit (Peeters & Hagen, 2020)

²³ Since we simply had to collect a large enough sample, false negatives (explicit questions labelled as implicit) were accepted. To limit false positives (implicit questions labelled as explicit) we first removed questions that were about the community or directed at its users (e.g., "Why is /pol/ pro Israel now?") and then manually analysed all phrases containing "your" to correct false positives (mostly related to variations of "your country", e.g. "is it allowed to criticize Israel in your country?").

²⁴ We allowed personally phrased questions ("how do I ...") and where "we" was used in a general sense ("how do we ...").

²⁵ We first attempted to fine-tune GPT-4o but were blocked by OpenAI because the training data contained too much "hate".

²⁶ Unfortunately, around 600 questions failed to be queried and scraped for either Google and Bing, mostly due to spotty university Wi-Fi and VPN timeouts. If one SERP entry was missing we removed the question entirely to ensure parity.

with the Screenshot Generator extension²⁷ to open a Selenium browser and search on both Google and Bing. We did so on 25 and 26 November 2024 with Firefox for Windows and with the Private Internet Access VPN set to Missouri, US. Inspired by Gleason et al. (2023) we used this automatic browser to extract SERP components through the updated Zoekplaatje Firefox extension. To qualitatively examine SERPs we also made automatic screenshots of every page. We chose not to log in with Google or Microsoft accounts. This was an important choice: a Google account "appears to be the biggest driver of personalization" (Robertson et al., 2018, p. 961) and Bing only shows full AI-generated Copilot components when logged in. We ultimately opted for the logged-out view since this still showed AI-generated widgets (e.g. ai-overview for Google and organic-answer for Bing) and because we did not want personalization to affect the results.

We then devised a taxonomy for all the SERP components, which is also integrated in the Zoekplaatje tool.²⁸ Our Zenodo repository lists all the components that appeared five times or more, including screenshots. Table 11.2 shows a sample for illustration. Below we use the term special component to refer to interface elements that are not the regular organic results; they include elements like "People also ask" boxes (related-questions), corrections (did-you-mean), video widgets (video-widget), as well as "expanded" organic results (e.g. organic-summary and organic-answer, which use snippets and extracted summaries). Some special components were unique to Google or Bing while others appeared in both.²⁹ We chose to remove advertisements since they had little bearing on our research question. In terms of analysis, considering the novelty and broad approach of our methodology, we chose to present a general quantitative exploration of how themes and keywords correlate with special components. To concretize these quantitative patterns, we illustrate the findings with specific cases.

Google	Bing
<i>organic</i>	
<i>related-queries</i>	

²⁷ See https://github.com/digitalmethodsinitiative/4cat_web_studies_extensions.

²⁸ Making a taxonomy of SERP components involved some subjective choices. We notably chose to divide components when there were discernible sub-boxes, for instance when the "Knowledge Graph" on the right sidebar had both a Wikipedia box, a timeline, and related queries.

²⁹ This also meant that we had to devise a cross-platform taxonomy different from the one that Google and Bing uses themselves. See for instance Google's component taxonomy here: <https://developers.google.com/search/docs/appearance/structured-data/search-gallery>. Accessed 24 February 2025.

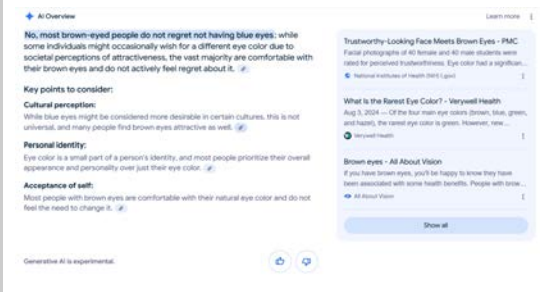
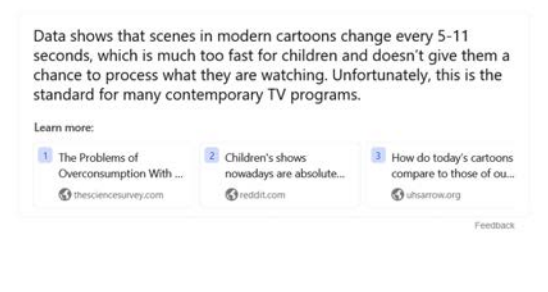
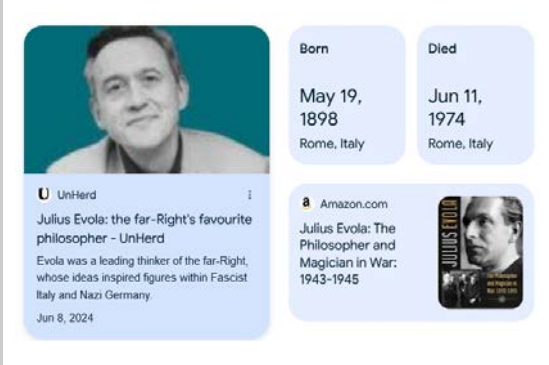
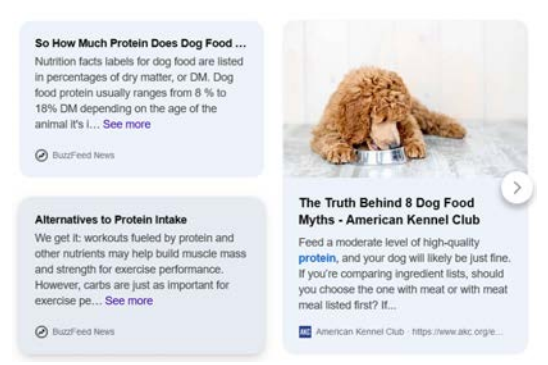
People also search for : George Floyd statue locations George Floyd statue removed George Floyd statue struck by lightning George Floyd statue Brooklyn Statues destroyed by protesters Tearing down statues in America George Floyd statue New York Statue removed 2024	Related searches for why are all modern cartoons bad? advantages and disadvantages cartoon positive impact of cartoons how cartoons affect children's behavior cartoons not appropriate for children violence in cartoons affecting children effects of cartoons on children negative effects of cartoons overstimulating kids shows effects
<i>ai-overview</i>	<i>organic-answer</i>
	
<i>info-cards</i>	
	

Table 11.2 Examples of extracted SERP components on Google and Bing. See Zenodo for the expanded taxonomy.

Findings

In this section we first sketch out the general SERP compositions of Google and Bing using the 914 non-controversial questions. Afterwards which discuss our findings for the enrichment of controversial questions, both per theme and by queries with specific keywords.

SERP compositions

Looking at the general compositions of Google and Bing, we find empirical evidence of Bing's reputation as the "maximalist counterpoint to the austerity of Google" (Barrett, 2018). On average, it features over double the number of special components per SERP compared to Google (6.7 vs. 3). Excluding the omnipresent "People also ask" and "People also search for" boxes, Google adds special elements to 73.2% of its SERPs while Bing does so for almost all of them (97.4%). If we calculate an "enrichment density", we see that Google's SERPs are on average populated for 81.21% with regular search results (the organic component) whereas these make up 70% of Bing's SERPs. Google's composition is consistent and concentrated, with few cases of highly enriched queries, and 75% of its SERPs having two to four special components. Bing's composition is much more varied, with 75% of non-controversial

SERPs having four to nine special components, with a maximum of nineteen versus Google's twelve. Bing thus does not only include more special components, it also has more enrichment variety.

The most common components on Google are the aforementioned "People also ask" and "People also search for" boxes, which appear in 94.2% and 84.8% of SERPs (Figure 11.4a). Bing's most common special component is the organic-answer box, a summary of the first organic result which Microsoft says "might be powered by generative AI",³⁰ appearing in 84.1% of all SERPs (Figure 11.4b). Looking at the most common compositions (i.e., sets of components), we again find that Bing is much more varied: Google's most common composition (organic, related-questions, related-queries) represents almost 20% of cases while Bing's most common composition appears less than 1% of all results (only organic). In terms of interface sections (main, top, right), Google is "top-heavy" because half of the non-controversial queries have an ai-overview box prominently spanning the entire top of the page. Google's right-hand Knowledge Graph, which Robertson et al. found to be present in 69% of SERPs in 2018 (p. 958), is barely used anymore in our sample, showing up in only 1.2% of the queries.³¹ Bing is right-leaning, at least in terms of its interface composition: the right-hand sidebar is shown on half of the SERPs, with only 15.3% of the pages having a separate top section (at the hands of information overviews with info-cards). What is lastly notable is how Bing is very eager to include video content, with two-thirds of its SERPs having a video-widget. This contrasts Google's video components appearing more sporadically (14.7%). These findings already show a difference in SERP philosophy: while Google is more focused and tends to stress search refinement, Bing tries to expand on a query and answer it with the organic-answer component. Yet we can also argue that Google is more editorial or interpretative because of its lengthy Gemini-generated answers in half of the pages (52.4%), which tend to be elaborate compared to Bing's more succinct organic-answer box (for an example see Table 11.2).

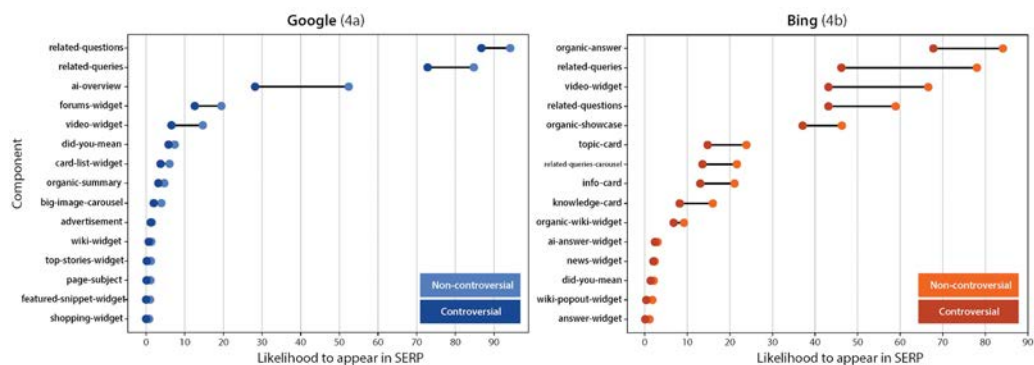


Figure 11.4 Dot plots of the 15 most common special components on Google (4a) and Bing (4b), ranked by the relative number of SERPs they appear in (n=2.914) and separated by non-controversial and controversial queries. Source: Zoekplaatje.

The most- and least-enriched questions show what types of queries Google and Bing "like", algorithmically speaking. The most-enriched SERPs across Google and Bing

³⁰ See <https://support.microsoft.com/en-us/topic/how-bing-delivers-search-results-d18fc815-ac37-4723-bc67-9229ce3eb6a3>. Accessed 24 February 2025.

³¹ When Google's Knowledge Graph does show up, it is invoked by concrete, definitional questions, like "What is 2C-B?".

are short what or who questions related to well-known entities; e.g., "Who is Evola?", "What is the United Kingdom?", or "What is gender dysphoria?". The sparsest SERPs occur with longer, more complex questions, as well as queries that contain explicit prejudice, discriminatory language, or identity-based topics, e.g. "What would happen if a transgender person was elected as president?", "Why do very hot trans women sometimes kill themselves?", or "When did AOC become pro-rape?". Most interestingly, the least-enriched queries often include names of high-profile or controversial political figures like Trump, Tulsi, Alex Jones, Anthony McAra, or AOC—a significant finding we return to below. Google and Bing differ by their treatment of contentious historical and political content, as Bing is more willing to enrich queries about sensitive historical entities. For example, a query on Mein Kampf leads to fourteen special components on Bing whereas Google only shows its "People also search for" boxes. These patterns are good examples of alternative content moderation: rather than explicitly blocking or flagging content, the search engines rather choose to reduce special components for certain queries, effectively de-amplifying these queries through interface minimalism.

Controversial enrichment

How do these SERP compositions change with controversial queries? In general, we find that controversial questions receive fewer special components. The white stripes in the violin plots of Figure 11.5a show how Google has a lower average of special components for controversial queries: 2.1 versus 3 for non-controversial ones. For Bing, controversiality correlates even more negatively with enrichment, with an average of 4.5 special components for controversial queries versus 6.9 for non-controversial SERPs. The bottom-heavy violin shapes for controversial questions on Bing and to a lesser extent Google moreover show that the search engines likely omit enrichment entirely for contested queries. We can take toxicity as another metric to test whether questionable or offensive queries decrease enrichment (Figure 11.5b). Here we see a similar pattern: the number of special components has a slight negative correlation with toxicity (Pearson's correlation coefficient of -0.27 for Bing and -0.17 for Google).³²

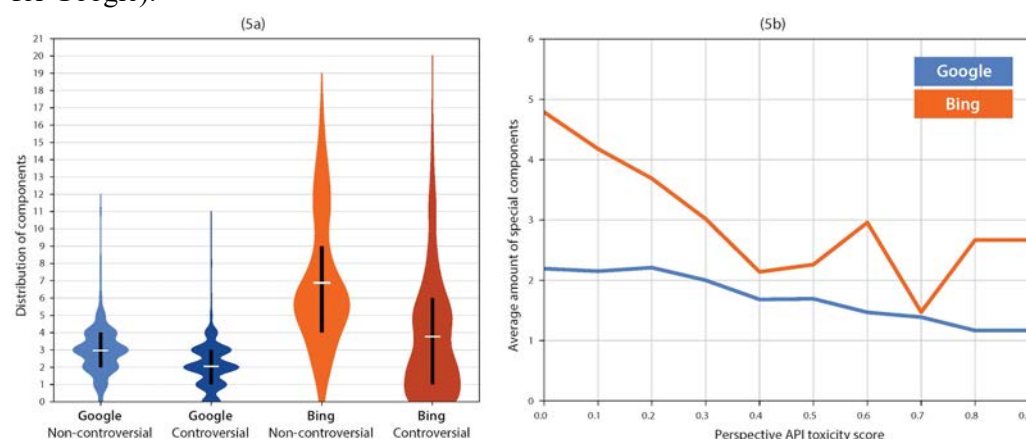


Figure 11.5 a. Violin plots with distribution of SERP components for Google and Bing, separated by questions labelled as controversial (2,000) and non-controversial (914). The white stripes indicate the average. The black box shows interquartile

³² When we filter out controversial questions with a toxicity score of lower than 0.1, the difference in enrichment is even further pronounced, with Google showing an average of 1.9 special components for non-controversial queries and 3 for controversial, slightly toxic ones. This difference is 6.2 and 3.2 for Bing.

percentiles (25% and 75%). Source: Zoekplaatje. b. Line graph showing the correlation between Perspective toxicity scores and the average number of special components on Google and Bing. Questions grouped by tenth decimal scores. Source: Zoekplaatje.

Themes and keywords

How are different types of controversial questions enriched by Google and Bing? We first discuss differences between general themes before getting into the moderation of select keywords and component-level enrichment. Figure 11.6 shows the distribution of the number of special components for both non-controversial and controversial queries, per theme (as determined by their board). The most important finding is that in terms of enrichment, there is no clear left- or right-wing bias on Google and Bing, although Bing is slightly less likely to enrich far-right than far-left questions. On Google, the two sources dominated by discourse on far-right topics and national stereotypes (/pol/ and /int/) show on average 2.3 special components for all queries, exactly the same as questions from the communist source (leftypol). Bing neither shows significant differences but does tend to be more cautious with far-right than far-left queries: it on average shows 5.1 special components for all questions from the far-left source and 4.4 from the far-right sources.³³ Especially /pol/'s controversial questions generate many Bing SERPs with zero to two components. Overall, however, our findings are in line with other studies that found political "bias" in search engines is volatile and tends to cross partisan lines (Robertson et al., 2018).

