Received: 7 September 2023 | Accepted: 11 October 2023

DOI: 10.1002/poi3.376

RESEARCH ARTICLE

P&I Policy & Internet PSO WILEY

Moderating borderline content while respecting fundamental values

Stuart Macdonald | **Katy Vaughan**

School of Law, Swansea University, Swansea, UK

Correspondence

Katy Vaughan, School of Law, Swansea University, Swansea, UK.
Email: k.e.vaughan@swansea.ac.uk

Abstract

As efforts to identify and remove online terrorist and violent extremist content have intensified, concern has also grown about so-called lawful but awful content. Various options have been touted for reducing the visibility of this borderline content, including removing it from search and recommendation algorithms, downranking it and redirecting those who search for it. This article contributes to this discussion by considering the moderation of such content, in terms of three sets of values. First, definitional clarity. This is necessary to provide users with fair warning of what content is liable to moderation and to place limits on the discretion of content moderators. Yet, at present, definitions of borderline content are vague and imprecise. Second, necessity and proportionality. While downranking and removal from search and recommender algorithms should be distinguished from deplatforming, tech companies' efforts to deamplify borderline content give rise to many of the same concerns as content removal and account shutdowns. Third, transparency. While a number of platforms now publish their content moderation policies and transparency data reports, these largely focus on violative, not borderline content. Moreover, there remain questions around access to data for independent researchers and transparency at the level of the individual user.

KEYWORDS

algorithms, borderline, content, content moderation, freedom of expression, recommendation, terrorist and violent extremist content, transparency

INTRODUCTION

Since the so-called 'Golden Age' of Islamic State on Twitter (Conway et al., 2018), much progress has been made in the identification and removal of online terrorist and violent extremist content (TVEC) - such that today the major platforms state that over 95% of TVEC is proactively removed by the platforms themselves before being flagged by Internet Referral Units or users. At the same time, concern has grown about 'borderline' content that falls just short of violating platforms' Terms of Service, and so is not liable to be removed, but nonetheless has the propensity to cause harm (Conway et al., 2021). This concern has been exacerbated by instances in which such content has been amplified, either through users resharing it in large numbers or the platforms' recommendation systems promoting it (or both together). Were it not for such amplification, much borderline content would only ever reach a small audience, or even none at all.

The visibility of borderline content may be reduced using a variety of methods, which we refer to as deamplificatory measures. These include removing the content from search and recommendation algorithms, downranking it, restricting users' ability to share it, redirecting those who search for it, nudging users towards authoritative information, geoblocking content to particular regions, and demonetising it. One of the key arguments advanced in support of deamplificatory measures is that they are more protective of the right to freedom of expression than content takedowns since they do not involve outright removal of the content from the platform. Nonetheless, deamplificatory measures still raise important human rights issues and legislators have moved to impose regulatory requirements, such as the UK's Online Safety Bill and the EU's Digital Services Act.

Like the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (Kaye, 2019), our premise in this article is that—while tech companies do not have the obligations of governments—their function and impact means that they should respect human rights standards. Indeed, there is a growing consensus that tech companies' policies on content moderation should be grounded in international human rights standards (BSR, 2021; Kaye, 2018). A human rights-based approach to content moderation offers an 'organising framework' to identify and assess the impact of moderation policies and develop a more structured, principled approach (Sander, 2020, p. 966). Aligning policies with international human rights standards also puts tech companies in a stronger position to respond to any State restrictions that threaten censorship of speech (Dias Oliva, 2020; Kaye, 2018). Moreover, one of the criteria for membership of the Global Internet Forum to Counter Terrorism (GIFCT[1]) is a public commitment to human rights, in accordance with the United Nations Guiding Principles on Business and Human Rights (UNGP).

The UNGP sets out a framework for State obligations and corporate responsibilities in respect of business-related human rights abuses (OHCHR, 2011). These principles establish a corporate responsibility for businesses to 'respect human rights' (Principle 11), to prevent or mitigate 'adverse human rights impacts' (Principle 13), and to have in place 'processes to enable the remediation of any adverse human rights impacts they cause or to which they contribute' (Principle 15). The corporate responsibility to respect extends to 'internationally recognised human rights' (Principle 12) and exists 'independently of States' abilities and/or willingness to fulfil their own human rights obligations' (OHCHR, 2011, p. 13). While the principles are nonbinding, they establish a 'global standard of expected conduct for all business enterprises wherever they operate' (OHCHR, 2011, p. 13). The argument for their adoption and implementation is particularly strong in the case of tech companies, given their 'overwhelming role in public life globally' (Kaye, 2018, p. 5).

The article accordingly employs an analytical framework that is based on how the right to freedom of expression is defined in international human rights treaties. It begins with a

description of this framework and methodology. The article then discusses in turn three sets of issues: definitional clarity; necessity and proportionality; and transparency. It concludes with some suggestions for how to improve the compliance of efforts to moderate borderline content with human rights standards.

Methodology and analytical framework

The article adopts a socio-legal approach. In broad terms, four types of materials were analysed. First, legislation. The focus here was on transnational legislation on international human rights standards and online regulation, though some reference was also made to relevant domestic legislation in both the UK and the US. To assist in the interpretation of these legislative texts, relevant jurisprudence was also considered. Second, to understand the wider context and operation of these legislative instruments, we examined official reports, including from United Nations committees and rapporteurs and UK parliamentary reports. This was supplemented with oral and written evidence submitted by interested organisations (in particular, human rights groups) as part of the legislative processes. Third, we examined the Terms of Service and transparency reports of relevant tech companies. Here, we focussed on companies that both: (1) employ algorithms to amplify/deamplify content (which excluded most smaller platforms); and, (2) make public their relevant policies and publish transparency reports. We also examined the transparency reports published by the GIFCT. Fourth, we examined relevant secondary literature, from a range of disciplines. Our analytical framework was derived from our reading of international human rights legislation, with the other materials organised in accordance with this framework.

