Protecting Young Users on Social Media: Evaluating the Effectiveness of Content Moderation and Legal Safeguards on Video-Sharing Platforms

Fatmaelzahraa Eltaher, Rahul Krishna Gajula, Luis Miralles-Pechuán, Patrick Crotty, Juan Martínez-Otero, Christina Thorpe, and Susan McKeever

Abstract. Video-sharing platforms such as TikTok, YouTube, and Instagram implement content moderation policies to reduce the exposure of minors to harmful videos. As video has become the dominant and most immersive form of online content, assessing how effectively these systems protect younger users is increasingly important. This study evaluates the effectiveness of video moderation for different age groups on TikTok, YouTube, and Instagram, based on a focused set of experimental accounts. Accounts were created for simulated users aged 13 and 18, and 3,000 recommended videos were analyzed in two interaction modes: *passive scrolling* and *search-based scrolling*. Each video was manually assessed for the severity of the harm using a unified harm classification framework. While low-severity harm was the most prevalent form encountered, the results show that accounts configured as 13-year-olds encountered harmful videos more frequently and rapidly than accounts configured as 18-year-olds. On YouTube, 15% of videos recommended to 13-year-old accounts during passive scrolling were classified as harmful, compared to 8.17% for adult accounts, with exposure occurring within an average of 3:06 minutes. This exposure appeared without user-initiated searches, highlighting weaknesses in algorithmic filtering. Results from our targeted study point to gaps in video moderation systems, suggesting the need for more effective safeguards to better protect minors from harmful online content.

doi:10.54501/jots.v3i2.251

1 Introduction

Social media has changed how most people connect, communicate, and share content, reaching over 5 billion users in 2024 and projected to surpass 6 billion by 2028 (Statista, n.d.). Leading platforms such as Facebook, YouTube, Instagram, TikTok, and Snapchat collectively engage billions of monthly users, making social networks one of the most popular online activities (Dixon, n.d.).

In England, 91% of teenagers aged 13 to 18 years and 65% of children aged 8 to 12 actively participate in social networks (Children's Commissioner 2022). In the United States, approximately 60% of teens aged 13 to 17 frequently use platforms including Instagram, Snapchat, and TikTok (Anderson, Faverio, and Gottfried 2023), while even younger children regularly access platforms such as YouTube and TikTok, usually with the knowledge and consent of their parents (Beresford et al. 2023; WeProtect Global Alliance 2023).

A significant portion of this engagement is now driven by video content, which has become the dominant and most immersive medium on social platforms. The rise of short-form video, popularized by TikTok and quickly mirrored by platforms like Instagram Reels and YouTube Shorts, represents a major shift in online engagement. These easily consumable and delivered videos are especially appealing to younger audiences, who increasingly prefer quick and visually engaging content over text or static images (Violot et al. 2024). This shift has led to a notable increase in the use of social media among children and adolescents, many of whom now use platforms such as TikTok and Snapchat as their primary sources of entertainment, social interaction, and even news (Liu et al. 2024; Ge et al. 2021).

Social media platforms use algorithms to personalize content recommendations based on user interests to keep users engaged. Although personalization improves user experiences, it has raised significant concerns about the amplification of potentially harmful content (Gorwa, Binns, and Katzenbach 2020). UNICEF defines harmful content as anything—be it an image, video, or text—that offends, upsets, or causes harm to individuals (Unicef, n.d.). To address these concerns, platforms implement various moderation strategies, ranging from simple content warnings to content removal or suspension of user accounts (Fiesler et al. 2018; Seering et al. 2019; Gillespie 2018).

Content moderation generally falls into three approaches: volunteer moderation in smaller online communities, commercial moderation on larger platforms, and automated moderation powered by artificial intelligence (Gillespie 2018). Given the enormous volume of daily content, AI-driven moderation systems have become increasingly important for detecting and filtering inappropriate material (Chandrasekharan et al. 2019). Furthermore, on January 7, 2025, Meta announced substantial revisions to Facebook and Instagram's content moderation procedures. Meta intends to terminate its fact-checking program in the United States and implement a community-driven approach similar to X's

Community Notes feature (Competition Policy International 2025).

Platforms typically classify content into three distinct categories: (1) permitted content, which is fully compliant with community guidelines; (2) prohibited content, such as hate speech, violent threats, or harassment, which must be removed immediately; and (3) restricted content, which remains accessible under specific conditions, such as through age verification processes or explicit warning screens (Meta, n.d., n.d., n.d.; YouTube, n.d., 2020; TikTok, n.d.; X, n.d.).

In response to growing concerns regarding child safety, platforms have introduced age-based restrictions to limit exposure to potentially harmful content. TikTok, for example, provides a parental-controlled "Restricted Mode" and automatically restricts access to certain age-sensitive content for users between 13 and 17 (TikTok, n.d., n.d.). Similarly, YouTube uses machine learning to classify age-restricted videos and encourages content creators to label their uploads correctly (YouTube 2019, n.d.). Facebook, Instagram, and Twitter have implemented similar restrictions designed to protect minors (Meta, n.d.; X, n.d.). However, social networks primarily rely on *self-declared* age at account setup time, making it easy for minors to fake their age or bypass controls. This absence of robust age verification raises significant doubts about the effectiveness or reach of age-related controls. Previous works indicate that content moderation alone cannot effectively protect minors unless supported by strict and verifiable age verification mechanisms (Eltaher et al. 2025).

Despite these safeguards, evidence shows that children encounter harmful content online. A survey by the Children's Commissioner for England indicated that nearly 45% of children aged 8 to 17 years have encountered harmful content on social networks (Children's Commissioner 2022). Moreover, a study in twenty-five European countries revealed that 20% of children aged 9 to 16 years were exposed to sexual content online (Staksrud, Ólafsson, and Livingstone 2013). Further research highlights algorithmic risks, demonstrating that toddlers watching child-friendly videos on YouTube have a 3.5% chance of encountering inappropriate content within ten recommended videos (Papadamou et al. 2020).

As video continues to dominate online engagement, it becomes critical to assess how effectively major video-sharing platforms protect younger users from harmful content. Although previous research has typically focused on specific harm categories (e.g., misogyny, mental health), few studies have undertaken systematic, cross-platform analyses that compare age-based differences in exposure across multiple interaction modes. This study contributes to addressing this gap by empirically evaluating the effectiveness of content moderation for minors on TikTok, YouTube, and Instagram for a set of targeted accounts. The research specifically examines: (1) how moderation performance differs for minors and adults; (2) how harmful content exposure varies between passive scrolling and active searching; (3) what types and severities of harmful content appear most often; and (4) how published moderation policies of platforms align

with actual outcomes.

In this work, we present a small-scale study using a controlled, reproducible experimental design to measure how leading video-sharing platforms moderate harmful content for users of different ages. We created accounts that simulated 13-year-old and 18-year-old user accounts on TikTok, YouTube, and Instagram, each subjected to two distinct behavioral modes: *passive scrolling* (algorithmic video feeds) and *search-based scrolling* (keyword-driven queries). Two accounts per age-group/platform/interaction mode were created, with each of the 20 accounts created viewing a fixed sample of 100 videos per session. All videos were manually annotated and classified by type of harm and severity (low, medium, high). We quantified overall exposure rates, time to the first harmful video, and discrepancies between observed content and stated moderation policies.

Our findings reveal that accounts configured as 13-year-olds encounter harmful content more frequently and more rapidly than adult users on all platforms, with the lowest level of severity the most prevalent level of harm encountered. Minor-designated accounts were exposed to harmful videos within three to five minutes of use, compared to approximately nine minutes for adults, and their exposure rates ranged from 7.8% to 15% (versus 4.7%–8.3% for adults). The most prevalent harmful categories included *Sensitive & Mature Themes* and *Hate Speech and Hateful Behavior*, especially on Instagram. Our results suggest that current moderation mechanisms remain inadequate for shielding minors from exposure to harmful content, underscoring the need for ongoing checks on moderation effectiveness and more age-specific safeguards in video recommendation systems.

2 Literature Review

This section reviews the current literature on content moderation limitations, focusing on protecting minors from harmful online content. It begins by analyzing empirical studies that reveal the shortcomings of existing moderation practices on popular social media platforms. This is followed by an overview of the legal frameworks in the United States, the European Union, and the United Kingdom that attempt to address these risks.

2.1 Content Moderation on Video-Sharing Platforms

Effective content moderation is critical for safeguarding minors online, particularly those who bypass age verification mechanisms. Between October and December 2024, TikTok removed more than 153 million videos for policy violations, while YouTube removed nearly 9.5 million. Despite efforts on the platform, research consistently shows that children can encounter harmful material through passive exposure and active interaction (Papadamou et al. 2020; Fibrilla, Fairus, and Raifah 2021). This has led to empirical investigations of how social network algorithms amplify potentially damaging content.

Similar challenges also appear in other areas of online risk. Research on fraud prevention shows that AI tools, such as machine learning models that detect suspicious behavior or deceptive activity, work best when combined with clear platform rules and enforcement actions (Xiao, Sellars, and Scheffler 2025). This similarity between harmful content moderation and anti-fraud efforts shows that both need flexible, transparent, and coordinated systems that blend technology with accountability to reduce online harm.

Some studies examined how passively watching suggested videos, without any interaction, can lead users to harmful content. A study by Amnesty International focused on the prevalence of mental health content on TikTok (Amnesty International 2023). Forty accounts were created, all simulating 13-year-old users. Within five to six hours of use, nearly half of the videos shown to accounts expressing interest in mental health were classified as potentially harmful.

Similarly, Regehr et al. (2024) conducted an experiment on TikTok to assess how easily users can access misogynistic and manosphere content. The findings revealed that the likelihood of encountering misogynistic content increased fourfold in five days, illustrating how rapidly ideologically charged material can become embedded in a user's feed.

Other studies examine the effect of engagement actions, such as liking and searching, on exposing users to certain types of dangerous videos. Baker, Ging, and Andreasen (2024) analyzed ten user accounts on YouTube Shorts and TikTok to explore the online experiences of boys and young men. They focused on exposure to problematic content, including manosphere ideologies, anti-feminist messages, racism, and anti-LGBTQ sentiments. Their findings revealed that YouTube Shorts featured a significantly higher proportion of such content (61.5%) compared to TikTok (34.7%).

In a related experiment, Williams, Farthing, and McIntosh (2021) created five TikTok accounts simulating 13-year-old girls. The researchers aimed to assess how quickly the platform's algorithm recommended questionable content, including ethnic or gender stereotypes, misinformation about COVID-19 or vaccines, or weight loss content. After viewing 600 videos, they found that one in four promoted harmful ethnic stereotypes and one in five reinforced damaging gender stereotypes. However, no weight-loss or COVID-19 misinformation was shown, possibly due to TikTok's content moderation efforts or limitations in the study design.