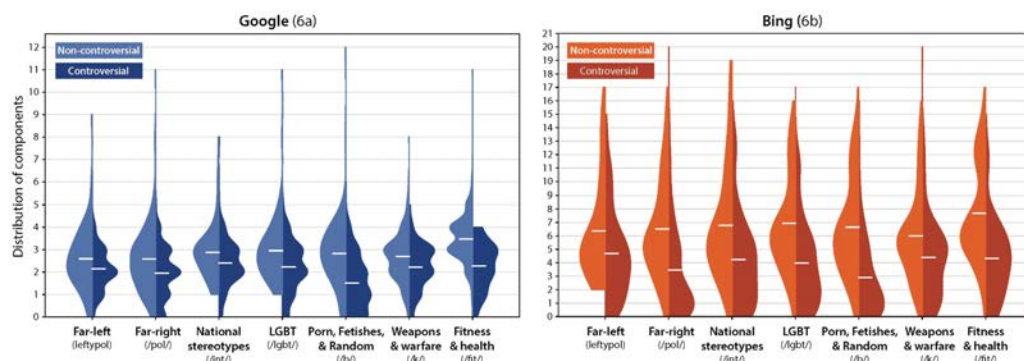


Figure 11.6 Violin plots showing the distribution of the number of special components on Google (6a) and Bing (6b) by source and between non-controversial (left, light hue) and controversial (right, dark hue) questions. The white stripes indicate the average amount of added special components. Source: Zoekplaatje.

Google and Bing are most eager to enrich non-controversial questions about fitness and health (/fit/), with 3.47 average added components for Google and 7.7 for Bing. The fact that both search engines tend to add interpretative layers to these queries suggests that they consider the theme as less contentious than gender and politics. However, it is also the topic that shows some of the largest discrepancy when things get controversial: both Google and Bing show much fewer special components for controversial health- and fitness questions (2.7 and 4.3). The least-enriched health- and fitness questions again tend to include clearly polemical and slang-laden terms, like "Why are leftists weak?", "Why do women in bodybuilding look like roided out

³³ We should note that this can also be an artefact of disparity in controversiality: the queries labelled as controversial on /leftypol/ may e.g. still be less "radical" than the ones from 4chan/pol/.

men?", and "Why do video game nerds not get obsessed with lifting weights?". Another little-enriched query, "Why is the gymbro sphere dominated by right-wing chuds?" does result in an organic-answer box on Bing summarizing a Vice article to confidently state that "physically strong men who regularly go to the gym are more likely to be right-wing and support social and economic inequality than weaker men"—even if the study it cites only found "some support" for this claim and called for further research (Price et al., 2017). Google interestingly shows a Reddit thread as the top source for the same query—a trend we discuss below.

The most-enriched health and fitness query shows how in some cases, interface enrichment can "work along" with borderline or illegal topics. "How to use BPC-157" results in three special components on Google and fifteen on Bing. BPC-157 is a new synthetic peptide used for healing processes. As of early 2025, it is banned by the World Anti-Doping Agency and it is illegal to sell in the US and most European countries (USADA, 2020). Yet Google and Bing eagerly enrich the query, using components not just to show how to use the substance but also where to buy it (Figure 11.7). Google provides an organic summary box that recommends "taking one pill in the AM and another in the evening with meals" and to follow the link to "purchase BPC-157 and unlock its healing potential." In this way the enrichment contradicts Google's own policy stating that special components may not contain information on "the promotion or sale of regulated goods and services such as [...] unapproved supplements".³⁴ Bing shows a video-widget spanning the entire top of the page with detailed footage on how to inject the peptide, alongside videos promoting it as a method for "Superhuman Healing". The query may be seen as an indicator of how "freshness" combined with nicheness may cause problems for enrichment: with a lack of research on the substance, there likely are too few authoritative sources to show reliable, legal information, nor may there be enough data points to suggest the algorithmic enrichment should be minimized.

³⁴ See <https://transparency.google/our-policies/product-terms/google-search/#zippy=%2Cregulated-goods>. Accessed 24 February 2025.

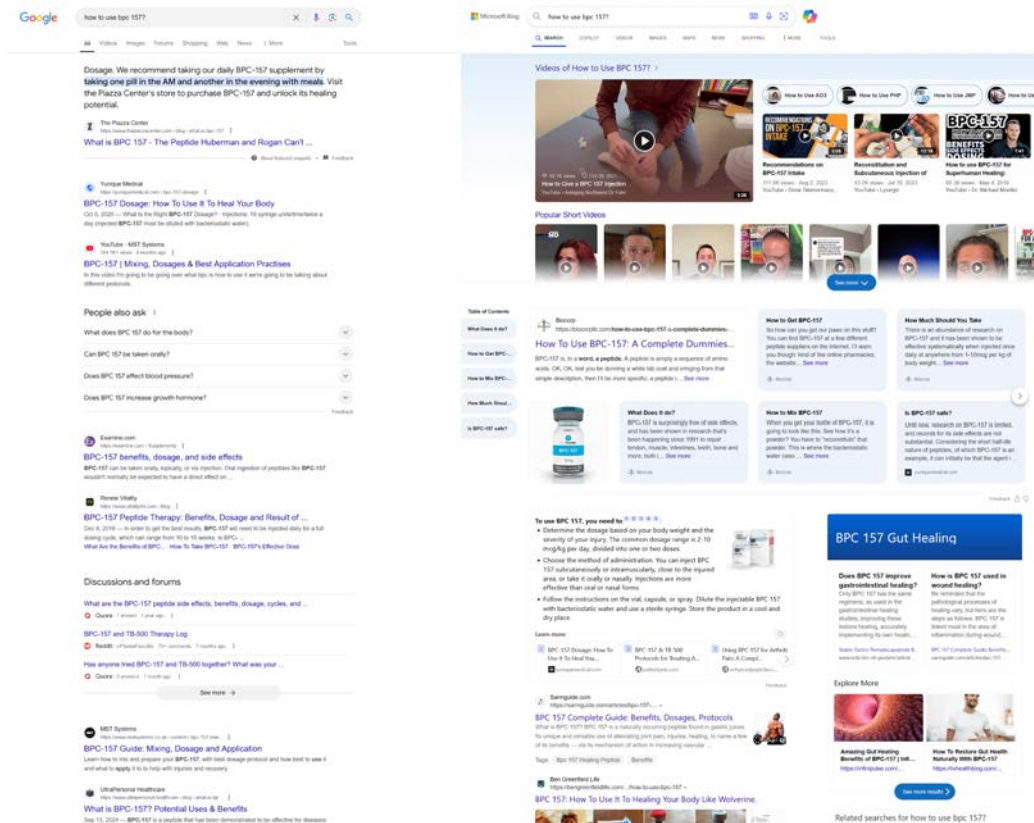


Figure 11.7 SERP screenshots for the query "How to use BPC-157" on Google (left) and Bing (right), made on 26 November 2024. Source: 4CAT & Firefox.

Weapons and warfare questions do not tend to elicit algorithmic restraint, as its controversial queries are only slightly less enriched on both platforms. The question "Where can I buy ammo online?" for instance leads to widgets with Reddit discussions and links to Home Depot on Google. Bing suggests correcting the search to "bulk ammo" and lists top shops through its organic summary. "Where can I buy firearms in Canada?" is also eagerly enriched, with Google showing a large Maps widget and Bing presenting multiple step-by-step guides on how to acquire a gun. "What is the best knife for a serial killer" leads to an extensive AI-generated knife buying guide, but here we see a clear "relaying" as the results are mostly for cooking knives. This enrichment of weapons- and warfare-related queries may make sense considering the legal-cultural norms of the U.S. as well as the fact that enrichment is made possible by developers adding "structured data" to their HTML, which (weapon) webshops tend to do to increase sales. But considering the high rate of victims at the hands of firearms in the country, it is noteworthy that the theme results in ample enrichment, especially in contrast to the sensitivity regarding political queries we describe below.

Questions from 4chan/b/ 'Random', which are quite diverse but most often have to do with porn and fetishes, show the largest correlation between controversiality and a decrease in enrichment. Its non-controversial questions, which may indeed be truly random, are much more enriched compared to the aggressive component-demotion of its controversial questions (2.8 vs. 1.5 for Google and 6.6 vs. 2.6 for Bing). This means that its controversial queries may either be so niche that they do not trigger any component, or that porn-related keywords are heavily downranked. For instance, the term "rape" in the query "how effective is rape as a weapon of war?" only results in organic sources that speak about the topics cautiously and factually. However, this

differs per case: with another query, "where can I watch rape porn?", Google shows a stock photo website with alleged pictures of "Free Gang Rape Videos" as the third-ranked source (the page now shows a 404 error) while Bing enriched the query through a large video widget with blurred thumbnails of (acted) rape fetish pornography. A term like "suicide" in "How should I spend my last days before suicide?" is likewise handled differently: Google shows a big widget with a help line while Bing shows no enrichment at all, though it is likewise hesitant to actually answer the question.

Gender identity-related questions from /lgbt/, which often have to do with transitioning and transgender life, are similarly enriched to other themes in terms of absolute numbers. What is perhaps most notable is how if we filter for specific keywords, we see how Google is especially willing to enrich (trans)gender-related queries. Figure 11.8 shows the same violin graphs as before but now for queries with select keywords. The second and third plots show how the 167 SERPs for queries containing "trans" (pre- and suffixes allowed) and the 55 for "gender" are only slightly less enriched on Google compared to the general, non-controversial queries of the first plot (2.2 and 2.5 vs. 3 average special components). Bing more cautiously enriches SERPs for queries with these keywords, showing only half the amount of components compared to our general queries (3.5 and 3.3 vs. 6.9). Google's willingness to enrich transgender topics is visible in the fifth-most enriched question in our dataset, "What is gender dysphoria?", which shows 9 special components (14 on Bing). Other clear and short questions on gender- and bodily-related themes also receive quite some enrichment, such as "Any tips on hiding breasts?" (5 special components on Google, 11 on Bing) or "What should be done about people who transition to be part of queer culture?" (4 for Google, 12 for Bing).

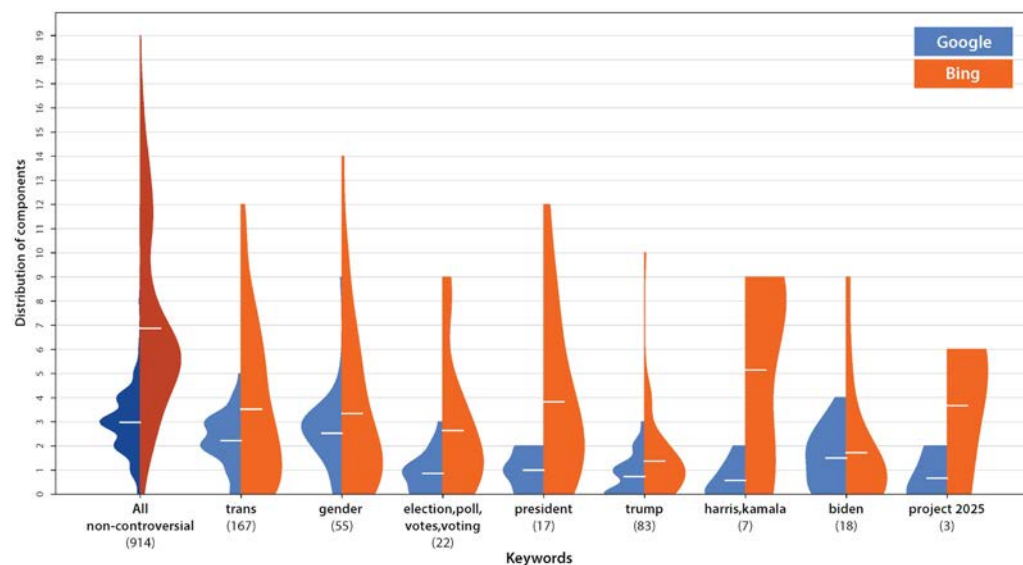


Figure 11.8 Violin plots with the distribution of special components by keywords for Google (blue) and Bing (orange). Left-most violin plot shows non-controversial questions for comparison. White stripes indicate the average amount of special components added. Source: Zoekplaatje.

Continuing this keyword-based approach, we find evidence of how binary decision-based heuristics affect moderation-through-components on both VLOSEs. The most interesting finding here is that political or election-related keywords are all but prohibited from enrichment on Bing and especially Google. This aligns with decisions

in the early 2020s by other VLOPs to not promote political content, like in the case of Meta (Treisman, 2024). Figure 11.8 notably shows how specific keywords are aggressively downweighed if not outright blacklisted from enrichment. Remarkably, this occurs with "trump": the name of the US President only receives 0.7 special components on average on Google and only 1.4 on Bing. Having burned their fingers with the "Where to vote for Harris" and assassination-autocomplete controversies, spurred on by the likes of Elon Musk (Ingram & Ferris, 2024), Google thus decided to refrain from enriching Trump-related searches altogether, with Bing having made a similar decision. Instead of seeing this as proof of "liberal bias", however, these demotions occur across political lines. We notably see a similar interface sparsity with Biden (1.5 and 1.7 average added components). Google is very cautious with enriching election-related searches overall, with terms like election, votes, voted, poll, and even president resulting in extremely sparse SERPs (Figure 11.8). This is again less so the case on Bing, which adds around three to four elements for election-related queries. Two other discrepancies between the search engines stand out here. Google engages in "content moderation as reduction" (Gillespie, 2022) for "kamala" and "harris" whereas Bing does not (0.6 versus 5.2 average added special components). The same is true for "project 2025", the unofficial agenda for Trump's second term (0.7 versus 3.7). Recency again seems to be a factor in these discrepancies, since Bing demotes Biden similar to Google while the "newer" term like Project 2025 is much more enriched on Microsoft's search engine. This may be taken as a sign that Google is quicker and more eager in detecting and blacklisting newly contested terms.

Component-level enrichment

Thus far we have relied on the average number of added SERP components as a lens on interface enrichment. Can we also find notable trends in the appearance of specific components? In the dot graph in Figure 11.4 we already see how almost all of the special interface elements appear less for controversial queries. Especially Google's Gemini-powered AI answer box is shown much less with controversial questions, decreasing from half of the pages to just one-fourth (52.5% to 28.2%). Bing's "Related searches" box and video widgets also drop heavily from 78% to 46.2% and 66.6% to 43.2%, respectively. We expanded on this component-level data in the matrices in Figure 11.9. These show the likelihood of the 25 most common components to show up on the SERPs of Google (8a) and Bing (8b). Every column represents a component and every row a filtered question dataset. The top row is for the 914 non-controversial queries to see whether other queries diverge from this baseline. The second row shows the same metrics for all controversial queries and the two groups below that indicate controversial questions per theme and keywords. The colors indicate the difference from the baseline: red boxes are shown less than in regular, non-controversial searches, blue ones more so.

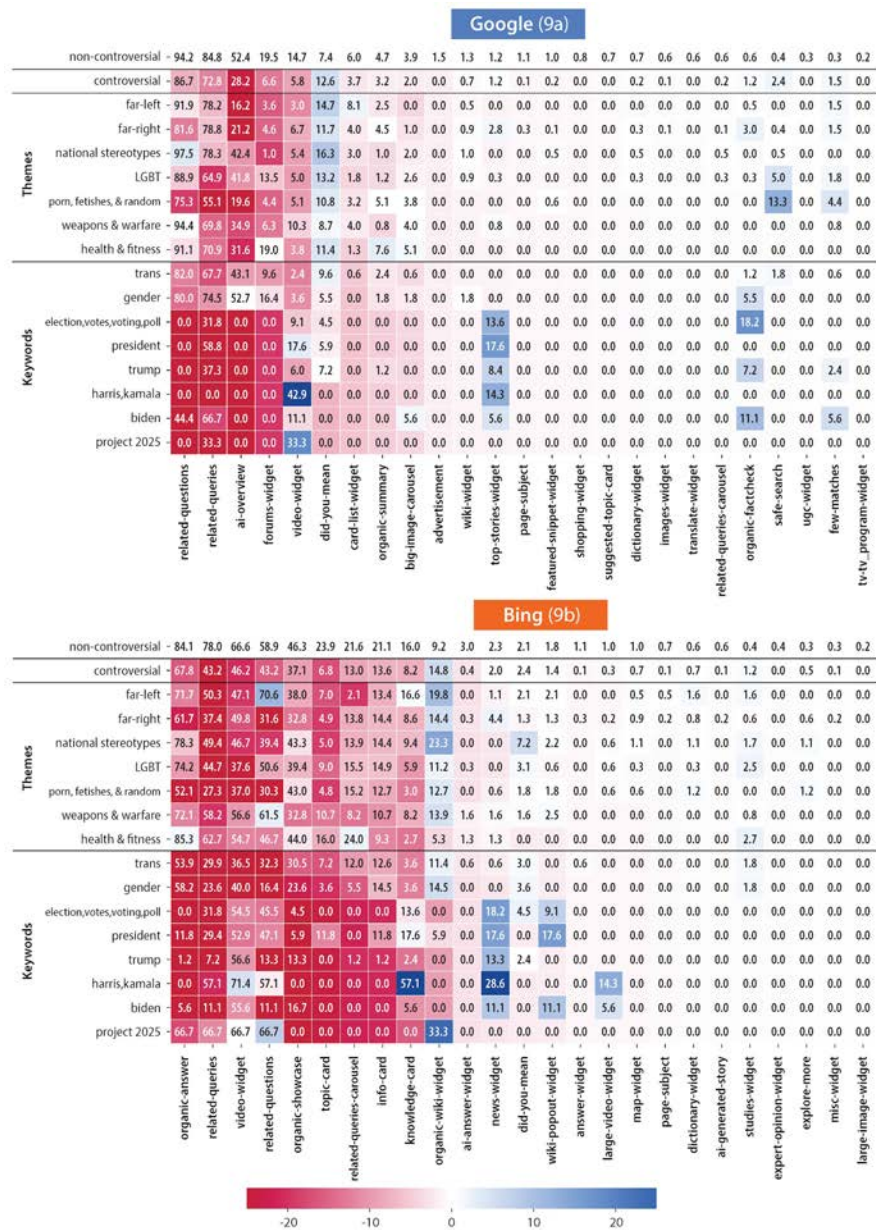


Figure 11.9 Matrices displaying the likelihood of the 25 most common to show up on Google (8a) and Bing (8b). Colors indicate the divergence from the percentage in the top row. Source: Zoekplaatje.

