Rogers, Ellie

Article

The need for greater transparency in the moderation of borderline terrorist and violent extremist content

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Rogers, Ellie (2025) : The need for greater transparency in the moderation of borderline terrorist and violent extremist content, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 14, Iss. 3, pp. 1-27, https://doi.org/10.14763/2025.3.2012

This Version is available at:
https://hdl.handle.net/10419/324160

INTERNET POLICY REVIEW

Volume 14 **Issue** 3

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

The need for greater transparency in the moderation of borderline terrorist and violent extremist content

Ellie Rogers *Swansea University*

DOI: https://doi.org/10.14763/2025.3.2012

Published: 3 July 2025
Received: 16 June 2024 **Accepted:** 20 February 2025

Funding: The author did not receive any funding for this research.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Rogers, E. (2025). The need for greater transparency in the moderation of borderline terrorist and violent extremist content. *Internet Policy Review*, *14*(3). https://doi.org/10.14763/2025.3.2012

Keywords: Transparency, Content moderation, Borderline content, Extremism, Social media

Abstract: Content moderation is becoming an increasingly prominent feature of legislation to increase the safety of online spaces. One aspect of this debate is moderating borderline content in the context of terrorism and violent extremism (borderline TVEC). For content moderation approaches to be proportionate in respecting users' rights whilst improving the safety of online spaces, transparency is crucial. This importance is recognised within recent legislation such as the Digital Services Act and the Online Safety Act. However, legislation does not provide direct requirements for transparency surrounding the moderation of borderline TVEC. As a result, there are concerns that transparency reporting will continue to focus on removals of content that is illegal or violative in nature. This article argues that there needs to be more transparency surrounding the moderation of borderline TVEC. Through a review of the literature, this article discusses the importance of increased transparency surrounding the moderation of borderline TVEC, and demonstrates the ways in which current legislation, tech company policies and content moderation processes are not conforming to transparent practices in the context of borderline TVEC moderation.

Introduction

Governments, academics and civil society have increasingly highlighted the need for tech companies to minimise online harms on their platforms (Ganesh, 2023). Most major tech companies are actively removing illegal content such as terrorist and violent extremist content (TVEC), as the e-Commerce directive makes them responsible for the content on their platform (Belova-Dalton, 2023; Rexhepi, 2023). Legislation including the Terrorist Content Online (TCO) Regulation, the Digital Services Act (DSA) and the Online Safety Act (OSA) also require in-scope platforms to remove illegal TVEC (Online Safety Act, 2023; EU Regulation 2021/784, EU Regulation 2022/2065)..

As well as removing illegal content, there have been increasing calls for platforms to address "borderline content" (UK Department for Digital, Culture, Media and Sport, 2021). Borderline content, also known as legal but harmful, lawful but awful, borderline violative and borderline illegal content, has become an established term used by tech companies, policymakers and academics, but there are definitional challenges. By nature, "borderline" is an intermediate position, where something is not fully classified as one thing or the other (Heldt, 2020; Liu, 2024). As such, the term "borderline content" is largely regarded as vague and definitionally complex, with no agreement on what it describes (Macdonald & Vaughan, 2023). This article focuses on borderline content in the context of TVE (borderline TVEC).

Borderline TVEC is typically protected from removal by freedom of expression rights, but there have been calls for platforms to reduce the prevalence of this content, due to its potential to cause harm (EU Counter-Terrorism Coordinator, 2020; Heldt, 2020; Mohan, 2022). Borderline TVEC can spread misinformation and harm and may increase the likelihood of individuals seeking out TVEC (EU Counter-Terrorism Coordinator, 2020). In certain contexts, borderline TVEC can be algorithmically amplified on platforms, increasing its visibility to users (Whittaker, 2022; Yesilada & Lewandowsky, 2022). Consequently, platforms claim to restrict users' access to borderline TVEC through reduction measures of downranking, demonetisation, warning labels and age restrictions (The YouTube Team, 2019; Díaz & Hecht-Felella, 2021).

An important part of content moderation is platforms providing meaningful transparency surrounding their moderation efforts. Transparency describes making resources of public or private powers visible and accessible, and the importance of this is recognised in recent legislation including the DSA and the OSA (Fisher, 2010). Current transparency reporting by tech companies is largely done on a vol-

untary basis and can take many forms, including documentation that offers explanations on moderation decisions and platform operations, data on user engagement and moderation metrics and research and audit findings on platforms, content and moderation approaches (Stray et al., 2022). Tech companies who produce transparency reports typically release them annually, focusing on content removals resulting from policy violations, so do not provide a full picture of all moderation efforts and often lack appropriate detail (Harling et al., 2023; Leerssen, 2023). This lack of transparency is seen particularly for reduction interventions that are increasingly being utilised to address borderline content (Access Now et al., n.d.).

This article argues that there needs to be greater transparency surrounding borderline TVEC moderation. To begin, section 1 explores the concepts of borderline content and borderline TVEC. Section 2 examines the importance of transparency. Section 3 outlines the transparency requirements within the DSA and OSA and highlights the gaps surrounding borderline TVEC moderation. Section 4 discusses where greater transparency is needed in the moderation of borderline TVEC. Finally, section 5 considers alternatives to transparency legislation.

Borderline content is increasingly becoming a policy and academic focus. Existing research and policy tend to discuss borderline content as a whole concept, rather than looking at a specific type, such as borderline TVEC (Macdonald & Vaughan, 2023). Content moderation and transparency are increasingly becoming connected and required within policy. As such, tech companies need to consider these requirements outside of the typical focus on violative and illegal content and focus on them in the context of borderline TVEC, which poses a significant challenge in online spaces. This article aims to address these gaps by exploring the importance of greater transparency in the moderation of borderline TVEC and the existing gaps where platforms can achieve more meaningful transparency surrounding the moderation of this content.

Section 1: Borderline content and borderline TVEC

In 2017, Google announced a stricter approach to content that does not clearly violate policies, but contains harmful themes (Walker, 2017). As part of this, YouTube announced their borderline content policy in January 2019, highlighting that they categorise harms on a spectrum (Gillespie, 2022). YouTube defines borderline content as “Content that comes close to — but doesn’t quite cross the line of — violating our Community Guidelines” (The YouTube Team, 2019). Aside from stating that borderline content is non-violative, YouTube’s definition provides limited information on identifying features of borderline content (Macdonald & Vaughan, 2023).

Facebook announced a borderline content policy in 2018 for non-violative content that is misleading, or harmful (Gillespie, 2022). Meta define borderline content as "content that are not prohibited by our Community Standards but that come close to the lines drawn by those policies" (Meta, 2025a). Meta have also released a list of examples that this "borderline" label may apply to, including: borderline adult nudity and sexual activity; borderline violence and graphic content; borderline bullying and harassment, hate speech and violence content; sensationalised and misleading vaccine information; and content posted to groups that is likely selling or offering services that are prohibited by Meta's Regulated Goods Community Standards (Meta, 2025a).

Academics and policymakers often describe borderline content as content that is usually protected by free speech parameters in a democratic environment, but that is inappropriate in public forums (Saltman & Hunt, 2023; Heldt, 2020). The UK Government (2020) highlighted that borderline content includes problematic misinformation, sexually suggestive content, graphic or gory imagery, hate content, online abuse, content promoting self-harm and eating disorders, misleading information about vaccines and content that risks delegitimising an upcoming election. These examples highlight that the term "borderline content" may be used to describe a variety of content that borders on multiple different policies, rather than one specific type of content (Macdonald & Vaughan, 2023).