An Australian study (Thomas and Balint 2022) investigated how YouTube and YouTube Shorts expose boys and young men to misogynistic content. Users were progressively recommended more radical content by liking and following mainstream and extreme influencers. YouTube Shorts, in particular, escalated the delivery of misogynistic and incel-related material.

Similarly, the Center for Countering Digital Hate (2022) reported that TikTok served suicide-related content in 2.6 minutes and eating disorder content in 8 minutes of use by simulated teen accounts. Every 39 seconds, TikTok displayed content about teens'

mental health and body image.

Ekō, an international advocacy organization (Ekō, n.d.), investigated harmful content on TikTok using an account registered with a 13-year-old user (Ekō 2023). Their focus included suicide, incel ideology, the manosphere, and drug-related content. The study concluded that a mere ten minutes of limited engagement with suicide-oriented content prompted TikTok's algorithm to push further self-harm and suicide-related material to these underage users.

Papadamou et al. (2020) explored YouTube's recommendation mechanisms. The study found that even when starting with safe content, there was a 3.5% chance of reaching inappropriate material within ten recommendations.

Research has explored how users from different cultures perceive the severity of harmful online content. Jiang et al. (2021) conducted a large-scale survey in eight countries, asking participants to classify and rank 66 types of online harm. The study revealed that perceptions of harm vary widely in cultural contexts and increase exponentially in perceived severity. For example, the "sale of marijuana" was rated highly severe in countries such as Vietnam and Indonesia, where it is strictly prohibited, but is considered relatively minor in the United States and Brazil, where attitudes and laws are more permissive. Such contrasts make it difficult to establish universal moderation standards and highlight the need for culturally informed frameworks that reflect regional norms, laws, and moral values.

2.1.1 Gaps in Current Research and the Present Study's Contribution

These studies highlight gaps in moderation policies that allow harmful videos to be served to young or otherwise vulnerable audiences. Although existing research often examines TikTok or YouTube with respect to specific harm categories, such as misogynistic narratives or mental health content, there is limited research on the full breadth of harmful videos. Accurate moderation hinges on well-defined categories of harm; however, key areas such as Child Safety in relation to grooming and exploitation, Sensitive and Mature Themes, and spam or fraud remain underexplored (Meta Platforms, n.d.; TikTok, n.d.; YouTube, n.d.).

Recent findings suggest that algorithmic audits of social media platforms, particularly TikTok, are often poorly reproducible and have short-term validity (Mosnar et al. 2025). This is because rapid changes in recommendation algorithms, content trends, and interface design make replication difficult and frequently make previous conclusions obsolete.

In addition, Instagram—one of the most widely used social networks—has been largely absent from previous examinations, leaving significant gaps in our understanding of its content moderation efficacy. To address these deficits, this work evaluates video moderation practices on multiple video-sharing platforms, including YouTube, Instagram,

and TikTok, and captures a broader range of harmful content categories. In addition, it investigates how search and scrolling behaviors influence the discovery of such material, offering a more comprehensive view of harmful content distribution across platforms.

2.2 Legislation on content moderation

Considering the importance that digital platforms have gained from a political (as a new public space), cultural (shaping the collective imagination), and social (as the primary source of entertainment for millions of people) standpoint, public authorities have progressively taken charge of their regulation. Although content moderation regulation was initially left to the platforms themselves, in most jurisdictions various legal instruments have started to address this matter, such as the Digital Services Act (DSA) of 2022 in the EU, the Online Safety Act of 2023 in the UK, and the Communications Decency Act in the US.

Moderation carried out by social networks, which can manifest in the deletion of the content, the reduction of items' visibility, or the suspension of users, no longer has an exclusively private scope but also carries a significant public dimension, affecting individuals' freedom of expression and the formation of a free public opinion. Rexhepi (2023) wrote "Social media networks have become forums and mediums of important conversation, and the responsibility to regulate it is too great for platforms to be left to deal with it alone."

We provide a brief overview of existing laws in the United States, the European Union (EU), and the United Kingdom to offer a broad perspective on regulating content moderation on social networks.

2.2.1 United States

At the federal level, the United States lacks specific regulations governing content moderation practices and policies of online platforms.

The First Amendment and Section 230 of the Communications Decency Act determine the key regulatory framework in content moderation. Section 230 shields interactive computer service providers, including social media platforms, and their users from liability for publishing—and, in certain instances, restricting access to or availability of—content posted by others. The purpose of Section 230 is to foster free speech while allowing interactive computer service providers to moderate content without government interference.

In alignment with the First Amendment, Section 230 establishes a notably hands-off regulatory framework, granting platforms significant autonomy and broad immunity in their content moderation practices (Kosseff 2019). This legal approach has been subject to two main criticisms: first, the wide discretion it affords providers in moderating and

removing content, which can potentially enable covert forms of censorship; second, the insufficient protections it offers to citizens against harmful content, particularly concerning the protection of minors from inappropriate material.

Several regulatory proposals have recently been introduced in response to growing concerns about content moderation, to explicitly encourage, discourage, prohibit, or mandate specific moderation practices on digital platforms (Cho and Zhu 2025). Although its final approval remains uncertain, one of the most notable legislative proposals is the Kids Online Safety Act (U.S. Congress 2024). Its provisions on the protection of minors and content moderation (Sections 4 and 6) are similar to those currently in force in the EU and the United Kingdom. In summary, KOSA aims to establish a protective and preventive framework that ensures content moderation adheres to standards of transparency, accountability, and procedural fairness.

Whether this regulation will be enacted or whether US authorities will pursue alternative public policies regarding content moderation remains unclear. Among the various proposals under consideration are: (1) maintaining the current hands-off regime; (2) encouraging changes to moderation systems through hearings or investigations; (3) regulating moderation to require specific conditions regarding speed and transparency, as KOSA proposes; (4) implementing federal advisory or regulatory oversight over social networks with certain powers over platforms; and (5) pursuing alternative measures such as digital education initiatives (Cho and Zhu 2025).

2.2.2 European Union

Within the EU, two main regulatory frameworks govern content moderation on digital platforms. Both frameworks impose a duty of care on platforms, requiring them to take reasonable steps to ensure user safety and to address illegal and harmful activities.

The first framework comes from the Audiovisual Media Services Directive (AVMSD) of 2018, which has been transposed into national legislation across Member States. It applies specifically to video-sharing platforms (VSPs) based in the EU, as well as to prominent users such as influencers. Article 28b of the AVMSD explicitly refers to video-sharing platforms. Paragraph 1 mandates measures to prevent minors from accessing harmful content and to prevent the public from accessing illegal content (e.g., hate speech). Article 28b, Section 3, outlines various potential measures that Member States can impose on VSPs—many of which pertain to content moderation mechanisms. These measures include establishing systems for rating and reporting content, as well as processes for handling user complaints.

The second key regulation is the Digital Services Act of 2022 (European Union 2022). This regulation applies to digital platforms that offer services to citizens in EU Member States, including significant platforms such as X, Instagram, YouTube, and many adult content websites. It imposes specific proactive obligations on providers to protect minors from

potentially harmful content.[1]

Concerning content moderation, the Digital Services Act aims to enhance transparency in content moderation systems by requiring platforms to publish moderation policies (Article 14), annual reports (Article 15), and the rationale behind their decisions (Articles 16–17). It also strengthens procedural due process guarantees by ensuring the right to appeal and the participation of human agents in the decision-making process (Articles 20, 21). Furthermore, large platforms and search engines must conduct impact assessments to identify risks to minors within their environments, including those posed by platform algorithms, and implement measures to mitigate them (Articles 34, 35).

In summary, the EU legal framework aims to ensure that content moderation systems are fair, transparent, and effective, contributing to a safer digital environment for all users, particularly minors. However, recent research suggests that many platforms do not fully meet these obligations in practice, particularly with regard to transparency and algorithmic risk assessments. In any case, when referring to content moderation, the regulation focuses on reporting and removing content rather than its "algorithmic" management, which may foster or reduce its visibility among users (Gillespie 2022).

2.2.3 United Kingdom

In the United Kingdom, the governing legislation is the Online Safety Act of 2023, which imposes a series of obligations on online service providers to enhance the protection of their users. The Act sets forth a framework of relatively broad provisions and guiding principles that will be further refined through secondary legislation (see Sections 12, 29, 30). Like EU regulations, the UK framework establishes general obligations and specific duties applicable to more prominent social media platforms. The Online Safety Act also contains provisions for the protection of minors and content moderation similar to those enacted in the EU. A brief overview of the most significant provisions is provided.

- Sections 11 and 12 address the protection of minors online, mandating that platforms conduct risk assessments and implement monitoring mechanisms to mitigate potential harms.
- Sections 20, 21, and 22 govern the systems for lodging and resolving complaints while also imposing a duty to uphold freedom of expression within social media environments. In particular, these sections require platforms to incorporate clear information on policies and procedures to handle and resolve relevant complaints into their terms of service, accessible to all users, including minors.
- Sections 71 and 72 regulate the processes by which users can contest actions taken against them or their content by the platform. Such procedures must be

1. See Recitals 71 and 89 RSD. Recital 89 states that large platforms "should take measures to protect minors from content that may harm their physical, mental, or moral development and provide tools that allow conditional access to such information."

explicitly outlined within the terms of use.

- Section 77 requires larger platforms to publish an annual transparency report containing detailed information on various operational aspects, including content moderation practices.

The provisions of the Online Safety Act are rather broad and therefore must be further specified by the UK's audiovisual authority, Ofcom. In one of its initial public consultations aimed at developing the Online Safety Act, Ofcom (2024) sets out ten specific measures regarding content moderation, underscoring the importance it intends to place on such protective techniques.

Although the British regulation is still in its initial phase of implementation, it is apparent that, in matters of content moderation, it aligns with the provisions outlined in the EU Digital Services Act: transparency in terms of use and specific moderation processes; active disclosure of annual actions in this area; and procedural due process safeguards—such as the right to appeal—for affected users. As we noted with European regulations on content moderation, UK legislation focuses more on reporting, assessing, and removing content than on platform algorithmic recommendations.

In conclusion, analyzing the legal frameworks in the United States, the European Union, and the United Kingdom reveals two key trends. First, the rise of regulatory initiatives—such as the EU's Digital Services Act (DSA), the UK's Online Safety Act, and the proposed Kids Online Safety Act (KOSA) in the US—reflects growing public and political pressure to oversee content moderation practices. In the EU, this has already been applied in concrete enforcement actions. In 2024, the European Commission launched formal investigations into platforms such as X, TikTok, AliExpress, Facebook, Instagram, and Temu for potential DSA violations (Fabbri 2025). Meanwhile, the US continues to debate whether and how to legislate in this area.