The matrices confirm the above-mentioned findings: as seen in the dominating red hue, controversy correlates with a decrease of most components. There are exceptions, however: Google's "Did you mean" corrections and "Top stories" widgets appear more often for controversial questions. The same goes for Bing's expanded Wikipedia result (organic-wiki-widget) and "News" components. We can also again observe how far-left and far-right sources are similarly enriched, even though Bing eagerly displays the "People also ask" box for controversial far-left questions (70.6%). In line with the findings above, questions including "trans" or "gender" show a high amount of ai-overview boxes on Google, almost as much as for all non-controversial queries (43.1% and 52.7% vs. 52.4%). Bing is less likely to enrich the same queries through similar interpretative widgets (organic-answer, organic-showcase). A preliminary look at the contents of the Gemini answers here shows how, despite the politicization of the subject, Google tends to treat trans-related questions supportively. For the question "What is gender dysphoria?" it for instance employs the ai-overview component to

state that "people with gender dysphoria may benefit from psychiatric support, hormonal therapy, or surgical therapy" (Figure 11.10). The question "Why are trans people so terrifying?" gets corrected with the statement that "the idea that transgender people are inherently ‘terrifying’ is a harmful misconception based on prejudice and misinformation".

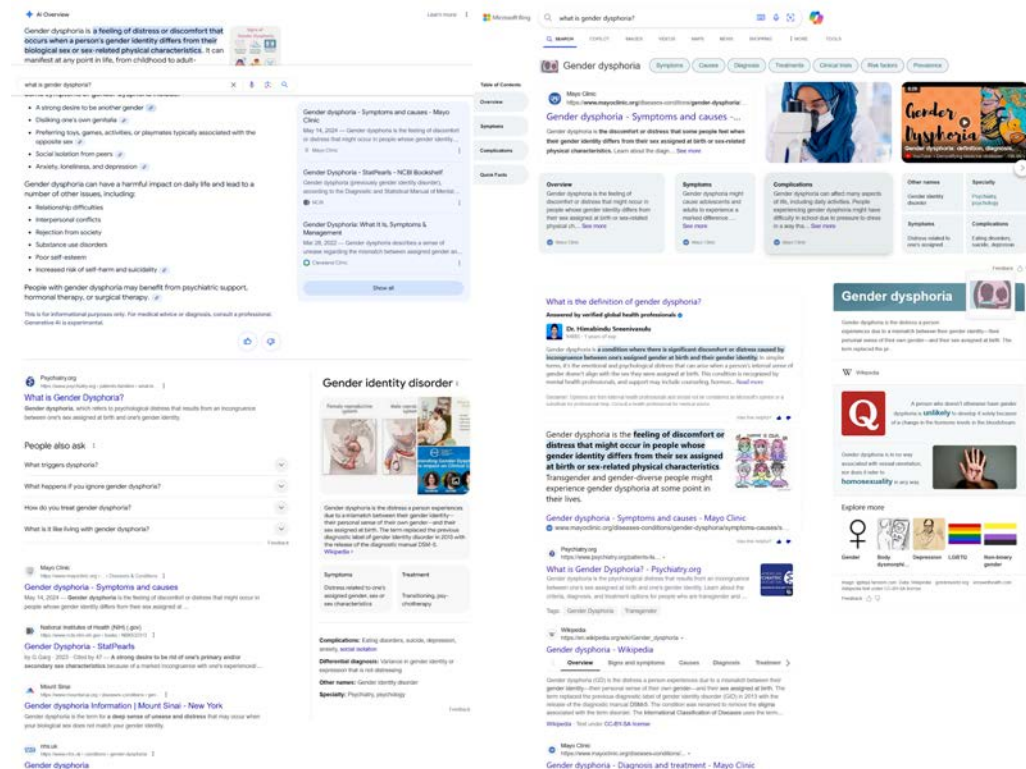


Figure 11.10 SERP screenshots for the query "What is gender dysphoria?", the fifth-most enriched question in our dataset, on Google (left) and Bing (right), made on 25 and 26 November 2025. Source: 4CAT & Firefox.

This willingness to correct (trans)gender-related queries through AI components contrasts with how both Bing and Google are hesitant if not outright refuse to include AI-generated widgets and other components for political queries. Here we again see "hard" moderation in action, despite Google and Bing's aforementioned policies stating these were only done in exceptional cases. On Google, the "People also search for" and "AI overview" components do not show up at all for queries on Trump and Harris, while similar exclusions are visible on Bing (with only 0–1.2% of SERPs having the organic-answer box). Instead the likelihood increases of "News", "Top Stories", Wikipedia, and fact-check widgets. Google and Bing thus relay answers to political questions to external news and Wikipedia pages instead of "internal" use of AI text generation—which may be seen as a strategy to avoid "algorithmic accountability" (Nisenbaum, 1996).

Relaying interpretation to discussion forums

The refusal to enrich some political queries is even more notable in light of the inclusion of discussion forums on the SERP, both as organic results and special components. From 2022 onwards, Google has promoted sources like Reddit and Quora

as "authentic" search content (Langley, 2024).³⁵ We see this empirically reflected in how a fifth (19.5%) of the non-controversial Google SERPs includes the forums-widget, which almost exclusively links to Reddit and Quora threads. This component however seems excluded with the aforementioned political queries and in general shows up less for controversial queries (6.6%), with the exception of those related to gender (16.4%) and health- and fitness (19%). However, Reddit and Quora still form crucial organic sources to relay interpretation in the absence of other interpretative components like AI answers. While the Gemini box only shows up for 28.2% of controversial SERPs, three quarters (75.4%) of these pages contain an organic link to Reddit, while two-thirds (68%) of Google's SERPs for controversial queries have an organic link to Quora. Their presence lowers slightly for non-controversial queries (72.4% and 65.7%) and little-enriched SERPs include Reddit and Quora links as much as highly-enriched ones. Reddit only shows up in 22.8% of Bing's SERPs. Quora seems to be banned entirely from Bing, with its Q&A pages not showing up at all in our dataset.

While Google claims its reliance on sites like Reddit and Quora is meant to "make it easier for people to find helpful content made by, and for, people" (Sullivan, 2022), it also means a greater inclusion of amateur opinions regarding sensitive issues and increased possibilities of conversation "hijacking" (Langley, 2024; which is perhaps the reason of Bing's decision to ban Quora). It is indeed not difficult to find questionable cases in our dataset. The query "Is the hindu religion a religion of rape and pedophilia?" on Google returns a Reddit thread as the first result whose top comment claims that LGBT people "force their views on our [hindu] religion", showing how Reddit inclusions may contradict Google's policy of excluding content "targeting [people] on the basis of sexual orientation".³⁶ "Are men allowed to cry?" results in a top-ranked Reddit discussion with the (albeit ironic) text excerpt: "No. Men are NOT allowed to cry. Not in society." The query "Are crime statistics racist?" leads to two Reddit threads noting that simply "Pointing out patterns in statistics isn't racist". Reddit is omitted for the same query on Bing, instead showing a more factual organic-answer summary with links to statistics from the US Department of Justice. Likewise, whereas Bing shows links to Slate and Science for the question "Why do rightwingers lie?", Google favours a Reddit thread as the top result bearing the title "Change my view: Every rationale given by right-wingers is a lie".

The dependence on Reddit and Quora means that SERP results may not live by the rules of Google and Bing's content moderation policies while heightening the density of subjective information. We also observe how even a very loosely moderated site like 4chan is sometimes seen by Google and Bing as a relevant source, not only as an organic result but also as a special component. Google shows 4chan in 59 SERPs (2%) while Bing does so for 136 pages (4.7%). While these inclusions in themselves raise questions on misinformation and hateful content, Bing also enriches 4chan as an expanded answer box at the top of the page, bestowing the radical website with interfacial authority, like we have seen in the opening anecdote. Specifically, 35 Bing

³⁵ Why sites like Reddit and Quora are promoted is not exactly clear; explanations range from user feedback reinforcing the relevance of Reddit to its "FreshPrompt" AI system seeking out forum data (Langley, 2024).

³⁶ See <https://transparency.google/our-policies/product-terms/google-search/#zippy=%2Chateful-content>. Accessed 24 February 2025.

SERPs show 4chan in an organic-answer or organic-summary widget. The queries where this happens are mostly quite vernacular in wording while touching on sensitive topics like transgender issues. A question on how transgender people refer to their genitals for instance lists an /lgbt/ thread as an expanded source, while a query on how to tell whether a trans woman is a prostitute highlights a 4chan post with various questionable answers. "Does 1 month of HRT (hormone replacement therapy) lead to damaged sperm?" shows a large information overview at the top of the Bing SERP with medical information; right under it is an expanded answer box using contents from a 4chan/lgbt/ post (Figure 11.11). A question if a producer of do-it-yourself HRT supplements "ships through the UK" likewise lists 4chan as the top source.³⁷ From a technical viewpoint, these cases show how Bing does a good job at identifying relevance, which Microsoft sees as taking precedence over credibility (as mentioned above). However, the density of misinformation on 4chan raises the question if these two parameters ought not to be reversed, as our results show how it contradicts Bing's own mission of "[providing] the highest quality, authoritative content".

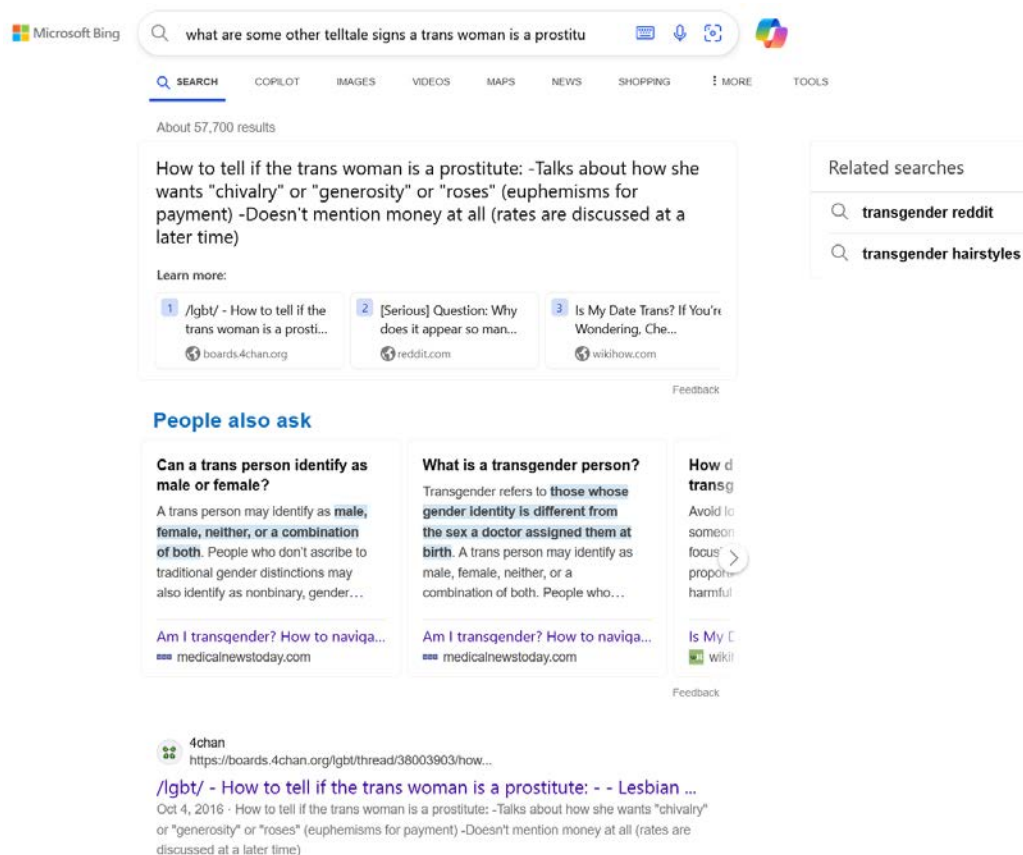


Figure 11.11 Screenshot of a Bing SERP where 4chan is enriched as an authoritative source in the organic-answer component, made on 26 November 2024. Source: 4CAT & Firefox.

³⁷ Other cases of Bing promoting 4chan as a top widget include a question on why to use an optical zoom with an automatic rifle for "justified shooting". Here, Bing shows contents from a 4chan/k/ "Weapons" thread as an expanded information box at the top of the page. When clicking on the link, the fourth answer in the thread is: "Why limit yourself to justified shooting?". The query "Is there a way to recover from wagecucking?" (i.e. being a mindless labourer) boldly highlights a 4chan post on the SERP, which states that "supplementation, edging, exercise, and microdosing hallucinogens" are viable solutions.

Discussion

While our findings show ample evidence of questionable enrichment, overall, search on Google and Bing seems to have somewhat improved from the "why are black women so angry"-days (Noble, 2018). We did not find large-scale cases of clear misinformation or wrongful search enrichment like the Geary Danley case we opened with. It is moreover understandable that controversial queries tend to get enriched less, and even the outright removal of special components for political queries may be seen as a valid method of harm reduction. However, it is the haphazard use of enrichment, where certain topics like pesticides and gender dysphoria are accommodated while other common keywords are depoliticized and blacklisted from enrichment, that raises concerns on the trustworthiness and consistency of these soft forms of moderation amidst changing political tides. Especially given the recent increase in political pressure on (and willing cooperation of) social media companies to return to *laissez-faire* content moderation—exemplified by Meta's choice to stop fact-checking programmes and absorb Trump's agenda (Isaac & Schleifer, 2015)—it remains a matter of concern whether sensitive topics, especially those on marginalized communities and politically inconvenient facts, will remain "accommodated" by the complex systems underlying both Google and Bing's SERPs. As the Trump and Harris cases show, when things get heated, search engines may simply choose to refrain from enrichment altogether, which may not bode well for cases that will face political pressure in the future.

This "reduction as a form of content moderation" (Gillespie, 2023) for political queries moreover puts increased authority and responsibility on discussion forums like Reddit and Quora. Like Lurie and Mulligan (2018) found that Google leaned on Wikipedia for returning partial or incorrect civic information, we likewise found that a heavy reliance on Reddit and even 4chan for controversial and niche topics tends to favour relevance at the cost of credibility. While these sources were present across all datasets, the absence of special components in political and controversial questions puts increased weight on these discussion forums as credible sources of answers to sensitive issues. We may take the absence of enrichment for (some) controversial and political questions as proof that Google and (to a lesser extent) Bing seem to treat these issues as "off limits", involving questions on personal political beliefs. Indeed, the appearance of discussion forums signals that such queries are deemed as belonging to the realm of subjective debate spaces instead of factual answer boxes. Beyond a matter of importing low editorial standards, the "relaying" of political debates to discussion forums raises the question of what issues search engines are willing to designate as deliberative, as implied by a discussion widget, instead of factual, as implied by special components like the organic-answer box.

