
“It Doesn’t Violate Our Community Standards”: Examining Social Media Platforms’ Moderation Practices for User-Reported Rule-Breaking Tobacco Content

George D. H. Pearson, Padmini Kucherlapaty, Nathan A.
Silver, Barbara A. Schillo, and Jennifer M. Kreslake

Growing concerns regarding harmful content on social media platforms underscore the need for systematic examination of these platforms’ advertised safeguards. User reports are a key mechanism through which platforms can be apprised of and address harmful content. This study uses tobacco content as an important public health context to examine how strictly social media platforms moderate content that violates their own community guidelines after it is reported through user-reporting mechanisms. Using a replicable observational study, we tested the responsiveness to user reports across the three platforms most popular with youth: Instagram, TikTok, and YouTube. We identified 294 cases (between 97 and 99 per platform) violating platforms’ community guidelines covering five content areas: insufficient age-gating, disallowed promotions, illegal sales, underage use, and dangerous activities. Cases were reported through the platform’s reporting mechanisms and monitored for at least two weeks for moderation actions. Of a total of 294 cases, 221 (75%) were unmoderated. Instagram moderated 51 of 98 (52%) cases, YouTube 14 of 97 (14%), and TikTok 8 of 99 (8%). Insufficient age-gating was the only category with more than 50% of cases moderated (29 of 56). Less than 10% of dangerous activity (1 of 29) and disallowed promotion (4 of 79) cases were moderated. Results show that obvious violations of community guidelines reported to the platforms through their own mechanisms were rarely addressed, subjecting users to harmful, and sometimes illegal tobacco-, nicotine-, and cannabis-related content. Findings call into question whether the platforms can be relied upon to self-regulate to protect users from harmful tobacco content online.

1 Introduction

Use of social media platforms is ubiquitous with over 250 million active social media identities in the United States (DataReportal 2025). Rates are especially high for youth, with a majority of 13 to 17 year olds in the U.S. stating they use the most popular social media platforms (YouTube, TikTok, and Instagram) at least once a day (Pew Research Center 2024), and with the average U.S. teen spending 4.8 hours a day on social media (Pew Research Center 2024).

This widespread use has drawn increasing public health attention, both as an avenue for campaign outreach (de Vere Hunt and Linos 2022; Jafar et al. 2023) and because of concerns about harmful content. Such content ranges from health misinformation (Armstrong and Naylor 2019; Bode and Vraga 2018; Chou, Oh, and Klein 2018) to content normalizing dangerous behaviors (Dane and Bhatia 2023; Donaldson et al. 2022; Molly Rose Foundation 2025b; O’Brien et al. 2020), and often seems to violate stated platform policies (Bryant 2024; Digitalt Ansvar 2024; Eltaher et al. 2026; Kong et al. 2024).

These findings raise concerns about how successfully platforms moderate harmful content. While a large body of work has studied moderation effectiveness (Eltaher et al. 2026; Gillespie 2018a; Gorwa, Binns, and Katzenbach 2020), much of this work has taken place in non-public health contexts, often focusing on politicized speech. Second, most work has been conducted by either examining the presence of content violating community guidelines (Lerman et al. 2025; Molly Rose Foundation 2025a; O’Brien et al. 2020) or monitoring it to examine possible removal (Alkiviadou 2025; Edelson et al. 2025; Global Witness 2023). Much less work has involved directly reporting content to the platforms via user reports (for example see Carlson and Rousselle 2020; Center for Countering Digital Hate 2022a, 2022b), and this has not been done before in a public health context.

Here, we use one of the most prominent public health contexts, tobacco (HHS 2024), to examine social media platforms’ moderation practices. We use a systematic, replicable framework to identify content violating platforms’ community guidelines on tobacco and report it through the platforms’ own reporting mechanisms, as a real user would. We then monitor reported content to examine moderation actions, as well as, in a novel approach, examine platform’s written adjudication of reported content. In so doing, we can analyze not only moderation actions (i.e. if the content is removed or not), but also the explicit written decision of the platform on whether content is deemed a violation.

1.1 Governing What’s Allowed on Public Platforms

Public platforms face generally few regulatory restrictions on what content they choose to allow. In the United States platforms are protected by Section 230, which states that except under a thin, extreme subset of content (e.g. child sexual abuse and copyrighted materials), platforms cannot be held liable for content posted by users (Klonick 2018). In

the European Union, the Digital Services Act (DSA) places more stringent requirements upon platforms. In Europe, large platforms must have a system to allow users to report content, and to remove all illegal content promptly once notified (European Commission, n.d.), although no specific timeframe for removal is given. However, these regulations only cover illegal content. Instead the strongest set of restrictions comes from platforms' own "Community Guidelines."

A primary function of these guidelines is part of "corporate social responsibility" efforts (Klonick 2018), which help protect users from potential harms (Gillespie 2018a) as well as instilling distinct community values and culture (Myers West 2018). Additionally, community guidelines act as public-facing communication to external stakeholders about what the platform finds permissible (Klonick 2018). Moderation is "the central service that platforms offer" (Gillespie 2018b, 202). A website with zero moderation would require few resources, but also be unappealing to users and investors (Gillespie 2018b). As such, the promise of what content is allowed and disallowed is a key representation platforms make. These claims can make the platform seem more appealing to advertisers (Klonick 2018; Madio and Quinn 2025); signal to potential regulators that the platforms are capable of self-regulating content (Flynn et al. 2025; Sawers 2022); and signal to potential users the ethics and safety of the platform (Green 2021). This last role of community guidelines has often been overlooked but remains key. Such representations appear in documents directed at parents and guardians of young users. For instance, TikTok's 'Guardian's Guide' states guidelines "help us to create a welcoming, safe, and positive experience. These guidelines apply to everyone and everything on our platform." (TikTok, n.d.-c). Instagram, in their guide for parents and guardians, states "everyone who uses Instagram must adhere to our Community Guidelines, which are designed to create a safe and open environment" (Instagram, n.d.). Lastly, YouTube's parental guide writes: "in order to protect our community, our Community Guidelines set out what's allowed and not allowed on YouTube" (YouTube, n.d.). In these instances, community guidelines act as an advertisement to parents and guardians of what their children can be expected to encounter online.

Community guidelines' role in protecting youth is especially notable given growing public health concerns about the content youth encounter on platforms (Bryant 2024; Kong et al. 2024; Molly Rose Foundation 2025a). Assessing how well platforms enforce these guidelines is therefore paramount.

1.2 Reviewing Moderation on Platforms

Platforms release moderation transparency reports (e.g., Google, n.d.-d; Meta, n.d.-b; TikTok, n.d.-b), however these involve relying on platform data and are often limited in scope (Urman and Makhortykh 2023). External attempts to assess moderation effectiveness face challenges. Platforms have internal documents for assessing content that are more detailed than public-facing community guidelines (Angwin and Grassegger 2017; Langvardt 2017; Myers West 2018), and automated moderation tools further

opaqueness (Gorwa, Binns, and Katzenbach 2020). For instance, Meta use the virality of content to assess whether it needs moderating, but offer little information on how it is used (King 2020).

A large amount of work examining content moderation has found an array of harmful content, much of which violates community guidelines (Eltaher et al. 2026; Gerrard 2018; Lerman et al. 2025; Molly Rose Foundation 2025a; O'Brien et al. 2020) that platforms fail to remove (Digitalt Ansvar 2024; Global Witness 2023; Lerman et al. 2025), as well as finding moderation actions are often too slow to stop spread (Edelson et al. 2025). Alternatively, some work finds some success in platforms' moderation capabilities (Papakyriakopoulos, Medina Serrano, and Hegelich 2020; Schneider and Rizoïu 2023), and others find, especially under the DSA in Europe, that platforms can over moderate (Alkiviadou 2025; The Future of Free Speech 2024), with particular concern that platforms may over moderate minorities and other disadvantaged groups (Flynn et al. 2025). In a public health context, excessive moderation may also inadvertently remove legitimate health promotion messaging by trusted campaigns (Gomes and Sultan 2024).

Such mixed results speak to the challenges of moderation, including: creating protocols capable of sorting through nuanced content (Angwin and Grassegger 2017; Gillespie 2018a); differing cultural and regional norms (Gillespie 2021; Shahid and Vashistha 2023; Waldon 2023); users' attempts to flout moderation (Gerrard 2018); and debates over freedom of expression (Alkiviadou 2025; Young 2022). However, beyond these challenges, the willingness of platforms to remove content has also been called into question (Zenone, Kenworthy, and Maani 2023; Zhong et al. 2026). Many platforms have cut back on moderation infrastructure (Goggin 2024), with moderation work being done by low-pay workers (Jackson 2025; Roberts 2016; Solon 2017), or by automated systems that are easily flouted (Gerrard 2018). Additionally, internal documents showing platforms ignoring internal reports showing harms to youth (Duffy 2025; Horwitz 2025a; Kerr 2024; Scherer and Tiffany 2026) and profiting off advertisements for scam products (Horwitz 2025b) raise further concerns about content moderation practices.

These findings have led to scrutiny from the U.S. Surgeon General (Murthy 2024) and work arguing social media is a commercial determinant of health (Lafontaine, Hanson, and Wild 2025; Zenone, Kenworthy, and Maani 2023). The commercial determinants of health framework analyzes the "strategies and approaches used by the private sector to promote products and choices that are detrimental to health" (Kickbusch, Allen, and Franz 2016, 895); for social media, Zenone, Kenworthy, and Maani (2023) argues the profit motivations of social media lead to policies and actions, such as failing to remove harmful content, that have substantial impacts on health.

User reports, where users use in-built features to flag potentially violating content to the platform, add two useful data points for assessing platform willingness to moderate content. First user reports can alleviate some of the burden on platforms to surveil all content, as violating content is already identified and categorized by the user, reducing a

considerable challenge in moderation. Second, on some platforms (e.g., Instagram and TikTok) the platform responds to user reports with their adjudication. In so doing, rather than implicit approval or disapproval of the content's presence on the platform through inaction or moderation, they directly label the content as complicit or contravening community guidelines.

Most research on user reports has focused on the qualitative experience of reporters, noting the reporting options presented to users: lack alignment with community guidelines (Cover et al. 2026; Flynn et al. 2025; Vilk 2023; Zhang, Montague, and Jhaver 2024); can be difficult to navigate (Flynn et al. 2025; Zhang, Montague, and Jhaver 2024; Zhong et al. 2026); or are overly simplistic (Cover et al. 2026; Crawford and Gillespie 2016; Shim and Jhaver 2024). In terms of effectiveness, users have mixed perceptions of user reports (Flynn et al. 2025; Vilk 2023; Zhang, Montague, and Jhaver 2024; Zhong et al. 2026). Research has also noted user concerns over the potential for abuse of user reports to attack content producers (Crawford and Gillespie 2016; Samermit et al. 2023; Take et al. 2024).

Attempts to quantitatively study the effectiveness of user reports are rare, with much work focusing on hate speech. While some findings show 48% (Carlson and Rousselle 2020) and 61% (European Commission 2025) of user-reported hate speech was removed, others find only 16% of antisemitic (Center for Countering Digital Hate 2022a), and 11% of anti-Muslim reported hate speech (Center for Countering Digital Hate 2022b) were moderated. However, to date, no work has systematically tested the effectiveness of user reports in a public health context.

Consequently, in this study we conduct the first quantitative assessment of platforms' enforcement of their own community guidelines when responding to user reports in a public health context.

1.3 Moderation, Public Health, and the Context of Tobacco Control

Social media has become of paramount importance in public health (Terry et al. 2023). While scholars have recognized platforms can open up new possibilities for public health messaging (Jafar et al. 2023), concerns remain not only over the direct impacts of the platforms themselves (Huang 2022; Ortutay and Huamani 2026; Sharp 2026), but also the potential for platforms to rapidly spread harmful content from health misinformation (Armstrong and Naylor 2019; Bode and Vraga 2018; Chou, Oh, and Klein 2018) to content normalizing eating disorders, (Dane and Bhatia 2023; Molly Rose Foundation 2025a), self-harm, (Bryant 2024; Dyson et al. 2016; Molly Rose Foundation 2025a; Susi et al. 2023) and substance use (Donaldson et al. 2022; Moreno and Whitehill 2016).