Second, by 2025, the regulation primarily targets transparency and accountability in reporting and content removal. The DSA requires Very Large Online Platforms (VLOPs) with over 45 million EU users to publish annual risk assessment reports addressing systemic risks such as disinformation and harm to minors. As of February 17, 2024, all intermediary service providers must release yearly transparency reports on their content moderation practices. Platforms must also disclose their moderation policies, provide clear complaint mechanisms, and ensure that users have a fair due process. However, while the DSA improves transparency in algorithmic recommendation systems, specific strong regulations to protect minors from harmful content via these systems remain limited.

Although these frameworks mark significant progress, a regulatory focus on post-exposure responses may be insufficient. Future efforts must more directly tackle algorithmic amplification of harmful content to protect young users in a meaningful way.

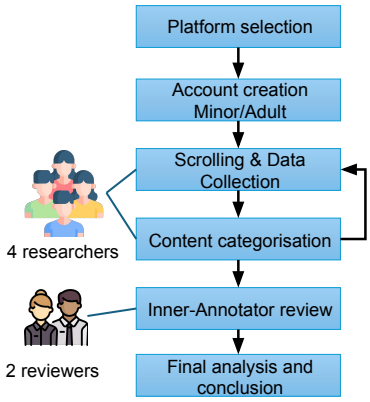


Figure 1: High-level methodology diagram illustrating the experimental sequence, including account creation and hierarchical video labeling (with third and, if needed, fourth annotator review to resolve disagreements).

3 Study Methodology

Our study examines how effectively social media video-sharing platforms protect minors from harmful or inappropriate content. Our objectives are to:

1. Compare the efficacy of content moderation systems for minors versus adults on leading video-sharing platforms.
2. Measure the speed at which harmful material surfaces under different interaction styles (passive scrolling versus explicit searching).
3. Assess whether the community guidelines of each platform align with their real-world enforcement practices.

Four researchers participated separately in the video annotation process. Each researcher managed a distinct set of accounts under standardized conditions; subsequently, two additional experts reviewed the labeled content to verify inter-annotator agreement. Figure 1 offers a visual overview of our experimental flow, from platform selection to final data analysis.

3.1 Platform Selection

To capture highly relevant mainstream user experiences, we focused on social media platforms where short-form videos and infinite scrolling feeds are the primary ways content is delivered. Two main criteria guided this selection:

1. The platform must provide algorithm-driven endless scrolling, continuously expos-

ing users to recommended videos.

2. The platform must have a sufficiently large user base to represent broader social media usage patterns.

Based on these two points, YouTube, Instagram, and TikTok were chosen due to their large user base and robust continuous-scrolling mechanisms for short videos. Consequently, our study focuses on Instagram Reels, TikTok's "For You" feed, and YouTube Shorts.

3.2 Creation of a Unified Framework for Harmful Content

This study aims to identify the types of harmful content present on the platforms. Therefore, we need a list of harmful content categories, definitions, and examples that act as a guideline.

Each selected platform operates under its own Community Guidelines (Meta Platforms, n.d.; TikTok, n.d.; YouTube, n.d.), which outline the allowed content and enforcement measures. However, these policies vary considerably. To conduct a consistent cross-platform evaluation, we established a *Unified Harmful Content Framework* that reconciles divergent content categorizations into a single taxonomy.

We started by reviewing the policies of each platform, capturing every category of harmful content. Categories present on one platform but absent on others were integrated into a single list, thus ensuring broad coverage. For example, Meta explicitly mentions "Privacy Violations," whereas TikTok's guidelines highlight "Dangerous Challenges and Activities" as a distinct category. Incorporating the unique categories of each platform produced the framework [2] summarized in Table 1.

Differences between how platforms moderate content for minors versus adults are readily observable in their published Community Guidelines, particularly regarding explicit or violent material. For example, YouTube (n.d.) and Meta (n.d.) specify a blanket prohibition on sexual content involving minors (i.e., illegal CSAM). In contrast, adult content deemed acceptable, such as artistic depictions, educational resources, or breastfeeding imagery, can be age-restricted or labeled rather than removed entirely. TikTok (2024) similarly applies tighter filters to minor accounts, restricting violent or sexual content, which, for adult users, can carry only warnings or age gates. Collectively, these policies highlight a shared commitment to age-sensitive moderation strategies that aim to protect younger audiences.

To further outline the spectrum of harmful content, our study is based on the enforcement policies of each platform (Meta Platforms, n.d.; Meta, n.d.; TikTok, n.d., n.d.; YouTube, n.d.; Google, n.d.), assigning severity levels to capture varying levels of risk. Low-severity

2. This taxonomy is publicly accessible at https://github.com/luismirallesp/OnlineContentCategories/blob/main/Social%20Media%20Platforms%20Community%20Guidelines%20New.xlsx.

Table 1: Summary of Content Categories. *Child Safety* excludes illegal child sexual abuse material (CSAM) and covers policy-relevant content that may sexualize or endanger minors without constituting CSAM.

Category	Subcategory	Definition
Violent and Criminal Behavior	Coordinating Harm	Organizing or inciting harmful actions, including real-world and digital crimes.
	Dangerous Organizations	Groups promoting violence, hate, or large-scale harm.
	Violence and Incitement	Calls for violence, threats, or glorification of harmful acts.
	Criminal Organizations	Large-scale criminal groups such as cartels or terrorist networks.
Hate Speech and Hateful Behavior	Hate Speech	Statements inciting hatred based on protected attributes.
	Hateful Groups	Organizations promoting hate-based ideologies.
	Cyberbullying	Repeated online harassment targeting individuals.
	Harassment	Intimidation, doxxing, or persistent unwanted behavior.
Abuse and Exploitation	Sexual and Physical Abuse	Non-consensual acts, coercion, or intimate violence.
	Human Exploitation	Trafficking, forced labor, or illegal adoption.
	Privacy Violations	Unauthorized sharing of sensitive personal data.
Mental and Behavioral Health	Self-Harm	Content encouraging self-harm or suicide.
	Eating Disorders	Promoting unhealthy eating habits.
	Dangerous Challenges	Urging participation in hazardous activities.
Sensitive and Mature Themes	Nudity	Regulation of sexual or explicit content.
	Graphic Content	Depictions of violence, mutilation, or harm.
	Sexual Services	Promotion or facilitation of sex work.
	Animal Abuse	Cruel or harmful treatment of animals.
Dis/Misinformation	General Misinformation	False or misleading content undermining public trust.
	Election Misinformation	Inaccurate claims compromising democratic processes.
	Health Misinformation	Content contradicting scientific or medical consensus.
	AI-Generated Media	Manipulated content such as deepfakes.
	Conspiracy Theories	Unsubstantiated allegations targeting institutions or individuals.
Regulated Goods	Illegal Sales	Promotion or trade of prohibited items (drugs, firearms).
	Gambling	Promotion of gambling, alcohol, or drugs.
Child Safety	Child Exploitation	Content sexualizing or endangering minors.
	Grooming	Deceptive interactions aimed at minors.
Spam and Fraud	Fake Engagement	Artificially boosting likes, views, or follows.
	Impersonation	Misleading accounts mimicking real users or brands.
	Fraud	Scams or deceptive financial schemes.
Privacy and Security	Personal Data	Sharing sensitive data, leading to risks.
	Cybersecurity	Breaches or unauthorized access to systems.
Legal Issues	IP Violations	Copyright or trademark infringement.
	Account Integrity	Measures preventing repeated policy violations.
	Authentic Identity	Ensuring users represent real individuals.
Enforcement Actions	Protection of Minors	Removal of content harmful to children.
	Local Laws	Compliance with jurisdiction-specific regulations.
	User Requests	Account removals upon user request.

material includes mild stereotypes or insensitive humor, while high-severity examples comprise direct violence or explicit hate speech.

3.3 User Interaction Modes

We examine two different models of user interaction—*passive* and *search-based*—to determine how different behaviors influence the probability of encountering harmful content. By comparing these modes, we aim to determine whether user intent, such as purposefully searching for specific topics, affects exposure rates, thereby informing content moderation strategies tailored to diverse usage patterns. Each experiment takes about 90–120 minutes on average.

Account 1: Passive Scrolling In this scenario, an account scrolls through 300 consecutive videos, each viewed for approximately 20 seconds, without performing additional actions such as searches, likes, comments, or shares. Any non-English videos are immediately skipped to simplify the annotation process. This setup approximates the experience of a new, largely passive user, offering insight into how recommendation algorithms push potentially harmful material to individuals with minimal engagement.

Account 2: Search-Based Scrolling In contrast, search-based scrolling adopts a *multi-stage* approach. First, the account views 100 videos passively, watching each one for 20 seconds. It then executes a search using "normal" keywords intended to simulate basic searches that a typical user might perform on the platform (see Table 2), briefly hovering over the first ten results (5–10 seconds each) without playing them in full. After this, the account proceeds with another 100-passive-video scroll. Subsequently, the user performs a second search using "low-risk" keywords, intending to simulate searches that are curious and potentially harmful, but not explicitly so, again hovering over the initial ten results. Finally, the account scrolls through the last set of 100 videos, maintaining the 20-second watch duration per video. As before, non-English videos are skipped immediately. By examining these sequential steps, we can see how even searching for seemingly benign or "low-risk" terms can unintentionally increase exposure to problematic or harmful content.

3.4 Account Creation

To systematically compare content moderation results across age groups and user behaviors, we created accounts for two age groups (13 and 18 years) on each platform. On TikTok and YouTube, we created two 13-year-old passive-scrolling accounts, two 13-year-old search-based scrolling accounts, two 18-year-old passive-scrolling accounts, and two 18-year-old search-based scrolling accounts, resulting in eight accounts per platform. This design allows each age and interaction-mode condition to be represented by two accounts, reducing the impact of random recommender effects.

Table 2: Keywords used for search-based scrolling.

Normal Keywords	Low-Risk Keywords
football	Fake news
make-up	Free game codes
games	Buy followers
happy	prank call
sad	calories
	flirting
	depressed
	Hackers
	Conspiracy
	Challenge
	Fake ID
	swatting
	fasting
	grope
	unalive

On Instagram, the web-based interface primarily returns user profiles rather than content-based search results. This means that instead of surfacing individual posts, videos, or trending content in response to queries, Instagram emphasizes user accounts that match the search terms. As a result, the ability to explore broader topics or topics via search is limited, especially for new or passive accounts. Consequently, we established only two pairs of Instagram accounts for each age group, which were limited to passive scrolling. In contrast, YouTube and TikTok supported the complete four-account pairs model (passive and search-based, each for ages 13 and 18), as their search functionalities return diverse content types (e.g., videos, channels, hashtags), making them more conducive to studying the impact of searching for keywords. In total, 20 accounts were created, each linked to a unique email address; however, the same email addresses could be reused on different platforms.