While the moderation of online content can impact a number of human rights, including the rights to privacy and nondiscrimination, the analytical framework employed in this article focuses on the right to freedom of expression. This is a fundamental human right that is enshrined in international and regional treaties, including the International Covenant on Civil and Political Rights and the European Convention on Human Rights, as well as the constitutions of many liberal democratic nations, such as the First Amendment to the US Constitution. International human rights treaties stipulate that, while restrictions on the right to freedom of expression are permissible, such restrictions must be prescribed by law, pursue a legitimate objective and meet the demands of necessity and proportionality.

For a restriction on the right to freedom of expression to be prescribed by law, it must comply with the principle of legality. Often referred to as the rule of law, the principle of legality stipulates that laws should be defined with a sufficient degree of clarity.[2] In terms of content moderation, it follows that a tech company's polices (as set out in its Terms of Service or community standards/guidelines) should be drafted with sufficient precision to 'enable users to predict with reasonable certainty what content places them on the wrong side of the line' (Kaye, 2018, p. 15). Such definitional clarity is important for several reasons. It provides users with fair warning of what content is not permissible, enabling them to make informed decisions about their use of the platform (Kaye, 2019).[3] It limits the discretion of content moderators, ensuring greater consistency in decision-making while guarding against potential misuse of power (Howard, 2018) and censorship creep (Citron, 2018). It also helps ensure that users have an effective opportunity to appeal moderation decisions, should their content be taken down (Macdonald et al., 2019). Definitional clarity is discussed further in part 2.

A restriction on the right to freedom of expression must pursue a legitimate objective, such as the prevention of crime or the protection of national security, public health or the rights of others. Tech companies must ensure that any interference of a content moderation policy with the speech of users is based on one of these legitimate aims. Any such restriction

must also be necessary to achieve the stated objective. One ingredient of the test of necessity is an assessment of whether the restriction on the right is proportionate to the aim sought to be achieved. So, tech companies must ensure that any actions taken on content (including deamplificatory measures) are both necessary and proportionate (Kaye, 2018). Together, the requirements of a legitimate objective, necessity and proportionality work to limit the restrictions that may be imposed on the right to freedom of expression, ensuring that the right is 'practical and effective', not 'theoretical or illusory'.[4] This three-step test—which provides a 'methodology and vocabulary for platforms to analyse whether their content policies and decisions are reasonable' (Dias Oliva, 2020, 617)—is discussed further in part 3.

In part 4 we discuss transparency. Numerous reasons have been identified for the importance of transparency, including preventing corruption (Audit Commission, 2010), uncovering mistakes (Stiglitz, 2002), building trust (Mol, 2008), improving public debate (Callon et al., 2009) and enhancing democracy (Gupta, 2008). Of especial importance for present purposes is the role of transparency in enabling oversight and promoting accountability (Obama, 2009). The UNGPs require companies to 'track the effectiveness of their response' to adverse human rights impacts (Principle 20), and to be 'prepared to communicate' externally how they address human rights impacts (Principle 21). Moreover, the clarity of definitions of terms such as 'borderline content' can only be assessed if companies make their content moderation policies publicly available, and the necessity and proportionality of moderation activity can only be evaluated if the details and objectives of such efforts are disclosed. In short, a lack of transparency will impede efforts to assess whether restrictions on the right to freedom of expression comply with human rights standards.

Definitional clarity

'Borderline content' has quickly become an established term. It is used by tech companies, policymakers and researchers alike. Notwithstanding this ready acceptance and usage of the term, there is also open acknowledgement of the definitional challenges associated with it. In a series of group discussions and interviews conducted on behalf of GIFCT, for example, it was 'clear that there is no overarching agreement between different sectors or geographies' on what borderline content is (Saltman, 2022, p. 11). An attempt at definition can be found in two blogs posted by senior members of YouTube. The first blog defined borderline content as 'videos that don't quite cross the line of our policies for removal but that we don't necessarily want to recommend to people' (Mohan, 2022). The second defined it as 'content that comes close to but doesn't quite violate our Community Guidelines' (Goodrow, 2021), offering conspiracy theory videos and other content that spreads 'problematic misinformation' as examples. Aside from saying that borderline content is not violative of a platform's Terms of Service, however, these definitions shed little light on the identifying features of borderline content. What, precisely, is it about this content that means it 'comes close to' violating Terms of Service and is not necessarily something that should be recommended to others?

Another attempt at definition can be found in Meta's Transparency Centre. This describes borderline content as 'types of content that are not prohibited by our Community Standards but that come close to the lines drawn by those policies', adding that such content is 'sensationalist or provocative and can bring down the overall quality of discourse on our platform'. Illustrative examples are offered, including: 'sexually suggestive' images that focus 'on the person's bottom or cleavage and the person is wearing minimal clothing'; 'content that depicts gory or graphic imagery of some kind'; and, 'misleading or

sensationalised information about vaccines […] that would be likely to discourage vaccinations' (Meta, 2023a). Meta's announcement of the reinstatement of the Facebook account of former US President Donald Trump also described as borderline content 'that contributes to the sort of risk that materialised on January 6, such as content that delegitimizes an upcoming election or is related to QAnon' (Clegg, 2023).