Here, we focus on tobacco content as an ideal context to study how successfully social media platforms enforce community guidelines in response to user reports. Not only have parallels been drawn between the tobacco industry and tech industry as commercial

determinants of health (Lafontaine, Hanson, and Wild 2025; Zenone, Kenworthy, and Maani 2023), but social media has also been a growing area of concern in tobacco control, with research documenting large volumes of tobacco content online (Kong et al. 2021; O'Brien et al. 2020; Silver et al. 2024) and finding associations between platform use and tobacco initiation among youth (Donaldson et al. 2022). Additionally, since tobacco and cannabis are regulated consumable goods, identifying content promoting or showcasing use requires less inference than may be required with other harmful content. Lastly, given tobacco is the single largest preventable cause of death worldwide (World Health Organization 2017), the collective will to limit tobacco's harm, especially among youth, is high (Al-Shawaf et al. 2023; Kruger et al. 2015).

In the United States, in 2016, the Food and Drug Administration's authority to regulate tobacco products was expanded to cover all nicotine products including e-cigarettes and oral nicotine pouches (Food and Drug Administration 2023). Additionally, the emergence of cannabis e-cigarettes (Chadi, Minato, and Stanwick 2020; Civiletti, Aslam, and Hutchison 2024) and dual-use with tobacco (Morean et al. 2015) mean cannabis e-cigarettes and tobacco products contain overlap in their cultural placement and are covered similarly in the community guidelines of platforms (Google, n.d.-c; Meta, n.d.-c; TikTok, n.d.-a). Consequently, we define tobacco as any product covered by the FDA's tobacco remit, as well as cannabis vaporizers.

While a large body of literature has examined harmful tobacco content on platforms (Kong et al. 2021; O'Brien et al. 2020; Silver et al. 2024), there is little understanding as to how successfully tobacco content is moderated. No work has looked directly at tobacco moderation, and transparency reports from the platforms lack precise data. For instance, TikTok's transparency reports only give figures for "regulated goods and services," reporting 61 million video removals in this category for the first three months of 2026 (TikTok, n.d.-b). Other platforms are even less clear. Meta reports removals for "drugs and firearms," reporting 3.4 million actions on drugs content for the first three months of 2026 (Meta, n.d.-b); however it is unclear if tobacco is included in this category. YouTube's transparency reports, meanwhile, do not seem to offer a clear category for regulated substances content (Google, n.d.-d).

Here, we examine five types of tobacco content violations seen on social media. For each, we discuss the public health rationale behind concerns and describe the relevant policies of the three platforms most popular with youth—Instagram, TikTok and YouTube (Pew Research Center 2024). Relevant sections of platform policies can be seen in Appendix A and full policy language can be seen in our Open Science Framework (OSF) archive: https://osf.io/wzjge/overview?view_only=774ea82d389f4c4da5645dd7318038ab.

1.3.1 Insufficient age-gating

Based on three decades of research, the U.S. Surgeon General declared a causal link between exposure to smoking in movies and youth tobacco initiation (HHS 2012). A

meta-analysis also finds evidence of a relationship between exposure to tobacco content on social media and tobacco initiation (Donaldson et al. 2022). Consequently, reducing youth exposure to tobacco media is a key priority (HHS, n.d.) and age-gating tobacco media is a key step to reduce this exposure (Kong et al. 2024). All three platforms have language requiring age-gating on some tobacco content (Google, n.d.-c; Meta, n.d.-c; TikTok, n.d.-a).

1.3.2 Disallowed promotions

The tobacco industry has a history of youth-appealing advertising strategies (Food and Drug Administration 2019; Freeman, Watts, and Astuti 2022). Tobacco brands have now moved online (Kong et al. 2021; O'Brien et al. 2020) where they continue to employ youth-appealing marketing (Pearson et al. 2025b; Silver et al. 2024). There are few regulations around tobacco marketing online (Silver et al. 2023; Silver et al. 2024) and those that do apply are largely unenforced (Silver et al. 2024; Sharp et al. 2025; Donaldson et al. 2026; Silver et al. 2023). TikTok (n.d.-a) and YouTube (Google, n.d.-c) prohibit promoting tobacco content, while Instagram requires paid-for collaboration posts to use their branded content tool, but prohibits using this for tobacco promotions (Meta, n.d.-a), essentially forbidding paid-for tobacco influencer posts.

1.3.3 Illegal sales

Concerns have been raised that increased anonymity online could facilitate the illegal sale of tobacco products to youth (Nali et al. 2024) with research documenting youth using social media to source products (Singer et al. 2025). Furthermore, recent work has examined “discreet shipping” (Dobbs, Schisler, and McCormick 2025; Pearson et al. 2025a), where the product is shipped as either regular mail not disclosing its content, or hidden inside another product, such as cosmetics kits. These discreet techniques may aid youth access. All three platforms forbid all tobacco sales (Google, n.d.-c; Meta, n.d.-c; TikTok, n.d.-a).

1.3.4 Underage use

Content explicitly showing youth or young adults using tobacco products can present tobacco use as aspirational and desirable to adolescents whose developing identities make them impressionable to such depictions (Food and Drug Administration 2019). While TikTok (n.d.-a) and YouTube (Google, n.d.-c) prohibit depictions of underage use, Instagram does not have a specific policy.

1.3.5 Dangerous activities

Youth are vulnerable to learning tobacco use via social learning (Ranker et al. 2024), as well as learning specific behavior (Harakeh et al. 2007). Hence, content depicting the dangerous use of tobacco products may be replicated by youth, leading to more

immediate health risks. These behaviors can include attempting to recharge disposable e-cigarette batteries (SAMS, n.d.), excessive use of tobacco products, or e-cigarette pranks. TikTok (n.d.-a) and YouTube (Google, n.d.-a) do not have policies about dangerous tobacco activities specifically, but do prohibit dangerous acts more broadly. Instagram does not have a specific policy.

2 Methods

To assess the platforms' responsiveness to user reports of violating content, we developed and implemented a replicable protocol to identify, report, and document platform responses to nicotine- and cannabis-related content prohibited by the platforms' own community guidelines. To allow greater replicability and transparency, we have added additional documents to an Open Science Framework (OSF) project, viewable at: https://osf.io/wzjge/overview?view_only=774ea82d389f4c4da5645dd7318038ab. The data used in the study are also available on request from the authors.

2.1 Procedure

To identify and report content violating the platforms' community guidelines, an account was created on each of the three platforms. The accounts were created specifically for this project, and used exclusively for all searches, browsing and reporting, meaning accounts had minimal history on the platforms. Furthermore, the accounts posted no content, had minimal followers (between zero and one) and followed few accounts (between 1 and 16). Accounts were accessed from a variety of locations including researchers' onsite work location and homes; however, they were only accessed from two machines.

Content was identified using both search terms and on the platform's feeds (e.g. TikTok's For You Page). Given this sampling method, all identified content was publicly available on the platforms at the time of reporting. Cases the lead author deemed met the reporting criteria (see Reporting Categories and Criteria, Section 2.2) were subsequently reported by a second researcher using the same account. Screen recordings and screenshots of the reporting process were captured to ensure successful reporting and were later used to record video metadata. All reports were made between July 29 and October 16, 2025.

All three platforms have an interface to view submitted reports (see the OSF archive: Supplementary File 2). This was monitored for at least two weeks after the report was made to record any moderation actions. Finally, after all data were collected, three members of the research team reviewed all the reported but unmoderated content to remove cases that failed to meet specified criteria (see Identifying Content, Section 2.3).

Table 1: The five categories of reported content, with descriptions.

Category	Description
Insufficient Age-Gating	Content which, while allowed on the platform, requires age restriction. Posts viewable while logged out of the social media platform were deemed to be non age-gated.
Disallowed Promotions	Content promoting or marketing products in a manner contravening the policies of the social media sites.
Illegal Sales	Content advertising services illegally selling or instructing users how to illegally buy products (usually for underage users).
Underage Use	Content that depicts an underage user using tobacco.
Dangerous Activities	Content depicting inappropriate or dangerous use of a product, such as excessive or dangerous use (e.g. consuming a whole can of Zyn in one go) or instructions on how to “recharge” a disposable battery.

2.2 Reporting Categories and Criteria

Based on a review of site policies as well as our pre-existing knowledge of social media content and tobacco control regulation derived from being researchers in the field, we initially identified five common categories of problematic tobacco content: insufficient age-gating, disallowed promotions, illegal sales, underage use and, dangerous activities. A definition for each of these categories can be seen in Table 1.

Three of these five categories had relevant community guidelines for all three platforms. Two—dangerous activities and underage use—while having relevant language on TikTok and YouTube, do not appear to be explicitly prohibited on Instagram. However, to allow better consistency across platforms in the categories of reported content, and given the potential harmful impact of this content, we included cases in this category for Instagram as well.

Despite broad similarities, language from the community guidelines varied across platforms. Additionally, guidelines lacked clear definitions for numerous terms (see sample policy language in Appendix A and full language on the OSF archive: Supplementary Folder, Platform Policies). For instance, regarding disallowed promotions, YouTube states they forbid content “that promotes a product that contains drugs, nicotine, or a controlled substance” (Google, n.d.-c), but does not define “promotes.” To overcome this, we created clear, explicated criteria for reporting content for each category and platform. These definitions were designed to be conservative so that reported content would be clear and obvious violations. For instance, for YouTube, we defined promotion as content “that promotes or advertises a tobacco product (i.e., gives a favorable impression of) that is posted by someone with a clear vested interest in people buying the product; specifically, either the manufacturer of the product or a retailer.” The reporting criteria and relevant language from community guidelines can be seen in Appendix A.

All reported cases also had to (a) feature or be about a product or substance covered by the U.S. Food and Drug Administration's (FDA) tobacco products remit (Food and Drug Administration 2023) or a cannabinoid e-cigarette; and (b) promote (or normalize) the substance or product (i.e., the content could not be part of cessation or prevention efforts).

2.3 Identifying Content

To construct a sample of content to examine content moderation effectiveness, the research team aimed to report over 300 cases, with over 100 per site, while keeping the number of cases within each category between sites approximately equal. While perfect parity within categories across sites could not be achieved, all sites had over 100 cases initially reported ($N = 314$).

A single 'case' could mean we reported: a post (where the post violated community guidelines); an account (where an account's purpose violated community guidelines, e.g. an account set up to illegally sell e-cigarettes, but where no individual posts violated community guidelines); or a post/account pair (where a post violated community guidelines and the account's clear purpose was also a violation; see the caveat for reporting accounts on YouTube under Reporting Process, Section 2.4). For a breakdown of the number of cases of each type reported as well as in the final sample, see the OSF archive: Supplementary File 6.

Content was identified using a snowball-like procedure, wherein the lead researcher used both the platform's own search and Google Search's within-site function (Google, n.d.-b) using keywords likely to generate violating content. Search terms were varied but based on keywords or themes identified in violating videos. Examples included: product brand names, "vape tricks", or "vape review" (insufficient age-gating); e-cigarette brand names, "vape paid partnership" or "vape shop" (disallowed promotions); "discreet vape" and "buy nic no id" (illegal sales); "blinkers at school", or "school vape" (underage use); and "zyn challenge", "nose blinkers", or "pool noodle blinker" (dangerous activities). To aid transparency, a full archive of search terms used for YouTube is available on the OSF archive: Supplementary File 5. An archive of search terms was not available for other platforms. Searches also led to the platforms' algorithms (e.g., TikTok's "For You Page") recommending violating content, which could also then be reported.

Every attempt was made to ensure that each case belonged to a unique social media account. Upon cleaning the data, we identified eight cases that involved four duplicate accounts. However, as no moderation action was taken in any of these cases, implying no cross-case interference, we have kept them in the dataset. Cases were removed due to errors with the reporting system (e.g., where we failed to record the violating content for review) or content being removed by the uploader ($n = 7$).