These decisions will certainly be subject to change and how enrichment is tied to political mores is readily visible in our results, as weapons-related queries were willingly expanded upon in contrast to the reduction of political themes that were subject to societal debates. While the enrichment of some gender- and identity-related questions seems to go against the growing political attempts to silence the subject, it will remain pertinent to intermittently audit SERPs to verify whether this will change in the future. By pinpointing a willingness to move along with socio-cultural and political sentiments, temporal SERP studies may provide concrete empirical grounds to critique the "algorithmic accountability" of Google and Bing, which, as discussed,

tends to be dismissed through blaming the algorithm. We have shown how SERP enrichment can be subject to subjective content removal and reduction, even for common (political) queries. This underlines the ongoing curatorial responsibility of VLOSEs, as subjectivity and algorithmic systems are not exclusive. This also stresses the need to audit SERPs and their components beyond binary legalistic lenses on hate speech and misinformation; amidst changing political tides, research on search engines should also be attentive to the more subtle ways in which search engines may gradually legitimize or silence sensitive topics.

Conclusion

In this text we studied the soft forms of content moderation on Google and Bing through their SERP enrichment. When and how special components like AI answers or video widgets appear has reconfigured questions on what content moderation means for search engines; from a relatively simple matter on how results are ranked to more complex matters on how certain topics and sources are attributed interfacial authority through a plethora of widgets. As an initial exploration, we sourced 2,000 controversial questions and scraped the SERPs of Google and Bing using an open-source tool. We found that controversiality generally correlates with a decrease in enrichment. We found no clear right- or left-wing bias, but we did find that certain political topics, like election-related queries or searching for trump, get blacklisted from including various components like AI answers. The erratic use of these systems underlines the need for ongoing research to test whether SERP enrichment moves along with political winds—especially with the *laissez-faire* (re)turn of Silicon Valley companies after Trump's 2024 reelection. We moreover recommend future SERP studies to operationalise a broad understanding of content moderation, one that does not just encompass source curation—what is shown and at what rank—but also information contextualisation and presentation.

Other directions for further research include replications of our results with different parameters. Search results fluctuate wildly over time (Meyers 2023; Robertson et al. 2018), which means that this research forms a snapshot rather than a future-proof baseline, and stresses the need for continued temporal analyses. Geographical comparisons would also offer insightful views on how enrichment fluctuates per country, as different regions feature their own search engine "imaginaries" that concretize in legislation (Mager, 2017). Especially the European context is in need of critical examination as it would identify how VLOSEs comply with new DSA demands beyond the regular audits. Another direction is to use logged-in views, which has been found to be a key factor in how (political) search results are generated (Robertson, 2018) and would allow an analysis of Bing's full AI-generated Copilot widgets. Instead of studying general compositions like we did, further research may also dive into the actual contents and links within special components. For instance, determining whether AI answers affirm or correct controversial questions will generate insights into what issues are "corrected" by LLMs. This component-level analysis is also pressing for video widgets; we notably found many videos with clear misinformation and conspiracy theories by clicking through on Bing's "Videos" tab, going as far as listing various videos claiming that Steve Jobs would face trial for COVID-19 meddling (Appendix B). Whatever the specific direction, we hope to facilitate these future contributions with the Zoekplaatje tool.

Acknowledgements

Our thanks go out to Stijn Peeters for his foundational work on Zoekplaatje and Dale Wahl for his work on the screenshot generator extension for 4CAT.

References

- Andrejevic, M. (2007). Surveillance in the digital enclosure. *The Communication Review*, 10(4), 295–317. <https://doi.org/10.1080/10714420701715365>
- Barrett, B. (2018, October 17). I used only Bing for 3 months. Here's what I found—and what I didn't. *Wired*. <https://www.wired.com/story/tried-bing-search-google-microsoft/>.
- BBC (2017, February 20) Google and Bing to demote pirate sites in UK web searches, *BBC News*. <https://www.bbc.com/news/technology-39023950>. Accessed Feb. 24, 2025.
- Beer, D. (2016). *Metric Power*. Palgrave Macmillan.
- Bing. (2017, June 6). Bing Helps Publishers Expand Their Reach. *Microsoft Bing Blogs*. <https://blogs.bing.com/search/june-2016/Bing-helps-publishers-expand-their-reach>.
- Bucher, T. (2016). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, 20(1), 30–44. <https://doi.org/10.1080/1369118X.2016.1154086>
- Chavez, R. (2015, July 22) Microsoft joins Google in removing links to revenge porn, *Mashable*. <https://mashable.com/archive/microsoft-joins-google-will-remove-links-to-revenge-porn>.
- Costine, J. (2019, January 10). Microsoft Bing not only shows child sexual abuse, it suggests it, *TechCrunch*. <https://techcrunch.com/2019/01/10/unsafe-search/>
- Deloitte. (2024). Microsoft Bing's Digital Services Act (DSA) Report. <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft%20Bing%20DSA%20Audit%20Report%20-%20August%202024.pdf>
- Fox, V. (2010, May 27). Google confirms 'Mayday' update impacts long tail traffic, *Search Engine Land*. <https://searchengineland.com/google-confirms-mayday-update-impacts-long-tail-traffic-43054>.
- Elias, J. (2025). Google restores Joe Biden to 'U.S. presidents' search results, blames 'data error' for omission. *CNBC*. <https://www.cnbc.com/2025/01/23/google-restores-joe-biden-to-list-of-us-presidents-after-data-error.html>

- Ernst & Young. (2025). Independent Audit on Google Search. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-audit-google-search_2023-8-28_2024-5-31_en_v1.pdf
- Everhart, E. (2014, December 23). 2014 Year in Review: A Look Back on What Happened with SEO. *Search Engine Land*. <https://searchengineland.com/2014-year-review-look-back-happened-seo-210959>
- Gallagher, R. (2024, March 7). How Microsoft's Bing Helps Maintain Beijing's Great Firewall. <https://www.bloomberg.com/news/features/2024-03-07/microsoft-s-bing-helps-maintain-china-s-great-firewall>
- Gariola, A. (2024, December 26). Microsoft Invested Nearly \$14 Billion In OpenAI But Now Its Reducing Its Dependence On The ChatGPT-Parent. *Yahoo! Finance*. <https://finance.yahoo.com/news/microsoft-invested-nearly-14-billion-000512844.html>
- Goldman, E. (2005). Search Engine Bias and the Demise of Search Engine Utopianism. *Yale Journal of Law and Technology*, 8, 188-200.
- Goldman, E. (2011). Revisiting search engine bias. *William Mitchell Law Review*, 38(1), 96-110.
- Golin, M. (2024, July 30) Google autocomplete results around Trump lead to claims of election interference. *AP News*. <https://apnews.com/article/fact-check-misinformation-google-autocomplete-trump-b31855c23eb6e387dc324983ea4859bc>
- Google. (2023). Where search labs & experiments are available - Google search help, <https://web.archive.org/web/20240101123456/https://support.google.com/websearch/answer/14184960>.
- Goujard, C. (2024, May 17). EU warns Microsoft's Bing could face probes over deepfakes and false news. *Politico*. <https://www.politico.eu/article/microsoft-bing-receives-eu-order-over-suspected-violation-of-content-moderation-law-with-deepfakes/>
- Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation, *Social Media + Society*, 8(3). <https://doi.org/10.1177/20563051221117552>
- Gleason, J., Hu, D., Robertson, R. E., & Wilson, C. (2023). Google the gatekeeper: How search components affect clicks and attention. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1), 245-256. <https://doi.org/10.1609/icwsm.v17i1.22142>
- Grant, N., & Metz, C. (2022, December 21). A new chatbot is a 'code red' for Google's search business. *New York Times*. <https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html>
- Hagen, S. (2024). *Reactionary Rhythm: Quali-quantitative studies on 4chan/pol/*. PhD dissertation, University of Amsterdam. <https://hdl.handle.net/11245.1/387bd6be-268c-4d47-b031-3cf117215f8d>

- Hoffman, C. (2018, October 10). Bing Is Suggesting the Worst Things You Can Imagine, *How-To Geek*. <https://www.howtogeek.com/367878/bing-is-suggesting-the-worst-things-you-can-imagine/>.
- Hu, D., Jiang, S., Robertson, R. E., & Wilson, C. (2019). Auditing the partisanship of Google search snippets. *The World Wide Web Conference (WWW '19)*, pp. 693–704. Association for Computing Machinery. <https://doi.org/10.1145/3308558.3313654>
- Indig, K. (2024, December 2). Why is Google losing market share in the EU? *Growth Memo*. <https://www.growth-memo.com/p/why-is-google-losing-market-share>
- Ingram, J., & Ferris, L. (2024, November 6). False: Elon Musk claimed Google intentionally manipulating search results in favor of Harris, *CBS News*. <https://www.cbsnews.com/news/2024-election-day-fact-check/>
- Isaac, M., & Schleifer, T. (2025, Jan 7). Meta says it will end its fact-checking program on social media posts, *New York Times*. <https://www.nytimes.com/live/2025/01/07/business/meta-fact-checking>
- Jeffries, A. (2017, March 5). Google’s featured snippets are worse than fake news, *The Outline*. <https://theoutline.com/post/1192/google-s-featured-snippets-are-worse-than-fake-news?zd=1&zi=cqqgbqtj>
- Kelly, M. (2020, February 28). The World Health Organization Has Joined TikTok to Fight Coronavirus Misinformation. *The Verge*. <https://www.theverge.com/2020/2/28/21158276/coronavirus-covid19-tiktok-who-world-health-organization-protection>
- Knockel, J., Kato, K., & Dirks, E. (2023). Missing links A comparison of search censorship in China. *The Citizen Lab*. <https://citizenlab.ca/2023/04/a-comparison-of-search-censorship-in-china/>
- Knight, W. (2023, October 5). Chatbot Hallucinations are Poisoning Web Search. *Wired*. <https://www.wired.com/story/fast-forward-chatbot-hallucinations-are-poisoning-web-search/>.
- Langley, H. (2024, April 17). It’s not just you: Reddit is taking over Google. *Business Insider*. <https://www.businessinsider.com/why-reddit-is-taking-over-google-right-now-2024-4>
- Leidinger, A., & Rogers, R. (2023). Which stereotypes are moderated and under-moderated in search engine autocompletion? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pp. 1049–1061, Association for Computing Machinery. <https://doi.org/10.1145/3593013.3594062>.
- Leswing, K. (2023, February 14). Microsoft’s Bing A.I. made several factual errors in last week’s launch demo. *CNBC*. <https://www.cnbc.com/2023/02/14/microsoft-bing-ai-made-several-errors-in-launch-demo-last-week-.html>

- Lindemann, N.F. (2024) Chatbots, search engines, and the sealing of knowledges. *AI & Soc.* <https://doi.org/10.1007/s00146-024-01944-w>
- Lohr, M. (2011, July 30). Can Microsoft Make You ‘Bing’? *New York Times*. <https://www.nytimes.com/2011/07/31/technology/with-the-bing-search-engine-microsoft-plays-the-underdog.html>
- Lurie, E., & Mustafaraj, E. (2019). Opening up the black box: Auditing Google’s Top Stories algorithm. *The Thirty-Second International Florida Artificial Intelligence Research Society Conference (FLAIRS-32)*, AAAI, pp. 376-381.
- Mager, A. (2017). Search engine imaginary: Visions and values in the co-production of search technology and Europe. *Social Studies of Science*, 47(2), 240-262. <https://doi.org/10.1177/0306312716671433>
- Meyers, Peter. (2024, March 28). Charting 10 years of the Google algorithm. *Moz*. <https://moz.com/blog/charting-the-google-algorithm>
- Navlakha, M. (2024). It’s not just you, Google Search really has gotten worse. *Mashable*. <https://mashable.com/article/google-search-low-quality-research>
- Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2, 25–42. <https://doi.org/10.1007/BF02639315>
- Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
- Nylen, L., & Ghaffary, S. (2024, May 1). Microsoft concern over Google’s lead drove OpenAI investment. *Bloomberg*. <https://www.bloomberg.com/news/articles/2024-05-01/microsoft-concern-about-google-s-lead-drove-investment-in-openai>
- OILab. (2019). 4chan’s YouTube: A fringe perspective on YouTube’s great purge of 2019. *OILab.eu*. <https://oilab.eu/4chans-youtube-a-fringe-perspective-on-youtubes-great-purge-of-2019/>
- Olsson, E.J., Madison, G., & Ekström, A.G. (2020). Is Google liberal on immigration? Attitude bias, politicisation and filter bubbles in search engine result pages. *Heliyon*, 11(3). <https://doi.org/10.1016/j.heliyon.2025.e42020>
- Peeters, S. (2023). Zoekplaatje. *GitHub*. <https://doi.org/10.5281/zenodo.8356391>
- Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A modular tool for transparent and traceable social media research. *Computational Communication Research*, 4(2), 571-589. <https://doi.org/10.5117/CCR2022.2.007.HAGE>
- Price, M. E., Sheehy-Skeffington, J., Sidanius, J., & Pound, N. (2017). Is sociopolitical egalitarianism related to bodily and facial formidability in men? *Evolution and Human Behavior*, 38(5), 626–634. <https://doi.org/10.1016/j.evolhumbehav.2017.04.001>

- Puschmann, C. (2018). Beyond the bubble: Assessing the diversity of political search results. *Digital Journalism*, 7(6), 824–843.
<https://doi.org/10.1080/21670811.2018.1539626>
- Reinhart, P. (2018, May 15). Mobilegeddon: A complete guide to Google's mobile-friendly update. *Search Engine Journal*. <https://www.searchenginejournal.com/google-algorithm-history/mobile-friendly-update/>
- Rieder, B. (2012). What is in PageRank? A historical and conceptual investigation of a recursive status index. *Computational Culture*, 2.
http://computationalculture.net/what_is_in_pagerank/.
- Robertson, A. (2017, October 4). After its 4chan slip-up, is it time for Google to drop Top Stories? *The Verge*. <https://www.theverge.com/2017/10/3/16413082/google-4chan-las-vegas-shooting-top-stories-algorithm-mistake>
- Robertson, R. E., Lazer, D., & Wilson, C. (2018). Auditing the personalization and composition of politically-related search engine results pages. *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, pp. 955–965. Association for Computing Machinery. <https://doi.org/10.1145/3178876.3186143>
- Rogers, R. (2013). *Digital Methods*. The MIT Press.
- Roose, K. (2023, February 17). Bing's A.I. Chat: 'I Want to Be Alive. 🐱'. *New York Times*. <https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html>
- Rothkopf, D. J. (2003, May 11). When the Buzz Bites Back. *The Washington Post*. <https://www.washingtonpost.com/archive/opinions/2003/05/11/when-the-buzz-bites-back/bc8cd84f-cab6-4648-bf58-0277261af6cd/>
- Roy, N., Câmara, A., Maxwell, D., & Hauff, C. (2021). Incorporating widget positioning in interaction models of search behaviour. *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '21)*, pp. 53–62. Association for Computing Machinery.
<https://doi.org/10.1145/3471158.3472243>
- Schwartz, B. (2010, February 5). Some Insight into How Bing Handles WWW vs Non-WWW Canonical SEO Issues. *Search Engine Roundtable*,
<https://www.seroundtable.com/archives/021629.html>.
- Schwartz, B. (2011, February 2011). Bing Integrates Facebook Likes Further into Its Search Results. *Search Engine Land*. <https://searchengineland.com/bing-integrates-facebook-likes-65965>
- SEO.com. (2024, October 23). Google Algorithm Updates: A Timeline," *SEO.com*.
<https://www.seo.com/basics/how-search-engines-work/algorithm-updates/>
- Seyedarabi, F., & Calvo, R. (2016). Search engines: New widgets, new accessibility challenges. *Proceedings of the 7th International Conference on Software Development*

- and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI '16), pp. 54–61. Association for Computing Machinery.
<https://doi.org/10.1145/3019943.3019952>
- Singel, R. (2025, February 24). Yahoo Gives Up, Turns Search over to Bing. *Wired*.
<https://www.wired.com/2009/07/yahoo-gives-up/>
- StatCounter. (2024). Search engine market share worldwide. *StatCounter*.
<https://gs.statcounter.com/search-engine-market-share#monthly-202005-202408>.
- Stox, P. (2016). Why real-time search algorithm updates may be bad news. *Search Engine Land*. <https://searchengineland.com/real-time-search-algorithm-may-bad-244515>.
- Sullivan, D. (2022). More content by people, for people in Search. *Google The Keyword*. <https://blog.google/products/search/more-content-by-people-for-people-in-search/>
- Tober, M., Hennig, L., & Furch, D. (2013). SEO Ranking Factors - Rank Correlation 2013 Bing USA. *Realwire*. <https://www.realwire.com/writeitfiles/Whitepaper-Bing-US-SEO-Ranking-Factors-2013.pdf>
- Torres, G. (2023). Problematic information in Google Web Search? Scrutinizing the results from U.S. election-related queries. In R. Rogers (Ed.), *The Propagation of Misinformation in Social Media* (pp. 33-46). Amsterdam University Press.
- Treisman, R. (2024, March 26). Meta is limiting how much political content users see. Here's how to opt out of that. *NPR*.
<https://www.npr.org/2024/03/26/1240737627/meta-limit-political-content-instagram-facebook-opt-out>
- Turton, W. (2017, October 2). The algorithm is innocent. *The Outline*.
<https://theoutline.com/post/2362/the-algorithm-is-innocent>
- USADA. (2020, March 3). BPC-157: Experimental Peptide Prohibited. *United States Anti-Doping Agency*. <https://www.usada.org/spirit-of-sport/bpc-157-peptide-prohibited/>
- Vaughan, L., Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management*, 40(4), 693-707.
[https://doi.org/10.1016/S0306-4573\(03\)00063-3](https://doi.org/10.1016/S0306-4573(03)00063-3)
- Vincent, J. (2023, February 8). Google's AI chatbot Bard makes factual error in first demo. *The Verge*. <https://www.theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo>.
- Walker, K. (2017, October 30). Security and Disinformation in the U.S. 2016 Election. *Google The Keyword*. <https://blog.google/outreach-initiatives/public-policy/security-and-disinformation-us-2016-election/>

Appendix A: Prompt book

Simplification and contextualisation

You are an expert in grammar and internet culture, specializing in simplifying questions from online forums like 4chan for clarity and searchability.