This feeling that there exists some content that is neither illegal nor in breach of a platform's Terms of Service, but which is nonetheless a cause for concern, was at the heart of the proposal, in the UK's Online Safety Bill, to impose a duty on some service providers to protect adults' online safety. The duty would have applied to 'legal but harmful' content (UK Government, 2020), requiring the provider to conduct a specific risk assessment (including before making any significant changes to the service) and state in its Terms of Service how it deals with such content.[5] While not as imprecise as vague references to what we might not want to recommend to others, the notion of harm is still 'an amorphous concept' (Elsom, 2020, p. 14). The Bill stated that 'harm' encompassed both physical and psychological harm, self-harm as well as harm resulting from the words or actions of others, and potential harm as well as actual harm. The harm might arise from the nature of the content itself, the simple fact of its dissemination (e.g., the malicious sharing of personal information) or the manner in which it was disseminated (e.g., repeatedly sending content to someone). Among the misgivings expressed by critics of the proposal was concern that the definition of harm was open-textured, thereby effectively delegating to service providers the decision of what content is harmful and leaving open the possibility of inconsistent and overly broad application (Joint Committee on the Draft Online Safety Bill, 2021). When the proposed duty was subsequently removed from the Bill during its Parliamentary passage, the reasons given included these definitional difficulties—and the potential knock-on effect that imprecise definition would have on freedom of speech (Elgot, 2022).

An underlying reason for these definitional challenges is that borderline content is an umbrella term. As well as problematic misinformation, sexually suggestive content, gory or graphic imagery, misleading information about vaccines and content that risks delegitimising an upcoming election (all mentioned above), the UK Government identified 'content promoting self-harm, hate content, online abuse that does not meet the threshold of a criminal offence, and content encouraging or promoting eating disorders' as examples of legal but harmful content (UK Government, 2020, p. 32). In fact, a study of eight official UK reports on platform regulation published between 2018 and 2020 found close to 100 different online harms that were said to need addressing (Schlesinger & Kretschmer, 2020). As well as needing to encompass this broad range of content types, the UK Government also insisted that any definition would need to be 'agile and flexible', not so narrowly drawn as to be unable to 'adapt to an ever changing societal and technological landscape' (Joint Committee on the Draft Online Safety Bill, 2021, p. 53). Self-evidently, it is not possible to concoct a concise, dictionary-style definition of borderline content that is general enough to encompass this wide variety of content and provide sufficiently flexibility to be able to respond to future developments, while simultaneously offering the level of precision required to avoid delegating decision-making responsibility to service providers. If the goals of consistent implementation and fair warning are to be achieved, and the use of moderation powers against unintended types of content (mission creep) is to be avoided, alternative approaches to engendering clarity are required.

One potential approach might be to compile a list of the types of content that are grouped together under the umbrella term borderline content, with each of these content types defined individually. This list-based approach also raises questions, however. The first is who would compile the list of content types deemed to be borderline. To leave this task to service providers would be to delegate the decision what content should be regulated, which was one of the key criticisms of the UK's Online Safety Bill. An alternative approach,

suggested by the UK's Joint Committee on the Draft Online Safety Bill (2021), is to construct a statutory list based on 'specific areas of law that are recognised in the offline world, or are specifically recognised as legitimate grounds for interference in freedom of expression' (55). Among the examples it offered were disinformation that is likely to endanger public health (which may include antivaccination disinformation), content and activity that promotes eating disorders and self-harm, and disinformation that is likely to undermine the integrity and probity of electoral systems. However, this approach is also not without its difficulties. Concerns would arise as to how it would be implemented in authoritarian states, as well as the practical difficulty of divergent lists in different countries.

The second question is whether a list of content types would be exhaustive or indicative. The difficulty with an exhaustive list is that it does not offer the desired flexibility. It would exclude content types that emerge only as society and technology evolve. On the other hand, while an indicative list would be nonexhaustive—and so could in the future be deployed against unforeseen types of content—its open-ended nature gives rise to other concerns. Most obviously, it would vest significant discretion in service providers. Open-ended, indicative lists are most effective when there is a clear theme that unifies the items on the list. This unifying theme acts as a constraint, by informing decisions as to what may justifiably be deemed to fall within the concept's scope (Macdonald, 2006). As we have shown, however, discussions of the assemblage of content types that are regarded as borderline content have not identified core unifying themes with a sufficient degree of clarity. Given that there are problems with both an exhaustive and an indicative list, an alternative approach is needed that reconciles the two sets of concerns. One possibility would be to create an exhaustive list of content types, with the possibility of adding new items to the list via a process containing safeguards to ensure independent oversight and multistakeholder consultation. This would provide some degree of flexibility, while also maintaining scrutiny and accountability.

Necessity and proportionality

Since borderline content does not violate platforms' Terms of Service, it is not liable to be removed. However, there are significant concerns that such content is not merely present on these platforms but that their algorithms are distributing and amplifying it to a larger audience (Cobbe & Singh, 2019; Guy, 2018; Keller, 2021; Lewis, 2019). These concerns are illustrated by the suicide of 14-year-old Molly Russell. The coroner's report on her death found that, after developing a depressive illness, Molly had begun viewing content online on self-harm and suicide. The platform's algorithms subsequently selected and provided her with further content that 'romanticised acts of self-harm by young people on themselves'. This content portrayed 'self- harm and suicide as an inevitable consequence of a condition that could not be recovered from' and 'contributed to her death in a more than minimal way' (Walker, 2022).

There are two, related sets of issues here. The first is how algorithmic recommendation systems prioritise content for *amplification*. By determining what content is given more visibility, algorithmic recommendation systems can have a significant effect on public awareness of matters of public interest (Cobbe & Singh, 2019). Tufekci (2015) highlights the example of the Ferguson protests in the US in response to the killing of unarmed black teenager, Michael Brown, by a white police officer in August 2014. Discussion dominated Twitter in the US, which brought the story to the attention of mainstream media. At the time, Twitter still operated an algorithmically unfiltered feed. In contrast, Facebook's algorithmic news feed was dominated by the ice bucket challenge, given its popularity among users and large volume of likes and shares. According to Tufekci, the effect of Facebook's algorithms

was to suppress news of the protests. Had Twitter also employed algorithmic filtering, the 'conversation about police accountability and race relations that has since shaken the country might never have made it out of Ferguson' (213).