This article is seeking a more focused discussion of borderline content in the context of TVE (borderline TVEC). Research suggests that users are posting coded harmful content that has less obvious links to TVE ideologies, to avoid violating platforms' policies (Ware, 2023; Ebner, 2023). As such, there are claims that extremist rhetoric is becoming more mainstream, in part, due to the presence of borderline TVEC on platforms (Ebner, 2023). For example, humour and satire can hide the radical position within content and allow it to be shared in more mainstream domains (Schwarzenegger & Wagner, 2018). These tools have been used as part of the normalisation of anti-immigration and racist rhetoric online, highlighting the importance of discussing borderline TVEC (Schwarzenegger & Wagner, 2018).

A GIFCT Working Group report defines borderline TVEC as "content that comes close to violating policies around terrorism and violent extremism and that shares some characteristics of hateful or harmful content" (Thorley et al., 2022). This content may be actioned under TVEC policies, or other policies that aim to mitigate related harms including hate speech, or misinformation (Saltman & Hunt, 2023). TVEC policies include content that praises, promotes, or aids TVE organisations or individuals, or that glorifies, depicts, facilitates, instructs on, or directs violent acts

(Ahmed & George, 2016; Holbrook, 2015; Google, 2025b; Meta, 2025b). Hate speech policies refer to content that threatens, dehumanises, attacks, degrades, or promotes violence or hatred against individuals or groups based on their protected characteristics such as age, ethnicity, race, disability, immigration status, nationality, religion, sex, gender, or sexual orientation (Google, 2025a; Meta, 2025c; TikTok, 2024). Misinformation policies cover content that is misleading or deceptive (YouTube, n.d.).

Platforms have different tolerance levels for content, which change over time as their policies develop (de Keulenaar et al., 2023). Content that one platform may classify as violative, another may classify as borderline, and content that may have previously been classified as borderline by a platform, may now be classified as violative and vice versa. These changes in tolerance are largely associated with changes in political contexts and real-world events (de Keulenaar et al., 2023). Tech companies often use state's terrorist designation lists to shape their TVE moderation policies, which are political in nature and often focus on foreign over domestic groups (Borelli, 2021; Ganesh, 2023). Consequently, there can be inconsistent moderation of content associated with TVE (Díaz & Hecht-Felella, 2021; Macdonald & Vaughan, 2023; Ganesh, 2023). The Christchurch attack highlighted these inconsistencies, as extreme-right content is often not classified as TVEC (Borelli, 2021). Whereas, marginalised communities reclaiming slurs used against them may be labelled as violative (Díaz & Hecht-Felella, 2021; Gorwa et al., 2020). Therefore, the content that is considered as TVEC, or borderline TVEC is highly context dependent, politicised and will vary between and within platforms.

Drawing on the existing literature, this article adopts the following working description for borderline TVEC: content that can contain harmful language or themes that are associated with extremist ideologies, but does not meet platforms' or legal thresholds for removal. Borderline TVEC may not violate platform policies due to an absence of incitements for violence, a clear association to a designated group, or explicitly hateful language towards an individual or group. However, it can still contain harmful rhetoric associated with extremist ideologies such as: anti-immigrant rhetoric, anti-LGBTQ+ narratives, harmful racist stereotypes, anti-Semitic rhetoric, or misogynistic language that is expressed in an implicit or discrete manner.

Section 2: The importance of transparency

Content moderation is largely considered to be an opaque process (Gorwa et al., 2020; de Keulenaar et al., 2023). In the late 2000s and early 2010s, companies

started to improve transparency practices, due to increasing pressure from various stakeholders (Albu & Flyverbom, 2016). In 2008, as part of efforts to combat censorship and protect human rights online, the Global Network Initiative (GNI) was created with Microsoft, Yahoo, Google and a number of civil society organisations (Gorwa & Ash, 2020). As part of a commitment to the GNI principles, Google introduced an annual transparency report in 2010, making it the first company to release data on content takedowns and government information requests (Gorwa & Ash, 2020). Other platforms then followed, with Twitter releasing their first transparency report in 2012 and Facebook releasing aggregate content moderation statistics in 2013 (Kessel, 2015; Gorwa & Ash, 2020).

Since then, companies have begun to provide more detail about their content moderation policies (Gorwa & Ash, 2020). In 2018, Google published their first community guidelines enforcement report, which provided statistics on the amount of removed violating material and the role of automated systems in detecting content (Gorwa & Ash, 2020). By 2020, Facebook was providing data on the number of identified community standards violations, the percentage of flagged content that was actioned, the amount of violating content found and flagged by automated systems and the speed at which the company's moderation infrastructure acted (Gorwa et al., 2020).

Despite some improvements to transparency reporting in recent years, the information provided by companies still lacks important detail. Transparency reports largely provide aggregate statistics on the quantity of removals, rather than information regarding the prevalence of various types of content, reduction efforts, or sufficient information to assess the effectiveness of, or biases within flagging systems (Díaz & Hecht-Felella, 2021; Suzor et al., 2019). Aggregate statistics can provide an overview of platform moderation, but are not always useful in reducing the overall opacity of the system, as key information remains hidden (Suzor et al., 2019; Gorwa & Ash, 2020; Ganesh, 2023). Often, platforms will identify that content has been removed for violating a general policy like hate speech without providing information on what part of the policy was violated (Díaz & Hecht-Felella, 2021). Even less transparency is provided surrounding reduction measures, making it impossible to fully assess the scope or effectiveness of these enforcement actions (Díaz & Hecht-Felella, 2021).

Whilst it is important that tech companies do not provide too much information to allow malign actors to exploit platforms and post harmful content in more discrete ways to avoid detection, greater transparency, particularly surrounding borderline TVEC, is needed. Greater transparency is important to: allow users to make in-

formed decisions about their online activity; address mistrust between users and platforms; ensure accountability of tech companies; and allow for research of platforms.

2.1 Supporting informed decisions

Respecting user agency and autonomy online is vital, which includes allowing users to make informed decisions about their online activity and content (Macdonald & Vaughan, 2023). For users to make fully informed decisions about their online activity, transparency surrounding platform operations and algorithms is necessary (Cobbe & Singh, 2019; Jhaver et al., 2023; Llansó et al., 2020; Gorwa & Ash, 2020). This information allows users to decide whether they want to use platforms, and to gain a deeper insight into the inner operations of algorithms and how they shape users' online spaces (Jhaver et al., 2023; Leerssen, 2023; Gorwa & Ash., 2020). For example, Facebook offers a tool that explains to users why they are seeing a certain post by giving an insight into how the recommendation system works (Stray et al., 2022).

Once users are aware of why they are seeing certain content, it is crucial that platforms are transparent about how users can control their online spaces, where these controls can be accessed and changed and what this will mean for shaping users' feeds. Transparency in this area can be beneficial for increasing users' understanding of platforms, their operational systems and providing users with the tools and information necessary to have more control over the type of content they see (Jhaver et al., 2023). To aid users' ability to control their online space, it is vital that platforms clarify definitions on harm categories to demonstrate what content users are opting out of viewing (Jhaver et al., 2023). For instance, providing clarity on borderline content definitions and examples, so that users understand what platforms consider to be harmful and why.