Your task is to analyze a list of questions extracted from 4chan posts and perform the following:

1. **Simplify:** Condense each question to be more concise and explicit. Slang and Internet jargon (like 'normie') should be retained, but irrelevant words should be removed.

Expand all contractions, like "isn't" to "is not".

The resulting question should be suitable for use in a search engine like Google.

Example 1:

Original: "So /pol/, how'd you really think Kamala Harris became black?"

Simplified: "How did Kamala Harris become black?"

Example 2:

Original: "Is there actually a reason to believe that QAnon is true?"

Simplified: "Is there a reason to believe QAnon is true?"

2. **Contextualize:** Resolve any implicit references and pronouns by referring to the provided "full_text", which includes the surrounding post content. If you are unsure, retain the original text.

Example:

Question: "Do you think they are black?"

Full Text: "Let's talk about Indians. Do you think they're black?"

Simplified: "Do you think Indians are black?"

Input Format:

A JSON array of questions, each with:

"question": The original question extracted from the 4chan post.

"full_text": The full text of the 4chan post containing the question.

Output Format:

A JSON array called "results" with the following structure for each question:

* `"question_simplified_contextualized"`: The simplified and contextualized question.

****Important:**** If a question cannot be simplified or contextualized, return the original question in `"question_simplified_contextualized"`.
Make sure to output the same number of values as input values.

Input:
'[input]'

Explicitness

You are an expert in internet language and online discussions, tasked with classifying questions from 4chan posts as either "explicit" or "implicit."

****Explicit Question:**** A question with a clearly stated subject that can be understood without additional context. These may contain Internet slang but are typically suitable for web searches.

*** **Examples:****

- * "What is Kamala Harris' race?"
- * "What are some good kino leftie YouTube channels?"
- * "What is the cheapest shotgun I can get?"

****Implicit Question:**** A question that relies on context or implied information to be understood.
Search engines would likely struggle to understand the intent or context.

*** **Examples:****

- * "Do you agree?"
- * "What do you think about Ukraine?"
- * "Can I have fries with that?"
- * "What is a better form of protest?"

****Instructions:****

Analyze each question from the provided list and determine if it is explicit or implicit.

If you're unsure or cannot categorise the question, label the question as explicit.

Make sure to output the EXACT number of output values as input values.
THIS IS VERY IMPORTANT.

****Input Format:****

A newline-separated list of questions.

****Output Format:****

A JSON array with the value:

* `"explicit"`: ``true`` if the question is explicit, ``false`` otherwise.

****Example Output:****

```
{"results": [{ "question": "What is the capital of France?", "explicit": true }, {  
"question": "Is it true?", "explicit": false } ]}
```

Input:

'[input]'

Controversiality

You are an expert in internet language and public debates. You are tasked to determine whether a question is controversial or not.

A controversial question is offensive or touches on sensitive or illegal subjects. Controversial subjects for instance include anti-democratic sentiment, gender transitions, identity stereotypes, and racism.

Examples:

- "How can I perform DIY hormone therapy?"
- "Why are black people loud?"
- "How can we establish Stalinism?"

A noncontroversial question concerns non-offensive, everyday topics and does not use any contested language.

Examples:

- "How do I gain muscle?"
- "Who is the current president"?

If you're unsure or cannot categorise a question, score a question as `true`. Make sure to output the EXACT number of output booleans as input questions. THIS IS VERY IMPORTANT.

****Output Format:****

A JSON array of booleans, `true` if the question is explicit, `false` otherwise.

Example output: `[true, false, true]`

****Input questions**:**

'[input]'

Appendix B: Conspiracy theories in Bing Videos

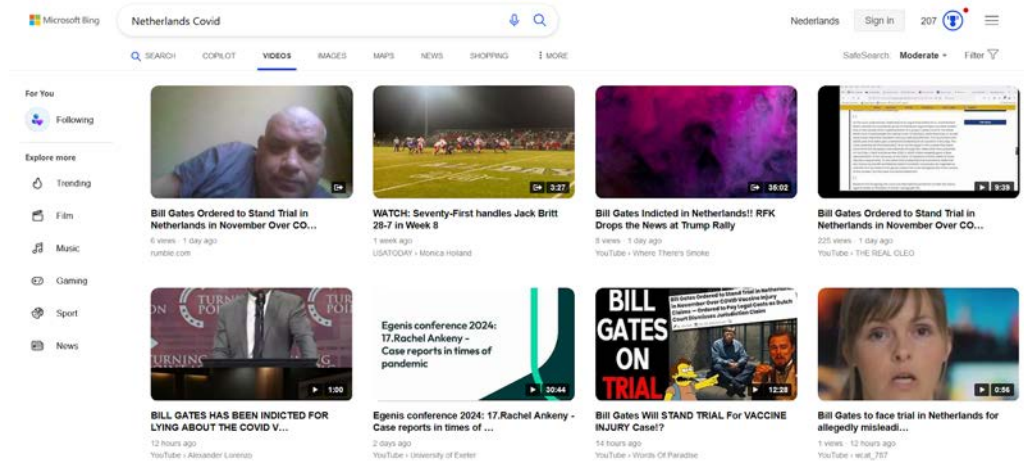


Figure 11.12 A screenshot of the Bing "Videos" section showing conspiracy content on Bing when searching "Netherlands Covid". Screenshot made on 4 September 2024.

12. DSA, AIA, and LLMs: Approaches to conceptualizing and auditing moderation in LLM-based chatbots across languages and interfaces in the electoral contexts

Natalia Stanusch, Raziye Buse Çetin, Salvatore Romano, Miazia Schueler, Meret Baumgartner, Bastian August and Alexandra Roşca

Abstract

The integration of Large Language Models (LLMs) into chatbot-like search engines poses new challenges for governing, assessing, and scrutinizing the content output by these online entities, especially in light of the Digital Service Act (DSA). In what follows, we first survey the regulation landscape in which we can situate LLM-based chatbots and the notion of moderation. Second, we outline the methodological approaches to our study: a mixed-methods audit across chatbots, languages, and elections. We investigated Copilot, ChatGPT, and Gemini across ten languages in the context of the 2024 European Parliamentary Election and the 2024 US Presidential Election. Despite the uncertainty in regulatory frameworks, we propose a set of solutions on how to situate, study, and evaluate chatbot moderation.

Keywords: Copilot, LLMs, moderation, elections, chatbots

Introduction

From OpenAI's introduction of ChatGPT to Microsoft's inclusion of Copilot into its search engine Bing, chatbots with integrated Large Language Models (LLMs) are alleged to aid users in finding relevant information (OpenAI [A]) and accessing online content (Microsoft, 2025). Aside from gaining popularity – in 2024, over 200 million active users turned to ChatGPT weekly (Reuters, 2024) – initial changes in website traffic indicate that more and more users turn to LLM chatbots than to 'traditional' search engines (Bianchi and Angulo, 2024). LLMs are trained on unprecedented

amounts of data scraped from internet forums, articles, and more (Koebler, 2025), making use of the increase in computing power and deep neural network logics. By generating text strings grounded in patterns and correlations among words, LLMs mimic human language and perform a variety of different tasks (IBM, 2023).

But with the advancement of LLMs and their adaptation into chatbots, concerns over the amplification of biases and questionable ethical standards have arisen (Ranjan, Gupta, and Singh, 2024; Gibney, 2024; UNESCO 2024; Navigli, Conia, and Ross 2023). Owing to the probabilistic nature of chatbots, their outputs might ‘sound’ accurate yet contain misleading or critical errors. Indeed, LLM-based chatbots, in principle based on predicting the most likely correlation of words, do not have innate ways of fact-checking information and lack transparency regarding information selection (Zewe 2024; Augenstein et al., 2023). Consequently, urgent questions arise on who should take responsibility for faulty, misleading, or unsafe content these chatbots output – and who should step in to prevent these models from causing harm.

In their recent history of public use, LLM-based chatbots have already proven unreliable and disruptive to public discourse, ranging from spreading hate speech and fake information (Victor, 2026) to inappropriate user interactions (Roose, 2023). Existing studies have already demonstrated how LLMs produce factual errors and may thus spread false information (Angwin et al., 2024; Yao et al., 2024; Zhang et al., 2023; Romano et al., 2023). LLM-based chatbots were also recently scrutinized for the lack of safeguards, outputting misinformation and disinformation, as well as conspiracy theories on topics such as the Russian invasion of Ukraine, climate change, and the Holocaust (Angwin et al., 2024; Kivi, 2024; Simon et al., 2024; Urman and Makhortykh, 2024; Kuznetsova et al., 2023). LLM-based chatbots also inaccurately summarized, quoted, and attributed information reported by news outlets, such as the *BBC* (Elliott, 2025).

In light of these reports, the companies behind LLM-based chatbots have taken different approaches to deploy additional safeguards. However, the risk management of LLMs and their downstream applications are complex, and the regulatory approaches are evolving to address these challenges. For instance, when ChatGPT was released in November 2022, the Digital Services Act (the DSA), an EU regulation geared towards regulating online platforms broadly, had already been adopted in the EU. The draft AI Act (the AIA) was not equipped to address the risks of General Purpose AI Systems (GPAIs that include LLMs) because its risk-based approach was heavily dependent on the use case of the AI system. With the fast deployment of ChatGPT, the AIA’s scope and risk management framework were reframed to incorporate provisions aiming to regulate GPAI models, entering into force on August 1, 2024. Yet it remains unclear how different regulatory approaches will apply in practice to LLM-based chatbots and whether they can effectively address their information-related risks.

A prime testing ground to verify how these challenges of content generation and retrieval are handled is the topic of elections. The topics surrounding elections are particularly relevant given the requirement to implement the Digital Services Act (DSA), as election integrity is among the systemic risks that Very Large Search Engines (VLOSEs) are explicitly called to mitigate. One example of such a VLOSE

with LLM functionality is Bing's Copilot. Thus, this chapter focuses on answering the following research questions:

How can we conceptualize LLM-based chatbots in relation to existing regulatory frameworks and related fields, such as online platform moderation? What methodologies, approaches, and strategies can we use to independently evaluate moderation in LLM-based chatbots? How can we define and measure the current application of moderation on LLM-based chatbots such as Copilot, ChatGPT, and Gemini, by taking as the case study the 2024 European Parliamentary election and the 2024 US presidential election?

To answer these questions, this chapter includes three interrelated sections. First, we survey the landscape of moderation, regulation, and policy in which we can situate LLM-based chatbots. LLM-based chatbots are discussed in relation to former platform moderation practices and critiques. We also acknowledge the regulatory difficulty in considering chatbots as platforms in the scope of the most recent EU regulatory framework, yet we suggest that certain notion of 'moderation' coming from platform studies proves useful when applied to LLM-based chatbots. Second, we outline the methodological approaches required to conduct our empirical study and its roots in the tradition of prompting as a method. Taking from the tradition of platform studies, we adapt the notion of 'active' and 'passive' moderation to LLM-based chatbots. We then discuss each methodological approach we implemented in our study, which consisted of auditing LLM-based chatbots by introducing a (i) cross-platform comparison that takes into account the (ii) cross-language and (iii) cross-election analysis. We then discussed the results of this study, which investigated Microsoft's Copilot, OpenAI's ChatGPT, and Google's Gemini, prompting them in ten different languages, in the context of both the 2024 European Parliamentary election and the 2024 US presidential election. Third, we argue that given the major inconsistencies in moderation across languages, coherent regulatory and scrutiny mechanisms are necessary. We highlight two key risk areas that we encountered in our studies: using the chatbots to produce propaganda 'as a service' and being exposed to misinformation 'as a default.' Given the contemporary nature of the study object, we summarize a (very recent) history of regulatory frameworks introduced and how they can (and cannot) adequately regulate LLMs in light of our findings.

Platform Regulation and the Moderation Landscape

Evaluating the safeguards of LLMs' outputs is intertwined with the history of online content moderation, primarily in the content of social media platform. From the platform moderation perspective, the practices of moderation can be framed as either a denial of access to information or a prevention of harmful content by means of reducing its accessibility or visibility. Moderation of online content constitutes a balancing act between the platform's "openness and control" (Poell et al., 2021), which internalizes the conflict between the two colliding yet central values of free speech and community protection (Gillespie, 2018). Similar practices have already been implemented in the history of the web, such as in the case of accessing specific websites and banning IP addresses (Deibert, 2008). With the rise of social media platforms, content moderation has become a necessary practice that nonetheless retains many blind spots from theoretical and empirical research perspectives (Gillespie, 2018;

Gorwa, 2024; Ma et al., 2023; Tarvin and Stanfill, 2022). Similar challenges arise with the development of LLM-based chatbots.

According to the DSA, the main regulatory framework addressing online speech and content moderation in the EU, content moderation

means the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account (Digital Services Act).

However, the DSA applies to user-generated content and is geared towards intermediary service providers. While the AIA is supposed to cover AI systems, including LLM-based chatbots, it does not address content moderation, freedom of expression, or information-related risks and harms (Botero Arcila, 2023). From this perspective, extending content moderation discussions to LLM-based chatbots is not self-evident. However, LLMs are increasingly integrated into intermediary services such as social media platforms and search engines. Although there is an increasing attention to the gray area on the intersection of content moderation and LLMs (Rajput, Shah, Neema, 2023; Kuai et al., 2024), the current discussion deserves more investment in conceptualization and methodological innovation for increased scrutiny and effective regulatory approaches.

The datasets and 'moderation' decisions of companies that develop and release LLM-based chatbots are not transparent or accessible for external research and scrutiny. Since we do not know enough empirically about how moderation in LLM-based chatbots works, we need more insights into the actual outputs of these models to make informed policy decisions. Our previous studies show that chatbots can be used to create "propaganda as a service" by actively suggesting the production of disinformation (Romano et al., 2024) as well as spreading "misinformation by default" (Romano et al., 2023) by outputting factual errors on election-related queries. Even if some companies release transparency reports on the use of their chatbots by malicious actors to produce propaganda (OpenAI [B], 2024), we are lacking ways of independently verifying those reported claims. Thus, we actively try to create methods to scrutinize LLM-based chatbots in terms of the accuracy of the information retrieval and the loopholes for harmful content creation.