At the end of the study, three lead authors independently reviewed all reported but

Table 2: Cases removed from the initial sample to reach the final sample.

Site	Instagram	TikTok	YouTube	Total
# of Items Reported	104	107	103	314
Dropped ^a	4	0	3	7
Removed During Review ^b	2	8	3	13
Final Sample	98	99	97	294

^a Dropped because the item was removed by the uploader, or other problems in reporting the case.

^b During final review, it was deemed the content did not violate the respective social media site's policies and was removed.

unmoderated cases and assessed them against reporting criteria. As our aim was to create a conservative estimate of the volume of content left unmoderated, rather than seeking reliability across coders, we decided to remove any cases that any researcher felt did not reach the criteria for reporting (in other words, only keeping cases where there was unanimous agreement among coders that content violated community guidelines). Moderated content was not re-reviewed as the platform had found it in violation. This made estimates of unmoderated content more conservative, by excluding potential unmoderated cases while crediting platforms for all moderated content.

Cases were often removed on technicality grounds, for instance, a TikTok account called “Cloud Connected” selling products with ‘no ID required’ was removed because it was never expressly stated the products sold were nicotine or cannabis products. In total 13 of 234 unmoderated cases (5.6%) were flagged by at least one researcher and were removed from the study (the number flagged by site can be seen in Table 2). For all other cases ($n = 221$) all three researchers deemed the content met the reporting criteria, leaving a final sample of $N = 294$. A breakdown of the sample is shown in Table 2.

2.4 Reporting Process

Consistent with previous research (Cover et al. 2026; Flynn et al. 2025; Zhang, Montague, and Jhaver 2024), we found platforms’ community guidelines often did not align with the options shown when reporting content (the full menu of options when reporting content is available in the OSF archive: Supplementary File 1). As such, we had to create rules for what options to use when reporting content. For instance, as Instagram has no options to report content for tobacco-related violations, we reported content under the “Other drugs” category, the closest category we could find.

Additionally, when reporting accounts on YouTube, we could find no reporting options matching the relevant community guidelines. Consequently, we only reported posts on YouTube and did not report accounts. The reporting options used for each site are available in the OSF archive: Supplementary File 1.

YouTube allows comments to be submitted with reports. Template responses were

created for each category (see OSF archive: Supplementary File 1) and used to submit comments with all reports to YouTube.

2.5 Response Options and Measures

All three platforms offer a page to monitor responses to submitted reports (directions to access these pages are available in the OSF archive: Supplementary File 2). Each case was monitored for at least two weeks after the report was made. Instagram and TikTok both usually provide a direct response to reports specifying their verdict. For instance, for posts deemed not in violation, a sample response from TikTok writes “We found that the reported content doesn’t violate our Community Guidelines,” while Instagram’s response reads “Our team reviewed [CREATOR’S] post. They found that it does not go against our Community Standards.” For removed content, a sample TikTok response reads “We’ve reviewed the video you reported and found that it violates our Community Guidelines. The video has been removed,” while Instagram writes “Thanks for reporting [CREATOR’S] post. We removed it from Instagram because it goes against our Community Standards.” Examples of all template responses received by the platforms are available in the OSF archive: Supplementary File 3.

Not all cases received a response from Instagram and TikTok. Additionally, on YouTube, the relevant page merely shows the current status of reported videos as either “Live,” “Restricted,” or “Removed,” and we were only able to monitor this page for changes to the status. Consequently, we monitored each case for at least two weeks for either a response from the platform (TikTok, Instagram) or a change in the status of the video (all platforms). After this time, all cases were settled into one of three groups.

2.5.1 Moderated

For insufficient age-gating, we considered the content moderated if it was age-gated or removed from the site. For all other categories, since the platforms stated the content was not allowed on the site, the content had to be removed within the two-week monitoring period to count as moderated. In cases where both the post and account were reported, we counted it as moderated if either was removed.

2.5.2 Suspended

After data collection, we noticed cases in which the platform stated it had removed content, but the content was later viewable on the site. In these instances, we considered the content “Suspended,” though we note the platform merely listed these as removed.

2.5.3 Unmoderated

Cases could be classed as unmoderated by one of three ways: (1) cases where the platform responded to say that the content did not violate their community guidelines; (2) for any category except insufficient age-gating, where the case received a minor moderation action (e.g. removed from the algorithm) but was not removed; and (3) where the platform failed to respond or take action on the case within the two-week monitoring period.

Examples of all types of responses received from the platforms are available in the OSF archive: Supplementary File 3.

2.6 Post and Account Metadata

To analyze the metadata of reported posts and accounts we reviewed the screen recordings made when submitting reports and recorded the relevant metadata viewable alongside the content at the time the report was made. For insufficient age-gating, metadata (e.g. likes) would be from users who should be able to view the content (i.e. not all likes would come from underage users). Consequently, we removed insufficient age-gating cases from analyses of metadata, leaving $n = 238$ cases (81%). As these 238 cases comprised $n = 11$ account cases; $n = 141$ post cases; and $n = 86$ post/account pair cases (see OSF archive: Supplementary File 6), we attempted to collect metadata for $n = 227$ posts and $n = 97$ accounts.

For reported posts we noted the like and comment count of the post. For accounts, we noted the number of followers the account had. If the full value was not shown and was truncated (e.g. 76.5k) we used the logical whole number (e.g. 76,500). For some posts, the metadata was not viewable in the reporting video: for likes $n = 209$ of 227 reported posts (92.1%) were available (Instagram = 59 of 69, 85.5%; TikTok = 79 of 79, 100%; YouTube = 71 of 79, 89.9%); for comments $n = 165$ of 227 (72.7%) were available (Instagram = 36 of 69, 52.2%; TikTok = 79 of 79, 100%; YouTube = 50 of 79, 63.3%). There was no missing data for account followers.

The view counts of reported posts were not recorded. However, we did make attempts to estimate this. The engagement rate (percentage of likes and comments per view) for Instagram has been measured at 0.48% (Cucu 2026a) and TikTok, 3.85% (Cucu 2026b). Consequently, we estimate views by taking the cumulative likes and comments and multiplying by 208.33 for Instagram ($100/0.48$) and 25.97 ($100/3.85$) for TikTok per video. On YouTube, an average view-to-like ratio is estimated to be between 4% and 6% (Rai 2025). We take an estimate of 5%, multiplying a video's likes by 20 to estimate views on YouTube.

Table 3: Items moderated per site and reporting category. Percentages are of the row total.

	Unmoderated	Suspended	Moderated ^a	Total Reported
Instagram				
Insufficient Age-Gating	4 (20%)	0 (0%)	16 (80%)	20
Disallowed Promotions	26 (92.9%)	1 (3.6%)	1 (3.6%)	28
Illegal Sales	4 (11.1%)	1 (2.8%)	31 (86.1%)	36
Underage Use	6 (100%)	0 (0%)	0 (0%)	6
Dangerous Activities	7 (87.5%)	0 (0%)	1 (12.5%)	8
Total	47 (48%)	2 (2%)	49 (50%)	98
TikTok				
Insufficient Age-Gating	13 (72.2%)	0 (0%)	5 (27.8%)	18
Disallowed Promotions	27 (100%)	0 (0%)	0 (0%)	27
Illegal Sales	35 (97.2%)	0 (0%)	1 (2.8%)	36
Underage Use	6 (75%)	0 (0%)	2 (25%)	8
Dangerous Activities	10 (100%)	0 (0%)	0 (0%)	10
Total	91 (91.9%)	0 (0%)	8 (8.1%)	99
YouTube				
Insufficient Age-Gating	10 (55.6%)	0 (0%)	8 (44.4%)	18
Disallowed Promotions	22 (91.7%)	0 (0%)	2 (8.3%)	24
Illegal Sales	27 (93.1%)	0 (0%)	2 (6.9%)	29
Underage Use	13 (86.7%)	1 (6.7%)	1 (6.7%)	15
Dangerous Activities	11 (100%)	0 (0%)	0 (0%)	11
Total	83 (85.6%)	1 (1%)	13 (13.4%)	97

^a For insufficient age-gating, for an item to be moderated it had to be placed behind an age-gate or removed. For all other categories, moderated means the item was removed.

3 Results

3.1 Moderation Responsiveness

A detailed breakdown of the total number of cases reported and moderated per site and per reporting category is shown in Table 3. To aid transparency and understanding, examples of unmoderated content are available in the OSF archive: Supplementary File 4.

Of the total 294 reported cases, 221 (75%) remained unmoderated. Instagram moderated nearly half of reported cases ($n = 49$; 50%), while moderation rates were much lower on YouTube ($n = 13$, 13.4%) and TikTok ($n = 8$; 8.1%). Instagram's rate of moderation would rise to 48 of 84 cases (57.1%) if dangerous activity and underage use reports were removed, which were not directly prohibited by Instagram's Community Guidelines.

Moderation varied by reporting category for each platform. Instagram regularly moderated illegal sales ($n = 32$; 88.9%) and content with insufficient age-gating ($n = 16$;

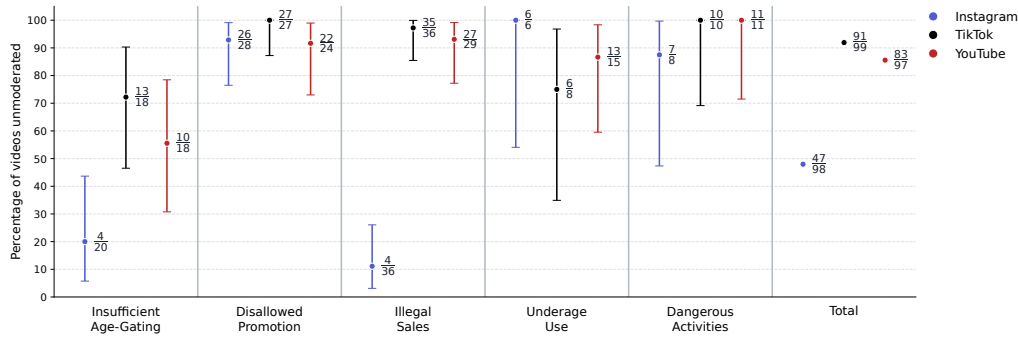


Figure 1: Percentage of content unmoderated with confidence intervals for each category. *Note.* Labels next to points indicate the total number of posts unmoderated over the total number of posts reported for that category and site. Confidence intervals calculated with Clopper-Pearson Interval. Confidence intervals are not provided for the “Total” column as this would contravene the assumptions of the test.

80%), but rarely moderated disallowed promotion ($n = 2$; 7.1%). On TikTok and YouTube, moderation was low across all categories, but content with insufficient age-gating was the highest moderated on both TikTok ($n = 5$; 27.8%) and YouTube ($n = 8$; 44.4%). As a result, insufficient age-gating was the only category where over 50% of cases were moderated in total ($n = 29$ of 56, 51.8%). Across all platforms, about a third of illegal sales content was moderated, and less than 10% of dangerous activity posts and disallowed promotions were moderated.

To create a confidence interval around the unmoderation rate (the rate at which content was not moderated) we used the Clopper-Pearson Interval (Wheeler 2020). The results of this analysis can be seen in Figure 1.

While, due to the low number of reports, confidence intervals for insufficient age-gating, underage use and dangerous activities are large, confidence intervals for illegal sales and disallowed promotions are much smaller, and suggest that even at a lower-bound estimate, 85.5% of illegal sales content on TikTok and 77.2% on YouTube go unmoderated.

We also find that Instagram left posts lacking age-gating unmoderated (LCI-UCI = 5.7%–43.7%) less often than TikTok (LCI-UCI = 46.5%–90.3%), and left illegal sales unmoderated (LCI-UCI = 3.1%–26.1%) less often than TikTok (LCI-UCI = 85.5%–99.9%) and YouTube (LCI-UCI = 77.2%–99.2%).