Educating users and allowing them to make more informed decisions about their online activity can help to reduce instances of accidental policy violations. Myers West (2018) surveyed users that had experiences of their content being moderated and found that the users asked for more transparency regarding what policy a removed piece of content violates, to help them to avoid inadvertently violating platforms' policies in the future. A lack of transparency surrounding content moderation processes makes it difficult for users to understand the rules, learn from these experiences and make informed decisions when using social media platforms (Suzor et al., 2019).

2.2 Addressing user mistrust

A lack of transparency surrounding content moderation approaches can lead to mistrust, backlash and uncertainties surrounding platforms' policies and operations (Leerssen, 2023; Llansó et al., 2020). Users frequently express confusion about how violations of rules are detected and enforced, and the role of algorithms, other users, law enforcement agencies and internal decision makers in content moderation processes, as they have insufficient information to understand moderation decisions (Suzor et al., 2019). Survey responses of users who had been adversely affected by content moderation suggested that they displayed confusion about the exact content that triggered a sanction from the platform and only half of the users expressed confidence in their understanding of the platforms' moderation action (Suzor et al., 2019).

User confusion can result in mistrust towards platforms and concerns about conspiracies and system biases, as users develop their own explanations about why their content was actioned, in the absence of explanations from platforms (Suzor et al., 2019). One example of the uncertainties a lack of transparency creates is shadowbanning, where users believe that they can still post content but their content is no longer visible to other users (Savolainen, 2022; Radsch, 2021; Leerssen, 2023). Whilst there may be some overlap between definitions and the terms may be used interchangeably, shadowbanning can differ slightly from downranking (Leerssen, 2023). According to original definitions, downranked content is typically still available to others, it is just made less visible, whereas shadowbanned content is no longer visible to other users at all (Radsch, 2021). Twitter (X) and Instagram's automated moderation systems have been accused of shadowbanning content that expresses political views, despite this content being permitted under their terms of service (Stoycheff, 2023). Shadowbanning has been denied by platforms as a moderation method, but due to a lack of transparency on what moderation processes are taking place, users have expressed suspicions of being shadowbanned (Savolainen, 2022).

One consequence of user mistrust in platforms is chilling effects, particularly when users feel that errors have occurred, or that they are being disproportionately targeted (Stoycheff, 2023). A chilling effect describes when a constitutionally protected activity such as speech, is discouraged or deterred due to a fear of sanctions (Schauer, 1978). In the context of online moderation, the hidden nature of moderation efforts can result in users attempting to avoid these sanctions by not posting content online and being less likely to post their political attitudes, share creative expression, disclose their religious identity, and deviate from the social norm

(Stoycheff, 2023). Chilling effects are thought to disproportionately impact marginalised communities, as they are at greater risk for adverse content moderation decisions, due to biases within algorithmic moderation (Díaz & Hecht-Felella, 2021). For example, there have been incidents of the over-removal of non-harmful Arabic language content during attempts to reduce content associated with TVEC (Díaz & Hecht-Felella, 2021). Consequently, users may avoid posting Arabic-language content due to fears of disproportionate removal and a lack of transparency on how this is monitored and rectified by platforms.

To reduce user mistrust in platforms, increased transparency on how content is identified, classified and actioned is required, for users to understand why a decision has been reached (EU Counter-Terrorism Coordinator, 2020; Suzor et al., 2019). Clarity on moderation approaches and content classification is particularly necessary for reduction measures which are harder for users to detect. It is important for platforms to notify users when reduction measures are used and to disclose information on reductions within transparency reports. Part of this should include clearer information on how platforms are defining harmful content such as borderline TVEC, what parameters are used to classify this content and how more harmful and less harmful content are classified in comparison, so users are clear what content is being actioned and how it is identified (Rexhepi, 2023).

Increased transparency on accessing and navigating appeals processes is also necessary to improve users' trust in platforms and reduce confusion. During the study by Myers West (2018), one user shared that they had received a notification stating that their content had been removed and that they had seven days to appeal the decision, but there was no information on how the appeals process works or how to access it (Myers West, 2018). Instances such as these are likely to contribute to user mistrust and frustration with platforms. Providing information on how appeals processes can be accessed and what this process entails can mitigate user frustrations and ensure due process by allowing users to appeal moderation decisions.

2.3 Ensuring accountability

Accountability is essential for ensuring that platforms are taking responsibility for their content moderation decisions and any biases within these decisions, or the moderation systems that are being used (Barberá, 2020). Transparency has become one of the main accountability mechanisms that platforms are using to regain users' trust (Gorwa & Ash, 2020). Greater transparency does not always mean that accountability is achieved, as it is often more dependent on the quality of information that is provided (Suzor et al., 2019). However, increased transparency can al-

low platforms to build accountability into their systems (Ananny & Crawford, 2018). Platforms taking accountability is important to further mitigate user mistrust towards platforms, by providing them with confidence that platforms will rectify biased or disproportionate moderation decisions and that moderation is consistent across the platform (Lévesque, 2021; Ganesh, 2023; Suzor et al., 2019; Macdonald & Vaughan, 2023).

2.4 Allowing for research

Increased transparency allows independent stakeholders such as researchers and journalists to have access to data to study and audit social media platforms (European Commission, 2024; Meßmer & Degeling, 2023). In addition to internal audits conducted by platforms, external research is necessary, but this requires access to the platform and certain data. Platforms' limited transparency reports currently restrict empirical research on the moderation of harmful content (Ganesh, 2023). Transparency surrounding access to data ensures that there is a fair vetting process to determine who has access to this data, how to gain access to this data and what data can be accessed.

Independent researchers having this access to platforms and data is crucial to allow for studies on the long-term impacts of moderation efforts and the prevalence and amplification of harmful content (Gillespie, 2022b; Meßmer & Degeling, 2023; Ganesh, 2023). These findings can aid platforms in monitoring and understanding evolving risks online and to assess how content moderation can be improved to be more effective, proportionate and minimise harm (Douek, 2021). Audits can assess whether the information that platforms release about their systems and content moderation processes are upheld in-practice and align with external research findings, to hold platforms accountable and strengthen the trust between platforms and users (Stray et al., 2022).

Section 3: Transparency legislation and the gaps surrounding borderline TVEC moderation

Given the importance of transparency, it has become an increasing requirement within legislation that mandates the moderation of harmful content. In 2016, the EU agreed a countering illegal hate speech Code of conduct with Facebook, Microsoft, Twitter and YouTube, which required companies to review removal notifications of illegal hate speech within 24 hours, work with trusted flaggers and strengthen platform cooperation on countering hate speech (Ganesh, 2023; European Commission, 2016). Since then, additional platforms have signed and the re-

vised Code was integrated into the DSA framework in January 2025 (European Commission, 2025).

In 2021, the European Parliament approved the TCO regulation (EU Regulation 2021/784) that came into effect in June 2022. This regulation requires platforms to remove the most harmful and illegal terrorist content within one hour of receipt of a removal order and includes transparency requirements. Given its focus on illegal terrorist content that meets the threshold for removal, this legislation is out of scope for this article, which focuses on borderline TVEC, content which is not illegal, nor meets the threshold for removal.