A particular challenge arose with the YouTube requirements for minors under 16. During account creation, the platform automatically requested parental supervision, requiring a parent or guardian to configure content restrictions. These options range from highly restrictive (only approved content) to nearly adult-like access. For this study, both child accounts were configured with the least restrictive parental settings. This approach was chosen to realistically simulate a scenario in which a minor sets up their account with minimal oversight, reflecting real-world behavior (Eltaher et al. 2025) and demonstrating how these safeguards can be bypassed or loosely enforced. By doing so, our objective was to assess the actual exposure risks faced by underage users who may not have parents who actively engage with them. Table 3 summarizes the final configurations of the accounts.

Table 3: Overview of account configurations across YouTube, Instagram, and TikTok for 13-year-old and 18-year-old accounts.

Platform	Age	Method
TikTok	13 (Minor)	Passive scroll
		Search-based scroll
	18 (Adult)	Passive scroll
		Search-based scroll
YouTube	13 (Minor)	Passive scroll
		Search-based scroll
	18 (Adult)	Passive scroll
		Search-based scroll
Instagram	13 (Minor)	Passive scroll
	18 (Adult)	Passive scroll

By consistently defining user ages and scrolling behaviors during account setup, we ensured reproducible results and minimized confounding variables across experiments.

3.5 Labeling Methodology

We adopted a multi-tiered labeling procedure, with particular attention to inter-annotator reliability and conflict resolution mechanisms.

Our primary classification process begins by tagging each video as *Not Harmful* (video does not fall under any of the categories defined in the framework of harmful content) or *Harmful* (video falls under any of the categories listed in the framework of harmful content defined earlier). Clips labeled *Harmful* are then further assigned to one of the harmful content categories from our framework (see Section 3.2) and given a severity rating: *Low*, *Medium*, or *High*.

Low-severity content typically includes material that engages with sensitive topics in a humorous or satirical manner. For instance, we encountered stand-up comedy clips joking about politics and identity in a self-aware, non-malicious tone. Similarly, several parody podcast segments mimicked controversial internet personalities for a comedic effect, without promoting the views being mocked.

High severity was used when videos contained explicit and severe forms of harm. In some cases, we observed people using a racial slur (i.e., the N-word) aggressively during rants, making blatantly sexist remarks about women's roles in society, or expressing homophobic views without irony or criticism. These videos demonstrated a clear intent to demean or provoke.

Medium severity was assigned in cases where interpretation was mixed. For example, one clip featured a creator using inflammatory language in a comedic skit, but the satirical

Table 4: Examples of Content by Severity and Category

Category	Severity	Examples (as observed during annotation)
Violent and Criminal Behavior	Low	A stand-up comedy set where the comedian joked about political revolutions, using irony to highlight the absurdity of extremist views.
	High	One video about how you can hit women if you transition into a woman.
Hate Speech and Hateful Behavior	Low	A parody podcast in which the host mimicked a controversial online figure using exaggerated speech and tone, seemingly to mock rather than promote hateful ideology.
	High	In one clip, a creator used the N-word repeatedly during an aggressive rant, alongside misogynistic and homophobic comments delivered without satire or critique.
Mental and Behavioral Health	Low	A video skit referenced a viral "breaking into dorm" circulating online, but presented it with a humorous tone, not clearly encouraging participation.
	High	A video in which a person was drunk and was flaunting the fact that they were drunk while driving.

framing was not immediately obvious. While one annotator viewed it as a parody, another found the content to be indistinguishable from serious hate speech. In such cases, where perceived intent or level of harm varied significantly between the annotators, we defaulted to a medium severity rating.

Table 4 categorizes some examples of videos according to their severity classification into a few categories. Initially, a single annotator (Labeler 1) evaluates each video for harm status, category, and severity. If Labeler 1 flags the video as Harmful, a second annotator (Labeler 2) reviews it to confirm or dispute the label. If Labeler 2 disagrees, either by labeling it Not Harmful or by assigning a different category or severity, a third annotator (Labeler 3) is used for a final review.

For example, suppose Labeler 1 classifies a video as "Harmful" under the category "Violent Behavior" with a Low severity. Labeler 2, upon review, disagrees and marks the video as *Not Harmful*, triggering the need for Labeler 3's judgment. If Labeler 3 agrees with Labeler 1 or 2, either confirming that the video is harmful and violent or agreeing that it is not harmful at all, the majority vote (2:1) is accepted as the final decision.

However, if Labeler 3 introduces a new perspective, for instance, labeling the content as *Harmful* but under a different category, such as "Mental and Behavioral Health," then there is no majority consensus. In such a case, a fourth annotator (Labeler 4) is brought in to make a final decision.

This hierarchical process addresses several challenges inherent in content moderation. First, some videos span multiple categories (e.g., a self-harm video might also feature explicit or graphic themes). In such cases, the final label selected is the one agreed upon by at least two annotators. Second, severity can be subjective. For instance, mild hate speech might be classified as *Low* severity by one annotator and *High* severity by another; in these instances, we apply a rule that if there is a disagreement between two levels (e.g., Low vs. High), the rating defaults to Medium.

Two categories from our Harmful Content Framework, *Privacy and Security* and *Enforcement Actions*, were excluded from manual annotation. These categories address content beyond immediate view (e.g., personal data leaks, platform-level enforcement decisions) and thus could not be reliably observed through passive or minimal interactive engagement alone.

3.6 Methodology and Ethical Considerations

Harm is a subjective and context-dependent construct, influenced by the cultural background, lived experiences, and levels of vulnerability of individuals (Gerrard 2019). In addition, the assessment of the harm of the content in the study must, due to the nature of the research task, be carried out by adult researchers. This introduces a potential disconnect: adults can interpret tone, nuance, or severity differently than adolescents, especially in borderline or ambiguous cases (Smith et al. 2024; Kumar et al. 2021). Indeed, even between adolescents, the level of harm or the perception of harm can differ according to individual vulnerabilities, culture, and other personal factors. Awareness of these discrepancies is important when evaluating harm thresholds and can influence how certain videos are classified and interpreted between age groups. However, while being aware of this subjectivity is important when interpreting results, manual labeling against a defined unified framework for harmful content helps to alleviate the impact of subjectivity, providing a published reference point of content definition and harm levels alongside the results.

Some content flagged as “low-risk,” such as mental health or sexual health conveyed in a humorous or satirical way, tread a fine line between harmful and informative. In certain contexts, these forms of expression could in theory increase the relatability and engagement among youth, particularly for topics that are typically stigmatized or avoided in more formal messages (Corrigan et al. 2014; Albers et al. 2022). Although such content raises moderation concerns, particularly when unconstrained by age, its presence in recommendation streams could also reflect how young audiences seek out accessible, peer-like discourse on sensitive issues in an adult-free environment.

We deliberately restricted user interaction to an absolute minimum, avoiding actions such as liking, commenting, following, or rewatching videos, to observe the default content served by the platform algorithms without behavioral influence. This passive approach was chosen to simulate the experience of a newly created account or minimally active

user and to isolate how platforms push content based solely on the user's age.

All non-English videos were skipped during the annotation phase to simplify evaluation and reduce linguistic ambiguity among annotators. Although this choice improved consistency, it also excludes a substantial portion of global content and may have resulted in the omission of potentially harmful material presented in other languages.

The creation of fake minor profiles to obtain data for our research warrants some ethical reflections. First, it is important to note that honesty in stating a user's age on social media serves two main purposes: (1) to prevent exposure to inappropriate content for their age and (2) to make it harder for adults to contact minors. Neither of these risks applies to our research, since (1) adults are exposed to content intended for minors, and (2) no interaction has been made with these profiles with other users, whether adults or minors.

The lack of sincerity in expressing the researcher's identity and age could be considered a form of deception with potential negative consequences for the platform, possibly violating its Terms of Use. To determine whether the creation of fake profiles is ethically justified, it is necessary to weigh the harm caused to the platform against the benefits of research (Elovici et al. 2014; Xiao, Sellars, and Scheffler 2025).

The harm is minimal, as evidenced by the platform's limited diligence in verifying the identity of new users. Meanwhile, the benefits are significant, as the study seeks to understand how content recommendation algorithms work and to contribute to the protection of minors on the internet—a goal that undoubtedly serves the public interest. Therefore, it can be concluded that the creation of fake profiles for this purpose is justified. Additionally, the research technique used is further justified by the difficulty of obtaining permission from social networks to carry out a study of this nature. Finally, the *Child Safety* category excludes child sexual abuse material (CSAM). No CSAM was encountered or retained during the study; any such material would have been immediately reported to law enforcement and excluded from the dataset.

4 Analysis of the Results

This section presents key findings on how age, interaction mode, and platform policies jointly shape users' exposure to harmful content. Our analysis integrates several data points from the experiments—including overall exposure rates to harmful content across all the categories defined, severity ratings per video deemed as harmful, time to first harmful video for every account, differences in total harmful content exposure for the 13-year-old and the 18-year-old accounts, and differences in exposure to harmful content based on user behavior (passive scrolling vs. searching and scrolling)—and reveals patterns that underscore the need for more robust age-specific moderation strategies.

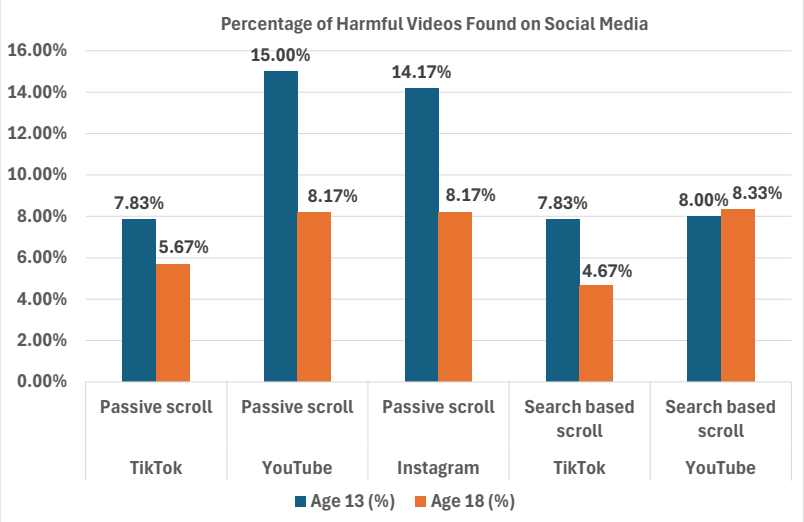


Figure 2: Comparison of age-based harmful content trends.