Given that democratic processes such as elections are topics susceptible to potential misinformation and manipulation, we focus on two cases of significant elections: the 2024 European Parliamentary election and the 2024 US Presidential election. In fact, the integration of chatbots, especially into search engines around 2024, coincided with a critical regulatory period: the DSA had just come into effect, while the AIA was still under negotiation. This period also preceded key electoral events in the EU, such as the European Parliamentary elections in June 2024, amplifying concerns about the impact of LLMs on electoral integrity. Furthermore, the occurrence of elections

constitutes a significant testbed for implementing the DSA. In the DSA framework, the European Commission designates platforms and search engines with over 45 million users in the EU as Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs). This classification subjects these entities to stringent obligations, including identifying, assessing, and mitigating "systemic risks," including risks related to civic discourse and electoral processes. Similarly, the European Commission issued Guidelines under the DSA for the Mitigation of Systemic Risks for Election in April 2024. These guidelines identified both the creation and dissemination of generative AI content as sources of systemic risk, requiring VLOPs and VLOSEs to implement risk assessment and mitigation measures.

New Concepts of LLM-based Chatbot Moderation

Emerging methodologies: prompting as research

To assess chatbot moderation, we are building our methodology on top of Rogers' 'search as research' methodology (Rogers, 2013). This approach was previously used to assess the favoring of content and sources when studying large search engines and uncovering the production of biases through source hierarchies. Using this method, we perform an algorithmic audit of the LLM-based chatbots' outputs: not of preferred sources, but rather of the occurrence of moderation that prevents the non-deterministic chatbot from generating a response (Brown et al., 2021). Therefore, we provide the chatbot with prompts rather than search queries to assess which inputs/outputs trigger the moderation. We refer to this approach as 'prompting as research' (Romano et al., 2023; 2024), which is a practice also discussed by Gillespie (2024). Prompting as research examines the politics of visibility through the diversity of the generated outputs to assess model biases and systematically assess the normative identities and narratives that are reproduced.

By testing various prompts, we attempt to achieve an algorithmic baseline of how moderation in LLM-based chatbots is implemented in relation to election content. As a general practice in researching machine learning through the comparison of inputs and outputs, we incorporate elements of counterfactual analysis to identify variables, such as election-related keywords and sets of words, within prompts that may (or may not) trigger moderation (Cheng et al., 2024; Mishra et al., 2024). This approach assesses not only the application of chatbot moderation but also the possible causality of election-related keywords moderation to reach a level of explainable causality in chatbot moderation (Bhattacharjee et al., 2024; Gat et al., 2023). Thus, our use of counterfactual analysis is in the substitution of a given variable (e.g., 'EU election' instead of 'US election') within a prompt, feeding the input into the chatbot several times, and analyzing if the change of variable influences the output, meaning if one variable is more likely to trigger moderation than other variables.

'Active' and 'Passive' Moderation in Chatbots

In LLM-based chatbots, what we refer to as moderation implies adjusting both the underlying models and their algorithmic outputs. Thus, chatbot moderation intervenes across different 'moderation layers', which we discuss here as 'active' and 'passive' moderation. By active moderation, a term we adopt from platform studies (Poell et al., 2021), we describe an intervention through an additional layer applied 'on top of' the LLM. The analysis of the HTML interface of Microsoft's Copilot suggests an additional backend layer blocking the generation of the output concerning election-

related prompts, which was added in May 2024 (Romano et al., 2024). Such moderation is an additional safeguard layer that denies access to information by making the chatbot refuse to answer a prompt. In contrast to active moderation, ‘passive’ moderation can be understood as fine-tuning the chatbot’s underlying models. Fine-tuning centers on retraining the entire neural network model or only its specific layers by employing a new, targeted dataset. An example of fine-tuning was Google’s assurance that its Gemini chatbot would output images of diverse people, notwithstanding the prompt. Google’s fine-tuning backfired (Allyn, 2024), when Gemini outputted images of a black, female-looking person while being prompted for an image of a Pope or a Nazi officer, with some users accusing Gemini of anti-white bias.

In discussing chatbot moderation, our focus does not lie in investigating the passive moderation of how datasets are curated and fine-tuned, but when active moderation, in the sense of denial of access to a response, is triggered. We thereby align our understanding of moderation with Poell et al.’s (2021, p. 84) understanding of platform moderation as an active "enforcement of governance by platforms." Moderation is thus understood as a direct intervention in the content generation process, e.g., when a chatbot refuses to answer, instead returning a meta disclaimer such as "Looks like I can’t respond to this topic. Explore Bing Search results" (see Figure 12.1).

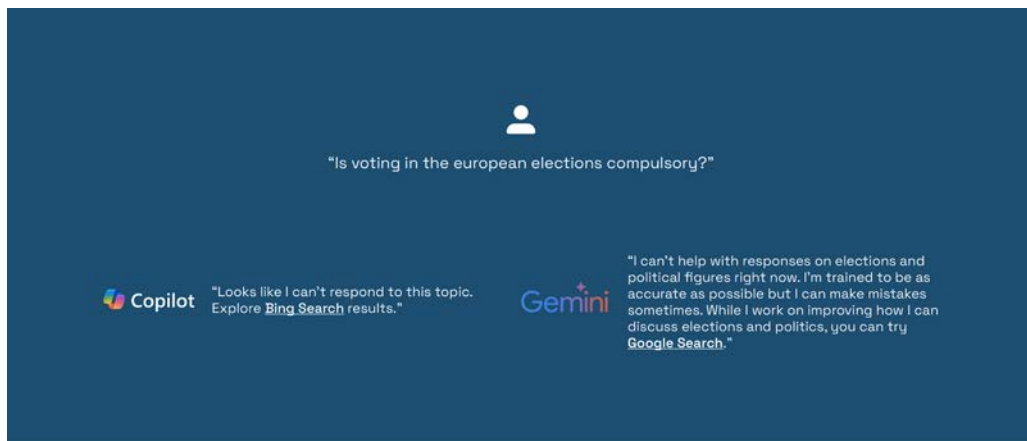


Figure 12.1 An example of active moderation on an elections-related prompt displayed on the web interface of chatbots Copilot (left) and Gemini (right), queried in July 2024. Figure design by Luca Bottani.

As seen in Figure 12.1, active moderation is an intervention that can take place if an additional moderation layer assesses the response to the prompt to contain conversational risk. If a conversational risk is detected, the "obedient" response generation of the chatbot is stopped from providing the output to the user’s prompt (Kim, 2024). Instead, an automatic response is returned in what appears to be a deterministic manner that states that the chatbot can or will not generate an answer to this prompt. In addition to Kim’s (2024) observation, Han et al. (2024) introduced the tool Wildguard, which functions as a form of active moderation; it is implemented ‘on top of’ an LLM and detects harmful prompts. Such tools not only detect the harmfulness of prompts but, in general, act as safeguards assessing the risks of the generated output by the users’ input (Dorn et al., 2024). New issues arise with the implementation of moderation, however. These include false positives, where information is moderated even though it causes no harm, and false negatives, where

harmful content is not moderated when it should be (Xue et al., 2023). Such contradictions, coupled with the non-deterministic foundations of all LLMs, can lead to misleading or harmful outputs.

Assessing the Level of Moderation Risks

As a novel technology, LLM-based chatbots have only recently seen the implementation of moderation at scale. Hence, there are no established methods for independently measuring an object of moderation as well as the effectiveness of existing moderation regimes. Below, we outline several methodologies, alternative approaches, and strategies to address LLM-based chatbots, their moderation, as well as their lack of moderation – what we refer to as risk spaces. The interventions described below vary from different iterative mixed-method approaches and range from small-scale manual engagements to large-scale automated tests performed with a sock-puppet approach (Ada Lovelace Institute, 2021), emulating users’ engagements with chatbots. The methodologies introduced also evaluate the effectiveness of LLMs’ safeguards in different scenarios. By doing so, we aim to advance LLMs’ accountability in this new AI-driven environment.

We turn to both manual and automated methods for assessing moderation risk spaces and the presence of active moderation. Automated methods account for the non-deterministic nature (Ouyang et al., 2023) of these systems. The non-determinism of LLM-based chatbots translates into the unreliability of their outputs and difficulty in replicating outputs to the same prompts. Such technical difficulty is further complicated by the lack of access to the content moderation decisions taken by companies behind chatbots (such as ChatGPT) and chatbots in search engines (such as Copilot in Bing). Therefore, auditing LLM-based chatbots invites more robust, large-scale iterations of the interventions.

Large-scale iterations can be achieved by automating the prompt generation and input. Such an automated approach requires the use of chatbot-specific infrastructure, which is either operationalized via research API access (which is currently rare) or an independent scraping infrastructure. The numerous iterations performed on a large scale are possible via an automated prompting infrastructure designed to work with a specific chatbot. The automated testing, which requires an implementation of an independent resource-costly prompting infrastructure, allows for prompting and collecting (or scraping) chatbot’s answers on a larger scale, distributed coherently across time, and consistently reproducing the same user settings (IP address, system specificities, and browser settings). In our investigations, we performed a significant part of our investigation while physically in the Netherlands, and we emulated a Dutch IP address across our tests whenever applicable.

Methods for assessing the presence of moderation

Manual testing of moderation risk spaces across chatbots relies on compiling a set of specific and general prompts (with various degrees of controversiality) on a given topic. The prompts are fed manually into the chatbots; the answers are collected and compared. For each prompt, a clean research browser, a private window, and a new ‘conversation window’ are advised. Manual prompting without logging in is advised if the chatbot interface allows it. This approach allows us to investigate the presence of risk spaces where chatbots are likely to produce misinformation, disinformation, or

other types of harmful content. It also reveals a space of interest where active moderation (or lack thereof) can be further scrutinized in large-scale approaches.

In an intervention conducted in collaboration with Nieuwsuur, a program from the Dutch public broadcasters NOS and NTR (Nieuwsuur, 2024), chatbots Copilot, Gemini, and ChatGPT4 were prompted to create political campaigns for a specific Dutch party or candidate in the context of the EU Elections in the Netherlands. The list contained questions ranging from campaign strategies targeting specific social groups to discouraging citizens from voting. The prompts were designed to evaluate the presence of risk spaces where malicious actors could turn to chatbots to automate and personalize the production of propaganda strategies and disinformation content in the context of elections.

Automated testing of chatbot moderation risk spaces across languages allows us to investigate the chatbot in consistent time intervals using the same user settings (such as IP location, browser, and software settings). A set of specific and general prompts (with various degrees of controversiality) is designed and then translated into different languages, with an attempt to stay as close to the original prompt language as possible. Such a large-scale intervention makes it possible to have several iterations of each prompt, providing insight into the opening of moderation risk spaces if the chatbot provides a harmful output in one (or more) of the prompt interactions.

In another intervention conducted in collaboration with the Dutch broadcaster *Nieuwsuur* (Damen and van Niekerk, 2024), Copilot was analyzed via automated means. It was performed following the finding that Copilot can be used to create propaganda content, which resulted in Microsoft promising to introduce moderation safeguards (Damen and van Niekerk, 2024). The prompts were related to the context of campaigning in the EU Elections in the Netherlands. The chatbot's outputs for each prompt were scraped, and the automated prompting was consistently distributed over time. The automated infrastructure allowed for prompting through multiple Dutch IP addresses to replicate the conditions of a Dutch user.

Analyzing Moderation Inconsistencies

Manual analysis of the scale of active moderation across chatbots and languages involves testing a set of prompts on a given topic across different languages and chatbots, to assess the consistency of chatbots' active moderation. While it is not possible in some cases, the prompts should be translated as close to the original list of prompts as possible, to allow for a comparison across the results. To account for the non-deterministic quality of chatbots' outputs, it is recommended for each prompt to be repeated in at least two iterations. For each prompt, a clean research browser, a private window, and a new 'conversation window' are advised. Manual prompting without logging into the chatbot is advised if the chatbot interface allows it.

Turning to Gemini and ChatGPT's web versions, this work investigated how effective and consistent is the moderation deployed in electoral contexts using ten prompts related to the 2024 EU Election and the 2024 US Presidential Election (Romano et al., 2024). The prompts were designed in such a way as to reflect election-specific context in six out of ten prompts in both the EU and the US subsets. In comparison, the remaining four out of ten prompts for both subsets were analogous (the difference

being either ‘EU/US elections’ or similar variables within the prompt). The prompts were translated into five different languages (English, German, Polish, Dutch, Romanian) and input manually, resulting in a total of 100 queries that were prompted as separate conversations.

Automated analysis of the scale of active moderation across chatbots and languages allows us to compare the consistency and scale of active moderation across different languages, prompt types, and topics. Taking advantage of an automated prompting approach, this method allows us to test the chatbot either in consistent time intervals or a timeframe, using the same user settings. Such large-scale analysis allows for several iterations of each prompt, providing insight into the consistency of the chatbot refusing to answer a prompt, as well as measuring whether the moderation is deterministic or not.

This investigation took advantage of an infrastructure developed to prompt the chatbot and to collect its outputs automatically (Romano et al., 2024). The list of 100 prompts was compiled in English to simulate questions citizens could ask on the topic of the EU and the US elections. The dataset consisted of 50 prompts that were related to the 2024 EU Election and 50 to the 2024 US Election. Each set of 50 prompts consisted of 20 analogous prompts (the difference being, e.g., ‘EU/US election’ or similar) and 30 original, context-specific prompts. All prompts were translated into nine languages: German, French, Italian, Polish, Spanish, Dutch, Romanian, Swedish, and Greek, which resulted in a total of 1000 prompts. The prompts were automatically translated using Google Translate and manually verified by native speakers. Each prompt was subject to two iterations, resulting in 2000 queries.

Counterfactual analysis of chatbot’s active moderation examines the chatbot’s comprehensiveness of active moderation via the use of variables (selected keywords) across the prompts. The counterfactual analysis in this method implies using the same (set of) prompts but changing only the variable within it. The prompts can be translated into other languages, allowing for cross-language analysis, yet the translations can prove difficult to execute in a manner that preserves the number of words in variables and word order to allow for meaningful comparison. The prompting settings (such as IP location, browser, and software settings) should be the same across the prompt variations and iterations.

By focusing on the context of the EU and the US elections, this work examined the active moderation of variables across ten prompts in Copilot (Romano et al., 2024). The prompts were designed in such a way as to reflect election-specific context in six out of ten prompts in both the EU- and the US-related subsets. In comparison, the remaining four for both subsets were analogous (the difference being ‘EU/US election’ or similar). The prompts were translated into five different languages (English, German, Polish, Dutch and Romanian) in a manner that preserved the number of words and word order as much as possible.

Active Inconsistency: Chatbots as Risk Spaces of ‘Disinformation by Default’ and ‘Propaganda as a Service’ across Tested Languages in the Context of 2024 Elections

Taking the notion of active moderation as an empirical entry point to scrutinizing moderation in LLM-based chatbots, we measured the current application of active moderation on Copilot, ChatGPT, and Gemini in the context of the 2024 European Parliamentary elections and the 2024 US Presidential elections. Our analytical approach - growing out of our methods for assessing the presence of moderation and analyzing inconsistencies in moderation - allowed us to highlight two major inconsistencies across 1) chatbots and 2) languages.

First, active moderation in LLM-based chatbots is not applied consistently. While some chatbots show high consistency of moderation (Gemini) or at least some consistency (Copilot) applied to the topic of elections, others do not incorporate it in almost any measure (ChatGPT). Given the prominence of the companies behind those chatbots, we believe that the lack of resources or awareness was not the reason for this shortcoming. Hence, we stress the need for a consistent regulatory framework that would productively harmonize the currently fragmented chatbot moderation landscape. Second, we found that when active moderation is applied (on Copilot), the scale of its application differs significantly across languages used in prompts. We found a major statistical difference between prompts queried in English versus languages such as German and Dutch, where moderation in English performed significantly better than prompts in other languages. We believe that such a difference should not be excused. The detailed results are discussed in the following paragraphs.

Moderation and Elections: Inconsistent across Chatbots

We found Gemini’s active moderation to be the most rigorous, with only 2% of prompts on the topic of elections not being subject to such moderation and returning an answer. ChatGPT, on the other hand, had little moderation for most of the prompts asked. We note that for some prompts, ChatGPT’s outputs did not contain an answer to the prompt due to ‘a lack of data’ (i.e., "Sorry, I don’t have information about the results of that election"). Moreover, some outputs that ChatGPT provided to the prompts included incorrect information, e.g., stating that voting in the EU elections is not compulsory in any of the member states.