3.2 Analyzing Post and Account Metadata

Metadata for reported posts can be seen in Table 4, while metadata for reported accounts can be seen in Table 5. Unmoderated posts that were, by the platforms’ own community guidelines, prohibited from the platform accounted for 688,784 likes and an estimated 15,714,012 views. The estimated view totals are extrapolated estimates from platform-wide engagement benchmarks (0.48% for Instagram, 3.85% for TikTok, 5% for YouTube).

Table 4: Metadata values for all reported posts and unmoderated posts per site. N is the number of posts for which the metadata was recorded; percentages are of the total number of posts reported.

	Likes			Comments			Views ^a		
	N (%)	Min–Max	Sum	N (%)	Min–Max	Sum	N (%)	Min–Max	Sum
Instagram									
All	59 (85.5%)	0–55,996	126,740	36 (52.2%)	0–238	553	36 (52.2%)	0–1,661,250	2,761,667
Unmoderated	35 (83.3%)	3–55,996	125,506	12 (28.6%)	0–238	474	12 (28.6%)	833–1,661,250	2,488,125
TikTok									
All	79 (100%)	0–53,200	376,255	79 (100%)	0–1,183	5,601	79 (100%)	0–1,412,545	9,918,338
Unmoderated	76 (100%)	0–49,800	309,420	76 (100%)	0–886	4,306	76 (100%)	0–1,300,987	8,148,727
YouTube									
All	71 (89.9%)	0–86,000	254,200	50 (63.3%)	0–940	3,821	71 (89.9%)	0–1,720,000	5,084,000
Unmoderated	66 (90.4%)	0–86,000	253,858	47 (64.4%)	0–940	3,797	66 (90.4%)	0–1,720,000	5,077,160

^a Views are an estimate based on likes and engagement rates (Cucu 2026a, 2026b; Rai 2025).

Reported accounts whose entire purpose contravened platform community guidelines but not moderated had a cumulative follower count of 482,058, almost all of which came from unmoderated accounts on TikTok.

4 Discussion

Though moderating platforms with millions of daily uploads is challenging, our analysis shows that even when clearly violating content is reported directly to the platforms using their own reporting mechanisms, tobacco-related content is rarely moderated. Our ambition was to identify and report content that was unambiguously in violation. As such, our threshold for reporting content was intentionally stringent. Yet, content that was easy to identify as violating the site’s policies (as well as, in some cases, U.S. federal law) was left unmoderated, including depictions of underage use and accounts openly advertising the sale of tobacco products to youth. The fact that some reported content was removed shows that sites are capable of finding rule-breaking content in violation. However, it appears moderation was often applied rarely and sporadically.

Notably, the platforms not only failed to remove the content, but often responded by saying the content did not violate their policies. Instagram’s policies state the branded content tool cannot be used to sell tobacco products (Meta, [n.d.-a](#)), but when we reported posts doing this, the response stated they found it “does not go against our Community Standards.” On TikTok, we reported an account noted in a case study for openly illegally selling e-cigarettes to youth (Pearson et al. 2025a). Yet, TikTok informed us the account “doesn’t violate our Community Guidelines” (see OSF archive: Supplementary File 4). These responses are a key addition to existing literature on content moderation practices. While not removing content can be seen as an implicit message that the content does not violate community guidelines, we record Instagram and TikTok writing to users to confirm that content was reviewed and that they deemed it not in

Table 5: Follower counts for all reported accounts and unmoderated accounts.

	N (%) ^a	Min–Max (Sum)
Instagram		
All	26 (100%)	0–4,333 (11,235)
Unmoderated	3 (100%)	27–120 (187)
TikTok		
All	60 (100%)	0–128,200 (482,024)
Unmoderated	59 (100%)	0–128,200 (481,871)

^a Number is the total number of accounts for which the metadata was recorded, percentage is of the total number of accounts reported.

violation despite unambiguous violations. This suggests a deeper misalignment between published community guidelines and platforms’ own internal moderation practices. Such misalignment, and the platforms’ practice of directly notifying users that content is not in violation, may have increased the lack of trust many users have in reporting mechanisms (Flynn et al. 2025; Vilks 2023; Zhong et al. 2026).

We also note problems with the reporting mechanisms themselves. In one instance, we reported content to Instagram (verified by reviewing screen recordings), but the report never showed up on the profile’s list of reports. Additionally, there were instances where reports initially were shown but later disappeared. Perhaps more concerning, we noted instances where the platforms informed us that content had been removed but was later viewable. While we classed these instances as “suspended,” the platform only described them as “removed” and in some instances, this content was viewable less than 48 hours after receiving the “removed” message. Informing users that content was removed when it is still viewable may create a false confidence in the moderation actions of the platform.

The options presented to users when reporting content also poorly matched the platforms’ community guidelines; in the case of YouTube leaving us unable to report accounts completely. These findings corroborate qualitative research that finds the confusing options when reporting content were a barrier to user reports (Flynn et al. 2025; Zhang, Montague, and Jhaver 2024). Improving reporting mechanisms will aid users and make it easier for platforms to understand why content has been reported. Simple options, such as better alignment between reporting categories and community guidelines, as well as more advanced features, such as the ability to search guidelines by keyword while reporting, may aid users. Allowing reporters to submit comments with their reports (such as on YouTube) can aid users to communicate rule-breaking content (Shim and Jhaver 2024) provided these comments are used by moderators.

There was variation in how successfully platforms moderated violations; particularly for insufficient age-gating and illegal sales. Regulators may wish to pay closer scrutiny to

certain platforms as a result. Variation existed within platforms as well. For instance, while Instagram was successful at removing illegal sales content, few posts breaking Instagram's rules on branded content were removed. Such variation within platforms suggests that successful moderation is not universal—a platform may be successful at removing one type of harmful content, but unsuccessful in others. Additionally, the moderation rates we find here are low compared to some studies in different contexts (Carlson and Rousselle 2020; European Commission 2025; Papakyriakopoulos, Medina Serrano, and Hegelich 2020). A possible explanation could be that substance use may receive less attention for moderation (compared to issues which receive greater attention in moderation literature, such as hate speech, e.g., Alkiviadou 2025; Carlson and Rousselle 2020; Center for Countering Digital Hate 2022a; The Future of Free Speech 2024). However, further research is needed to identify the factors contributing to this variation. Regardless of the cause of such variations, audits of content moderation—both by external organizations and by the platforms themselves—should consider multiple contexts, and literature should avoid solely focusing on specific moderation contexts.

This is especially the case given differing priorities and risks for over moderation in public health or substance use versus political contexts. While much moderation work is concerned with how moderation may impact freedom of speech (Alkiviadou 2025; The Future of Free Speech 2024), in some public health contexts, especially substance use, over moderation is often less concerning. While public health concerns exist that over moderation may inadvertently censor public health messaging (Gomes and Sultan 2024) or cut vulnerable populations off from supportive messaging (Zhang et al. 2024), the speech rights of, for instance, drug dealers to reach customers (Fuller et al. 2024; Li et al. 2019), alcohol brands to market products (Carah and Brodmerkel 2021), or both legal (Kong et al. 2021) and illegal (Pearson et al. 2025a) tobacco sellers to market products are less of a concern. Consequently, differing contexts may require different moderation standards.

Attempts to understand different content contexts are hindered by platforms' transparency reports lacking alignment with community guidelines. Although all three platforms have tobacco-related policies, platform transparency reports (which often report high numbers of content removals; Google, n.d.-d; Meta, n.d.-b; TikTok, n.d.-b) do not highlight which community guideline rule the content violated. Consequently, it is hard to meaningfully compare our results here to existing transparency reports.

These findings call into question the ability of social media platforms to self-regulate and moderate the content disseminated on their platforms, highlighting a need for policy and infrastructure to facilitate third-party monitoring and removal of egregiously harmful content. The European Union's (EU's) Digital Services Act (DSA) has imposed such a framework on these platforms, with provisions relevant to our findings, including an obligation to enable third-party audits, guidelines around reporting, expeditious

action by the platforms when notified of illegal content, and explicit focus on consumer protection (European Commission, n.d.). Evidence of overzealous moderation continues to raise concerns regarding online censorship (The Future of Free Speech 2024). Despite these limitations, the framework set up by the DSA enables promising opportunities for refinement of dubious criteria and transparency to evaluate and improve enforcement (Chapman 2025; Griffin 2025). While the DSA's impact on countries outside of the EU, such as the U.S. remains both contentious (Armangau and Tadeo 2025) and hypothetical (Nunziato 2023), future comparative research is needed to determine whether our results are specific to the U.S., due to differences in moderation practices between topics, or indicative of an over-correction of moderation practices in response to early DSA findings.

Previous work using the commercial determinants of health framework has argued that the profit motivations of platforms may lead to actions that impact public health (Zenone, Kenworthy, and Maani 2023). Moderation is both an expensive endeavor requiring substantial resources (Gillespie 2018a) and one that can decrease user engagement (Liu, Yildirim, and Zhang 2022). The results here suggest platforms may need to increase efforts and the resources exhausted to moderate in line with their community guidelines.

Other options may increase the efficiency of moderation. While automated tools are prone to failure (Gerrard 2018; Gorwa, Binns, and Katzenbach 2020), direct and clear language such as “buy vape no ID needed” should be identifiable with AI tools, removing the most obvious violators. While removing all substance trading from platforms may be impossible, simple steps can make sure content is harder to find and spreads less easily. Platforms could also partner with organizations in key disciplines (such as mental health, substance use, etc.) to identify common trends and slang.

Community guidelines act as a representation for what regulators, advertisers, users and guardians can expect on the platform and what actions the platform will take to remove violations. Our results suggest such guidelines may not be an accurate portrayal of content moderation actions. Platforms either need to improve their moderation capabilities or do a greater job of acknowledging their limitations in protecting users from harmful content.

4.1 Limitations

We can only view the content that is reported and subsequent moderation decisions. Other unknown factors may influence moderation, including both content features—for instance whether content received other reports—and features of the reporting account. We used a new account with minimal history, made over 300 reports to the platforms, and we cannot be certain what impact network or device fingerprinting may play. Using a brand-new account with minimal history to report content found via search may look like inorganic behavior, or potential brigading, leading to our reports being ignored or

deprioritized. However, some of our reported cases were moderated throughout the study period, suggesting, at the very least, that our account was not shadow banned from making reports.

Due to the imprecise language used in the policies, we had to create our own rules for identifying violations. The platform's own internal procedures are likely more specific than public-facing community guidelines (Angwin and Grassegger 2017; Langvardt 2017; Myers West 2018) and the lack of removal actions for some cases may be due to this mismatch. However, to counter this, our guidelines sought to restrict our reporting to only unambiguous violations (for instance, when reporting content that 'promoted' tobacco products on YouTube, we only reported promotional content made specifically by commercial tobacco entities).

Our focus was on assessing content moderation actions after content was reported, and we did not focus on the volume or proportion of overall content that was violating. Other work has documented the variety and volume of tobacco content on social media (Dobbs, Schisler, and McCormick 2025; Kong et al. 2021; O'Brien et al. 2020; Sharp et al. 2025). We cannot make claims to how prevalent rule-breaking content is, how regularly users encounter it, and can only make estimates of violating contents' reach and viewership. Future work may wish to focus on real-world behavior; creating a naturalistic feed and only reporting content that organically appeared. Virality can impact moderation decisions (King 2020), and such an approach may lead to more recent, viral content being reported with greater moderation success. Previous research has noted minor changes in the frequency of harmful content shown in platforms' algorithmic feeds compared to active searching (Eltaher et al. 2026).

While TikTok and YouTube have specific policies forbidding underage use of tobacco products and dangerous acts, Instagram does not. Given the potential concerning effects of this type of content (Silver et al. 2026; HHS 2012), and to have greater consistency in the reported categories between sites, we reported violations in these categories. Violations on Instagram for these categories should be considered with this in mind.