The second set of issues focuses on deliberately reducing the visibility of borderline content by *deamplifying* it (Gillespie, 2022). Companies have taken steps (both overtly and covertly) to stop their recommendation systems from spreading potentially harmful content. For example, Google has publicly stated that it gives harmful, untrustworthy and spammy content the lowest priority in its search algorithms system, meaning that such content appears lower in Google search results. Examples include the homepages of Stormfront and the Flat Earth Society (Google, 2022a). Meanwhile, according to the Twitter files released by Elon Musk, the platform utilised 'visibility filtering', without informing users, to limit the visibility of tweets from certain users (Nava & Golding, 2022).

A popular slogan in discussions of algorithmic recommendation systems is DiResta's (2018) statement that 'free *speech* does not mean free *reach*. There is no right to algorithmic amplification'. In a similar vein, Douek (2021) has observed, 'De-amplification does not reduce the amount of speech and does not directly impede the ability to speak. Whether speech is amplified by platforms' algorithms is a separate question from whether it can be posted in the first place' (816). At the same time, however, as the US Supreme Court has stated, the difference between banning speech and burdening speech is 'but a matter of degree'.[6] Like deplatforming, promoting some items of content, while curbing the dissemination of others, amounts to an important gatekeeping function (Llanso et al., 2020). It is therefore unsurprising that platforms' ability to influence the reach of content gives rise to similar concerns as their ability to deny access to the platform in the first place. As the example of the Ferguson protests illustrates, moderating the reach of content can have a great influence on collective awareness of politics, current affairs, and scientific consensus (Cobbe & Singh, 2019).

Some concerns stem from the lack of definitional clarity described earlier. The ambiguity of existing definitions of borderline content means that the power to deamplify content is relatively unconstrained. This presents a risk of the power being applied inconsistently, in a discriminatory manner, or even deliberately misused. There is also the potential for 'censorship creep', where powers that were designed for certain purposes or situations are used in other ways and other contexts (Citron, 2018).

While deamplification does not involve the removal of content or shutdown of accounts, its practical effect may still be to suppress speech and prevent users from 'effectively making their voices heard' (Llanso et al., 2020, p. 20). Steps taken to protect marginalised or disempowered groups can in some instances have the opposite effect (Bromwell, 2022). For example, by not considering context, algorithms have been shown to disproportionately flag online content from LGBTQ communities presenting the danger that legitimate speech from members of these communities is suppressed (Dias et al., 2021). This may be exacerbated by the enactment of legislative or regulatory regimes requiring companies to deamplify borderline content. In the same way that some companies have taken an overly cautious approach to tight deadlines for the removal of terrorist content, they might also resort to over-enforcement to ensure compliance with regulatory requirements on borderline content, deamplifying content that cannot properly be regarded as harmful (Keller, 2021).

A rights-based framework for assessing efforts to deamplify content would help to mitigate these concerns, by requiring the identification of a legitimate objective for the deamplification measures and an assessment of their necessity and proportionality. The legitimate objectives specified in Article 19(3) of the International Convention on Civil and Political Rights are respect for the rights or reputations of others and the protection of national security and public order, health and morals. The most common categories of restricted online content correspond to one or more of the Article 19(3) legitimate objectives

(Global Partners Digital, 2018). For example, restricting content supporting terrorism corresponds with the goal of protecting national security or public order. A focus on these objectives would help solicit answers to some of the questions that existing definitions of borderline content leave unanswered, such as the reasons why it would not be desirable to recommend the content to others and what is the nature and likelihood of the feared harm.

Having identified a legitimate objective, the next question would be whether the steps taken to deamplify the content are necessary to achieve this objective. This means that the aim of the measure cannot be achieved by another means (UNHRC General Comment No. 34 2011, para 33). Applying the necessity test is arguably one of the more challenging aspects of applying international human rights law standards to companies' content moderation practices. This is in part a result of a lack of clear and consistent guidance from the UN Treaty bodies; and, also due to the distinct nature of platforms in comparison to States (Sander, 2020). Nevertheless, it remains valuable for reasons that will be discussed further, and it is possible to identify principles from human rights law to assist platforms in its application in this context. In particular, the test of necessity includes an assessment of whether measures are a proportionate restriction on speech. Applying the principle of proportionality requires that restrictive measures, such as actions reducing the visibility of content must, 'be appropriate to achieve their protective function (UNHRC General Comment No. 27 1999, para 14).[7] Factors to be taken into account include whether there is a sufficient basis for believing a particular interest is in danger (McBride, 1999) and whether less intrusive means were available to achieve the same legitimate objective (Arai-Takahashi, 2002). A proportionality approach would also look at any countermeasures that have been implemented, such as the information provided to affected users and any appeals process (The Santa Clara Principles 2021).

A proportionality-based approach along these lines would have several benefits. First, it 'explicitly acknowledges interests other than the individual speech right, and thereby dignifies those interests and the importance of evaluating them in their particular context' (Douek, 2021, p. 785). Second, the least intrusive means principle requires consideration of the range of deamplificatory options. This encourages a more nuanced approach than the simple binary of remove-or-remain, with the action taken tailored to the specific context. Third, it explicitly recognises what Douek (2021) has described as the move from taxonomist to grocer: that is, a move away from an approach that is based on categorising types of content towards one that is focused on weighing competing interests. At the same time, it is important to note that an approach based on balancing competing interests leaves 'the legitimacy of rules contingent on the rule maker providing reasons that articulate the purpose of rules, the reason why they pursue legitimate aims, and what interests have been recognised and evaluated' (ibid, 821). This raises issues of transparency, to which we now turn.