Building on the TCO, the DSA (EU Regulation 2022/2065) was approved by the European Parliament in October 2022 and came into effect in February 2024. The DSA covers content that poses systemic risks to the Union and includes transparency requirements for platforms. As such, the DSA is in-scope for this article, as borderline TVEC likely meets platform criteria of posing systemic risks.

The OSA received Royal Assent in October 2023 and is expected to be fully enforced by 2026. The OSA focuses on illegal content, protecting young people from legal but harmful content and outlines transparency requirements for platforms. Therefore, the OSA is also in-scope for this discussion on borderline TVEC transparency, as platforms are required to identify and reduce the visibility of this content for young users.

This section will outline the key transparency requirements within the DSA and OSA and highlight where requirements for borderline TVEC transparency remain unclear.

3.1 The DSA

The DSA applies to platforms that offer services to citizens of EU member countries (Meßmer & Degeling, 2023; Rexhepi, 2023). Providers of very large online platforms (VLOPs) are required to assess systemic risks to the Union that arise from the design or function of their service and its systems (EU Regulation 2022/2065). Systemic risks include: the dissemination of illegal content; and any actual or foreseeable negative effects for the exercise of fundamental rights, on civic discourse and electoral processes and in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being (EU Regulation 2022/2065). In 2025, the revised EU Code of conduct on countering illegal hate speech was integrated into the DSA framework (European Commission, 2025).

The DSA also outlines transparency requirements, to address distrust between platforms and users and address harms associated with algorithms (European Commission, 2024; Leerssen, 2023; Rexhepi, 2023; Pirang, 2020). Article 14 requires platforms to disclose the restrictions that they impose on content within their terms and conditions, including information on moderation policies, procedures and tools such as algorithmic and human decision making and their internal complaint handling procedures (EU Regulation 2022/2065). Article 15 requires platforms to release annual reports outlining any content moderation that took place, categorised by: the type of illegal content or violation of the service's terms and conditions; the detection method; and the type of restriction applied (Rexhepi, 2023; EU Regulation 2022/2065). Article 17 states that platforms must provide a statement of reasons to the user for each content moderation decision (i.e., restrictions to the visibility of content through removal, demotion, disabling access, suspension or demonetisation, the violation the approach is responding to and the information used to make the moderation decision), to allow users to challenge decisions through internal appeals and external dispute resolution (European Commission, 2024; Gorwa et al., 2020; Leerssen, 2023; Pirang, 2020; EU Regulation 2022/2065). Platforms are required to provide transparency on the operation of algorithms and how this impacts the content that is surfaced to users as part of Article 27 (Belova-Dalton, 2023; European Commission, 2024; Meßmer & Degeling, 2023).

The DSA also outlines provisions in Article 40 to allow researchers more access to data on social media platforms and requires platforms to allow for independent audits on content moderation and algorithmic systems (European Commission, 2024; Stray et al., 2022; Ganesh, 2023). These provisions allow vetted researchers access to VLOP data for the purpose of research that contributes to the detection, identification and understanding of harmful content and to the assessment of the effectiveness and impacts of content moderation efforts (Regulation 2022/ 2065).

3.2 The OSA

The OSA applies to user-to-user and search services that have a significant number of UK users or that can be used in the UK. The OSA focuses on the moderation of illegal content such as TVEC and protecting young people from legal but harmful content including: pornographic content; content that promotes, instructs on or encourages suicide, self-injury, eating disorders, or acts of violence against others; content that is abusive or incites hatred against individuals protected characteristics (i.e., race, religion, sexual orientation, disability, sex, or gender reassignment); and bullying content (Online Safety Act, 2023).

The OSA also outlines mandatory transparency reporting standards in Schedule 8, that the regulating body Ofcom will issue in the form of annual transparency notices (Harling et al., 2023; Online Safety Act 2023). These notices will outline the information that platforms must include in their transparency reports (Online Safety Act, 2023). Transparency within the OSA is focused on empowering users to control their online spaces by making more informed decisions about how they spend their time online and adjusting requirements for transparency reporting (Ofcom, 2023, 2024). There are three dimensions of transparency that these notices are proposed to focus on (Ofcom, 2023).

The first dimension requires more transparency on the actions that services take to protect users such as platform safety features and decision-making processes (Ofcom, 2023). This dimension may include a requirement for platforms to provide information on the incidence, dissemination and number of users who encountered illegal content and content that is harmful to children and how they enforce community guidelines and content moderation tools (Harling et al., 2023; Online Safety Act, 2023). Platforms may be required to make information on removed or restricted content and suspended and banned users clear and accessible and justify why a specific action was used (Online Safety Act 2023). Transparency reports may have to include information on features that allow users to report illegal content and content harmful to children and that help users manage risks on platforms. Platforms may also be required to provide accessible and transparent complaints and appeals procedures for users to question moderation decisions (Online Safety Act, 2023).

The second dimension surrounds Ofcom's transparency as a regulator, including greater clarity on how changes are implemented across in-scope services and the impact of regulation (Ofcom, 2023). This requirement includes assisting platforms in complying with transparency requirements and publishing guidance for platforms (Online Safety Act 2023). Ofcom will be required to produce annual transparency reports based on the information provided by platforms' transparency reports (Online Safety Act 2023).

The third dimension involves greater transparency to allow for open engagement with other sectors including law enforcement, civil society and academia (Ofcom, 2023). This final requirement aims to ensure regulation is informed by experts and to hear from users, to assist service providers' understanding of the effectiveness of their moderation efforts and determine where improvements can be made (Ofcom, 2023, 2024).

3.3 The gaps surrounding borderline TVEC moderation

Whilst these regulations require transparency in areas that could include the moderation of borderline TVEC, such as further transparency on restriction measures, there remains a lack of guidance on how platforms can ensure they are applying transparency requirements to the moderation of non-violative content such as borderline TVEC.

In the DSA, the phrasing of the requirements places more emphasis on transparency surrounding the moderation of illegal and violative content. The DSA requires that "the information reported shall be categorised by the type of illegal content or violation of the terms and conditions of the service provider, by the detection method and by the type of restriction applied" (Regulation 2022/2065). Requiring platforms to categorise transparency reporting by the "type of illegal content or violation" and how this content was detected and actioned, may result in platforms not disclosing any additional categories of content moderation including restrictions imposed on non-violative content such as borderline TVEC and how this content is detected and actioned.

The transparency requirements in the OSA are not yet fully enforced. Schedule 8 of the OSA outlines "matters about which information may be required" (Online Safety Act, 2023), suggesting that it is not yet determined what information from this list platforms will be required to disclose within transparency reports. This uncertainty leaves the potential that transparency surrounding borderline TVEC, which is only actioned for children under the OSA, may not be a requirement for platforms. If the requirements to provide transparency surrounding content that is harmful to children are included within Ofcom's transparency notices, there may still be information gaps about borderline TVEC moderation. For example, there may be cases where platforms are restricting borderline TVEC for adults as well as children, but they may not include this information within transparency reports, as the proposed OSA transparency requirements only cover disclosing information about the moderation of borderline content in the context of children.

The DSA and OSA both include some scope for borderline TVEC and transparency reporting about this content, but without explicit requirements for platforms to provide information on borderline TVEC moderation, it is likely that tech companies will continue to focus their transparency reporting on content that violates their terms of service (Macdonald & Vaughan, 2023). As such, borderline TVEC remains largely unaddressed within transparency requirements and tech companies often fail to moderate borderline TVEC with sufficient transparency as a result.