We must exercise caution as to the generalizability of the results. Sample sizes were small, and therefore categories had large confidence intervals. However, in some instances well over 90% of cases were left unmoderated suggesting systematic under moderation. Differences in the moderation rates between sites must also be treated with caution. Not only do community guidelines differ between platforms, but the content reported also varied. Still, the different rates at which some categories of violations were moderated between platforms may reflect underlying differences. Lastly, while we make an estimate of video views counts, these numbers are an estimate and should the actual engagement-to-view ratio differ, actual view counts may vary considerably.

4.2 Conclusions

This study offers a replicable methodology for measuring the moderation actions of social media platforms in response to user reports. This can be used in other areas of key public health interest, such as self-harm (Bryant 2024; Dyson et al. 2016; Molly Rose Foundation 2025b; Susi et al. 2023), eating disorders (Dane and Bhatia 2023; Molly Rose Foundation 2025a) and substance use (Carah and Brodmerkel 2021; Moreno and Whitehill 2016).

Social media platforms continue to make claims about the steps taken to protect users from harmful content (Domingo 2024; Kindelan 2024; Paul 2024) and create policies that prohibit certain content from the site (Google, n.d.-c; Meta, n.d.-c; TikTok, n.d.-a). However, our findings show that even when users report content that clearly violates community guidelines, social media platforms rarely take moderation action, countering claims that platforms can successfully safeguard their platforms.

References

- Alkiviadou, Natalie. 2025. *Moderating Hate or Moderating Rights? The Paradox of the European Approach to Online Hate Speech and Platform Liability*, CELE Research Paper 69. CELE. <https://doi.org/10.2139/ssrn.5432875>.
- Angwin, Julia, and Hannes Grassegger. 2017. "Facebook's Secret Censorship Rules Protect White Men from Hate Speech but Not Black Children." ProPublica, June. Accessed July 14, 2026. <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>.
- Armangau, Romane, and Maria Tadeo. 2025. "Europe Defends Its Digital Rules after US Targets Breton with Visa Ban." euronews, December 24, 2025. Accessed July 31, 2026. <https://www.euronews.com/my-europe/2025/12/24/europe-defends-its-digital-rules-after-trump-administration-targets-breton-with-visa-ban>.
- Armstrong, Paul W., and C. David Naylor. 2019. "Counteracting Health Misinformation: A Role for Medical Journals?" *JAMA* 321 (19): 1863–64. <https://doi.org/10.1001/jama.2019.5168>.
- Bode, Leticia, and Emily K. Vraga. 2018. "See Something, Say Something: Correction of Global Health Misinformation on Social Media." *Health Communication* 33 (9): 1131–40. <https://doi.org/10.1080/10410236.2017.1331312>.
- Bryant, Miranda. 2024. "Instagram Actively Helping Spread of Self-Harm among Teenagers, Study Finds." *Guardian*, November 30, 2024. <https://www.theguardian.com/technology/2024/nov/30/instagram-actively-helping-to-spread-of-self-harm-among-teenagers-study-suggests>.
- Carah, Nicholas, and Sven Brodmerkel. 2021. "Alcohol Marketing in the Era of Digital Media Platforms." *Journal of Studies on Alcohol and Drugs* 82 (1): 18–27. <https://doi.org/10.15288/jsad.2021.82.18>.
- Carlson, Caitlin Ring, and Hayley Rousselle. 2020. "Report and Repeat: Investigating Facebook's Hate Speech Removal Process." *First Monday*, <https://doi.org/10.5210/fm.v25i2.10288>.
- Center for Countering Digital Hate. 2022a. *Failure to Protect: How Tech Giants Fail to Act on User Reports of Antisemitism*. <https://counterhate.com/wp-content/uploads/2022/05/Failure-to-Protect.pdf>.
- . 2022b. *Failure to Protect: Social Media Platforms Are Failing to Act on Anti-Muslim Hate*. <https://counterhate.com/wp-content/uploads/2022/05/Anti-Muslim-Hate-Failure-to-Protect.pdf>.
- Chadi, Nicholas, Claudia Minato, and Richard Stanwick. 2020. "Cannabis Vaping: Understanding the Health Risks of a Rapidly Emerging Trend." *Paediatrics & Child Health* 25 (Suppl. 1): S16–S20. <https://doi.org/10.1093/pch/pxaa016>.

- Chapman, Peter. 2025. "Advancing Platform Accountability: The Promise and Perils of DSA Risk Assessments." Tech Policy Press, January. Accessed July 31, 2026. <https://techpolicy.press/advancing-platform-accountability-the-promise-and-perils-of-dsa-risk-assessments>.
- Chou, Wen-Ying Sylvia, April Oh, and William M. P. Klein. 2018. "Addressing Health-Related Misinformation on Social Media." *JAMA* 320 (23): 2417–18. <https://doi.org/10.1001/jama.2018.16865>.
- Civiletto, Cody W., Sunny P. Aslam, and Julia Hutchison. 2024. "Electronic Vaping Delivery of Cannabis and Nicotine." In *StatPearls*. Treasure Island (FL): StatPearls Publishing. <https://www.ncbi.nlm.nih.gov/books/NBK545160/>.
- Cover, Rob, Jennifer Beckett, Benedetta Brevini, Catharine Lumby, Rhyle Simcock, and Jay Daniel Thompson. 2026. "Reporting Online Abuse to Platforms: Factors, Interfaces and the Potential for Care." *Convergence: The International Journal of Research into New Media Technologies* 32 (1): 142–58. <https://doi.org/10.1177/13548565251324508>.
- Crawford, Kate, and Tarleton Gillespie. 2016. "What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint." *New Media & Society* 18 (3): 410–28. <https://doi.org/10.1177/1461444814543163>.
- Cucu, Elena. 2026a. "2026 Instagram Organic Engagement Benchmarks." Social Insider, February. Accessed July 21, 2026. <https://www.socialinsider.io/social-media-benchmarks/instagram>.
- . 2026b. "2026 TikTok Benchmarks." Social Insider, April. Accessed July 21, 2026. <https://www.socialinsider.io/social-media-benchmarks/tiktok>.
- Dane, Alexandra, and Komal Bhatia. 2023. "The Social Media Diet: A Scoping Review to Investigate the Association between Social Media, Body Image and Eating Disorders amongst Young People." Edited by Rakesh Singh. *PLOS Global Public Health* 3 (3). <https://doi.org/10.1371/journal.pgph.0001091>.
- DataReportal. 2025. "Digital 2025: The United States of America," February. Accessed July 26, 2026. <https://datareportal.com/reports/digital-2025-united-states-of-america>.
- de Vere Hunt, Isabella, and Eleni Linos. 2022. "Social Media for Public Health: Framework for Social Media–Based Public Health Campaigns." *Journal of Medical Internet Research* 24 (12). <https://doi.org/10.2196/42179>.
- Digitalt Ansvar. 2024. "InstaHARM: En undersøgelse af Instagrams manglende indholdsmoderation af selvskadeindhold," November. Accessed September 28, 2026. <https://www.digitaltansvar.dk/aktuelt/instagrams-moderation-af-selvskade>.

- Dobbs, Page D., Eric D. Schisler, and Charlotte McCormick. 2025. “#Discreetshipping: Selling E-Cigarettes on TikTok.” *Nicotine & Tobacco Research* 27 (4): 748–52. <https://doi.org/10.1093/ntr/ntae081>.
- Domingo, Aldohn. 2024. “Meta, TikTok, Big Tech CEOs, Set to Defend Social Media Safety for Kids.” *Tech Times*, January 30, 2024. Accessed April 29, 2026. <https://www.techtimes.com/articles/301172/20240130/meta-tiktok-big-tech-ceo-defend-social-media-safety.htm>.
- Donaldson, Scott I., Amanda DeJesus, Emily Wang, Ollie Ganz, Julia Chen-Sankey, and Jon-Patrick Allem. 2026. “Evaluating Policy Compliance among Leading Tobacco Brands on Instagram.” *Tobacco Control*, <https://doi.org/10.1136/tc-2025-059609>.
- Donaldson, Scott I., Allison Dormanesh, Cindy Perez, Anuja Majmundar, and Jon-Patrick Allem. 2022. “Association between Exposure to Tobacco Content on Social Media and Tobacco Use: A Systematic Review and Meta-analysis.” *JAMA Pediatrics* 176 (9). <https://doi.org/10.1001/jamapediatrics.2022.2223>.
- Duffy, Clare. 2025. “‘You Never Want to Leave’: TikTok Employees Raise Concerns about the App’s Impact on Teens in Newly Unsealed Video.” *CNN Business*, August 20, 2025. Accessed April 29, 2026. <https://www.cnn.com/2025/08/20/tech/tiktok-north-carolina-lawsuit-unsealed-employee-video>.
- Dyson, Michele P., Lisa Hartling, Jocelyn Shulhan, Annabritt Chisholm, Andrea Milne, Purnima Sundar, Shannon D. Scott, and Amanda S. Newton. 2016. “A Systematic Review of Social Media Use to Discuss and View Deliberate Self-Harm Acts.” Edited by Soraya Seedat. *PLOS ONE* 11 (5). <https://doi.org/10.1371/journal.pone.0155813>.
- Edelson, Laura, Borys Kovba, Hanna Yershova, Austin Botelho, Damon McCoy, and Tobias Lauinger. 2025. “Measurement and Metrics for Content Moderation: The Multi-Dimensional Dynamics of Engagement and Content Removal on Facebook.” *Journal of Online Trust and Safety* 2 (5). <https://doi.org/10.54501/jots.v2i5.220>.
- Eltaher, Fatmaelzahraa, Rahul Krishna Gajula, Luis Miralles-Pechuán, Patrick Crotty, Juan Martínez-Otero, Christina Thorpe, and Susan McKeever. 2026. “Protecting Young Users on Social Media: Evaluating the Effectiveness of Content Moderation and Legal Safeguards on Video-Sharing Platforms.” *Journal of Online Trust and Safety* 3 (2). <https://doi.org/10.54501/jots.v3i2.251>.
- European Commission. 2025. “2025 Monitoring Exercise under the Code of Conduct on Countering Illegal Hate Speech Online+: Factsheet.” Accessed September 24, 2026. <https://ec.europa.eu/newsroom/dae/redirection/document/126826>.
- . n.d. “User Rights under the Digital Services Act.” *Shaping Europe’s Digital Future*. Accessed July 20, 2026. <https://digital-strategy.ec.europa.eu/en/factpages/user-rights-under-digital-services-act>.

- Flynn, Asher, Zarina Vakhitova, Lisa Wheildon, Bridget Harris, and Brady Robards. 2025. "Content Moderation and Community Standards: The Disconnect between Policy and User Experiences Reporting Harmful and Offensive Content on Social Media." *Policy & Internet* 17 (3). <https://doi.org/10.1002/poi3.70006>.
- Food and Drug Administration. 2019. *The Public Health Rationale for Recommended Restrictions on New Tobacco Product Labeling, Advertising, Marketing, and Promotion*. <https://www.fda.gov/media/124174/download>.
- . 2023. "Regulation and Enforcement of Non-Tobacco Nicotine (NTN) Products." Accessed April 21, 2026. <https://www.fda.gov/tobacco-products/products-ingredients-components/regulation-and-enforcement-non-tobacco-nicotine-ntn-products>.
- Freeman, Becky, Christina Watts, and Putu Ayu Swandewi Astuti. 2022. "Global Tobacco Advertising, Promotion and Sponsorship Regulation: What's Old, What's New and Where to Next?" *Tobacco Control* 31:216–21. <https://doi.org/10.1136/tobaccocontrol-2021-056551>.
- Fuller, Ashly, Marie Vasek, Enrico Mariconti, and Shane D. Johnson. 2024. "Understanding and Preventing the Advertisement and Sale of Illicit Drugs to Young People through Social Media: A Multidisciplinary Scoping Review." *Drug and Alcohol Review* 43 (1): 56–74. <https://doi.org/10.1111/dar.13716>.
- Gerrard, Ysabel. 2018. "Beyond the Hashtag: Circumventing Content Moderation on Social Media." *New Media & Society* 20 (12): 4492–511. <https://doi.org/10.1177/1461444818776611>.
- Gillespie, Tarleton. 2018a. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press. <https://doi.org/10.12987/9780300235029>.
- . 2018b. "Platforms Are Not Intermediaries." *Georgetown Law Technology Review* 2:198–216. <https://georgetownlawtechreview.org/platforms-are-not-intermediaries/GLTR-07-2018/>.
- . 2021. "'Napalm Girl.'" Yale University Press, September 20, 2021. Accessed April 21, 2026. <https://yalebooks.yale.edu/2021/09/20/napalm-girl/>.
- Global Witness. 2023. "Hate Speech Targeting Women Journalists Approved by Platforms." Accessed April 22, 2026. <https://globalwitness.org/en/campaigns/digital-threats/violent-and-sexualised-hate-speech-targeting-women-approved-for-publication-by-social-media-platforms/>.
- Goggin, Ben. 2024. "Big Tech Companies Reveal Trust and Safety Cuts in Disclosures to Senate Judiciary Committee." NBC News, March 29, 2024. Accessed April 17, 2026. <https://www.nbcnews.com/tech/tech-news/big-tech-companies-reveal-trust-safety-cuts-disclosures-senate-judicia-rcna145435>.