Transparency

Transparency may be defined as 'the decision to make visible, or provide access to, the resources on which an exercise of public or private power may be based' (Fisher, 2010, p. 274). While it is sometimes described in terms of opening the doors or turning on the lights, these metaphors are somewhat misleading. They imply the continuation of business as usual, but with others simply being able to see what is going on. In fact, as Fisher (2010) observes, 'transparency often requires quite significant changes to how institutions operate. At the very least, documents must be found, data produced, and analysis carried out. These activities take resources in themselves and thus will have implications for organisations'

(279). She states that the creation of transparency mechanisms involves at least seven different questions:

> why something should be transparent; what should be transparent; when should it be transparent; what is the trigger for transparency; what is the institutional apparatus needed for transparency; who are the end users of transparency; and what are the consequences of transparency (Fisher, 2010, p. 273).

Here we focus on the first two of these questions: why and what? Before turning to these questions, though, it is important to emphasise that the creation and operation of transparency mechanisms require difficult decisions to be made. Transparency cannot mean the full disclosure of everything. For a start, in the case of most companies—not just those the size of Google or Meta—to release *everything* would be too much. Locating the desired information would be extremely difficult, if not impossible, and those outside the company may require some explanation or interpretation of the information in order for it to be intelligible. There may also be good reasons for limiting disclosure, including protecting the privacy of individual users and their data, maintaining the competitive advantage of companies that have invested in technological development, and not providing insights that could help malevolent actors circumvent automated content moderation tools. These are all relevant considerations in the design of transparency mechanisms.

The first question to ask is: why should the efforts of tech companies to regulate borderline content be transparent? The answer to this question is important because it informs the answers to the questions that follow: the reasons why a person performs an activity will influence how that activity is approached and performed. According to GIFCT, the value of transparency is that it promotes multistakeholder collaboration. GIFCT's, 2022 transparency report states that transparency reporting 'encourages an open and inclusive internet and multistakeholder approach' and offers 'our cross-sector community the ability to understand our approach, the efforts we take to address the current threat landscape online, and how we are advancing our mission' (GIFCT, 2022, p. 3). Similarly, the human rights review of GIFCT conducted by BSR stated that transparency has value in spreading expertise, insight and learning on how to prevent terrorist exploitation of online platforms, but also highlighted its importance in enabling enhanced accountability and correcting misunderstandings (BSR, 2021). Accountability is also identified by Meta (2023b) as the reason for its transparency reporting, while Google (2022b) says it aims to inform discussions about online content regulation and Twitter (2022a) says it wants policymakers and the general public to be better informed.

The emphasis here is on ensuring key stakeholders and the general public are well-informed, to promote cross-sector collaboration, improve policymaking and ensure accountability. While each of these objectives is, of course, important, it is also worth noting possible reasons for transparency that are not explicitly mentioned. For example, another possible motivation for transparency mechanisms is respect for the autonomy and agency of individual users, improving their ability to make informed decisions about how they use the platform and challenge moderation decisions with which they disagree. Transparency mechanisms can also improve the quality of an organisation's own decision-making, such as ensuring that content moderation decisions are accurate and consistent.

The second question to consider is, in the context of moderation of borderline content, what should be transparent? The tech sector has largely focused on two transparency mechanisms. The first is the publication of policies on moderation of (borderline) content. For example, Facebook has outlined a three-pronged approach: the removal of content that violates its Community Standards; reducing the spread of 'problematic content' that is nonviolative but nonetheless 'misleading or harmful'; and informing users by providing

additional context, so that they can decide whether to read, trust or share it (Lyons, 2018). It has also stated that groups and movements that do not meet the criteria to be designated as a dangerous organisation and banned from the platform, but which nonetheless pose a demonstrable risk to public safety, will be downranked in search results (Meta, 2020). Twitter's policy on hateful conduct includes as enforcement options the downranking of tweets in replies and making tweets ineligible for amplification in search results and on the timelines of users who don't follow the user that posted the tweet (Twitter, 2022b). Meanwhile, Reddit has a policy of quarantining subreddits, to prevent their 'content from being accidentally viewed by those who do not knowingly wish to do so' (Reddit, 2022). Quarantined subreddits display a warning requiring users to explicitly opt in to viewing the content. They do not appear in nonsubscription-based feeds and are not included in search or recommendations.

Perhaps the most detailed description is offered by YouTube. YouTube's 'Four Rs of Responsibility' include raising up 'authoritative voices' and reducing the spread of content that 'brushes right up against our policy line' (YouTube, 2019). The company reports a 70% drop-in watch time of nonsubscribed, recommended borderline content in the US since it began demoting borderline content in recommendations (Goodrow, 2021). Classifiers are used to determine whether a video is 'authoritative' or 'borderline'. These classifications are generated by human evaluators using a publicly available set of guidelines.[8] Key questions in determining authoritativeness include the topic of the video, the reputation of the speaker and the channel the video is on and whether expertise is needed to achieve the video's goal. Key questions in determining borderline status include whether the content is inaccurate, misleading, deceptive, insensitive, or intolerant and whether the video is harmful or has the potential to cause harm (Goodrow, 2021).

The second transparency mechanism is reports containing statistical data and breakdowns. This is reflected in the membership criteria of both GIFCT ('regular, public data transparency reports' are required) and Tech Against Terrorism (the prospective member must 'commit to improve transparency reporting') (GIFCT, 2022; Tech Against Terrorism, 2022). While there is variation in the content of these reports, there 'appears to be a trend toward … reporting on a company's own content moderation policies and aggregate impacts on specific types of content', as well as requests for data from governments and law enforcement (Radsch, 2022, p. 20). The EU's Digital Services Act will formalise transparency reporting obligations for all platforms (other than small or micro ones),[9] requiring annual, publicly available, easily comprehensible reports containing information on: content moderation policies and practices; any use made of automated means for the purpose of content moderation; and data on, among other things, orders received from authorities in Member States, referral notices and complaints received and responses to these.