4.1 Age-Based Trends in Harmful Content Exposure

To assess whether platforms effectively adjust their moderation practices for younger audiences, we compared user accounts configured as 13-year-olds and 18-year-olds under identical conditions. As shown in Figure 2, *13-year-old accounts generally received higher levels of harmful content on most platforms and interaction modes*. The only exception was YouTube in the search-based scenario, where both age groups faced nearly identical exposure rates.

A key finding is that harmful content for the 13-year-old accounts ranged from 7.83% to 15%, while older accounts typically experienced lower exposure rates (4.67% to 8.33%). Even modest percentages for younger audiences are concerning, since no platform should recommend damaging material to minors. These results reinforce the importance of *age-specific moderation protocols* designed to provide stronger safeguards for younger users.

4.2 Impact of User Interaction Modes on Harmful Content Recommendations

We then examined how different interaction styles affect harmful content recommendations. Specifically, we compared *passive scrolling*, which involves no active engagement (such as likes, searches, or comments), with *search-based scrolling*, which includes targeted keyword searches alongside passive viewing. Section 3.3 details these procedures.

Figure 3 shows the proportion of harmful content encountered by the simulated 13-year-

old and 18-year-old users on YouTube and TikTok in both modes. For adults (Figure 3b), TikTok exhibited a slight reduction in harmful content when users actively searched (from 5.67% to 4.67%), while YouTube's rates remained basically unchanged (8.17% vs. 8.33%).

For 13-year-olds (Figure 3a), exposure to harmful content on TikTok remained unchanged at 7.83% when users searched, while YouTube's dropped significantly (from 15% to 8%). Together, these mixed outcomes suggest that *search-based interactions* can dampen or exacerbate exposure to harmful material, depending on the platform's recommendation algorithms.

4.3 Impact of Search Behavior on Harmful Content Exposure

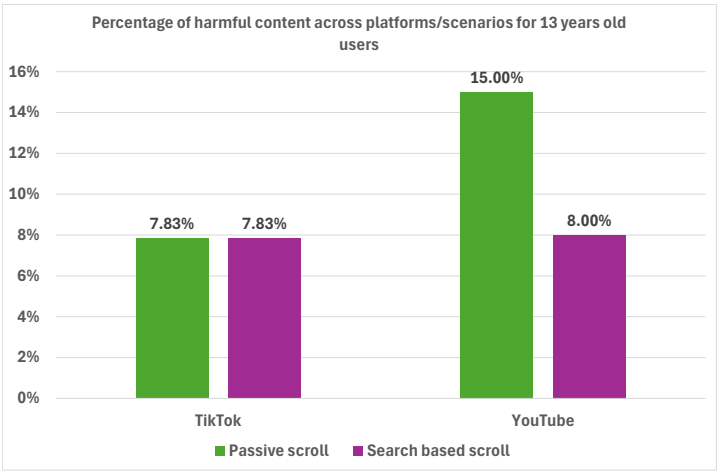
To further unravel how specific keyword searches influence exposure, we tracked recommendations in three sequential "rounds" (Figure 4). Each round involved scrolling through 100 videos: (1) purely passive scrolling; (2) searching with neutral keywords (e.g., "football"); and (3) searching with riskier keywords (e.g., "fasting").

In TikTok, the amount of harmful content for the younger accounts increased following the neutral search, but decreased by the final round. On YouTube, exposure for younger accounts remained broadly stable across the first two rounds, but increased markedly after low-risk keyword searches. This pattern implies that specific keyword searches may not necessarily push vulnerable users toward more harmful material in all cases, but can substantially amplify exposure depending on the platform and search context, highlighting potential pitfalls in platform-based content filtering.

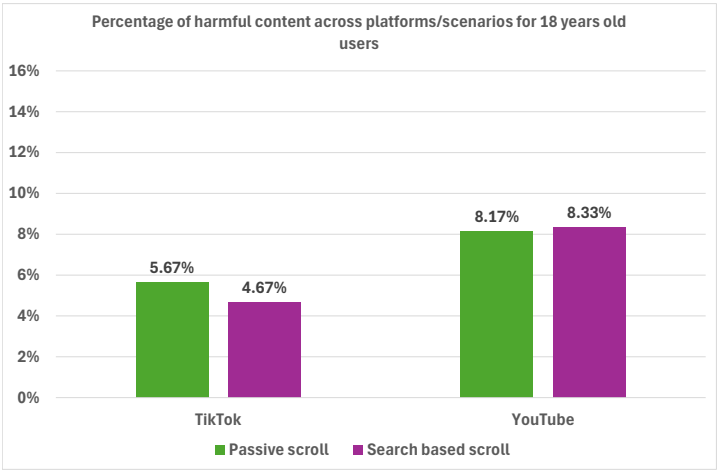
4.4 Time to First Harmful Video Across Platforms

Another measure of moderation quality is how quickly new users encounter potentially harmful videos. Figure 5 illustrates the time that elapsed before a user's first exposure to harmful content, broken down by platform, age group, and interaction mode.

As shown in Figure 5, accounts configured as 13-year-olds usually encountered harmful content faster than accounts configured as 18-year-olds. In passive scrolling, YouTube (3:06 minutes) and TikTok (3:49 minutes) enable near-immediate exposure for minors. However, search-based scrolling trends differ by platform: on TikTok, it slightly *increases* the time to the first harmful clip (4:43 minutes vs. 3:49 minutes), whereas on YouTube, it *reduces* it significantly (1:28 minutes vs. 3:06 minutes). Overall, child-configured accounts in our study commonly encountered harmful content in under five minutes, whereas adult-configured accounts showed longer and more variable times, ranging from approximately three minutes to just over eight minutes, raising concerns about the current safeguards' capacity to prevent early, potentially harmful exposure.



(a)



(b)

Figure 3: Analysis of harmful content recommendations across platforms and scenarios for accounts configured to (a) 13-year-old users and (b) 18-year-old users.

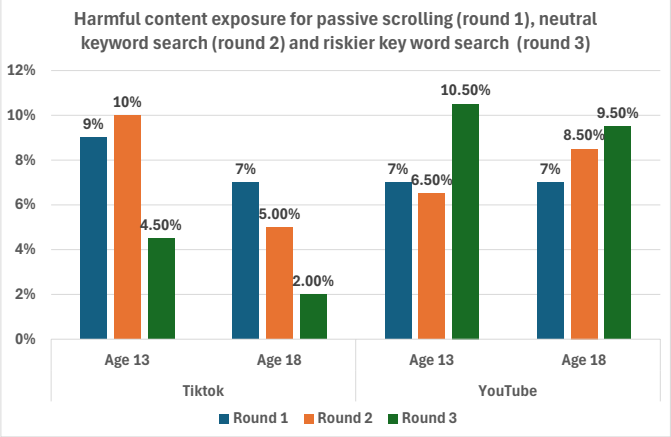


Figure 4: Impact of search behavior on harmful content exposure.

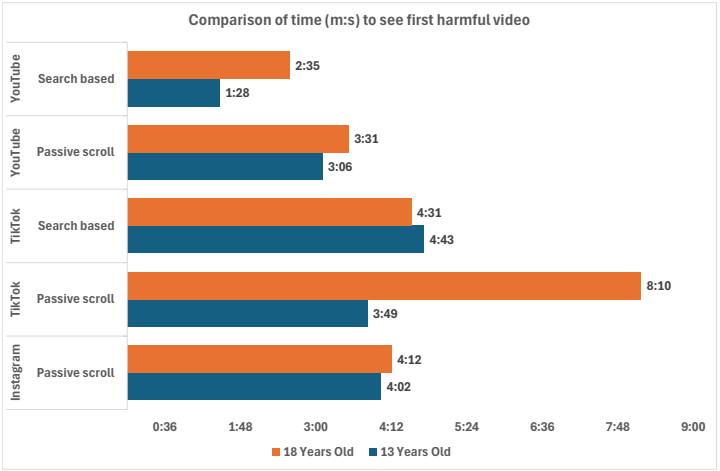


Figure 5: Time (*minutes:seconds*) to the first harmful video across platforms and age groups.

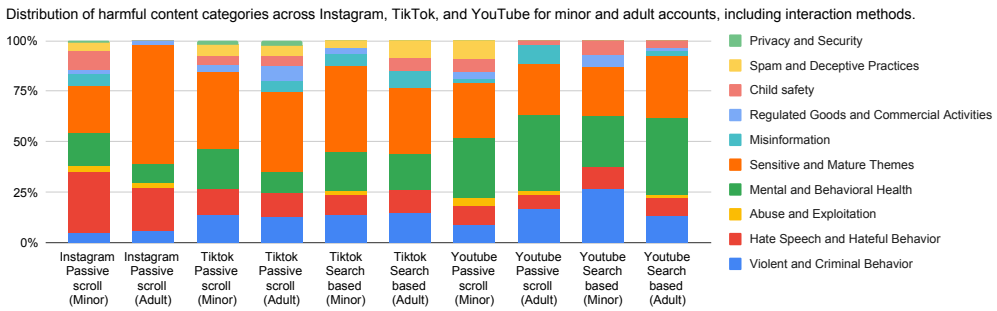


Figure 6: Percentage Distribution of harmful content categories across different platforms and interaction methods for accounts configured as (a) 13-year-old users and (b) 18-year-old users.

4.5 Types and Severity Levels of Harmful Content Encountered

Next, we examine the types of harmful content that the accounts encountered most frequently, along with their severity. Figure 6 shows category-level distributions for 13- and 18-year-old user accounts on YouTube, TikTok, and Instagram, each with passive or search-based scrolling. In this figure, all category distributions are normalized so that the total equals 100%, allowing a clearer comparison between platforms and user types.

Across all account groups, *Sensitive & Mature Themes* consistently dominate recommended harmful content. In addition, Instagram accounts configured as 13-year-olds exhibit a significantly higher exposure (5%) to *Hate Speech and Hateful behavior*.

Harmful videos also vary in severity. Table 5 uses a heatmap color scheme to represent severity levels, where yellow denotes low severity, orange shades indicate medium severity, and red signifies high severity. The intensity of each color varies according to the corresponding percentage, effectively illustrating both the severity and prevalence of harmful content categories across different scrolling methods. On TikTok, 13-year-old accounts encountered more low-severity *Sensitive & Mature Themes* under both passive (3%) and search-based (4.17%) scrolling. These younger users still faced more harmful content than older users, whose exposure rates decreased across most categories.