- Gomes, André Belchior, and Aysel Sultan. 2024. "Problematizing Content Moderation by Social Media Platforms and Its Impact on Digital Harm Reduction." *Harm Reduction Journal* 21. <https://doi.org/10.1186/s12954-024-01104-9>.
- Google. n.d.-a. "Harmful or Dangerous Content Policy." YouTube Help. Accessed January 13, 2026. <https://support.google.com/youtube/answer/2801964?hl=en>.
- . n.d.-b. "How to Use the Site Search Operator." Google for Developers. Accessed July 30, 2026. <https://developers.google.com/search/docs/monitor-debug/search-operators/all-search-site>.
- . n.d.-c. "Illegal or Regulated Goods or Services Policies." YouTube Help. Accessed January 13, 2026. <https://support.google.com/youtube/answer/9229611?hl=en#zippy=%2Cexamples-of-age-restricted-content>.
- . n.d.-d. "YouTube Community Guidelines Enforcement." Google Transparency Report. Accessed July 15, 2026. <https://transparencyreport.google.com/youtube-policy/removals>.
- Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1). <https://doi.org/10.1177/2053951719897945>.
- Green, Ben. 2021. "The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice." *Journal of Social Computing* 2 (3): 209–25. <https://doi.org/10.23919/JSC.2021.0018>.
- Griffin, Rachel. 2025. "Governing Platforms through Corporate Risk Management: The Politics of Systemic Risk in the Digital Services Act." *European Law Open* 4 (2): 223–53. <https://doi.org/10.1017/elo.2025.17>.
- Harakeh, Zeena, Rutger C. M. E. Engels, Rick B. Van Baaren, and Ron H. J. Scholte. 2007. "Imitation of Cigarette Smoking: An Experimental Study on Smoking in a Naturalistic Setting." *Drug and Alcohol Dependence* 86 (2–3): 199–206. <https://doi.org/10.1016/j.drugalcdep.2006.06.006>.
- Horwitz, Jeff. 2025a. "Meta Buried 'Causal' Evidence of Social Media Harm, US Court Filings Allege." Reuters, November 23, 2025. Accessed July 31, 2026. <https://www.reuters.com/sustainability/boards-policy-regulation/meta-buried-causal-evidence-social-media-harm-us-court-filings-allege-2025-11-23/>.
- . 2025b. "Meta Is Earning a Fortune on a Deluge of Fraudulent Ads, Documents Show." Reuters, November 6, 2025. Accessed April 29, 2026. <https://www.reuters.com/investigations/meta-is-earning-fortune-deluge-fraudulent-ads-documents-show-2025-11-06/>.
- Huang, Chiungjung. 2022. "A Meta-analysis of the Problematic Social Media Use and Mental Health." *International Journal of Social Psychiatry* 68 (1): 12–33. <https://doi.org/10.1177/0020764020978434>.

- Instagram. n.d. *A Parent and Guardian's Guide to Instagram*. https://scontent-iad3-2.xx.fbcdn.net/v/t39.8562-6/10000000_135331025861966_430020200632791153_2_n.pdf.
- Jackson, Jasper. 2025. "Social Media Moderators' Lives Are Getting Worse. Big Tech Needs to Take Responsibility." The Bureau of Investigative Journalism, April 27, 2025. Accessed April 17, 2026. <https://www.thebureauinvestigates.com/stories/2025-04-27/social-media-moderators-lives-are-getting-worse.-big-tech-needs-to-take-responsibility>.
- Jafar, Zain, Jonathan D. Quick, Heidi J. Larson, Verner Venegas-Vera, Philip Napoli, Godfrey Musuka, Tafadzwa Dzinamarira, Kolar Sridara Meena, T. Raju Kanmani, and Eszter Rimányi. 2023. "Social Media for Public Health: Reaping the Benefits, Mitigating the Harms." *Health Promotion Perspectives* 13 (2): 105–12. <https://doi.org/10.34172/hpp.2023.13>.
- Kerr, Dara. 2024. "Snapchat Brushed Aside Warnings of Child Harm, Documents Show." NPR, October 4, 2024. Accessed July 31, 2026. <https://www.npr.org/2024/10/04/nx-s1-5137218/snapchat-brushed-aside-warnings-of-child-harm-documents-show>.
- Kickbusch, Ilona, Luke Allen, and Christian Franz. 2016. "The Commercial Determinants of Health." *The Lancet Global Health* 4 (12): e895–e896. [https://doi.org/10.1016/S2214-109X\(16\)30217-0](https://doi.org/10.1016/S2214-109X(16)30217-0).
- Kindelan, Katie. 2024. "Instagram Introduces Mandatory 'Teen Accounts' with Built-In Limits, Parental Controls." ABC News, September 17, 2024. Accessed April 16, 2026. <https://abcnews.com/GMA/Family/instagram-introduces-teen-accounts/story?id=113734038>.
- King, Jeff. 2020. "How We Review Content." Meta Newsroom, August. Accessed July 15, 2026. <https://about.fb.com/news/2020/08/how-we-review-content/>.
- Klonick, Kate. 2018. "The New Governors: The People, Rules, and Processes Governing Online Speech." *Harvard Law Review* 131 (6): 1598–670. <https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/>.
- Kong, Grace, Karissa E. Kuguru, Harmanpreet Bhatti, Isha Sen, and Meghan E. Morean. 2021. "Marketing Content on E-Cigarette Brand-Sponsored Facebook Profile Pages." *Substance Use & Misuse* 56 (4): 442–48. <https://doi.org/10.1080/10826084.2021.1878223>.
- Kong, Grace, Linnea Laestadius, Julia Vassey, Anuja Majmundar, Andrea M. Stroup, Helen I. Meissner, Ziyad Ben Taleb, Tess Boley Cruz, Sherry L. Emery, and Dan Romer. 2024. "Tobacco Promotion Restriction Policies on Social Media." *Tobacco Control* 33 (3): 398–403. <https://doi.org/10.1136/tc-2022-057348>.

- Kruger, Judy, Roshni Patel, Michelle C. Kegler, Nancy D. Brener, and Brian A. King. 2015. "National and State Attitudes of US Adults toward Tobacco-Free School Grounds, 2009–2010." *Preventing Chronic Disease* 12. <https://www.ncbi.nlm.nih.gov/pmc/articles/4699742>.
- Lafontaine, Jesse, Isabel Hanson, and Cervantée Wild. 2025. "The Impact of the Social Media Industry as a Commercial Determinant of Health on the Digital Food Environment for Children and Adolescents: A Scoping Review." *BMJ Global Health* 10 (2). <https://doi.org/10.1136/bmjgh-2023-014667>.
- Langvardt, Kyle. 2017. "Regulating Online Content Moderation." *SSRN Electronic Journal*, <https://doi.org/10.2139/ssrn.3024739>.
- Lerman, Kristina, Minh Duc Chu, Charles Bickham, Luca Luceri, and Emilio Ferrara. 2025. "Safe Spaces or Toxic Places? Content Moderation and Social Dynamics of Online Eating Disorder Communities." *EPJ Data Science* 14 (1). <https://doi.org/10.1140/epjds/s13688-025-00575-5>.
- Li, Jiawei, Qing Xu, Neal Shah, and Tim K. Mackey. 2019. "A Machine Learning Approach for the Detection and Characterization of Illicit Drug Dealers on Instagram: Model Evaluation Study." *Journal of Medical Internet Research* 21 (6). <https://doi.org/10.2196/13803>.
- Liu, Yi, Pinar Yildirim, and Z. John Zhang. 2022. "Implications of Revenue Models and Technology for Content Moderation Strategies." *Marketing Science* 41 (4): 831–47. <https://doi.org/10.1287/mksc.2022.1361>.
- Madio, Leonardo, and Martin Quinn. 2025. "Content Moderation and Advertising in Social Media Platforms." *Journal of Economics & Management Strategy* 34 (2): 342–69. <https://doi.org/10.1111/jems.12602>.
- Meta. n.d.-a. "Branded Content Policies." Meta Business Help Center. Accessed January 13, 2026. <https://www.facebook.com/business/help/221149188908254>.
- . n.d.-b. "Community Standards Enforcement." Meta Transparency Center. Accessed July 15, 2026. <https://transparency.meta.com/reports/community-standards-enforcement/regulated-goods/>.
- . n.d.-c. "Restricted Goods and Services." Meta Transparency Center. Accessed January 13, 2026. <https://transparency.meta.com/policies/community-standards/restricted-goods-services/>.
- Molly Rose Foundation. 2025a. *Children's Exposure to Suicide, Self-Harm, Depression, and Eating Disorder Content Online*. <https://mollyrosefoundation.org/wp-content/uploads/2025/10/Molly-Rose-Foundation-childrens-exposure-to-harmful-content.pdf>.
- . 2025b. *How Effectively Do Social Networks Moderate Suicide and Self-Harm Content?* https://mollyrosefoundation.org/wp-content/uploads/2025/01/DSA_Transparency_report_MRF-1.pdf.

- Morean, Meghan E., Grace Kong, Deepa R. Camenga, Dana A. Cavallo, and Suchitra Krishnan-Sarin. 2015. "High School Students' Use of Electronic Cigarettes to Vaporize Cannabis." *Pediatrics* 136 (4): 611–16. <https://doi.org/10.1542/peds.2015-1727>.
- Moreno, Megan A., and Jennifer M. Whitehill. 2016. "#Wasted: The Intersection of Substance Use Behaviors and Social Media in Adolescents and Young Adults." *Current Opinion in Psychology* 9:72–76. <https://doi.org/10.1016/j.copsyc.2015.10.022>.
- Murthy, Vivek H. 2024. "Surgeon General: Why I'm Calling for a Warning Label on Social Media Platforms." *New York Times*, June 17, 2024. <https://www.nytimes.com/2024/06/17/opinion/social-media-health-warning.html>.
- Myers West, Sarah. 2018. "Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms." *New Media & Society* 20 (11): 4366–83. <https://doi.org/10.1177/1461444818773059>.
- Nali, Matthew C., Vidya Purushothaman, Zhuoran Li, Meng Zhen Larsen, Raphael E. Cuomo, Joshua Yang, and Tim K. Mackey. 2024. "Identification and Characterization of Illegal Sales of Cannabis and Nicotine Delivery Products on Telegram Messaging Platform." *Nicotine & Tobacco Research* 26 (6): 771–79. <https://doi.org/10.1093/ntnr/ntad248>.
- Nunziato, Dawn C. 2023. "The Digital Services Act and the Brussels Effect on Platform Content Moderation." *Chicago Journal of International Law* 24 (1): 115–28. <https://cjl.uchicago.edu/print-archive/digital-services-act-and-brussels-effect-platform-content-moderation>.
- O'Brien, Erin Keely, Leah Hoffman, Mario Antonio Navarro, and Ollie Ganz. 2020. "Social Media Use by Leading US E-Cigarette, Cigarette, Smokeless Tobacco, Cigar and Hookah Brands." *Tobacco Control* 29:e87–e97. <https://doi.org/10.1136/tobaccocontrol-2019-055406>.
- Ortutay, Barbara, and Kaitlyn Huamani. 2026. "What's Next in Social Media Legal Battles after a New Mexico Jury Finds Meta Platforms Harm Children." PBS News, March 25, 2026. Accessed April 16, 2026. <https://www.pbs.org/newshour/nation/whats-next-in-social-media-legal-battles-after-a-new-mexico-jury-finds-meta-platforms-harm-children>.
- Papakyriakopoulos, Orestis, Juan Carlos Medina Serrano, and Simon Hegelich. 2020. "The Spread of COVID-19 Conspiracy Theories on Social Media and the Effect of Content Moderation." *Harvard Kennedy School Misinformation Review*, <https://doi.org/10.37016/mr-2020-034>.
- Paul, Kari. 2024. "Zuckerberg Tells Parents of Social Media Victims at Senate Hearing: 'I'm Sorry for Everything You've Been Through.'" *Guardian*, January 31, 2024. <https://www.theguardian.com/us-news/2024/jan/31/tiktok-meta-x-congress-hearing-child-sexual-exploitation>.