Various concerns have been expressed about current transparency reporting practices, including: selective use of metrics; lack of contextual information to enable a full understanding of the data provided; the use of proportional metrics that fail to give an accurate indication of the scale of harm; and, the difficulty in making cross-platform comparisons when companies use different metrics (Harling et al., 2023; Joint Committee on the Draft Online Safety Bill, 2021). It is also noteworthy that companies' transparency reports focus—often exclusively—on content that violates the platform's Terms of Service. By definition, borderline content is excluded from their scope. As a result, it is unclear how much content is classified as borderline, which types of content receive this classification, and what actions are most frequently deployed against such content—a lack of clarity that is exacerbated by the imprecision of definitions of borderline content. At a qualitative level, companies' published policies offer some insight into the actions that platforms take, but questions remain, in particular in respect of the operation of algorithms that amplify or downrank content (Whittaker et al., 2021).

A further concern is the lack of access for independent researchers (Brown, 2023). Such access—which is necessary for independent evaluation and validation of internal company studies (MacCarthy, 2022), particularly given concerns about companies' ability to self-police (Tiku, 2020)—has been opposed for reasons including user privacy (Joint Committee on the Draft Online Safety Bill, 2021). The present lack of access 'hinders much-needed scientific progress towards understanding the prevalence, impact, causes, and dynamics of online activity that creates a risk of harm' (ibid, 120). The Digital Services Act will impose a requirement on providers of very large online platforms and search engines to provide access to vetted researchers at academic institutions for 'the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union'.[10] Here, the process for vetting researchers will be critical. There are difficult questions about inclusivity, diversity, and independence, on the one hand, and confidentiality and security of potentially highly sensitive data, on the other (Engler, 2021; Leerssen, 2020).

There are also strong reasons for improving transparency at the level of the individual user, not least enhanced accountability, and correcting mistakes. This applies not only to those consuming content, who it has been argued should be offered an explainable rationale for why they have been recommended content (Luria, 2022; Whittaker, 2022).[11] It also applies to those posting content that has been classified as borderline (Costanza-Chock et al., 2022). Respect for users' autonomy requires that they be told that their content has been so classified and what action has been taken to reduce the spread of their content. As noted above, since a proportionality assessment involves consideration of countermeasures an appeals process will also normally be required. As mandated by the Santa Clara principles, users should be provided with information on the appeals process, as well as sufficient information about the reasons for the decision in their specific case for them to be able to make meaningful representations.

CONCLUSION

This article has discussed the moderation of borderline content from the perspective of definitional clarity, necessity and proportionality, and transparency. It has offered suggestions for how to improve the compliance of these moderation efforts with international human rights standards. Concocting a sufficiently clear and unambiguous, yet flexible and futureproof, definition of the umbrella term borderline content is an impossible task. A more promising approach would be to create an exhaustive list of types of content regarded as borderline—with individual definitions for each of these—along with a separate procedure for adding new items to the list that engages stakeholders and is subject to independent oversight. As well as clearer definitions, the objective of any deamplificatory measures should be made clear and the necessity and proportionality of these measures assessed. This would involve scrutiny of the empirical basis for the belief that the measures are necessary to achieve the stated objective, as well as an assessment of whether other, less intrusive means could be deployed and the adequacy of any countermeasures that have been implemented, in particular, an appeals process. All of these suggestions require a commitment to transparency. While there has been progress in recent years in respect of publishing content moderation policies and regular reporting of data, much of this has focused on violative, rather than borderline, content. A commitment to accountability and correcting errors requires that independent researchers be granted access to data, while respect for the autonomy and agency of individual users entails informing those whose content has been deamplified of the reasons for this and granting them the opportunity to appeal.

These suggestions raise numerous questions about implementation. Which content types should appear on the list of borderline content and how should each of these types of content be defined? What steps can companies take to ensure restrictions on the right to freedom of expression have a legitimate objective and are necessary and proportionality? And how will consistency be ensured, within companies and across different companies and jurisdictions? How will inclusivity, diversity and independence be ensured when vetting researchers for access to data? And how can platforms of all sizes, but particularly smaller one, build the capacity to ensure greater transparency at the individual user level? The further questions raised by this article revolve around institutional responsibility. There are sound reasons why tech companies should not be left to produce definitions of borderline content, be the ultimate arbiters of necessity and proportionality, and vet researchers. Equally, there are sound reasons not to vest these tasks in national governments, given the potential for divergent approaches and abuse of such powers by authoritarian regimes. To address these institutional questions, a fruitful way forward would be to examine the remediation mechanisms of the UNGP.[12] In this sense, the UNGP not only provides the foundation for the rights-based analysis advanced in this article, but perhaps also assistance in answering some of the questions that the article has posed.

ORCID

Katy Vaughan https://orcid.org/0000-0001-5025-5747

ENDNOTES

[1] GIFCT is an NGO that exists to prevent terrorist exploitation of digital platforms. It currently has 22 members, including Facebook, YouTube, Microsoft and Twitter (GIFCT, 2022).

[2] For example, in *R v Rimmington* [2005] UKHL 63 Lord Bingham described the principle of legality as follows: 'There are two guiding principles: no one should be punished under a law unless it is sufficiently clear and certain to enable him to know what conduct is forbidden before he does it; and no one should be punished for any act which was not clearly and ascertainably punishable when the act was done'.

[3] For some, offering such clarity to malevolent actors is counterproductive. On this view, vaguer definitions have a deterrent effect: malevolent actors will be less certain what they can get away with and as a result will be more cautious. This is discussed further below in part 4 on transparency.

[4] *Airey v Ireland* (1979-80) 2 EHRR 305.

[5] The proposal applied to 'priority content' (i.e., content that had been designated by the Government as harmful to adults) and other content that presented 'a material risk of significant harm to an appreciable number of adults in the United Kingdom' (clause 55(3) of the Bill, as amended on report. Available at: https://publications.parliament.uk/pa/bills/cbill/58-03/0209/220209.pdf).