Section 4: A lack of transparency in the moderation of borderline TVEC

Addressing borderline TVEC begins with platforms identifying and classifying content. Classification tools often involve a combination of automated and manual methods such as machine learning and human moderators (Goodrow, 2021; Gorwa et al., 2020; Lévesque, 2021). Human moderators are necessary as context is important for identifying and categorising this content, but these individuals must have appropriate training on the subject area, given the vague and subjective definitions provided for borderline content (Goodrow, 2021; Thorley et al., 2022; van der Vegt et al., 2019).

Once content has been identified and classified as borderline TVEC, platforms have disclosed that they utilise reduction measures to address it, which aim to limit the visibility of this content and reduce its searchability (The YouTube Team, 2019; Gillespie, 2022a; Ganesh, 2023). Reduction measures include: removing it from algorithms, or downranking it, to reduce the distribution of this content; demonetising the content, to prevent the user making money from the content; and providing restrictions or warnings such as fact-checking labels, age restrictions, geoblocking and temporary holds (Gillespie, 2022a; Heldt, 2020; Llansó et al., 2020; Pirang, 2020; Díaz & Hecht-Felella, 2021; Ganesh, 2023). Platforms such as YouTube, Facebook, Twitter (X), LinkedIn, TikTok and Instagram have announced that they utilise reduction efforts for borderline content (Gillespie, 2022a; The YouTube Team, 2019).

Platforms may also allow users to control the content they view on platforms themselves by utilising platforms' in-built features (Jhaver et al., 2023). Simple user-led moderation can include users blocking certain accounts to avoid seeing content posted by them and users following certain accounts to view their content within their newsfeed or timeline (Gillett et al., 2022; Jhaver et al., 2023). Engaging with content can also allow users to shape their online spaces. Positively engaging with content through views, likes and comments is likely to result in users seeing more of this content whereas, ignoring, reporting and negatively engaging with content through dislikes is likely to result in users seeing less of this content (Jhaver et al., 2023).

There are also more complex moderation tools that individuals can use to select their preferences of content (Jhaver et al., 2023). Users can choose to change the content that is algorithmically amplified to them by muting keywords and setting up sensitivity controls (Cobbe & Singh, 2019; Jhaver et al., 2023; Llansó et al.,

2020; Panday, 2021; Sander, 2020). There are also some platform specific preference and personalisation controls. On YouTube, users can pause, edit, or delete search and watch history, to change personalisation recommendations (Goodrow, 2021) and Instagram's sensitive content control measure allows users to choose how much sensitive content is filtered out of their recommendations on the explore page (Meta, 2021).

A vital part of content moderation is transparency from platforms on their policies and terms of service, their content moderation processes and actions and oversight and appeal mechanisms to address moderation errors and biases. This section will highlight where platforms do not currently moderate borderline TVEC with sufficient transparency, as they focus on removals within transparency reports, leave gaps surrounding the moderation processes that they use to address borderline TVEC and provide limited information surrounding appeals and oversight measures.

4.1 Transparency reports focus on removals

The majority of platforms who release transparency reports, focus on removals (Díaz & Hecht-Felella, 2021). Google's Transparency Report 2024 on YouTube's Community Guidelines enforcement includes information on video, channel and comment removals, the policy that the content or account violated, the method by which the removed content was detected, the view count of the removed videos, appeal and reinstatement figures, the country that removed videos were posted from and human flagger figures. Meta's Community Standards Enforcement Report 2024 includes information on the prevalence of violative content, content actioned by removal, labelling or disabling accounts, appealed and restored content, and proactive flagging. TikTok's Community Guidelines Enforcement Report 2024 discloses information on content, comment and account removals, account and livestream suspensions, restored content, the use of automation, view counts of removed content, and policy reasons for removal. Twitter (X)'s 2025 Transparency Report includes information on account suspensions, posts removed or labelled, the policies that were violated by content and numbers of posts that were moderated by humans and automated systems.

Despite this focus on removals and violative content within transparency reports, platforms have acknowledged that they reduce borderline content. In 2019, YouTube launched changes to reduce recommendations of borderline content and harmful misinformation (The YouTube Team, 2019). Meta states that they demote borderline content to reduce its distribution in users feeds (Meta, 2023). TikTok

claims to remove borderline content from recommendations (TikTok, n.d.). Twitter (X) disclose that they limit the visibility of posts through downranking, removing the post from algorithms and restricting engagement opportunities with posts (X, n.d.). In terms of transparency surrounding these reduction efforts, platforms issue vague statements about the effectiveness of their reduction measures, without providing the information to allow for independent verification (Díaz & Hecht-Felella, 2021). For example, YouTube has claimed that there was a 70 percent decrease in the watch time of borderline content coming from non-subscribed recommendations in the US, as a result of reduction efforts (The YouTube Team, 2019).

The focus on removals within existing transparency reporting leaves gaps surrounding the reduction measures that platforms have disclosed that they are using. No information is provided on how much content is demonetised, downranked, or otherwise restricted and the reasons why this content is reduced are not provided (Díaz & Hecht-Felella, 2021). Some reduction measures are more visible to users than others, as users are likely aware that their content has been demonetised, but may never be aware that their content has been downranked (Díaz & Hecht-Felella, 2021). As such, the absence of information on these reduction measures within transparency reports has left a lot of uncertainty surrounding the moderation of borderline TVEC, as the scope and effectiveness of these enforcement actions cannot be assessed (Díaz & Hecht-Felella, 2021).

4.2 Uncertainty surrounding borderline TVEC moderation

Given the focus on the moderation processes involved in the detection and removal of violative content within transparency reports, there remains large gaps when it comes to the processes that are involved in reduction efforts that platforms use to moderate borderline TVEC.

Platforms should deal with borderline TVEC on an individual basis, as a one-size-fits all approach risks not respecting freedom of expression rights due to the subjective nature of this content (Rexhepi, 2023). It is unclear to what extent an individualised approach is utilised to moderate borderline TVEC due to a lack of transparency from platforms. YouTube disclosed that they use machine learning to detect and limit the spread of borderline content (The YouTube Team, 2019). The nature and operational features of this system are unknown, making it unclear whether YouTube deals with borderline content on a case-by-case basis, or a one-size-fits all strategy. Meta disclosed that the majority of their reduction procedures are applied equally to each piece of content, but in certain situations, they cannot

adopt a one-size-fits-all approach (Meta, 2023). Aside from Meta highlighting that in certain regions and during critical events, their enforcement processes need to be adjusted, it is uncertain in what exact situations and for what exact pieces of content they adopt a case-by-case approach compared to a one-size-fits-all approach.

Since borderline TVEC must be identified by platforms before it can be actioned, the processes that platforms utilise to identify and classify content are important for users to be aware of (Buntain et al., 2021; Gillespie, 2022a). Despite there being a range of definitions for borderline content provided across sectors there remains uncertainty about the term. Borderline content is described as coming close to violating platform policies, but it is uncertain how this closeness is measured, as tech platforms do not provide measurable standards to judge content against (Macdonald & Vaughan, 2023). Meta provides a list of examples of content that they categorise as borderline (see Section 1), with some explanations as to what this may look like (Meta, 2025a). For example, borderline bullying and harassment, hate speech and violence and incitement content may include content that dehumanises individuals or groups who are not defined by their protected characteristics (Meta, 2025a). This explanation implies that the content is considered to be non-violative because it does not target individuals' protected characteristics, as if it did, it would be classified as hate speech (see Section 1). However, this is not explicitly stated by Meta, creating uncertainty surrounding the identification and classification of this content and it is uncertain what other platforms consider to be borderline content.