- Pearson, George D. H., Diana L. Davidson, Barbara A. Schillo, and Jennifer M. Kreslake. 2025a. "Investigating an E-Cigarette Brand's Use of Music Festivals for Social Media Campaigns and Experiential Marketing." *Tobacco Prevention & Cessation* 11:1–4. <https://doi.org/10.18332/tpc/199607>.
- Pearson, George D. H., Diana Lane Davidson, Barbara Schillo, and Jennifer M. Kreslake. 2025b. "'Discreet Shipping' on TikTok Enables Selling of E-Cigarettes to Youth." *Tobacco Control* 34 (2): 274–75. <https://doi.org/10.1136/tc-2023-058315>.
- Pew Research Center. 2024. "Teens, Social Media and Technology 2024," December 12, 2024. Accessed July 31, 2026. <https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/>.
- Rai, Shivani. 2025. "What Is a Good Like to View Ratio on YouTube?" Veefly Blog, February. Accessed July 21, 2026. <https://blog.veefly.com/youtube/what-is-a-good-like-to-view-ratio-on-youtube/>.
- Ranker, Lysie R., Jiaxi Wu, Traci Hong, Derry Wijaya, Emelia J. Benjamin, Aruni Bhatnagar, Rose M. Robertson, Jessica L. Fetterman, and Ziming Xuan. 2024. "Social Media Use, Brand Engagement, and Tobacco Product Initiation among Youth: Evidence from a Prospective Cohort Study." *Addictive Behaviors* 154. <https://doi.org/10.1016/j.addbeh.2024.108000>.
- Roberts, Sarah T. 2016. "Digital Refuse: Canadian Garbage, Commercial Content Moderation and the Global Circulation of Social Media's Waste." *Wi: Journal of Mobile Media* 10 (1). <https://doi.org/10.65968/zjdw4316>.
- Samermit, Patrawat, Anna Turner, Patrick Gage Kelley, Tara Matthews, Vanessa Wu, Sunny Consolvo, and Kurt Thomas. 2023. "'Millions of People Are Watching You': Understanding the Digital-Safety Needs and Practices of Creators." In *32nd USENIX Security Symposium (USENIX Security 23)*, 5629–45. USENIX Association. <https://www.usenix.org/conference/usenixsecurity23/presentation/samermit>.
- SAMS. n.d. "The Hidden Fire Risks Posed by Disposable Vape Pens." Accessed April 29, 2026. <https://www.samsltd.co.uk/hidden-fire-risks-disposable-vape-pens/>.
- Sawers, Paul. 2022. "Big Tech's Push to Self-Regulate Harmful Content in New Zealand Is 'Weak Attempt to Preempt Regulation,' Critics Say." TechCrunch, July 25, 2022. Accessed July 17, 2026. <https://techcrunch.com/2022/07/25/big-techs-push-to-self-regulate-harmful-content-in-new-zealand-is-weak-attempt-to-preempt-regulation-critics-say/>.
- Scherer, Michael, and Kaitlyn Tiffany. 2026. "How Meta Executives Talked about Child Safety behind the Scenes." *The Atlantic*, February 27, 2026. Accessed March 12, 2026. <https://www.theatlantic.com/technology/2026/02/meta-child-safety-documents-instagram/686163/>.

- Schneider, Philipp J., and Marian-Andrei Rizoïu. 2023. "The Effectiveness of Moderating Harmful Online Content." *Proceedings of the National Academy of Sciences*, <https://doi.org/10.1073/pnas.2307360120>.
- Shahid, Farhana, and Aditya Vashistha. 2023. "Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?" In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. <https://doi.org/10.1145/3544548.3581538>.
- Sharp, Kellen, Marzieh Babaeianjelodar, Dhiraj Murthy, Rachel R. Ouellette, Juhan Lee, Amanda de la Noval, Neil Kamdar, and Grace Kong. 2025. "Remediated Marketing: Leveraging Computer Vision and Rule-Based Classification Models to Detect E-Cigarette Warning Labels across Social Media." *Information, Communication & Society*, 1–20. <https://doi.org/10.1080/1369118X.2025.2500485>.
- Sharp, Sonja. 2026. "Landmark L.A. Jury Verdict Finds Instagram, YouTube Were Designed to Addict Kids." *Los Angeles Times*, March 25, 2026. <https://www.latimes.com/california/story/2026-03-25/social-media-lawsuit-trial-meta-google-verdict>.
- Al-Shawaf, Maeh, Kya N. Grooms, Margaret Mahoney, Natasha Buchanan Lunsford, and Deirdre Lawrence Kittner. 2023. "Support for Policies to Prohibit the Sale of Menthol Cigarettes and All Tobacco Products among Adults, 2021." *Preventing Chronic Disease* 20. <https://doi.org/10.5888/pcd20.220128>.
- Shim, Yunhee, and Shagun Jhaver. 2024. "Incorporating Procedural Fairness in Flag Submissions on Social Media Platforms." *arXiv*, <https://doi.org/10.48550/arXiv.2409.08498>.
- Silver, Nathan A., Adrian Bertrand, Padmini Kucherlapaty, and Barbara A. Schillo. 2023. "Examining Influencer Compliance with Advertising Regulations in Branded Vaping Content on Instagram." *Frontiers in Public Health* 10. <https://doi.org/10.3389/fpubh.2022.1001115>.
- Silver, Nathan A., Brenda Dimaya, Elexis C. Kierstead, Madison Iskra, Maeh Al-Shawaf, Michael A. Tynan, and Jessica M. Rath. 2026. "Exposure to Tobacco Imagery in Streaming Television Is Associated with Increased Intentions to Smoke and Vape." *Addictive Behaviors Reports* 23. <https://doi.org/10.1016/j.abrep.2026.100668>.
- Silver, Nathan A., Padmini Kucherlapaty, Adrian Bertrand, Robert N. Falk, and Jessica Miler Rath. 2024. "Improving Enforcement Measures and Establishing Clear Criteria: A Content Analysis of Tobacco-Brand-Owned Instagram Accounts." *Nicotine & Tobacco Research* 26 (9): 1175–82. <https://doi.org/10.1093/ntr/ntae052>.
- Singer, Jill M., Elizabeth G. Klein, Micah L. Berman, Jeffrey J. Wing, and Megan E. Roberts. 2025. "'I'm Sure within 5 Minutes I Can Find Someone That Will Be Selling It': Qualitative Insights into How Social Media Facilitates the Sale of Banned Flavoured Tobacco Products." *Tobacco Control* 35 (4): 549–51. <https://doi.org/10.1136/tc-2024-058937>.

- Solon, Olivia. 2017. "Underpaid and Overburdened: The Life of a Facebook Moderator." *Guardian*, May 25, 2017. <https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content>.
- Susi, Karima, Francesca Glover-Ford, Anne Stewart, Rebecca Knowles Bevis, and Keith Hawton. 2023. "Research Review: Viewing Self-Harm Images on the Internet and Social Media Platforms: Systematic Review of the Impact and Associated Psychological Mechanisms." *Journal of Child Psychology and Psychiatry* 64 (8): 1115–39. <https://doi.org/10.1111/jcpp.13754>.
- Take, Kejsi, Victoria Zhong, Chris Geeng, Emmi Bevensee, Damon McCoy, and Rachel Greenstadt. 2024. "Stoking the Flames: Understanding Escalation in an Online Harassment Community." *Proceedings of the ACM on Human-Computer Interaction* 8 (CSCW1): 1–23. <https://doi.org/10.1145/3641015>.
- Terry, Kirsty, Fei Yang, Qiang Yao, and Chaojie Liu. 2023. "The Role of Social Media in Public Health Crises Caused by Infectious Disease: A Scoping Review." *BMJ Global Health* 8 (12). <https://doi.org/10.1136/bmjgh-2023-013515>.
- The Future of Free Speech. 2024. *Preventing "Torrents of Hate" or Stifling Free Expression Online? An Assessment of Social Media Content Removal in France, Germany, and Sweden*. <https://futurefreespeech.org/wp-content/uploads/2024/05/Preventing-Torrents-of-Hate-or-Stifling-Free-Expression-Online-The-Future-of-Free-Speech.pdf>.
- TikTok. n.d.-a. "Community Guidelines." Accessed January 13, 2026. <https://www.tiktok.com/community-guidelines>.
- . n.d.-b. "Community Guidelines Enforcement Report: January–March 2026." Accessed September 24, 2026. <https://www.tiktok.com/safety/en/transparency/cg-report>.
- . n.d.-c. "Guardian's Guide." TikTok Safety Center. Accessed April 16, 2026. https://www.tiktok.com/safety/en/tools-and-guides/guardians-guide?gad_source=1&gad_campaignid=20471031496&gclid=EAIaIQobChMI__q355vzkwMV-DIIBR2zcRiKEAAYASAAEgIxY_D_BwE.
- U.S. Department of Health and Human Services. 2012. *Preventing Tobacco Use among Youth and Young Adults: A Report of the Surgeon General*. Centers for Disease Control et al. <https://www.ncbi.nlm.nih.gov/books/NBK99237/>.
- . 2024. *Eliminating Tobacco-Related Disease and Death: Addressing Disparities*. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Office of the Surgeon General. <https://www.hhs.gov/sites/default/files/2024-sgr-to-bacco-related-health-disparities-full-report.pdf>.

- . n.d. “Reduce the Proportion of Adolescents Exposed to Tobacco Marketing (TU-22).” Healthy People 2030. Accessed July 21, 2026. <https://odphp.health.gov/healthypeople/objectives-and-data/browse-objectives/tobacco-use/reduce-proportion-adolescents-exposed-tobacco-marketing-tu-22>.
- Urman, Aleksandra, and Mykola Makhortykh. 2023. “How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting across Online Platforms.” *Telecommunications Policy* 47 (3). <https://doi.org/10.1016/j.telpol.2022.102477>.
- Vilk, Viktorya. 2023. *Shouting into the Void*. PEN America. <https://pen.org/report/shouting-into-the-void/>.
- Waldon, Patricia. 2023. “One-Size-Fits-All Content Moderation Fails the Global South.” Cornell Chronicle, April. Accessed April 21, 2026. <https://news.cornell.edu/stories/2023/04/one-size-fits-all-content-moderation-fails-global-south>.
- Wheeler, Andrew. 2020. “Confidence Intervals around Proportions.” Andrew Wheeler, November 30, 2020. Accessed July 21, 2026. <https://andrewpwheeler.com/2020/11/30/confidence-intervals-around-proportions/>.
- World Health Organization. 2017. “Tobacco: WHO Framework Convention on Tobacco Control.” Accessed April 21, 2026. <https://www.who.int/news-room/questions-and-answers/item/tobacco-and-the-who-framework-convention-on-tobacco-control>.
- Young, Greyson K. 2022. “How Much Is Too Much: The Difficulties of Social Media Content Moderation.” *Information & Communications Technology Law* 31 (1): 1–16. <https://doi.org/10.1080/13600834.2021.1905593>.
- YouTube. n.d. *Exploring YouTube Confidently: A Family Guide to Supervised Experiences*. <https://static.googleusercontent.com/media/www.youtube.com/en//myfamily/parentguide.pdf>.
- Zenone, Marco, Nora Kenworthy, and Nason Maani. 2023. “The Social Media Industry as a Commercial Determinant of Health.” *International Journal of Health Policy and Management* 12. <https://doi.org/10.34172/ijhpm.2022.6840>.
- Zhang, Alice Qian, Kaitlin Montague, and Shagun Jhaver. 2024. “Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Content.” *arXiv*, <https://doi.org/10.48550/arXiv.2309.06688>.
- Zhang, Cindy C., Grayden Zaleski, Jaya N. Kailley, Katelyn A. Teng, Mahala English, Anna Riminchan, and Julie M. Robillard. 2024. “Debate: Social Media Content Moderation May Do More Harm than Good for Youth Mental Health.” *Child and Adolescent Mental Health* 29 (1): 104–6. <https://doi.org/10.1111/camh.12689>.
- Zhong, Victoria, Meghna Nair, Susan McGregor, Damon McCoy, and Rachel Greenstadt. 2026. “‘The Report Button Is Just for Decoration’: Moderation Practices and Needs of YouTube Creators.” *Journal of Online Trust and Safety* 3 (2). <https://doi.org/10.54501/jots.v3i2.266>.