[6] *US v Playboy* 529 U.S. 803, 812 (2000).

[7] Referred to in General Comment No. 34.

[8] The guidelines total 176 pages, which raises the question whether a simplified version should be made available for the purpose of informing the public.

[9] Defined in Recommendation 2003/361/EC. A small enterprise is one that employs fewer than 50 persons and whose annual turnover and/or annual balance sheet total does not exceed €10 million. For micro enterprises the respective figures are 10 persons and €2 million.

[10] Article 40. 'Very large' is defined as more than 45 million average monthly users of the service in the EU (Article 33).

[11] Article 27 of the Digital Services Act requires online platforms that use recommender systems to set out in their terms and conditions the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters.

[12] For example, the B-Tech project. See: https://www.ohchr.org/en/business-and-human-rights/b-tech-project (accessed 28 April 2023).

REFERENCES

Arai-Takahashi, Y. (2002). *The margin of appreciation doctrine and the principle of proportionality in the jurisprudence of the ECHR*. Intersentia.

Audit Commission. (2010). *The Truth Is Out There: Transparency in an Information Age*. Audit Commission.

Bromwell, D. (2022). *Regulating free speech in a digital age: Hate, harm and the limits of censorship*. Springer.

Brown, M. (2023, 1 March). *The problem with TikTok's new researcher API is not TikTok* [Tech Policy Press]. https://techpolicy.press/the-problem-with-tiktoks-new-researcher-api-is-not-tiktok/

BSR. (2021). Human rights assessment: Global internet forum to counter terrorism. https://gifct.org/wp-content/uploads/2021/07/BSR_GIFCT_HRIA.pdf

Callon, M., Lascoumes, P., & Barthe, Y. (2009). In G. Burchell, *Acting in an uncertain world: An essay on technical democracy*. MIT Press.

Citron, D. (2018). 'Extremist speech, compelled conformity, and censorship creep'. *Notre Dame Law Review*, *93*(3), 1035–1072.

Clegg, N. (2023). 'Ending suspension of trump's accounts with new guardrails to deter repeat offenses', Meta Newsroom, 25th January. https://about.fb.com/news/2023/01/trump-facebook-instagram-account-suspension/

Cobbe, J., & Singh, J. (2019). Regulating recommending: Motivations, considerations, and principles. *European Journal of Law and Technology*, *10*(3), 1–37. https://ejlt.org/index.php/ejlt/article/view/686/982

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A., & Weir, D. (2018). Disrupting daesh: Measuring takedown of online terrorist material and its impacts. *Studies in Conflict & Terrorism*, *42*(1–2), 141–160. https://doi.org/10.1080/1057610x.2018.1513984

Conway, M., Watkin, A. L., & Looney, S. (2021). Violent extremism and terrorism online in 2021: The year in review. *European Union*. https://www.voxpol.eu/presenting-vox-pols-2021-year-in-review/

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3531146.3533213

Dias, O. T., Antonialli, D. M., & Gomes, A. (2021). 'Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online'. *Sexuality & Culture*, *25*(2), 700–732.

Dias Oliva, T. (2020). Content moderation technologies: Applying human rights standards to protect freedom of expression. *Human Rights Law Review*, *20*(4), 607–640.

DiResta, R. (2018). Free speech is not the same as free reach, wired, 30th August. https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/

Douek, E. (2021). 'Governing online speech: From "Posts-As-Trumps" to proportionality and probability'. *Columbia Law Review*, *121*(3), 759–833.

Elgot, J. (2022). "UK minister defends U-turn over removing harmful online content". The Guardian. November 29. https://www.theguardian.com/technology/2022/nov/29/minister-defends-u-turn-over-removing-harmful-online-content-online-safety-bill

Elsom, C. (2020). Safety without Censorship: A better way to tackle online harms. London: Centre for Policy Studies.

Engler, A. (2021). *Platform data access is a lynchpin of the EU's Digital Services Act*. Brookings Institution. https://www.brookings.edu/blog/techtank/2021/01/15/platform-data-access-is-a-lynchpin-of-the-eus-digital-services-act/

Fisher, E. (2010). 'Transparency and administrative law: A critical evaluation'. *Current Legal Problems*, *63*(1), 272–314.

GIFCT. (2022). GIFCT Transparency Report. https://gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf

Gillespie, T. (2022). 'Do not recommend? Reduction as a form of content moderation'. *Social Media + Society*, *8*(3), 205630512211175.

Global Partners Digital. (2018). A Rights-Respecting Model of Online Content Regulation by Platforms. https://www.gp-digital.org/wp-content/uploads/2018/05/A-rights-respecting-model-of-online-content-regulation-by-platforms.pdf

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

Google. (2022a). Search Quality Evaluator Guidelines. https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

Google. (2022b). Google Transparency Report. https://transparencyreport.google.com/?hl=en_GB

Gupta, A. (2008). 'Transparency under scrutiny: Information disclosure in global environmental governance'. *Global Environmental Politics*, *8*(2), 1–7.