YouTube exhibited a larger overall share of harmful content, particularly among 13-year-old accounts. These users frequently encountered low-severity mental health-related content (5.17% through passive scrolling) and other sensitive topics (5%). Although active searching reduced exposure to some mental health-related harms, 18-year-olds still faced a nontrivial amount of such material (3.33% through search-based scrolling).

Instagram presented a different pattern: 13-year-old accounts encountered a relatively high percentage of low-severity hate speech (5%), sensitive themes (3.5%), and mental health content (2.5%). For 18-year-olds on Instagram, overall exposure to harmful

Table 5: Distribution of Harmful Content Categories by (Platform & Age), Method, and Severity (in %).

User Info		Severity	Harmful Content Categories (%)									
Platform & Age	Method (scroll/search)		Viol. &Crim.	Hate Speech	Abuse &Expl.	Mental Health	Sens. Themes	Misinfo.	Reg. Goods	Child Safety	Spam	Privacy &Sec.
TikTok 13	Passive scroll	Low	0.83	1.17	0.00	1.33	3.00	0.00	0.33	0.17	0.50	0.17
		Medium	0.00	0.00	0.00	0.33	0.17	0.00	0.00	0.00	0.00	0.00
		High	0.33	0.00	0.00	0.00	0.17	0.00	0.00	0.17	0.00	0.00
	Search based scroll	Low	1.33	1.00	0.17	1.67	4.17	0.50	0.33	0.00	0.33	0.00
		Medium	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00
		High	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TikTok 18	Passive scroll	Low	0.67	0.50	0.00	0.67	2.17	0.33	0.50	0.33	0.33	0.17
		Medium	0.17	0.33	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00
		High	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Search based scroll	Low	0.83	0.67	0.00	1.00	1.83	0.50	0.00	0.33	0.50	0.00
		Medium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		High	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
YouTube 13	Passive scroll	Low	1.67	1.67	0.83	5.17	5.00	0.33	0.50	1.17	1.67	0.00
		Medium	0.00	0.00	0.00	0.33	0.00	0.00	0.17	0.00	0.00	0.00
		High	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Search based scroll	Low	2.33	1.00	0.00	1.50	2.17	0.00	0.33	0.67	0.00	0.00
		Medium	0.17	0.00	0.00	0.83	0.17	0.00	0.00	0.00	0.00	0.00
		High	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00
YouTube 18	Passive scroll	Low	1.33	0.67	0.17	3.00	2.33	0.83	0.00	0.17	0.00	0.00
		Medium	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00
		High	0.17	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00
	Search based scroll	Low	1.17	0.83	0.17	3.33	3.00	0.17	0.17	0.17	0.00	0.00
		Medium	0.17	0.00	0.00	0.33	0.00	0.00	0.00	0.17	0.00	0.00
		High	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Instagram 13	Passive scroll	Low	0.83	5.00	0.33	2.50	3.50	1.00	0.33	1.33	0.67	0.17
		Medium	0.00	0.00	0.17	0.17	0.33	0.00	0.00	0.17	0.00	0.00
		High	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Instagram 18	Passive scroll	Low	0.33	1.33	0.17	0.67	4.67	0.00	0.17	0.00	0.00	0.00
		Medium	0.17	0.50	0.00	0.17	0.17	0.00	0.00	0.00	0.00	0.00
		High	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00

content was lower, except for sensitive themes, which increased to 4.67%.

Taken together, **low-severity harmful content** is the most prevalent form encountered across all platforms, categories, interaction types, and age brackets, suggesting a potential for normalization over time. Repeated exposure, even to less extreme harmful content, has been associated with desensitization, including reduced emotional response and increased accessibility to aggressive cognitions (Krahé et al. 2011). However, the evidence—particularly for children and adolescents—remains largely correlational, with desensitization reflected more consistently in attitudinal and empathic changes than in overt behavior (Funk 2005). More recent research suggests that desensitization is not uniform, but selective, with repeated exposure leading to reduced sensitivity to some forms of harm while increasing awareness of others (Meerson, Koban, and Matthes 2025).

4.6 Discussion and Key Findings

Our findings indicate that **13-year-old accounts in our study experienced higher and faster exposure to harmful videos** compared to 18-year-old accounts in the majority of scenarios. For accounts configured as minors, exposure rates range from 7.83% to 15%; for accounts configured as adults, they average between 4.67% and 8.33%. For

accounts configured as minors, that exposure usually came within five minutes.

Such disparities underscore the urgent need for *age-specific video moderation*: allowing minors to encounter harmful videos—particularly within minutes of use—is problematic from both ethical and safety standpoints.

One possible explanation for this pattern is that social media algorithms prioritize engagement, often recommending content that maximizes watch time and interaction. Younger users may be more likely to engage with extreme videos, even unintentionally, which could lead to recommendation systems pushing increasingly harmful videos. This raises critical questions about whether current moderation strategies effectively safeguard younger audiences.

Interaction modes also impact exposure, though in platform-specific ways. On YouTube, *search-based scrolling* significantly reduced harmful videos for accounts configured as minors compared to passive scrolling. However, in TikTok, the extent of harmful recommendations for accounts for minors is the same for both interaction modes. Moreover, risky keyword searches on YouTube amplified exposure, increasing from 7% to 10.5% in sequential rounds.

Beyond mere percentage rates, child-configured accounts also encountered harm more *quickly* (in five minutes) than adult-configured accounts (in under nine minutes). This finding raises serious questions about the real-time effectiveness of current safeguards in limiting early exposure of minors.

The dominant category of harmful videos for both sets of 13- and 18-year-old accounts is *Sensitive & Mature Themes* such as violence, shocking content such as car crashes or accidents, and some sexually suggestive content. However, Instagram accounts configured as 13-year-olds also faced a higher incidence of *Hate Speech and Hateful Behavior* like racial abuse and misogyny. Low-severity forms of harm, ranging from 0.17% to 5% across platforms, may appear minor, yet pose cumulative risks of desensitization and normalization.

In general, the data suggest that recommendation systems can amplify harmful videos rather than suppress them, exposing minors to risk. Our findings should be interpreted in the context of the study design. Because recommendation systems exhibit inherent variability, the results represent one rigorously measured point on a wider spectrum of possible system behaviors. Moreover, our analysis reflects platform dynamics at a specific moment in time; tracking changes longitudinally will be essential to determine how these risks evolve.

This highlights the need for more robust and transparent video moderation methods to effectively reduce such exposure.

4.7 Implications for Content Moderation

Our study represents findings for a small set of sample accounts. As further context to these particular results, strategies used by young people to manage their online content usage indicate that they employ active navigation strategies, rather than behaving as passive recipients. Duvekot et al. (2024) draw on 94 empirical studies from 35 countries over the past two decades, showing how children and young people look to curate their own custom news feed, by choosing particular apps, following selected outlets, and consuming short, fragmentary items that feel personally relevant. Similarly, Swart (2021) highlights how tactics used by young people for content navigation include managing privacy, selectively engaging with content, and relying on social verification. These findings point toward young people finding ways to sidestep untrustworthy or unwanted material and maintain a sense of safety online. This is encouraging when assessing the level of risk associated with using online platforms, but it nevertheless does not reduce the responsibility of the social media platforms for content moderation standards.

Based on our results, we propose the following recommendations to improve user safety, particularly for minors.

- **Strengthening Age-Specific Moderation:** Across our tests, nominal 13-year-old users encountered harmful material more frequently and at faster rates than older profiles. Although recommender performance may fluctuate, this consistent directional trend underscores the need for enhanced protective measures for younger users, particularly in the early minutes of platform engagement.

- **Improving Consistency in Enforcement:** Despite stated platform policies, *Sensitive & Mature Themes* and *Hate Speech* categories surfaced in our recommendations. Some users, including adults, may intentionally seek borderline or sensitive content. While not all such content violates platform policies, our findings indicate a clear gap between stated moderation standards and their practical enforcement, particularly for accounts registered as minors. This points to the need for clearer harm definitions, more consistent enforcement, and greater transparency in how moderation decisions affect recommendation systems, including through independent audits and public reporting.

- **Mitigating Normalization of Low-Severity Harm:** A growing body of evidence indicates that repeated exposure to harmful or emotionally intense media content can contribute to desensitization, particularly among adolescents. Experimental and neurophysiological studies show that even short-term exposure to violent or harmful digital media can dampen empathy-related responses. For example, adolescents demonstrate reduced neural sensitivity to others' pain following violent gameplay (Miedzobrodzka et al. 2023), while habitual exposure to violent media has been linked to broader emotional desensitization patterns (Fanti et al. 2009). Foundational longitudinal research similarly shows that emotional desensitization

during adolescence can predict later aggressive or harmful behaviors (Mrug, Madan, and Windle 2016). Although our study does not directly measure psychological impacts, this literature highlights the importance of considering the cumulative influence of repeated exposure to lower-severity harmful material within algorithmic recommendations.

By integrating these improvements, social media platforms can align daily recommendation algorithms with their expressed commitments to user safety. This alignment is particularly urgent for minors, given their increased vulnerability and rapid exposure timelines.

5 Conclusion

This study presents an exploratory assessment of how the leading video-sharing platforms handle young user accounts' exposure to potentially harmful content, based on passive and search-based engagements on TikTok, YouTube, and Instagram. Our comparison between 13- and 18-year-old user accounts showed that accounts set at 13 years old faced disproportionately higher levels of harmful videos, spanning 7–15% versus 4–8% for accounts configured for adults. In several cases, harmful videos were recommended to the child accounts within just three or four minutes, highlighting not only a greater volume but also a faster pace of exposure.

A severity-focused analysis reveals that *low-severity* harmful videos are particularly widespread, potentially normalizing harmful material and fostering desensitization over time. Moreover, our findings expose inconsistencies in content definitions: some platforms explicitly list categories such as "Privacy Violations" or "Dangerous Challenges," while others lack clarity, creating enforcement gaps and leaving younger users susceptible to incompletely regulated content.

Age-specific filtering and better alignment between platform policies and content moderation of video recommendation algorithms are therefore essential steps to reduce youth exposure to harmful material. Instead of prioritizing engagement metrics, platforms must hold minors' feeds to more stringent moderation standards. *Stronger regulatory oversight* is also needed to drive platforms to implement uniform protections for minors and ensure consistent enforcement of community guidelines.