It is often unclear how borderline TVEC is moderated by platforms, but there is increasing use of automated methods for content moderation in general, as it is a more time efficient way of dealing with large quantities of content, whilst reducing the workload for human moderators (Rexhepi, 2023; Macdonald & Vaughan, 2023). YouTube claims that a machine learning model is used to detect and reduce the visibility of borderline content (The YouTube Team, 2019). However, it is unclear how much human evaluation (if any) is used in addition to the automated system, how this system operates to detect and moderate borderline TVEC and what systems other platforms are using (Gorwa et al., 2020; Myers West, 2018).

Where automated methods have been used, there are criticisms that the systems contain biases and result in higher moderation errors compared to human moderation, as they cannot always accurately consider context (Rexhepi, 2023). Existing transparency reports do not include adequate information to assess who is most affected by content moderation errors and whether some communities are dispro-

portionately targeted, but research suggests that users from marginalised communities may be at greater risk for content moderation errors and system biases (Díaz & Hecht-Felella, 2021). One study found that automated models for detecting hate speech were 1.5 times more likely to flag tweets written by self-identified African American users as offensive (Sap et al., 2019). Current transparency reporting fails to provide the necessary information to evaluate how automated systems function and whether these tools are making mistakes or contain biases when assessing borderline TVEC (Díaz & Hecht-Felella, 2021).

4.3 A lack of transparency surrounding appeals and oversight mechanisms

Transparency surrounding moderation decisions and appeals is crucial to ensure accountability, due process and clarity for users to understand how their content is being moderated and what oversight measures are in place (Ganesh, 2023). More recently, certain platforms have provided greater accessibility to appeals processes, for users who feel their content has been wrongly actioned or require further explanation (Díaz & Hecht-Felella, 2021). On Meta platforms, users can appeal moderation decisions and submit their case to Meta's Oversight Board, where it will be reviewed by an external body (Meta, n.d.-c). On YouTube, users can submit appeals, which go to human reviewers (Google Transparency Report, 2024). TikTok claims to notify users when their content has been removed, where users will be given the opportunity to appeal the decision (TikTok, n.d.).

However, appeals processes remain inconsistent, focused on removed content and hidden from public view. There are inconsistent notifications when a users' content has been actioned and a lack of transparency on what appeals are available for users and how they navigate this process (Díaz & Hecht-Felella, 2021). Notifying users about their actioned content is important for reduction measures that are used to moderate borderline TVEC, as it is less obvious when these have occurred (Díaz & Hecht-Felella, 2021). Despite this, if platforms do notify users when their content has been actioned, it is typically for content removals, leaving users unable to appeal restrictions to their content if they are unaware they have happened (Díaz & Hecht-Felella, 2021).

The lack of appeals processes for reduced content can have a disproportionate effect on marginalised communities who may be subject to increased moderation errors, as they will not have meaningful access to appeal such decisions (Díaz & Hecht-Felella, 2021). As a consequence, chilling effects may be exacerbated, as users are not able to question content moderation decisions they perceive to be

unfair, or have the option to be provided with additional explanations. As such, they may limit, or fully stop their online posts, to avoid this frustration.

Chilling effects can be harmful by reducing users' freedom to express their beliefs online. According to UK and EU legislation (i.e., Article 10 of the UKs Human Rights Act 1998, Article 11 of the EU Charter of Fundamental Rights and Article 10 of the European Convention on Human Rights), individuals have the right to hold their own opinions and express these freely, without interference by public authority (provided these opinions are not expressed in a violent or hateful way). When content is wrongly actioned by platforms and users are not able to appeal these decisions, their freedom of expression may be limited for the piece of content that is wrongly restricted, but also for future posts, as they may limit or stop posting content, to reduce frustrations from moderation errors and a lack of access to appeals.

Section 5: Alternatives to legislation

Given the lack of clear legislative guidance on how platforms can achieve meaningful transparency in the context of borderline TVEC moderation, there is non-legislative guidance which may be beneficial for tech companies to adhere to. One example is the Santa Clara Principles, which are a collaborative civil society effort that aim to set transparency standards and promote freedom of expression and user agency (Access Now et al., n.d.; Suzor et al., 2019). The second iteration of the Santa Clara Principles detail how platforms can achieve meaningful transparency and accountability (Access Now et al., n.d.).

The Foundational Principles outline how companies can integrate human rights and due process into all moderation processes. The Principles highlight how tech companies can ensure their policies can be understood by users, how to ensure content moderation policies, tools and actions are considerate of different cultures and contexts, the importance of informing users of state involvement in content moderation decisions and how to ensure users are confident that their content is handled with respect to human rights (Access Now et al., n.d.).

The Operational Principles set out important elements for companies to include within transparency reports. The Principles highlight the importance of utilising statistics to reflect all content moderation decisions made in quarterly reports, to provide users and researchers with a better understanding of the systems in place (Access Now et al., n.d.). For example, the policies that content was actioned under, numbers related to content removals and restrictions, the number of appeals and

how many of these were successful or unsuccessful, the amount of state requests for content and how and when automated processes are used (Access Now et al., n.d.). The Principles also outline how companies should provide notice to a user whose content has been actioned (Access Now et al., n.d.). For example, the specific policies that were violated, how the content was detected and actioned, how users can access support channels, and information on how users can appeal the decision (Access Now et al., n.d.). The final Operational Principle outlines how companies should set out appeals processes for users and why these are important to have in place (Access Now et al., n.d.). For example, companies should ensure there are independent human reviewers dealing with appeals, the process is easily understandable for users and users are informed of the appeal outcome (Access Now et al., n.d.).

Conclusion

This article demonstrates that there is a need for greater transparency surrounding the moderation of borderline TVEC. The DSA and OSA require platforms to address borderline TVEC and they outline transparency requirements. However, there is a lack of explicit requirements for platforms to provide transparency for non-violative content such as borderline TVEC within legislation. As a result, transparency reports focus on removals of illegal and violative content, leaving transparency surrounding borderline TVEC moderation limited to statements from platforms claiming that they reduce this content (Macdonald & Vaughan, 2023).

There remains uncertainty surrounding the content moderation processes that are used to address borderline TVEC, the effectiveness of these systems and whether they contain biases that result in moderation errors. This uncertainty can result in chilling effects, restrictions to users freedom of expression and the disproportionate moderation of certain content, all of which are exacerbated by the opaque nature of appeals and oversight mechanisms.

Addressing borderline TVEC is becoming a large part of content moderation debates, due to its potential to spread harm. It is vital that this moderation goes hand-in-hand with increased transparency from platforms to improve definitional clarity surrounding borderline content and how this content is moderated by platforms, respect user rights and agency, hold platforms accountable for moderation decisions, increase users understanding and trust in platforms' moderation efforts and allow for research into the effectiveness and limitations of these moderation efforts.