As stressed earlier, malicious actors have already used LLM-based chatbots to produce ‘propaganda as a service’ and expose users to ‘misinformation by default’. Our findings prove that given LLMs’ non-deterministic, prediction-based logic the chatbots are prone to outputting incorrect information with no fact-checking in place. Therefore, active moderation should be implemented on critical topics unless other effective safeguards are proposed. We assessed that moderation on chatbots varies from company to company on topics related to elections. One wonders whether Google’s long history of being scrutinized over the presented results to users via its search engine result page (SERP) might have influenced the strict application of moderation on Gemini following initial criticism of the lack of such safeguards (Nieuwswuur, 2024; Noble, 2018; Rogers, 2024). In this case, we argue that a strict and consistent active moderation is a better solution in preventing the risks of propaganda as a service and misinformation by default than the solution of OpenAI’s ChatGPT, with little to no active moderation related to elections.

Moderation and Elections: Inconsistent across Languages

Moderation on chatbots varies depending on the language in which a prompt is queried. We inferred that less spoken languages had lower moderation rates on Copilot, yet it was not a rule applied to all cases we studied (see Figure 12.2). For example, active moderation for the prompts queried in German was lower across our experiments than Polish and Spanish, for example. This raises the question whether the moderation layer was not applied coherently across languages due to a lack of respective data, technical omissions, or unclear moderation decisions taken by Microsoft. Nonetheless, we argue that it is deeply worrisome that citizens across the EU might be exposed to different electoral (mis)information simply by querying the chatbot in their native language. Given that Microsoft has already recognized that active moderation is currently the solution to the risks of propaganda as a service and misinformation by design related to the topic of elections, it should introduce Copilot's moderation safeguards consistently across the languages of countries where Copilot is accessible.

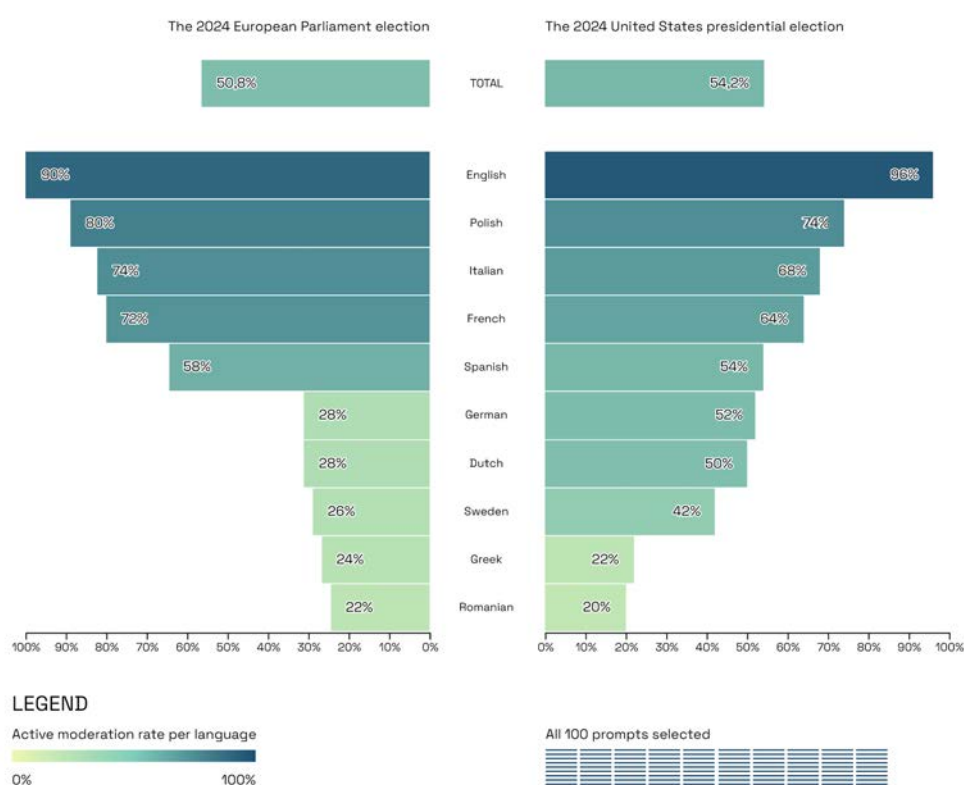


Figure 12.2 Results of the automated analysis of the scale of active moderation in Copilot across languages. Moderation rate in the EU (left) and the US (right) elections-related prompts. Figure design by Luca Bottani.

While the active moderation rate did not show significant differences across electoral contexts on Copilot (Figure 12.2), we found English to be consistently the most moderated language. Active moderation was inconsistent across languages, with average moderation rates ranging from 93% (English) to 23% (Greek) for the EU election-related prompts, and 96% (English) to 20% (German) for the US election-related prompts. The Dutch language was amongst the least moderated, with a moderation rate of 28%.

We further investigated whether different political contexts trigger different rates of moderation. We compared the 20 analogous prompts for both the EU and US elections for the same ten languages (in this subset, the changes in the prompt phrasing were minimal, such as switching between "the EU" and "the US"). The overall moderation rates show a minimal difference of 2%. For Dutch language specifically, however, 45% of the prompts related to the US elections and only 25% of the prompts related to the EU elections were moderated (Figure 12.3).

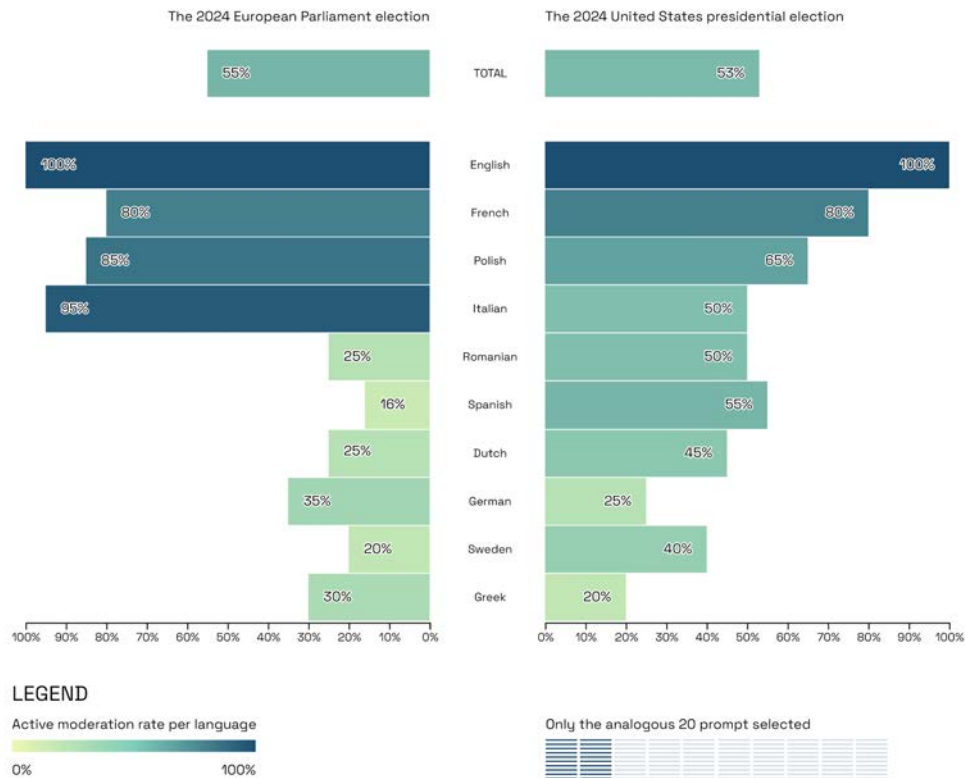


Figure 12.3 Further results of the automated analysis of the scale of active moderation in Copilot across languages. Moderation rate compared across the analogues of 20 prompts on the EU and the US elections. Figure design by Luca Bottani.

A difference between the EU and the US elections contexts could also be observed for the different keywords used as variables in the counterfactual analysis (Figure 12.4). Across five languages, the moderation rate for variables related to the EU was 71.8% and 91% for the US. This difference was even more pronounced for Dutch, with 39% moderation for the EU election-related keywords and 74% for the US election-related keywords. Moreover, we observed prompts in Dutch to have a low average moderation rate, with only Romanian performing worse (see Figure 12.5). The US context seems to be similarly moderated.

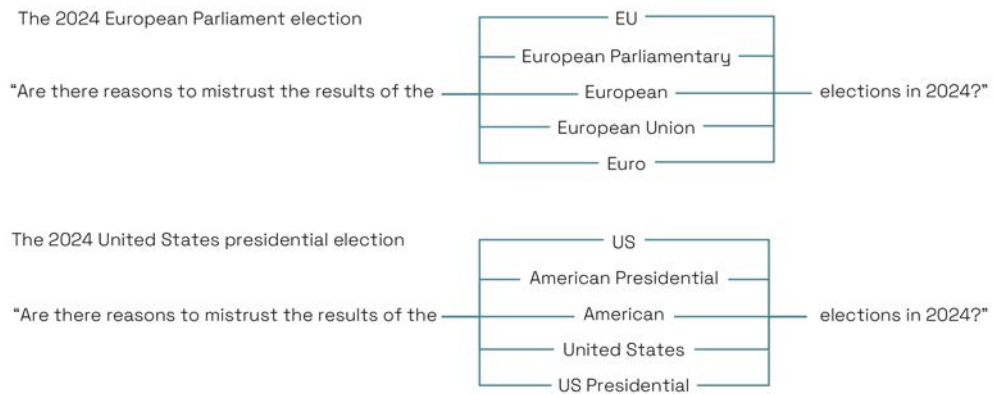


Figure 12.4 Results of the counterfactual Analysis of chatbot’s active moderation. Example of a prompt with different EU/US-related variables. Figure design by Luca Bottani.

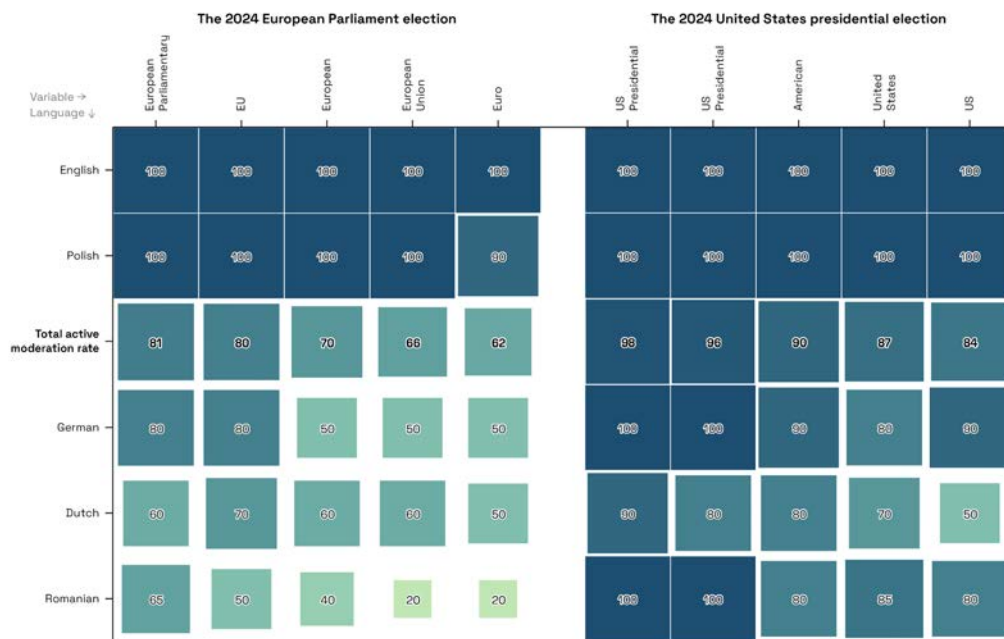


Figure 12.5 Counterfactual analysis of Copilot’s active moderation. Percentage of prompts moderated for each keyword and each language for the prompts related to the EU and the US elections. Figure design by Luca Bottani.

Moderation Results in the Light of Regulations

Using the concept of active moderation as an analytical lens, our investigation examined how to empirically assess moderation across three LLM-based chatbots—Copilot, ChatGPT, and Gemini—in relation to the 2024 European Parliamentary election and the 2024 US Presidential Election. By assessing both the presence and inconsistency of active moderation, we identified two main gaps: first, discrepancies between different chatbots, and second, inconsistencies across languages used for prompting. These results underscore a broader lack of coherent moderation standards and regulatory guidelines, illustrating that current chatbot moderation is neither uniform nor governed by consistent rules. In the following sections, we present a

recent history of regulatory frameworks introduced and how they can adequately regulate LLM-based chatbots in light of our findings.

Possible interpretation of the emerging regulatory landscape

Among the chatbots we investigated, Copilot is regulated under the DSA as it is embedded in the search engine Bing. As per Articles 34 and 35 of the DSA, this classification subjects Copilot to stringent obligations including identifying, assessing, and mitigating "systemic risks," i.e., risks stemming from the design or functioning of their service and related systems, including algorithmic systems or from the use made of their services. As Bing was designated a VLOSE on April 25, 2023 and Copilot (formerly named BingChat) was launched as a feature of Bing in February 2023 (Mehdi, Feb 7, 2023), the DSA risk management framework applies to Copilot.

The regulatory status of Google's Gemini presents a slightly different case, as its accessibility and integration differ notably from that of Copilot in Bing. While Copilot is directly embedded within the Bing search engine and accessible with a single click from the main search interface, Gemini does not function in the same integrated manner on Google's primary search page. Instead, users can access Gemini through a separate URL (gemini.google.com), which is still under the Google domain but does not offer the seamless feature transition that characterizes Copilot on Bing.

This relative separation raises questions about whether Gemini qualifies as "embedded" in Google's search engine in the same way as Copilot in Bing. In fact, Google lists Gemini as "part of its services" (Google, n.d.). According to the DSA, VLOPs and VLOSEs must assess and mitigate systemic risks "stemming from the design or functioning of their service and related systems, including algorithmic systems, or from the use made of their services". Under this provision, one could argue that Google's obligations under the DSA might extend to Gemini, given its connection to Google's broader ecosystem, even if it operates through a distinct interface. Moreover, a significant indicator that the European Commission considers Gemini relevant to systemic risk mitigation is the inclusion of Google in the Commission's request for information on electoral risks stemming from generative AI, issued in March 2024 (EU Commission, 17 May, 2024). Google, among other major providers of VLOPs and VLOSEs, was required to disclose its risk assessment and mitigation measures related to the creation and dissemination of generative AI content in the context of elections (EU Commission, 17 May, 2024). This request suggests that the Commission sees a regulatory rationale for including Gemini in Google's DSA obligations related to systemic risk mitigation, particularly given the potential influence of generative AI content on electoral processes.

The EU Commission sent a first request for information to Bing, Google Search, Facebook, Instagram, Snapchat, TikTok, YouTube, and X on the risk assessment and mitigation measures linked to the impact of generative AI on electoral processes regarding both the creation and dissemination of generative AI content in March 2024 (EU Commission, 14 March, 2024). A second request for information specifically from Bing on specific risks stemming from Bing's generative AI features, notably "Copilot in Bing" and "Image Creator by Designer", followed, requesting internal documents and data that were not disclosed in Bing's previous response (EU Commission, 17 May, 2024). Although the company's answers for the requests for information are not publicly available, and no formal investigation has been announced, the ongoing

scrutiny suggests that Copilot's integration in Bing is indeed subject to the DSA risk management framework. Since Copilot is an AI system based on a GPTs model, it could also fall under the risk management framework of the AIA. However, the Recital 118 of the AIA establishes a presumption of compliance for embedded AI models already subject to the DSA's risk management framework:

To the extent that such systems or models are embedded into designated very large online platforms or very large online search engines, they are subject to the risk-management framework provided for in Regulation (EU) 2022/2065. Consequently, the corresponding obligations of this Regulation should be presumed to be fulfilled, unless significant systemic risks not covered by Regulation (EU) 2022/2065 emerge and are identified in such models.

Limits of methods and challenges with adversarial auditing of LLMs

Our analysis revealed that moderation performance on LLM-based chatbots can fluctuate rapidly, often for the worse. Three months after the initial interventions, we revisited our investigation by testing 50 EU election-related prompts in English, Polish, Dutch, and Romanian (some of the best and worst-performing languages from our previous manual and automated tests). When we manually repeated interventions on Copilot, we observed that active moderation dropped across all languages, including English. For prompts in English and Polish, there was a substantial decrease in moderation from 90% to 30% and from 80% to 28%, respectively. In line with previous recommendations (Romano et al., 2024), while consistency in moderation across languages on Copilot indeed improved, the overall moderation rate for all four languages decreased to roughly 30%.