The findings of this study contribute evidence that minors experience exposure to potentially harmful or influential content on video-sharing platforms. Several avenues for future research may help build on this work. First, our findings warrant further investigation on a larger scale to increase the confidence of content moderation findings. Another potential direction involves repeating the study later to evaluate whether social media platforms have improved the effectiveness of their algorithms and moderation systems in protecting teenage users. Another promising area is the development and

evaluation of educational interventions, such as game-based tools, to help children and their parents or guardians recognize early signs of online grooming. Additionally, future research could explore the feasibility of device-level solutions designed to filter harmful content before it is displayed to minors.

Further investigation is also warranted on the subjectivity of content classification, particularly on low-level harm. This study relied on adult researchers to evaluate the content, although there may be notable differences in how adults and minors perceive and interpret this material. Content that adults perceive as low severity can be interpreted by younger users as highly distressing or threatening, due to their limited life experience or emotional maturity. Involving young people directly, through surveys, interviews, or focus groups, can help close the gap between how adults and minors understand harmful content. These methods allow young users to share how they feel about different types of content, making classification systems more accurate and better suited to their experiences.

It may also be valuable to examine the role of low-risk content that uses humor or satire to address sensitive topics such as mental health or sexual health. Although such content may raise moderating concerns, it may also facilitate engagement and understanding among youth, particularly when formal messaging is less effective. Finally, the hypothesis that repeated exposure to low-severity harmful content contributes to desensitization remains an important, though underexplored, research question. Empirical work in this area could help clarify the long-term impacts of such content on young audiences. Together, these directions offer opportunities to support a safer and more balanced digital environment for children and adolescents.

Additionally, future work could expand on the current methodology by incorporating a wider range of user behaviors, including engagement with specific content, to better understand how recommendation algorithms adapt and escalate exposure based on interaction, as well as including multilingual content, ideally with the support of annotators fluent in relevant languages, to ensure broader cultural and linguistic representation in harm detection.

References

Albers, Leondard F., Folkertje B. Bergsma, Hilda Mekelenkamp, Rob C. M. Pelger, Eveliene Manten-Horst, and Henk W. Elzevier. 2022. "Discussing Sexual Health with Adolescent and Young Adults with Cancer: A Qualitative Study Among Healthcare Providers." *Journal of Cancer Education* 37, no. 1 (February): 133–40. https://doi.org/10.1007/s13187-020-01796-0.

Amnesty International. 2023. *Driven into the Darkness: How TikTok Encourages Self-Harm and Suicidal Ideation.* Amnesty International. https://www.amnesty.org/en/documents/pol40/7350/2023/en/.

Anderson, Monica, Michelle Faverio, and Jeffrey Gottfried. 2023. "Teens, Social Media & Technology 2023." Pew Research Center, December 11, 2023. https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/.

Baker, Catherine, Debbie Ging, and Maja Brandt Andreasen. 2024. *Recommending Toxicity: The Role of Algorithmic Recommender Functions on YouTube Shorts and TikTok in Promoting Male Supremacist Influencers.* https://voxpol.eu/file/recommending-toxicity-the-role-of-algorithmic-recommender-functions-on-youtube-shorts-and-tiktok-in-promoting-male-supremacist-influencers/.

Beresford, O., A. Cooney, A. Keogh, E. Flynn, and M. Messena. 2023. *Keeping Kids Safer Online.* Research report. https://www.cybersafekids.ie/wp-content/uploads/2023/09/CSK_Data-Trends-Report-2023-V2-Web-Version.pdf.

Center for Countering Digital Hate. 2022. *Deadly by Design: TikTok Pushes Harmful Content Promoting Eating Disorders and Self-Harm into Young Users' Feeds.* https://counterhate.com/wp-content/uploads/2022/12/CCDH-Deadly-by-Design_120922.pdf.

Chandrasekharan, Eshwar, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. "Crossmod: A Cross-Community Learning-Based System to Assist Reddit Moderators." *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 1–30. https://doi.org/10.1145/3359276.

Children's Commissioner. 2022. "Digital Childhoods: A Survey of Children and Parents," September 29, 2022. https://www.childrenscommissioner.gov.uk/report/digital-childhoods-a-survey-of-children-and-parents/.

Cho, Clare Y., and Ling Zhu. 2025. *Social Media: Content Dissemination and Moderation Practices.* Report 46.662. https://www.congress.gov/crs_external_products/R/PDF/R46662/R46662.10.pdf.

Competition Policy International. 2025. "Meta's Moderation Shift and the TikTok Ban: What's Cooking in the US and What This Means for EU Platform Regulation," February 19, 2025. Accessed March 19, 2025. https://www.pymnts.com/cpi-posts/metas-moderation-shift-and-the-tiktok-ban-whats-cooking-in-the-us-and-what-this-means-for-eu-platform-regulation/.

Corrigan, Patrick W., Karina J. Powell, Joyce K. Fokuo, and Kristin A. Kosyluk. 2014. "Does Humor Influence the Stigma of Mental Illnesses?" *The Journal of Nervous and Mental Disease* 202, no. 5 (May): 397–401. https://doi.org/10.1097/NMD.0000000000000138.

Dixon, Stacy Jo. n.d. "Most Popular Social Networks Worldwide as of April 2024, by Number of Monthly Active Users." Statista. Accessed March 19, 2025. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Duvekot, Sophie, Camila Melícia Valgas, Yael de Haan, and Wiebe de Jong. 2024. "How Youth Define, Consume, and Evaluate News: Reviewing Two Decades of Research." *New Media & Society* 0, no. 0 (July 26, 2024): 14614448241262809. https://doi.org/10.1177/14614448241262809.

Ekō. 2023. *Suicide, Incels, and Drugs: How TikTok's Deadly Algorithm Harms Kids.* Ekō. https://s3.amazonaws.com/s3.sumofus.org/images/eko_Tiktok-Report_FINAL.pdf.

———. n.d. "About Us." Accessed March 19, 2025. https://www.eko.org/about/.

Elovici, Yuval, Michael Fire, Amir Herzberg, and Haya Shulman. 2014. "Ethical Considerations When Employing Fake Identities in Online Social Networks for Research." *Science and Engineering Ethics* 20, no. 4 (November 3, 2014): 1027–43. https://doi.org/10.1007/s11948-013-9473-0.

Eltaher, Fatmaelzahraa, Rahul Krishna Gajula, Luis Miralles-Pechuán, Christina Thorpe, and Susan McKeever. 2025. "The Digital Loophole: Evaluating the Effectiveness of Child Age Verification Methods on Social Media." In *Proceedings of the 11th International Conference on Information Systems Security and Privacy - Volume 2: ICISSP,* 213–22. INSTICC, SciTePress. ISBN: 978-989-758-735-1. https://doi.org/10.5220/0013248300003899.

European Union. 2022. "Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)." Accessed March 19, 2025. https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng.

Fabbri, Matteo. 2025. "The Role of Requests for Information in Governing Digital Platforms Under the Digital Services Act: The Case of X." *Journalism and Media* 6, no. 1 (March 12, 2025): 41. https://doi.org/10.3390/journalmedia6010041.

Fanti, Kostas A., Eric Vanman, Christopher C. Henrich, and Marios N. Avraamides. 2009. "Desensitization to Media Violence over a Short Period of Time." *Aggressive Behavior: Official Journal of the International Society for Research on Aggression* 35, no. 2 (March): 179–87. https://doi.org/10.1002/ab.20295.

Fibrilla, Firda, Martini Fairus, and Holiratul Raifah. 2021. "Exposure to Pornography Through Social Media on Sexual Behavior of High School Teenagers in Metro City." *IOSR Journal of Nursing and Health Science* 9, no. 6 (July): 1–8. https://doi.org/10.9790/1959-0906040108.

Fiesler, Casey, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. "Reddit Rules! Characterizing an Ecosystem of Governance." In *Proceedings of the International AAAI Conference on Web and Social Media,* vol. 12. 1. June 15, 2018. https://doi.org/10.1609/icwsm.v12i1.15033.

Funk, Jeanne B. 2005. "Children's Exposure to Violent Video Games and Desensitization to Violence." *Child and Adolescent Psychiatric Clinics* 14, no. 3 (July): 387–404. https://doi.org/10.1016/j.chc.2005.02.009.

Ge, Jiaoju, Yuepeng Sui, Xiaofeng Zhou, and Guoxin Li and. 2021. "Effect of Short Video Ads on Sales Through Social Media: The Role of Advertisement Content Generators." *International Journal of Advertising* 40, no. 6 (January 15, 2021): 870–96. https://doi.org/10.1080/02650487.2020.1848986.

Gerrard, Ysabel. 2019. "Behind the Screen: Content Moderation in the Shadows of Social Media." *New Media & Society* 22 (September 23, 2019): 579–82. https://doi.org/10.1177/1461444819878844.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press. https://archive.org/details/Custodians_of_the_Internet_by_Terleton_Gillespie.

———. 2022. "Do Not Recommend? Reduction as a Form of Content Moderation." *Social Media and Society* 8, no. 3 (August 19, 2022). https://doi.org/10.1177/20563051221117552.

Google. n.d. "YouTube Policy Removals." Accessed March 19, 2025. https://transparencyreport.google.com/youtube-policy/removals?hl=en.

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, no. 1 (February 28, 2020): 2053951719897945. https://doi.org/10.1177/2053951719897945.

Jiang, Jialun Aaron, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. "Understanding International Perceptions of the Severity of Harmful Content Online." *PLOS ONE* 16, no. 8 (August 27, 2021): 1–22. https://doi.org/10.1371/journal.pone.0256762.

Kosseff, Jeff. 2019. “First amendment protection for online platforms.” *Computer Law & Security Review* 35 (2): 199–213. ISSN: 2212-473X. https://doi.org/https://doi.org/10.1016/j.clsr.2018.12.002.

Krahé, Barbara, Ingrid Möller, L. Rowell Huesmann, Lucyna Kirwil, Juliane Felber, and Anja Berger. 2011. “Desensitization to Media Violence: Links with Habitual Media Violence Exposure, Aggressive Cognitions, and Aggressive Behavior.” *Journal of Personality and Social Psychology* 100 (4): 630. https://doi.org/10.1037/a0021711.

Kumar, Deepak, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. “Designing Toxic Content Classification for a Diversity of Perspectives.” In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021),* 299–318. June 4, 2021. https://doi.org/10.48550/arXiv.2106.04511.

Liu, Mingli, Aixia Zhuang, Jill M. Norvilitis, and Tian Xiao. 2024. “Usage Patterns of Short Videos and Social Media Among Adolescents and Psychological Health: A Latent Profile Analysis.” *Computers in Human Behavior* 151 (February): 108007. ISSN: 0747-5632. https://doi.org/10.1016/j.chb.2023.108007.