References

Access Now et al. (n.d.). *Santa Clara Principles on transparency and accountability in content moderation*. Santa Clara Principles. https://santaclaraprinciples.org

Ahmed, M., & George, F. L. (2016). *A war of keywords: How extremists are exploiting the Internet and what to do about it*. Tony Blair Institute for Global Change. https://www.institute.global/insights/geopolitics-and-security/war-keywords-how-extremists-are-exploiting-internet-and-what-do-about-it

Albu, O. B., & Flyverbom, M. (2019). Organizational transparency: Conceptualizations, conditions, and consequences. *Business & Society*, *58*(2), 268–297. https://doi.org/10.1177/0007650316659851

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Are, C. (2022). The shadowban cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, *22*(8), 2002–2019. https://doi.org/10.1080/14680777.2021.1928259

Barberá, P. (2020). Social media, echo chambers, and political polarization. In J. A. Tucker & N. Persily (Eds.), *Social media and democracy* (pp. 34–55). Cambridge Univerity Press. https://www.cambridge.org/core/books/social-media-and-democracy/social-media-echo-chambers-and-political-polarization/333A5B4DE1B67EFF7876261118CCFE19

Belova-Dalton, O. (2023). *The impact of new media and the internet on terrorism*. https://www.sisekaitse.ee/et/impact-new-media-and-internet-terrorism

Borelli, M. (2023). Social media corporations as actors of counter-terrorism. *New Media & Society*, *25*(11), 2877–2897. https://doi.org/10.1177/14614448211035121

Buntain, C., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). YouTube recommendations and effects on sharing across online social platforms. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–26. https://doi.org/10.1145/3449085

Cobbe, J., & Singh, J. (2019). Regulating recommending: Motivations, considerations, and principles. *European Journal of Law and Technology*, *10*(3). https://doi.org/10.2139/ssrn.3371830

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

Díaz, Á., & Hecht-Felella, L. (2021). *Double standards in social media content moderation*. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation

Douek, E. (2020). Governing online speech: From 'posts-as-trumps' to proportionality and probability. *Columbia Law Review*, *121*(3), 759–834. https://doi.org/10.2139/ssrn.3679607

Ebner, J. (2023). *From margins to mainstream: How extremism has conquered the political middle*. International Center for Counter-Terrorism. https://icct.nl/publication/margins-mainstream-how-extremism-has-conquered-political-middle

EU Counter-Terrorism Coordinator. (2020). *The role of algorithmic amplification in promoting violent*

and extremist content and its dissemination on platforms and social media (No. 12735/20). https://www.tandis.odihr.pl/retrieve/1143044

European Commission. (2016). *The EU Code of conduct on countering illegal hate speech online*. European Commission. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

European Commission. (2024). *Questions and answers on the Digital Services Act*. European Commission. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348

European Commission. (2025). *The code of conduct on countering illegal hate speech online +*. https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online

European Court of Human Rights & Council of Europe. (n.d.). *European convention on human rights*. https://www.echr.coe.int/documents/d/echr/convention_ENG

European Union. (2015). *EU Charter of Fundamental Rights*. FRA European Union Agency for Fundamental Rights. https://fra.europa.eu/en/eu-charter/article/11-freedom-expression-and-information

European Union. (2022). *Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online*. https://eur-lex.europa.eu/eli/reg/2021/784/oj/eng

European Union. (2024). *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*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065

Fisher, E. (2010). Transparency and administrative law: A critical evaluation. *Current Legal Problems*, *63*(1). https://doi.org/10.1093/clp/63.1.272

Ganesh, B. (2023). Content moderation: Social media and countering online radicalisation. In *The Routledge handbook on radicalisation and countering radicalisation*. https://www.taylorfrancis.com/reader/read-online/16814d6f-6e87-4869-9ca0-34587f607bc0/chapter/pdf?context=ubx

Gillespie, T. (2022a). Do not recommend? Reduction as a form of content moderation. *Social Media + Society*, *8*(3), 20563051221117552. https://doi.org/10.1177/20563051221117552

Gillespie, T. (2022b). Reduction / borderline content / shadowbanning. *Yale Journal of Law & Technology*, *24*, 476–492.

Gillett, R., Stardust, Z., & Burgess, J. (2022). Safety for whom? Investigating how platforms frame and perform safety and harm interventions. *Social Media + Society*, *8*(4), 20563051221144315. https://doi.org/10.1177/20563051221144315

Global Internet Forum to Counter Terrorism. (2024). *Membership*. GIFCT. https://gifct.org/membership/

Goodrow, C. (2021, September 15). *On YouTube's recommendation system*. YouTube Official Blog. https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/

Google. (2025a). *Hate speech policy*. YouTube Help. https://support.google.com/youtube/answer/2801939?hl=en-GB

Google. (2025b). *Violent extremist or criminal organisations policy*. YouTube Help. https://support.goo

gle.com/youtube/answer/9229472?hl=en-GB

Google Inc. (2024). *YouTube community guidelines enforcement – Google transparency report*. https://transparencyreport.google.com/youtube-policy/removals?hl=en

Gorwa, R., & Ash, T. G. (2020). Democratic transparency in the platform society. In *Social media and democracy: The state of the field and prospects for reform* (pp. 286–312). Cambridge University Press. https://books.google.com/books?hl=en&lr=&id=NEjzDwAAQBAJ&oi=fnd&pg=PA286&dq=info:vF5-2oCD240J:scholar.google.com&ots=5k2LduMi8L&sig=Thir8nlDuTULBUJ-_usqd8Fqj78

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945

Harling, A.-S., Henesy, D., & Simmance, E. (2023). Transparency reporting: The UK regulatory perspective. *Journal of Online Trust and Safety*, *1*(5). https://doi.org/10.54501/jots.v1i5.108

Heldt, A. (2020). Borderline speech: Caught in a free speech limbo? *Internet Policy Review*. https://policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510

Holbrook, D. (2015). Designing and applying an 'extremist media index'. *Perspectives on Terrorism*, *9*(5), 57–68.

Jhaver, S., Zhang, A. Q., Chen, Q., Natarajan, N., Wang, R., & Zhang, A. (2023). *Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor*. arXiv. https://doi.org/10.48550/ARXIV.2305.10374

Kessel, J. (2015, February 9). *Three years of increased #transparency... And counting*. X Blog. https://blog.x.com/en_us/a/2015/three-years-of-increased-transparency-and-counting

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

Lévesque, M. (2021). In the shadows of content moderation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3789311

Liu, D. (2024). Borderline content and platformised speech governance: Mapping TikTok's moderation controversies in South and Southeast Asia. *Policy & Internet*, *16*(3), 543–566. https://doi.org/10.1002/poi3.388

Llansó, E., Hoboken, J., Leerssen, P., & Harambam, J. (2020). Artificial Intelligence, content moderation, and freedom of expression (Transatlantic working group on content moderation online and freedom of expression). *Transatlantic Working Group*. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

Macdonald, S., & Vaughan, K. (2024). Moderating borderline content while respecting fundamental values. *Policy & Internet*, *16*(2), 347–361. https://doi.org/10.1002/poi3.376

Meßmer, A.-K., & Degeling, M. (2023). *Auditing recommender systems: Putting the DSA into practice with a risk-scenario-based approach*. Stiftung Neue Verantwortung. https://arxiv.org/ftp/arxiv/papers/2302/2302.04556.pdf

Meta. (n.d.-a). *Detecting violations*. Meta: Transparency Centre. https://transparency.meta.com/en-gb/enforcement/detecting-violations/