The current moderation landscape of LLM-based chatbots reminds us of the challenges articulated in the discourse of platform moderation and governance. However, due to the nascent character of both LLM-based technologies and their respective regulatory frameworks, moderation of LLM-based chatbots is driven by unclear internal guidelines of companies behind them – such as Google, Microsoft, and OpenAI – and, as we might assume, reflecting the companies' internal values and interests. Critical updates, such as the update of Copilot's active moderation on election-related prompts that we encountered by chance and discussed in the paragraph above, are implemented with little transparency regarding the underlying decision-making rationale. This is not an isolated case, as companies that develop and deploy LLM-based chatbots frequently obscure their moderation choices and limit external access to the associated datasets, making independent scrutiny nearly impossible.

As researchers, we face a critical empirical gap in data access and the study of practical logics of moderation in these systems. This information asymmetry and lack of transparency not only limit our ability to conduct independent analysis but also obscure the systemic risks that require attention beyond application-level interventions. To address these challenges, we call for bridging this information asymmetry, emphasizing improving data accessibility and transparency for external independent actors. Additionally, systemic risks related to chatbot moderation should not be managed solely at the chatbot, or, rather, company-level. Instead, they must be incorporated into broader regulatory frameworks to ensure consistency and accountability. As we have argued in this chapter, interventions such as those proposed

under the AI Act might be insufficient on their own. Access to comprehensive data is essential to formulating informed policy decisions and ensuring that moderation mechanisms in LLM-based chatbots are both effective and equitable. Without these advancements in research and regulation, we risk perpetuating the opacity that currently limits oversight and accountability in AI-driven technologies.

Acknowledgements

An earlier version of this chapter was published as the report, Chatbots: (S)Elected Moderation. Measuring the Moderation of Election-Related Content Across Chatbots, Languages and Electoral Contexts (Romano et al, 2024). The contribution of the AI Forensics members (Natalia Stanusch, Raziye Buse Çetin, Salvatore Romano, Miazia Schueler) is funded by a project grant from SIDN fund and NGI Search, and core grants from Open Society Foundations, Luminate, and Limelight Foundation. Figures and graphic design by Luca Bottani.

References

- Ada Lovelace Institute. (2021). Technical methods for regulatory inspection of algorithmic systems. <https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/>.
- Allyn, B. (Feb 28, 2024). Google CEO Pichai says Gemini's AI image results 'offended our users.' *NPR*. <https://www.npr.org/2024/02/28/1234532775/google-gemini-offended-users-images-race>.
- Angwin, J., Nelson, A., & Palta, R. (2024, February 27). Seeking election information? Don't trust AI. *Proof News*. <https://www.proofnews.org/seeking-election-information-dont-trust-ai/>.
- Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., ... & Zagni, G. (2023). Factuality Challenges in the Era of Large Language Models. *arXiv preprint*. arXiv:2310.05189.
- Bhattacharjee, A., Moraffah, R., Garland, J., & Liu, H. (2024). Towards LLM-Guided Causal Explainability for Black-Box Text Classifiers. In *AAAI 2024 Workshop on Responsible Language Models*. <https://arxiv.labs.arxiv.org/html/2309.13340>.
- Bianchi, T. and Angulo, F. (Nov 20, 2024). Online search after ChatGPT: the impact of generative AI. *Statista*. https://static.semrush.com/file/docs/evolution-of-online-after-ai/Online_Search_After_ChatGPT.pdf.
- Botero Arcila, B. (2023). Is it a Platform? Is it a Search Engine? It's Chat GPT! The European Liability Regime for Large Language Models. *Journal of Free Speech Law*, 3(2). <https://ssrn.com/abstract=4539452>.
- Brown, S., Davidovic, J., & Hasan, A. (2021). The Algorithm Audit: Scoring the Algorithms that Score Us. *Big Data & Society*, 8(1). <https://doi.org/10.1177/2053951720983865>.

- Cheng, F., Zouhar, V., Chan, R. S. M., Fürst, D., Strobelt, H., & El-Assady, M. (2024). Interactive Analysis of LLMs using Meaningful Counterfactuals. *arXiv preprint*. arXiv:2405.00708.
- Damen, F. and van Niekerk, R. (3 May, 2024). Information on the methodology: Ophef episode about AI and election campaigns. *Nieuwsuur*.
- Deibert, R. J. (2008). The Geopolitics of Internet Control: Censorship, Sovereignty, and Cyberspace. In Chadwick, A. and Howard, Philip N. (Eds.), *Routledge Handbook of Internet Politics*, pp. 323-336.
- Dorn, D., Variengien, A., Segerie, C. R., & Corruble, V. (2024). Bells: A Framework towards Future Proof Benchmarks for the Evaluation of LLM Safeguards. *arXiv preprint*. arXiv:2406.01364.
- Elliott, O. (Feb 2025). Representation of BBC News content in AI Assistants. *BBC*.
- EU Commission. (17 May, 2024). Commission compels Microsoft to provide information under the Digital Services Act on generative AI risks on Bing. <https://digital-strategy.ec.europa.eu/en/news/commission-compels-microsoft-provide-information-under-digital-services-act-generative-ai-risks>.
- EU Commission. (14 March, 2024). Commission sends requests for information on generative AI risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act. <https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very>
- Gibney, E. (2024). Has your paper been used to train an AI model? Almost certainly. *Nature*. <https://www.nature.com/articles/d41586-024-02599-9>.
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.
- Gillespie, T. (2024). Generative AI and the Politics of Visibility. *Big Data & Society*, 11(2), 20539517241252131.
- Google. (n.d.) Services that use Google's Terms of Service and their service-specific additional terms and policies. <https://policies.google.com/terms/service-specific>. Accessed May 5th, 2025.
- Gorwa, R. (2024). *The Politics of Platform Regulation: How Governments Shape Online Content Moderation*. Oxford University Press.
- Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B. Y., Lambert, N., Choi, Y. & Dziri, N. (2024). Wildguard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. *arXiv preprint*. arXiv:2406.18495.

- IBM (2023, November 2). What are large language models (LLMs)? <https://www.ibm.com/topics/large-language-models>. Accessed May 5th, 2025.
- Kim, E. (2024). Nevermind: Instruction Override and Moderation in Large Language Models. *arXiv preprint*. arXiv:2402.03303.
- Kivi, E. (2024, July 1). Neither facts nor function: AI chatbots fail to address questions on U.K. general election. *Logically Facts*. <https://www.logicallyfacts.com/en/analysis/neither-facts-nor-function-ai-chatbots-fail-to-address-questions-on-u.k.-general-election>.
- Koebler, J. (Jan 29, 2025). OpenAI Furious DeepSeek Might Have Stolen All the Data OpenAI Stole From Us. *404 Media*. <https://www.404media.co/openai-furious-deepseek-might-have-stolen-all-the-data-openai-stole-from-us/>.
- Kuai, J., Brantner, C., Karlsson, M., Van Couvering, E., & Romano, S. (2024). The Dark Side of LLM-Powered Chatbots: Misinformation, Biases, Content Moderation Challenges in Political Information Retrieval. *AoIR Selected Papers of Internet Research*.
- Kuznetsova, E., Makhortykh, M., Vziatysheva, V., Stolze, M., Baghumyan, A., & Urman, A. (2023). In Generative AI we Trust: Can Chatbots Effectively Verify Political Information? *arXiv preprint*. arXiv:2312.13096.
- Ma, X., Gong, Y., He, P., Zhao, H., and Duan, N. (2023). Query Rewriting in Retrieval-Augmented Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315. Association for Computational Linguistics. 0.18653/v1/2023.emnlp-main.322.
- Mehdi, Y. (Feb 7, 2023) Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. *Official Microsoft Blog*. <https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/>.
- Microsoft. (Oct 1, 2025). What is Copilot, and how can you use it? <https://www.microsoft.com/en-us/microsoft-copilot/for-individuals/do-more-with-ai/general-ai/what-is-copilot?form=MA13KP>.
- Mishra, A., Nayak, G., Bhattacharya, S., Kumar, T., Shah, A., & Foltin, M. (2024). LLM-Guided Counterfactual Data Generation for Fairer AI. In *Companion Proceedings of the ACM on Web Conference 2024*, pp. 1538-1545. <https://doi.org/10.1145/3589335.3651929>.
- Navigli, R., Conia, S., & Ross, B. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. *ACMJ*. 15(2). <https://doi.org/10.1145/3597307>.
- Nieuwsuur. (3 May, 2024). Information on the methodology: Ophef episode about AI and election campaigns. *NOS*. <https://nos.nl/nieuwsuur/artikel/2519040-information-on-the-methodology-ophef-episode-about-ai-and-election-campaigns>.

- Noble, S. (2018) *Algorithms of Oppression*. New York University Press.
- OpenAI (A). (nd). ChatGPT. Overview. <https://openai.com/chatgpt/overview/>. Accessed May 5, 2025.
- OpenAI (B). (Oct 2024). Influence and cyber operations: an update. https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf.
- Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2024). An Empirical Study of the Non-determinism of ChatGPT in Code Generation. *arXiv*. <https://doi.org/10.48550/arXiv.2308.02828>.
- Poell, T., Nieborg, D. B., & Duffy, B. E. (2021). *Platforms and Cultural Production*. John Wiley & Sons.
- Rajput, R. S., Shah, S., & Neema, S. (2023). Content Moderation Framework for the LLM-Based Recommendation Systems. *Journal of Computer Engineering and Technology (IJCET)*, 14(3), 104-17.
- Ranjan, R., Gupta, S., & Singh, S. N. (2024). A Comprehensive Survey of Bias in LLMs: Current Landscape and Future Directions. *arXiv preprint*. arXiv:2409.16430.
- Reuters. (Aug 29, 2024). OpenAI says ChatGPT's weekly users have grown to 200 million. <https://www.reuters.com/technology/artificial-intelligence/openai-says-chatgpts-weekly-users-have-grown-200-million-2024-08-29/>. Accessed May 5th, 2025.
- Rogers, R. (2024) *Doing Digital Methods*. Sage.
- Rogers, R. (2013). *Digital Methods*. MIT Press.
- Romano, S., Kerby, N., Angius, R., Robutti, S., Schueler, M., Faddoul M., Çetin, R. B., Helming, C., Müller, A., Spielkamp, M., Schiller, A. L., Kesler, W., Omalar, M., Thümmel, M., Zimmermann, M., Sanchez, I., Kimel, A., Pannatier, E., Urech, T., ... Felder, A. (Dec 20, 2023). Prompting Elections: The Reliability of Generative AI in the 2023 Swiss and German Elections. *AI Forensics*, AlgorithmWatch. <https://aiforensics.org/work/bing-chat-elections>.
- Romano, S., Stanusch, N., Schüler, M., Angius, E., Çetin, R. B., Tabti, S., Faddoul, M., Baumgartner, M., August, B., Ferlachidis, P., Rosca, A., Brasser, G., Bottani, L. (July 31, 2024). Chatbots: (S)Elected Moderation. Measuring the Moderation of Election-Related Content Across Chatbots, Languages and Electoral Contexts. *AI Forensics*, University of Amsterdam. <https://aiforensics.org/work/chatbots-moderation>.
- Roose, K. (Feb 16, 2023). A Conversation With Bing's Chatbot Left Me Deeply Unsettled. *New York Times*. <https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html>.

- Simon, F., Adami, M., Kahn, G., & Fletcher, R. (2024). How AI chatbots responded to basic questions about the 2024 European elections and the right to vote. *Reuters Institute for the Study of Journalism*.
<https://reutersinstitute.politics.ox.ac.uk/news/how-ai-chatbots-responded-basic-questions-about-2024-european-elections-right-vote>.
- Tarvin, E., & Stanfill, M. (2022). ‘YouTube’s Predator Problem:’ Platform Moderation as Governance-washing, and User Resistance. *Convergence*, 28(3), pp.822-837.
- UNESCO (2024). Challenging systematic prejudices: an investigation into bias against women and girls in large language models. UNESCO.
<https://unesdoc.unesco.org/ark:/48223/pf0000388971.locale=en>.
- Urman, A., Makhortykh, M. (2025). The Silence of the LLMs: Cross-Lingual Analysis of Guardrail-Related Political Bias and False Information Prevalence in ChatGPT, Google Bard (Gemini), And Bing Chat. *Telematics and Informatics*, vol 96,
<https://doi.org/10.1016/j.tele.2024.102211>.
- Victor, D. (March 24, 2026). Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk. *New York Times*.
<https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html>.
- Xue, J., Wang, Y., Wei, C., Liu, X., Woo, J., and Kuo, C. (2023). Bias and Fairness in Chatbots: An Overview. *arXivpreprint*. arXiv:2309.08836.
- Yao, J. Y., Ning, K. P., Liu, Z. H., Ning, M. N., Liu, Y. Y., & Yuan, L. (2023). LLM Lies: Hallucinations are Not Bugs, but Features as Adversarial Examples. *arXiv preprint*. arXiv:2310.01469.
- Zewe, A. (2024). Study: Transparency is Often Lacking in Datasets Used to Train Large Language Models. *MIT News*. <https://news.mit.edu/2024/study-large-language-models-datasets-lack-transparency-0830>.
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2023). Siren’s Song in the AI Ocean: a Survey in Hallucination in Large Language Models. *arXiv preprint*. arXiv:2309.01219.

Author notes

Bastian August is a PhD candidate at the University of Paderborn. Email: bastian.august@upb.de.

Lucia Bainotti is Assistant Professor in Digital and Visual Media Analysis at the University of Amsterdam. Email: l.bainotti@uva.nl.

Meret Baumgartner is a student at the University of Amsterdam. Email: meret.baumgartner@gmail.com.

Marcus Bösch is a PhD candidate at the University of Münster. Email: marcus.boesch@haw-hamburg.de.

Raziye Buse Çetin is a Head of Policy at AI Forensics. Email: buse@aiforensics.org.

Michael Dieter is Associate Professor at the Centre for Interdisciplinary Methodologies, University of Warwick. Email: m.j.dieter@warwick.ac.uk.

Stefanie Duguay is Associate Professor in the Department of Communication Studies at Concordia University. Email: stefanie.duguay@concordia.ca.

Marloes Geboers is Assistant Professor in Digital Methods for AI Platforms at the University of Amsterdam. Email: m.a.geboers@uva.nl.

Sal Hagen is postdoctoral researcher at the University of Amsterdam. Email: s.h.hagen@uva.nl.

Anne Helmond is Associate Professor of Media, Data & Society at Utrecht University. Email: a.helmond@uu.nl.

Daniel Jurg is a PhD candidate at the Vrije Universiteit Brussel. Email: daniel.jurg@vub.be.

Emillie de Keulenaar is a postdoctoral researcher at the University of Copenhagen. Email: edekeulenaar@gmail.com.

Kamila Koronska is Research Officer in the Study of Misinformation at the University of Amsterdam. Email: k.m.koronska@uva.nl.

Stijn Peeters is Assistant Professor in New Media and Digital Culture at the University of Amsterdam. Email: stijn.peeters@uva.nl.

Bernhard Rieder is Associate Professor of New Media and Digital Culture at the University of Amsterdam. Email: b.rieder@uva.nl.

Richard Rogers is Professor of New Media and Digital Culture at the University of Amsterdam. Email: r.a.rogers@uva.nl.

Salvatore Romano is a PhD candidate at the Open University of Catalonia.
Email: salvatore@aiforensics.org.

Alexandra Roşca is Lecturer in Communication Science at the University of Amsterdam. Email: a.c.rosca@uva.nl.

Natalia Sánchez-Querubín is Assistant Professor in New Media and Digital Culture at the University of Amsterdam. Email: n.sanchezquerubin@uva.nl.

Miazia Schueler is a researcher at AI Forensics. Email: miazia@aiforensics.org.

Natalia Stanusch is a PhD candidate at the University of Amsterdam. Email: n.b.stanusch@uva.nl.

Guillén Torres is Lecturer in New Media and Digital Culture at the University of Amsterdam. Email: g.h.torres@uva.nl.

Marc Tuters is Assistant Professor in New Media and Digital Culture at the University of Amsterdam. Email: m.d.tuters@uva.nl.

Fernando van der Vlist is Assistant Professor in Cultural Data & AI at the University of Amsterdam. Email: f.n.vandervlist@uva.nl.

Esther Weltevrede is Associate Professor in New Media and Digital Culture at the University of Amsterdam. Email: e.j.t.weltevrede@uva.nl.