Meerson, Rinat, Kevin Koban, and Jörg Matthes. 2025. “Too Much of What? Two-Wave Panel Evidence for Selective (De-) Sensitization Through Frequent Exposure to Different Kinds of Digital Hate.” *Journal of Computer-Mediated Communication* 30, no. 2 (March 13, 2025): zmaf002. https://doi.org/10.1093/jcmc/zmaf002.

Meta. n.d. “Child Sexual Exploitation, Abuse, and Nudity.” Accessed March 19, 2025. https://transparency.meta.com/policies/community-standards/child-sexual-exploitation-abuse-nudity/.

———. n.d. “Efforts to Protect Minors in Facebook.” Accessed March 19, 2025. https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps/.

———. n.d. “Facebook Community Standards.” Accessed March 19, 2025. https://transparency.meta.com/policies/community-standards/.

———. n.d. “Instagram Community Standards.” Accessed March 19, 2025. https://help.instagram.com/477434105621119/?helpref=faq_content.

———. n.d. “Instagram Warning Screen.” Accessed March 19, 2025. https://www.facebook.com/help/instagram/188848648282410.

———. n.d. “Transparency: Content Moderation.” Accessed March 19, 2025. https://transparency.meta.com/enforcement/.

Meta Platforms. n.d. “Community Standards.” Accessed March 19, 2025. https://transparency.meta.com/policies/community-standards/.

Miedzobrodzka, Ewa, Johanna C. van Hooff, Lydia Krabbendam, and Elly A. Konijn. 2023. "Desensitized Gamers? Violent Video Game Exposure and Empathy for Pain in Adolescents–an ERP Study." *Social Neuroscience* 18, no. 6 (December): 365–81. https://doi.org/10.1080/17470919.2023.2284999.

Mosnar, Matej, Adam Skurla, Branislav Pecher, Matus Tibensky, Jan Jakubcik, Adrian Bindas, Peter Sakalik, and Ivan Srba. 2025. "Revisiting Algorithmic Audits of TikTok: Poor Reproducibility and Short-term Validity of Findings." In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 3357–66. SIGIR '25. Padua, Italy: Association for Computing Machinery, July 13, 2025. ISBN: 9798400715921. https://doi.org/10.1145/3726302.3730293.

Mrug, Sylvie, Anjana Madan, and Michael Windle. 2016. "Emotional Desensitization to Violence Contributes to Adolescents' Violent Behavior." *Journal of Abnormal Child Psychology* 44, no. 1 (January): 75–86. https://doi.org/10.1007/s10802-015-9986-x.

Ofcom. 2024. "Protecting Children from Harms Online—A Summary of Our Consultation," May 8, 2024. Accessed March 19, 2025. https://www.ofcom.org.uk/online-safety/protecting-children/protecting-children-from-harms-online.

Papadamou, Kostantinos, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. "Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children." In *Proceedings of the International AAAI Conference on Web and Social Media,* 14:522–33. June 13, 2020. https://doi.org/10.1609/icwsm.v14i1.7320.

Regehr, Kaitlyn, Caitlin Shaughnessy, Minzhu Zhao, and Nicola Shaughnessy. 2024. *Safer Scrolling: How Algorithms Popularise and Gamify Online Hate and Misogyny for Young People.* https : //www.ascl.org.uk/ASCL/media/ASCL/Help % 20and % 20advice/Inclusion/Safer-scrolling.pdf.

Rexhepi, Rrita. 2023. "Content Moderation: How the EU and the US Approach Striking a Balance Between Protecting Free Speech and Protecting Public Interest." *Trento Student Law Review* 5, no. 1 (August 5, 2023). https://doi.org/10.15168/tslr.v5i1.2550.

Seering, Joseph, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. "Moderator Engagement and Community Development in the Age of Algorithms." *New Media & Society* 21, no. 7 (January 11, 2019): 1417–43. https://doi.org/10.1177/1461444818821316.

Smith, Kimberly E., Rosa Acevedo-Duran, Jennifer L. Lovell, Aliyah V. Castillo, and Valeria Cardenas Pacheco. 2024. "Youth Are the Experts! Youth Participatory Action Research to Address the Adolescent Mental Health Crisis." *Healthcare* 12, no. 5 (March 5, 2024): 592. https://doi.org/10.3390/healthcare12050592.

Staksrud, Elisabeth, Kjartan Ólafsson, and Sonia Livingstone. 2013. "Does the Use of Social Networking Sites Increase Children's Risk of Harm?" *Computers in Human Behavior* 29, no. 1 (January): 40–50. https://doi.org/10.1016/j.chb.2012.05.026.

Statista. n.d. "Number of Social Media Users Worldwide from 2017 to 2028." Accessed March 19, 2025. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

Swart, Joëlle. 2021. "Tactics of News Literacy: How Young People Access, Evaluate, and Engage with News on Social Media." *New Media & Society* 25, no. 3 (May 2, 2021): 505–21. https://doi.org/10.1177/14614448211011447.

Thomas, Elise, and Kata Balint. 2022. *Algorithms as a Weapon Against Women: How YouTube Lures Boys and Young Men into the 'Manosphere'.* https://au.reset.tech/uploads/algorithms-as-a-weapon-against-women-reset-australia.pdf.

TikTok. 2024. "Youth Sexual and Physical Abuse." Accessed March 19, 2025. https://www.tiktok.com/community-guidelines/en/safety-civility#4.

———. n.d. "Age-Restricted Content on TikTok LIVE." Accessed March 19, 2025. https://support.tiktok.com/en/safety-hc/account-and-user-safety/age-restricted-content-on-tiktok-live.

———. n.d. "Community Guidelines Overview." Accessed March 19, 2025. https://www.tiktok.com/community-guidelines/en/overview.

———. n.d. "Restricted Mode on TikTok." Accessed March 19, 2025. https://support.tiktok.com/en/safety-hc/account-and-user-safety/restricted-mode.

———. n.d. "TikTok Content Moderation." Accessed March 19, 2025. https://www.tiktok.com/transparency/en/content-moderation/.

U.S. Congress. 2024. *H.R.7891 - 118th Congress (2023-2024): Kids Online Safety Act.* Congress.gov, September 18, 2024. https://www.congress.gov/bill/118th-congress/house-bill/7891.

Unicef. n.d. *What is Harmful Content?* Accessed March 19, 2025. https://www.unicef.org.au/parent-teacher-resources/online-safety/harmful-content?srsltid=AfmBOoop2OVU6TzfZV4FVbLSqd7L8fSHxWhmffxSAQbG6EyulGBCM2rc.

Violot, Caroline, Tuğrulcan Elmas, Igor Bilogrevic, and Mathias Humbert. 2024. "Shorts vs. Regular Videos on YouTube: A Comparative Analysis of User Engagement and Content Creation Trends." In *Proceedings of the 16th ACM Web Science Conference,* 213–23. New York, NY, USA: Association for Computing Machinery, May 21, 2024. ISBN: 9798400703348. https://doi.org/10.1145/3614419.3644023.

WeProtect Global Alliance. 2023. *Global Threat Assessment 2023.* https://www.weprotect.org/wp-content/uploads/Global-Threat-Assessment-2023-English.pdf.

Williams, Dylan, Rys Farthing, and Alex McIntosh. 2021. *Surveilling Young People Online: An Investigation into TikTok's Data Processing Practices.* https://au.reset.tech/uploads/resettechaustralia_policymemo_tiktok_final_online.pdf.

X. n.d. "Apply Age Restrictions in Twitter." Accessed March 19, 2025. https://help.x.com/en/rules-and-policies/age-assurance.

Xiao, Madelyne, Andrew Sellars, and Sarah Scheffler. 2025. "When Anti-Fraud Laws Become a Barrier to Computer Science Research." In *Proceedings of the 2025 Symposium on Computer Science and Law,* 1–16. CSLAW '25. Munich, Germany: Association for Computing Machinery, March 25, 2025. ISBN: 9798400714214. https://doi.org/10.1145/3709025.3712206.

YouTube. 2019. "An Update on Our Efforts to Protect Minors in Youtube," June 3, 2019. https://blog.youtube/news-and-events/an-update-on-our-efforts-to-protect/?.

———. 2020. "YouTube Community Standards," September 22, 2020. Accessed March 19, 2025. https : //www.youtube.com/intl/ALL_ie/howyoutubeworks/policies/community-guidelines/#community-guidelines.

———. n.d. "Child Safety Policy." Accessed March 19, 2025. https://www.youtube.com/intl/ALL_ie/howyoutubeworks/our-commitments/managing-harmful-content/#remove.

———. n.d. "Community Guidelines." Accessed March 19, 2025. https://www.youtube.com/howyoutubeworks/policies/community-guidelines/.

———. n.d. "Using Technology to More Consistently Apply Age Restrictions." Accessed March 19, 2025. https://blog.youtube/news-and-events/using-technology-more-consistently-apply-age-restrictions/.

Authors

Fatmaelzahraa Eltaher is an Assistant Lecturer at the School of Computer Science, Technological University Dublin, Ireland. She has contributed to fields like machine/deep learning, computer vision, and natural language processing (NLP), specifically in projects that address real-world challenges. Email: fatmaelzahraa.eltaher@tudublin.ie

Rahul Krishna Gajula is a Research Assistant at the School of Computer Science, Technological University Dublin, Ireland.

Luis Miralles-Pechuán is a Lecturer at the School of Computer Science, Technological University Dublin, Ireland. He has worked extensively in applying machine learning to many domains such as COVID-19, accessibility, or AI-plagiarism detection.

Patrick Crotty is a Medical Student at the School of Medicine, Trinity College Dublin, Ireland.

Juan Martínez-Otero is a Lecturer at the School of Law, University of Valencia, Spain, specializing in the protection of children and youth in audiovisual media and the internet.

Christina Thorpe is Head of Cybersecurity at the School of Informatics and Cybersecurity, Technological University Dublin, Ireland. Her research focuses on child safety and wider cybersecurity challenges.

Susan McKeever is a Senior Lecturer at the School of Computer Science, Technological University Dublin, Ireland, specializing in applied artificial intelligence including the use of AI to protect children.

Acknowledgements

We would like to acknowledge the support of Safe Online (https://safeonline.global/who-we-are/) and the associated Tech Coalition Research Fund for their support of the N-Light project, of which this study is a part, and for their review comments on the work.

Data availability statement

Available upon request.

Funding statement

This paper is part of the N-Light project funded by the Safe Online Initiative of End Violence and the Tech Coalition through the Tech Coalition Safe Online Research Fund

(Grant number: 21-EVAC-0008-Technological University Dublin).

Ethical standards

The work of the N-Light project has been approved by the Technological University Dublin Research Ethics committee.

Keywords

Social media; content moderation; online harm; algorithmic transparency; child safety; age-restricted content; platform policies.