Authors

George D. H. Pearson is at the Schroeder Institute at Truth Initiative. Email: gpearson@truthinitiative.org.

Padmini Kucheralapaty is at the Schroeder Institute at Truth Initiative.

Nathan A. Silver is at the Schroeder Institute at Truth Initiative.

Barbara A. Schillo is at the Schroeder Institute at Truth Initiative.

Jennifer M. Kreslake is at the Schroeder Institute at Truth Initiative.

Acknowledgments

The authors wish to thank Claire Trevithick and Bamlack Tamire for their support with data collection and preparation of supplementary materials.

Data and Code Availability

The data used in the study is available on request from the authors. To allow greater replicability and transparency, we have added additional documents to an Open Science Framework (OSF) project, viewable at: https://osf.io/wzjge/overview?view_only=774ea82d389f4c4da5645dd7318038ab.

Ethics Statement

This study used publicly available content from Instagram, TikTok, and YouTube, and did not involve direct interaction with or recruitment of human participants. Therefore, Institutional Review Board determination was not applicable. Content violating policies was found on the platforms and reported using the same mechanisms available to real users. Consequently, the study posed no increased harms of exposing individuals to rule-breaking content and posed minimal risk to the platforms themselves, in fact auditing the effectiveness of a system the platforms already have in place. The study design held significant benefit for the understanding of content moderation.

Funding Statement

This study was funded by Truth Initiative.

Disclosure Statement

The authors declare no competing interests.

Keywords

Content moderation; user reporting; tobacco; youth safety; public health

Appendix

A. Platform Policies and Reporting Criteria

Here we provide sample text from each platform’s community guidelines and the criteria we used to decide whether content should be reported, for each of the five reporting categories described in this paper. The categories are covered in the order used in the body: insufficient age-gating (Section A.1), disallowed promotion (Section A.2), illegal sales (Section A.3), underage use (Section A.4), and dangerous activities (Section A.5).

A.1. Insufficient Age-Gating

Posts were considered to lack age-gating if they could be viewed without logging in.

A.1.1. Sample text from content guidelines

Instagram “For the following content, we restrict visibility to adults 18 years of age and older:

- Content posted by or promoting legitimate brick-and-mortar entities, including retail businesses, websites or brands, which attempt to buy, sell, trade, donate or gift of alcohol or tobacco products.
- Content depicting the consumption of tobacco, nicotine products, or “ENDS” products.
- Content that coordinates or promotes the use of tobacco, nicotine products, ‘ENDS’ products, or tobacco brands.” (Meta, [n.d.-c](#))

TikTok “FYF INELIGIBLE & AGE-RESTRICTED:

Showing or promoting cannabis-related items, or adult tobacco use.” (TikTok, [n.d.-a](#))

YouTube “Age-restricted content

Sometimes content doesn’t violate our policies, but it may not be appropriate for viewers under 18.

Examples of age-restricted content

- Content that promotes a cannabis dispensary.
- Content that reviews brands of nicotine e-liquid.” (Google, [n.d.-c](#))

A.1.2. Criteria for reporting

Instagram Content promoting or marketing a tobacco product posted by a legitimate retailer or by the manufacturer that is not age-restricted, despite platform policies requiring age restriction for such content.

TikTok Content that is not age-restricted, despite policies requiring age restriction for content that:

1. Showed adult use of a tobacco product
2. Shows an adult promoting or modeling a tobacco product

YouTube Content that is not age-restricted, despite policies requiring age restriction for content that:

1. Reviews a tobacco product
2. Promotes a tobacco product, but where the account promoting the content did not have a clear vested interest in the sale of the product (i.e. was not a manufacturer or retailer).

A.2. Disallowed Promotion**A.2.1. Sample text from content guidelines**

Instagram “We define branded content as a creator or publisher’s content that features or is influenced by a business partner for an exchange of value, such as monetary payment or free gifts.

Branded content may only be posted with the use of the branded content tool, and creators must use the branded content tool to tag the featured third party product, brand, or business partner.

[...]

Certain goods, services, or brands may not be promoted with branded content. We prohibit promotion of the following:

[...]

3. Tobacco products, vaporizers, electronic cigarettes, or any other products that simulate smoking.” (Meta, n.d.-a)

TikTok “Regulated Goods and Services

We don’t allow trading, marketing, or providing access to regulated, prohibited, or high-risk goods and services.

[...]

Trade: The sale, purchase, exchange, giveaway, or redirection of goods or services. This includes posting links, sharing contact or location information, or directing people off-platform.

Tobacco Products: Vaping products, synthetic nicotine, combustible tobacco (like cigarettes), smokeless tobacco (like chewing tobacco), E-cigarettes, and other electronic nicotine delivery systems.” (TikTok, n.d.-a)

“NOT ALLOWED

- Trading, marketing, or providing access to (except where otherwise specified below):
 - Alcohol, tobacco, drugs, or other regulated substances” (TikTok, n.d.-a)

YouTube “Here are some examples of content that’s not allowed on YouTube.

Keep in mind that this isn’t a complete list.

[...]

- Content that promotes a product that contains drugs, nicotine, or a controlled substance.” (Google, n.d.-c)

A.2.2. Criteria for reporting

Instagram Content that is either:

- A paid-for promotional post using the “branded content” tool on Instagram. (Determined by if the post says “Paid Partnership” next to the account posting the content.)
- A clear and open ‘paid for’ brand collaboration—(for instance, the user says in the post description the post is a “paid partnership” or uses “#ad” or “#paidpartnership”)

TikTok Content that promotes or advertises a tobacco product (i.e. gives a favorable impression of) that is posted by someone with a clear vested interest in people buying the product; specifically either the manufacturer of the product or a retailer. An additional clear indicator is a link to where the product can be bought or the post refers to a link in the bio where the product can be bought.

YouTube Content that promotes or advertises a tobacco product (i.e. gives a favorable impression of) that is posted by someone with a clear vested interest in people buying the product; specifically either the manufacturer of the product or a retailer.

A.3. Illegal Sales

A.3.1. Sample text from content guidelines

Instagram “We do not allow:

Content that:

Attempts to buy, sell or trade tobacco/nicotine related products, or products that simulate smoking, including all kinds of “ENDS” Electronic Nicotine Delivery Systems Products (e.g., electronic cigarettes, vapes, and nicotine free-vapes). Except when posted by a Page, Group, or Instagram profile representing legitimate brick-and-mortar entities, including retail businesses, websites, brands, or a private individual sharing content on behalf of legitimate brick-and-mortar entities, including offering delivery services and brand giveaways.” (Meta, n.d.-c)

TikTok “NOT ALLOWED:

Trading, marketing, or providing access to (except where otherwise specified below):

Alcohol, tobacco, drugs, or other regulated substances” (TikTok, n.d.-a)

YouTube “Don’t post content on YouTube if it aims to directly sell, link to, or facilitate access to any of the regulated goods and services listed below. Making the sale of these items or facilitating the use of these services possible by posting links, email, phone number or other means to contact a provider directly is not allowed:

[...]

Nicotine, including vaping products” (Google, n.d.-a)

A.3.2. Criteria for reporting

All platforms Content that facilitates the sale of tobacco products to underage users.

This can be either:

1. Posts instructing users where they can buy products underage
2. Posts advertising services to sell products to underage users
3. Specifically advertising a tobacco product to be used in school by youth

Any of the following can be used as evidence that the post/account is facilitating underage sales:

1. A clear offer that anyone of any age can buy the product. Example: “No id required”, “We sell to any age”, “No signature required”
2. An offer to ship the product discreetly—usually hidden as another type of product—so others would be unaware of the contents (see *Discreet Shipping*). Example: *Selling e-cigarettes hidden inside cosmetics kits*.
3. Suggesting the user contacts an account directly via DMs or through other anonymous platforms to purchase. Example: *Telling buyers to contact them on Telegram to buy. Telling buyers to DM them on Instagram to buy*.

A.4. Underage Use

A.4.1. Sample text from content guidelines

Instagram No relevant content guidelines.

TikTok “NOT ALLOWED:

[...]

Showing or promoting young people participating in gambling, using alcohol, tobacco, or fireworks, or having weapons.” (TikTok, [n.d.-a](#))

YouTube “Here are some examples of content that’s not allowed on YouTube:

[...]

Showing minors drinking alcohol, using vaporizers, e-cigarettes, tobacco or marijuana, or misusing fireworks.” (Google, [n.d.-c](#))

“We also don’t allow content showing dangerous acts by minors. This includes content that shows minors:

- Drinking alcohol
- Using vaporizers, e-cigarettes, tobacco, or marijuana” (Google, [n.d.-a](#))

A.4.2. Criteria for reporting

All platforms Content that shows an underage individual using a tobacco product. The post must not be designed to be satire, but show genuine usage. The post must also provide substantial evidence that the product was used. Merely holding the product is not enough; further evidence is required, for instance, holding the product while exhaling vapor, inserting the product into their mouth etc.

The age of the user can be determined either by:

1. Physical appearance: if the user is *very* clearly underage. For instance, clearly under 16.
2. Contextual clues: if the user is young, dressed as a student, and the video clearly states or shows the user in a school setting (note, must be school and not college), then there is enough contextual information to declare the user underage.

A.5. Dangerous Activities

A.5.1. Sample text from content guidelines

Instagram No relevant content guidelines.

TikTok “Dangerous Activity and Challenges: We don’t allow content that shows or promotes dangerous stunts, dares, or challenges that could lead to physical harm.” (TikTok, [n.d.-a](#))

YouTube “YouTube doesn’t allow content that encourages dangerous or illegal activities that risk serious physical harm or death.” (Google, [n.d.-a](#))

“[Not allowed to show] Risking serious harm or death: Behavior that shows adults risking serious bodily harm or death, particularly if someone watching could imitate the dangerous act or if the content encourages or praises the dangerous act” (Google, [n.d.-a](#))

A.5.2. Criteria for reporting

All platforms Content that shows tobacco/nicotine/e-cigarette products being used (or tutorials on how to use) in dangerous ways. Three specific types of content are reported:

1. *Incorrect Usage*: Using a product in a way that is more likely to cause harm. Example: A “double nose blinker” where two cannabinoid vapes are inhaled through each nostril simultaneously
2. *Dangerous Dosage*: Using high quantities of a product in one go. Examples: A “pool noodle blinker” where several cannabinoid vapes are inserted into a pool noodle so that they can all be inhaled at the same time. Users consuming an excessive amount of ONP products—such as a whole can—simultaneously.
3. *Disposable Battery Recharging*: Posts informing users of a dangerous method to recharge a disposable vape battery.