Guy, E. (2018, February 12). 'Inside the Two Years That Shook Facebook – and the World', *The Wired*. https://www.wired.com/story/inside-facebook-mark-zuckerberg-2-years-of-hell/

Harling, A. S., Henesy, D., & Simmance, E. (2023). 'Transparency reporting: The UK regulatory perspective'. *Journal of Online Trust and Safety*, *1*(5), 1–8.
Howard, J. (2018). Should we ban dangerous speech? *British Academy Review*, *32*, 19–21.
Joint Committee on the Draft Online Safety Bill. (2021). Draft Online Safety Bill. Report of Session 2021-22. https://committees.parliament.uk/publications/8206/documents/84092/default/
Kaye, D. (2018). Report of the Special Rapporteur on the promotion and protection of the freedom of opinion and expression, David Kaye. U.N. General Assembly, A/HRC/38/35.
Kaye, D. (2019). Report of the Special Rapporteur on the promotion and protection of the freedom of opinion and expression, David Kaye. U.N. General Assembly, A/74/486.
Keller, D. (2021). 'Amplification and its discontents: Why regulating the reach of online content is hard'. *Journal of Free Speech Law*, *1*(1), 227–268.
Leerssen, P. (2020). 'The soap box as a black box: regulating transparency in social media recommender systems'. *European Journal of Law and Technology*, *11*(2), 1–51.
Lewis, P. (2019). '"Fiction is outperforming reality": How YouTube's algorithm distorts truth', The Guardian, 2nd February. https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth
Llanso, E., van Hoboken, J., Leerssen, P., & Harambam, J. (2020). 'Artificial Intelligence, Content Moderation, and Freedom of Expression', The Transatlantic Working Group Paper Series. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf
Luria, M. (2022). "This is Transparency to Me": User Insights into Recommendation Algorithm Reporting. Centre for Democracy and Technology. https://cdt.org/wp-content/uploads/2022/10/algorithmic-transparency-ux-final-100322.pdf
Lyons, T. (2018). 'The Three-Part Recipe for Cleaning up Your News Feed', Facebook Newsroom, 22nd May. https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/
MacCarthy, M. (2022). 'Transparency is essential for effective social media regulation'. Brookings Institution. November 1, 2022. https://www.brookings.edu/blog/techtank/2022/11/01/transparency-is-essential-for-effective-social-media-regulation/
Macdonald, S. (2006). 'A suicidal woman, roaming pigs and a noisy trampolinist: Refining the ASBO's definition of Anti-Social behaviour. *Modern Law Review*, *69*(2), 183–213.
Macdonald, S., Correia, S. G., & Watkin, A. L. (2019). 'Regulating terrorist content on social media: Automation and the rule of law,'. *International Journal of Law in Context*, *15*(2), 183–197.
McBride, J. (1999). Proportionality and the European Convention on Human Rights. In E. Ellis (Ed.), *The Principle of Proportionality in the Laws of Europe* (pp. 23–36). Hart Publishing.
Meta. (2020). 'An Update to How We Address Movements and Organizations Tied to Violence', Meta Newsroom. https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/
Meta. (2023a). 'Content borderline to the Community Standards', Transparency Center. https://transparency.fb.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards
Meta. (2023b). Transparency Center. https://transparency.fb.com/en-gb/
Mohan, N. (2022). '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/
Mol, A. (2008). *Environmental Reform in the Information Age: The Contours of Informational Governance*. Cambridge University Press.
Nava, V., & Golding, B. (2022). 'Latest 'Twitter Files' reveal secret suppression of right-wing commentators', New York Post, 8th December. https://nypost.com/2022/12/08/suppression-of-right-wing-users-exposed-in-latest-twitter-files/
Obama, B. (2009) 'Transparency and Open Government: Memorandum for the Heads of Executive Departments and Agencies', The White House. Available at: https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government (Accessed: 4 February 2023).
OHCHR. (2011). Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, HR/PUB/11/04. United Nations. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf
Radsch, C. (2022). Transparency Reporting: Good Practices and Lessons from Global Assessment Frameworks. https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-ResearchAgendaScopingPaper-1.1.pdf
Reddit. (2022). Quarantined Subreddits. https://www.reddithelp.com/hc/en-us/articles/360043069012
Saltman, E. (2022). GIFCT Executive Summary and Discussion of Dr Jazz Rowa's Algorithms Research. https://gifct.org/wp-content/uploads/2022/09/GIFCT-22WG-ContextualityIntros-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), 939–1006.
Schlesinger, P., & Kretschmer, M. (2020) 'The changing shape of platform regulation', LSE. Available at: https://blogs.lse.ac.uk/medialse/2020/02/18/the-changing-shape-of-platform-regulation/ (Accessed: 1 December 2022).

Stiglitz, J. E. (2002). Information and the change in the paradigm in economics. *American Economic Review*, *92*(3), 460–501.
Tech Against Terrorism. (2022). 'Tech Against Terrorism's Eight Requirements of Membership', The Tech Against Terrorism Trustmark. https://www.techagainstterrorism.org/membership/trustmark/
Tiku, N. (2020). 'Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it', Washington Post, 23rd December. https://www.washingtonpost.com/technology/2020/12/23/google-timnit-gebru-ai-ethics/
Tufekci, Z. (2015). 'Algorithmic harms beyond facebook and google: emergent challenges of computational agency'. *Colorado Technology Law Journal*, *13*(2), 207–217.
Twitter. (2022a). Twitter Transparency Center. https://transparency.twitter.com/
Twitter. (2022b). 'Hateful conduct policy', Help Center. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
UK Government. (2020). Online Harms White Paper: Full Government Response to the consultation. *CP 354*. The Stationery Office. 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
Walker, A (2022). 'Regulation 28 Report to Prevent Further Deaths', The Coroner's Service, 13th October. https://www.judiciary.uk/wp-content/uploads/2022/10/Molly-Russell-Prevention-of-future-deaths-report-2022-0315_Published.pdf
Whittaker, J. (2022). Recommendation algorithms and extremist content: A review of empirical evidence. *Internet Policy Review*, *10*(2), 1565. https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TR-Empirical-1.1.pdf
Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). 'Recommender systems and the amplification of extremist content'. *Internet Policy Review*, *10*(2), 1–29. https://doi.org/10.14763/2021.2.1565
YouTube. (2019). The four Rs of responsibility, part 1: Removing harmful content [Inside YouTube]. https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/

How to cite this article: Macdonald, S., & Vaughan, K. (2023). Moderating borderline content while respecting fundamental values. *Policy & Internet*, 1–15. https://doi.org/10.1002/poi3.376