Meta. (n.d.-b). *Taking action*. Meta: Transparency Centre. https://transparency.meta.com/en-gb/enforcement/taking-action/

Meta. (n.d.-c). *The oversight board*. Meta: Transparency Centre. https://transparency.meta.com/en-gb/oversight/

Meta. (2021). Introducing sensitive content control. In *Meta: Transparency Centre*. https://about.fb.com/news/2021/07/introducing-sensitive-content-control/

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

Meta. (2024). *Community standards enforcement*. Meta: Transparency Centre. https://transparency.meta.com/reports/community-standards-enforcement/

Meta. (2025a). *Content borderline to the community standards*. Meta: Transparency Centre. https://transparency.meta.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards

Meta. (2025b). *Dangerous organisations and individuals*. Meta: Transparency Centre. https://transparency.meta.com/en-gb/policies/community-standards/dangerous-individuals-organizations/

Meta. (2025c). *Hateful conduct*. Meta: Transparency Centre. https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/

Mohan, N. (2022, February 17). *Inside responsibility: What's next on our misinfo efforts*. YouTube Official Blog. https://blog.youtube/inside-youtube/inside-responsibility-whats-next-on-our-misinfo-efforts/

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Ofcom. (2023). *Ofcom's approach to implementing the Online Safety Act*. https://www.ofcom.org.uk/__data/assets/pdf_file/0017/270215/10-23-approach-os-implementation.pdf

Ofcom. (2024). *Online safety transparency consultation: Consultation on transparency guidance*. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/consultation-draft-transparency-reporting-guidance/main-docs/consultation-on-transparency-guidance.pdf?v=371129

Panday, J. (2021, September 14). *CCP continues to seal off its digital economy*. Internet Governance Project. https://www.internetgovernance.org/2021/09/14/__trashed-2/

Pirang, A. (2020, April 24). *User attention is the new frontier in content regulation – Improving the accountability of soft content moderation by social media platforms*. https://www.juwiss.de/64-2020/

Radsch, C. C. (2021). Shadowban/shadow banning. In *Glossary of platform law and policy terms*. https://repositorio.fgv.br/server/api/core/bitstreams/1064b8ef-7d47-44be-9413-174681575745/content#page=296

Rexhepi, R. (2023). Content moderation: How the EU and the U.S. approach striking a balance between protecting free Speech and protecting public interest. *Trento Student Law Review*, *5*(1), 67–98.

Saltman, E., & Hunt, M. (2023). *Borderline content: Understanding the gray zone* (p. 16). Global

Internet Forum to Counter Terrorism. https://gifct.org/wp-content/uploads/2023/06/GIFCT-23WG-Borderline-1.1.pdf

Sander, B. (2020). Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation. *Fordham International Law Journal*, *43*(4), 939.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. https://doi.org/10.18653/v1/P19-1163

Savolainen, L. (2022). The shadow banning controversy: Perceived governance and algorithmic folklore. *Media, Culture & Society*, *44*(6), 1091–1109. https://doi.org/10.1177/01634437221077174

Schauer, F. (1978). Fear, risk and the First Amendment: Unraveling the chilling effect. *Boston University Law Review*, *58*, 685–732.

Schwarzenegger, C., & Wagner, A. (2018). Can it be hate if it is fun? Discursive ensembles of hatred and laughter in extreme right satire on Facebook. *Studies in Communication | Media*, *7*(4), 473–498. https://doi.org/10.5771/2192-4007-2018-4-473

Stoycheff, E. (2023). Cookies and content moderation: Affective chilling effects of internet surveillance and censorship. *Journal of Information Technology & Politics*, *20*(2), 113–124. https://doi.org/10.1080/19331681.2022.2063215

Stray, J., Thorburn, L., & Bengani, P. (2022, August 8). *A menu of recommender transparency pptions*. Tech Policy Press. https://techpolicy.press/a-menu-of-recommender-transparency-options

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*(0), 0.

The YouTube Team. (2019, December 3). *The four rs of responsibility, part 2: Raising authoritative content and reducing borderline content and harmful misinformation*. YouTube Official Blog. https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/

Thorburn, L. (2022, November 23). *What does it mean to give someone what they want? The nature of preferences in recommender systems*. Medium. https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157

Thorley, T., Llansó, E., & Meserole, C. (2022). *Methodologies to evaluate content sharing algorithms & processes* (GIFCT Working Groups Output 2022, pp. 8–40). Global Internet Forum to Counter Terrorism. https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TA-Evaluate-1.1.pdf

TikTok. (n.d.). *Content violations and bans*. Tiktok Support. https://support.tiktok.com/en/safety-hc/account-and-user-safety/content-violations-and-bans

TikTok. (2023, January 8). *Strengthening enforcement of sexually suggestive content*. TikTok Newsroom. https://newsroom.tiktok.com

TikTok. (2024a). *Community guidelines enforcement report*. https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2023-4/

TikTok. (2024b). *Countering hate speech & behavior*. TikTok Safety Centre. https://www.tiktok.com/safety/en/countering-hate

UK Department for Digital, Culture, Media and Sport. (2021). *Online safety bill*. https://assets.publishing.service.gov.uk/media/609a6f8b8fa8f56a3f720b63/Draft_Online_Safety_Bill_Bookmarked.pdf

UK Government. (2020). *Online harms white paper: Full government response to the consultation*. https://assets.publishing.service.gov.uk/media/5fd8af718fa8f54d5f67a81e/Online_Harms_White_Paper_Full_Government_Response_to_the_consultation_CP_354_CCS001_CCS1220695430-001__V2.pdf

UK Government. (2023). *Online Safety Act § 2023 chapter 50*. UK Legislation. https://www.legislation.gov.uk/ukpga/2023/50/section/77

Van der Vegt, I., Gill, P., Macdonald, S., & Kleinberg, B. (2019). Shedding light on terrorist and extremist content removal. *Global Research Network on Terrorism and Technology*, *3*. https://static.rusi.org/20190703_grntt_paper_3.pdf

Walker, K. (2017, June 18). *Four steps we're taking today to fight terrorism online*. Google Blog. https://blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/

Ware, J. (2023). *The third generation on online radicalization*. Program on Extremism at George Washington. https://extremism.gwu.edu/sites/g/files/zaxdzs5746/files/2023-06/third-generation-final.pdf

Whittaker, J. (2022). *Recommendation algorithms and extremist content: A review of empirical evidence* (GIFCT Working Groups Output 2022, pp. 8–34). Global Internet Forum to Counter Terrorism. https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TR-Empirical-1.1.pdf

X. (n.d.). *Our range of enforcement options for violations*. X Help Center. https://help.x.com/en/rules-and-policies/enforcement-options

X. (2025). *2025 Transparency Report*. X Transparency. https://transparency.x.com/en/reports/global-reports/2025-transparency-report

Yesilada, M., & Lewandowsky, S. (2022). A systematic review: YouTube recommendations and problematic content. *Internet Policy Review*, *11*(1). https://doi.org/10.14763/2022.1.1652

YouTube. (n.d.). *YouTube misinformation policies – how YouTube works*. YouTube. https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/

Published by

Alexander von Humboldt Institute for Internet and Society

CAIS Research for the Digital Age

in cooperation with

CREATe

cnrs centre internet et societe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies