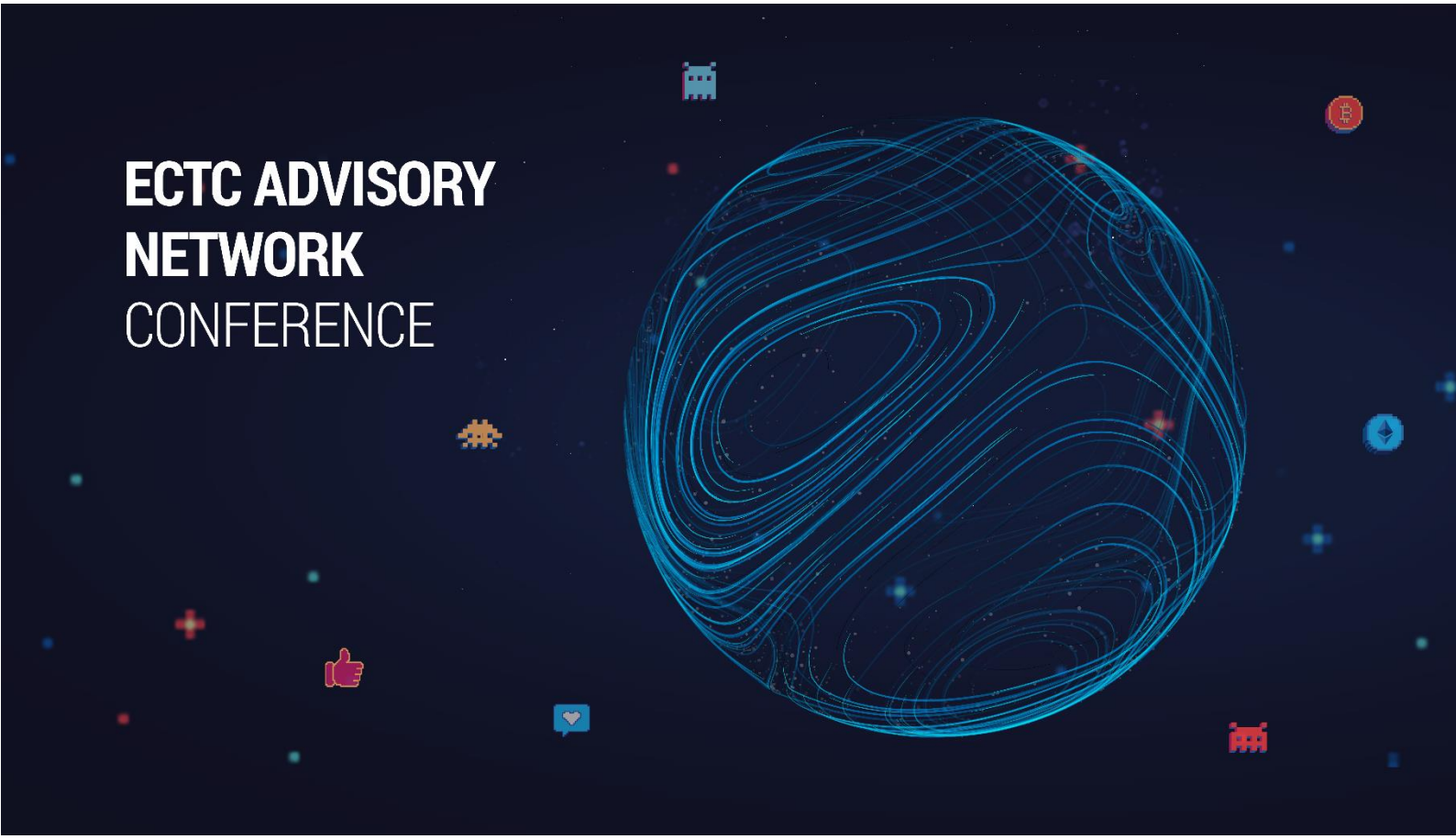
Moderating Borderline Content While Respecting Fundamental Values

CONFERENCE PUBLICATION



DATE

14-15/03/2023

AUTHORS

Stuart Macdonald

Katy Vaughan

This paper was presented at the 4th conference of the European Counter Terrorism Centre (ECTC) Advisory Network on terrorism and propaganda, 14-15 March 2023, at Europol Headquarters, The Hague. The views expressed are the authors' own and do not necessarily represent those of Europol.

Introduction

One of the key themes that emerged on the first day of the Fourth Annual Conference of the European Counter-Terrorism Centre (ECTC) Advisory Network on Terrorism and Propaganda was the increasingly amorphous nature of terrorism and the blurring of longstanding distinctions, such as between terrorism and extremism. Alongside this, concern has grown about so-called 'borderline content', including how terrorist and violent extremist actors are using this content strategically to evade detection online (Saltman and Hunt 2023). While the term borderline content is also amorphous, there appears to be a consensus that it is content that falls just short of violating platforms' Terms of Service – and so is not liable to be removed – but which nonetheless has the potential to cause harm (Conway, Watkin and Looney 2022). Hence, it is sometimes described as 'legal but harmful', or 'lawful but awful' content.

Given this feeling that borderline content is harmful – or, at least, potentially harmful – various options have been suggested for reducing its visibility and prominence, such as removing it from search and recommendation algorithms, downranking it and restricting users' ability to share it. In this paper, we use the term deamplification to refer to all these different options.

One of the key arguments 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. While this is true, our argument is that deamplificatory measures nonetheless raise significant human rights issues that need to be addressed. Indeed, legislators have moved to impose regulatory requirements, such as the UK's Online Safety Bill and the EU's Digital Services Act.

Our premise, like the one of the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (Kaye 2019), is that tech companies should respect human rights standards. Admittedly, private companies do not have the same obligations as governments. Nevertheless, there are several reasons why it is important that these private companies respect human rights in this context – not least the fact that respect for human rights is a key component of an effective counterterrorism strategy (Londras 2017). Indeed, one of the criteria for membership of the Global Internet Forum to Counter Terrorism (GIFCT) 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). Although the principles are non-binding, they establish a 'global standard of expected conduct for all business enterprises wherever they operate' (OHCHR 2011, 13). The argument for their adoption and implementation is particularly strong in the case of tech companies, given these companies' 'overwhelming role in public life globally' (Kaye 2018, 5). 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, 966).

We discuss these human rights issues under three headings. First, definitional clarity. This is a core feature of the rule of law. Laws and rules should be sufficiently clear to guide the actions and decisions of citizens. Second, international human rights treaties stipulate that any restrictions 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. Any such restriction must also be necessary to achieve this objective. One ingredient of the test of necessity includes an assessment of whether the measures are a proportionate restriction on speech. Third, transparency. This enables oversight and promotes accountability. It is also a prerequisite for the previous two requirements. Definitions can only be assessed if they are made publicly available. And the necessity and proportionality of moderation activity can only be assessed if the details and objectives of such efforts are disclosed.

Definitional Clarity

If the speed limit on a particular road is 50 kilometres per hour, there is nothing wrong with driving at a speed of 48, 49 or even 50 kilometres per hour. Your speed either exceeds the limit or it does not. This is how the law operates. A line is drawn and your conduct either crosses the line or it does not. You are either liable or not liable, guilty or not guilty.

Similarly, in respect of content moderation, content is either prohibited or permitted by a company's Terms of Service. Since there is consensus that borderline content does not violate Terms of Service, it is permitted. Moreover, being close to crossing the permitted/prohibited boundary is not in itself problematic, as the example of driving at 50 kilometres per hour illustrates. So, when we talk about borderline content, there must be something additional that influences how the content is perceived. Something that fuels the feeling that it should be deamplified. The difficulty has been articulating this.

These definitional challenges have been recognised. 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, 11). One attempt at definition is YouTube's statement that borderline content is content that does not 'quite cross the line of our policies for removal but that we don't necessarily want to recommend to people' (Mohan 2022). Yet this merely raises the further question: why exactly does YouTube not want to recommend it to people?

The UK's Online Safety Bill originally contained provisions that aimed to protect the online safety of adults. These provisions targeted borderline content, which was defined as content that is 'legal but harmful' (UK Government 2020, 32). The Bill defined harm in quite sweeping terms – including physical and psychological harm, self-harm as well as harm from others, potential and actual harm, harm arising from the nature of the content, harm arising from the simple fact of its dissemination, and harm arising from the manner of its dissemination (UK Government 2020). This was criticised for being so broad as to confer wide discretion on platforms, which left open the potential for inconsistent application (Joint Committee on the Draft Online Safety Bill 2021). Eventually, the Government removed the provisions from the Bill, with one of the reasons being the definitional challenges and their potential impact on the protection of freedom of speech (Elgot 2022).

There are a number of reasons why definitional clarity is important. It allows users to make informed decisions about how they use the platform (Kaye 2018, 15). It imposes limits on the discretion of content moderators and helps ensure consistency in decision-making (Howard 2018). It guards against censorship creep, where powers that were designed for certain purposes or situations are used in other ways and in other

contexts (Citron 2018). And it helps ensure users have an effective opportunity to appeal against moderation decisions (Macdonald, Correia & Watkin 2019). Of course, there are also competing considerations. In a fast-moving landscape, definitions need to be sufficiently flexible to encompass societal and technological developments (Joint Committee on the Draft Online Safety Bill 2021, 53). There is also concern about possible adversarial shifts (Tech Against Terrorism 2022). If content moderation policies are too open, then bad actors will be able to game the system. Some vagueness generates uncertainty and can act as a deterrent. But these competing considerations have not prevented us from working towards clearer definitions of difficult terms like terrorism – and so it is hard to see why they should stop us trying to define borderline content more clearly too.

At present, a fundamental problem and underlying reason for the definitional challenges is that borderline content is an umbrella term. It is used to refer to a variety of conduct. For example, it is used to encompass misinformation, sexually suggestive content and gory or graphic imagery (Meta 2023a). 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, 32). It is impossible to concoct a definition that embraces all these types of content and is clear and precise and is flexible enough to be future proof. If we want fair warning, consistent implementation and to avoid censorship creep, we need to move away from trying to develop a short dictionary-style definition of borderline content and develop an alternative approach.

A potential solution would be to take a list-based approach. This would entail compiling a list of the content types that are deemed to be borderline and developing individual definitions for each of these types of content. A process could be included for adding new items, with safeguards to ensure independent oversight and multi-stakeholder consultation. The aim being to provide the necessary flexibility whilst maintaining scrutiny and accountability.

Necessity and Proportionality

A popular slogan in relation to algorithmic recommendation systems is DiResta's (2018) statement that 'free speech does not mean free reach'. The article in which this statement appears talks about algorithmic amplification and was written in response to claims from President Trump that social media companies had rigged their platforms. The key point that the article was seeking to make was that arguments over whether or not social media companies were guilty of censorship was distracting from a more important question: what is going wrong with algorithmic amplification and how can it be fixed? It was in this context that DiResta wrote 'It would be good to remind them [the politicians and pundits complaining about censorship] that free speech does not mean free reach. There is no right to algorithmic amplification'. In other words, the right to free speech does not entail the right to have one's speech amplified.

The problem is that the slogan is now often taken out of its original context and used to make a different claim. It is sometimes used to argue that deliberate deamplification of content does not impinge on free speech rights. Others have made similar claims. For example, Douek (2021) has observed that 'de-amplification does not reduce the amount of speech and does not directly impede the ability to speak'. While this may be true, it is also the case that deamplification has much in common with deplatforming. Although deamplification does not remove the content altogether, it

does burden speech – and, as the US Supreme Court has stated, the difference between banning speech and burdening speech is 'but a matter of degree'.[1]

Deamplification raises many of the same concerns as deplatforming. Promoting certain content and controlling the dissemination of other items amounts to a gatekeeping function (Llanso et al., 2020). The practical effect of deamplification can be to suppress speech and prevent users from making their voices heard. There is a danger that deamplificatory measures are applied inconsistently, or in a discriminatory fashion. There is a danger of censorship creep (Citron 2018). And if companies have a legal duty to deamplify borderline content, there is a danger that they will adopt a cautious approach and resort to overenforcement to avoid liability (Keller 2021).

A rights-based framework would help guard against these dangers, by requiring platforms to identify a legitimate objective for any deamplificatory measures, and to assess such measures' necessity and proportionality (Kaye 2018). Two points about this test are worth highlighting. First, the assessment of proportionality would include consideration of whether the least intrusive means were employed to achieve the stated objective (UNHRC General Comment No. 34 2011, para 33). This would encourage consideration of the range of different deamplificatory options. Second, the proportionality assessment would also look at any countermeasures that have been implemented, such as information that is provided to affected users and any appeals process (The Santa Clara Principles 2021). This leads us to issues of transparency.

Transparency

There are various reasons for a commitment to transparency. Transparency has value in sharing expertise and insight on how to prevent terrorist exploitation of online platforms (BSR 2021). It can promote multi-stakeholder collaboration (GIFCT 2022), inform policymaking and raise public awareness (Twitter 2022), and it enables accountability (Meta 2023b). Transparency at the level of the individual user is also important. It respects the autonomy of users, improving their ability to make informed decisions about how they use online platforms and challenge decisions.

One transparency mechanism is transparency reporting – that is, quantitative, statistical data. This is reflected in the membership criteria of both GIFCT ('regular, public data transparency reports' are required) and Tech Against Terrorism (prospective members must 'commit to improve transparency reporting') (GIFCT 2022; Tech Against Terrorism 2023, respectively). The EU's Digital Services Act will formalise transparency reporting obligations for all platforms (other than small or micro-ones).[2] However, transparency reports generally focus on content that violates terms of service – and so, by definition, do not include data on borderline content. This means that it is unclear how much content is classified as borderline, which types of content receive this classification, and what actions are most frequently taken.

A further transparency mechanism is the publication of content moderation policies. Some companies do provide some information on deamplification measures. Perhaps the most detailed information comes from YouTube (YouTube 2019). Human evaluators use a publicly available set of guidelines to decide whether a video is authoritative or borderline. Content that is classified as borderline is demoted in recommendations, in order to reduce its spread. Key questions in determining borderline status include whether the content is inaccurate, misleading, deceptive,

[1] *US v Playboy* 529 U.S. 803, 812 (2000).
[2] 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 EUR 10 million. For micro enterprises, the respective figures are 10 persons and EUR 2 million.

insensitive or intolerant, and whether the video is harmful or has the potential to cause harm (Goodrow 2021).

While this provides some insight, it also raises the earlier questions and concerns about definitional clarity. Moreover, it is focused on the generic, policy level. This is of course important. But there are also strong reasons for transparency at the level of the individual user, not least enhanced accountability and correcting mistakes. If content is deamplified, will the user that posted the content be told how and why it has been deamplified? If so, will there be an opportunity to appeal against the deamplification? If there is no appeals process, this is a factor that may suggest that the measure taken was disproportionate. On the other hand, if there is an appeal process, there remains the question as to whether a user will have an effective opportunity to appeal. 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 paper has discussed the moderation of borderline content from the perspective of definitional clarity, necessity, proportionality and transparency. It has offered suggestions for how to improve the compliance of these moderation efforts with international human rights standards. First, creating an exhaustive list of defined content types considered borderline and defining these, accompanied by a process for adding new items to the list that is subject to independent oversight and multi-stakeholder consultation. This would go some way to improve definitional clarity. In addition, the objective of deamplification measures should be made clear and the necessity and proportionality of these measures assessed. This includes an assessment of whether alternative, less intrusive measures could be utilised and the adequacy of countermeasures, in particular the availability of an appeals mechanism. These suggestions require a commitment to transparency in respect of the moderation of borderline, in addition to violative, content. At the level of the individual user, those whose content has been deamplified should be informed, with an explanation of reasons and the opportunity to appeal.

Reference List

BSR (2021). Human Rights Assessment: Global Internet Forum to Counter Terrorism. Available at: gifct.org/wp-content/uploads/2021/07/BSR_GIFCT_HRIA.pdf (Accessed: 20 December 2022).

Citron, D. (2018). 'Extremist Speech, Compelled Conformity, and Censorship Creep,' Notre Dame Law Review, 93(3), pp. 1035-1072.

Conway, M., Watkin, A. and Looney, S (2021). 'Violent Extremism and Terrorism Online in 2021: The Year in Review' (RAN Policy Report, 2021).

DiResta, R. (2018). 'Free Speech is Not the Same As Free Reach', Wired, 30th August. Available at: wired.com/story/free-speech-is-not-the-same-as-free-reach/ (Accessed: 4 February 2023).

Douek, E. (2021). 'Governing Online Speech: From "Posts-As-Trumps" to Proportionality and Probability', Columbia Law Review, 121(3), pp. 759-833.

Elgot, J. (2022). 'UK minister defends U-turn over removing harmful online content', The Guardian. 29 November. Available at: theguardian.com/technology/2022/nov/29/minister-defends-u-turn-over-removing-harmful-online-content-online-safety-bill (Accessed: 19 December 2022).

Goodrow, C. (2021). 'On YouTube's recommendation system', YouTube Official Blog, 15th September. Available at: blog.youtube/inside-youtube/on-youtubes-recommendation-system/ (Accessed: 30 November 2022).

GIFCT (2022). 2022 GIFCT Transparency Report. Available at: gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf (Accessed: 19 December 2022).

Howard, J. (2018). 'Should we ban dangerous speech? British Academy Review, 32, pp. 19 – 21.

Joint Committee on the Draft Online Safety Bill (2021). Draft Online Safety Bill. Report of Session 2021-22. Available at: committees.parliament.uk/publications/8206/documents/84092/default/ (Accessed: 1 December 2022).

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), pp. 227 – 268.

Llanso, E., van Hoboken, J., Leerssen, P. and Harambam, J. (2020). 'Artificial Intelligence, Content Moderation, and Freedom of Expression', The Transatlantic Working Group Paper Series. Available at: ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf (Accessed: 4 February 2023).

Londras, F. (2017). 'Tearing up human rights law won't protect us from terrorism' The Conversation (7 June 2017).

Meta (2023a). 'Content borderline to the Community Standards', Transparency Center. Available at: transparency.fb.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards (Accessed: 24 April 2023).

Meta (2023b). Transparency Center. Available at: https://transparency.fb.com/en-gb/ (Accessed: 20 December 2022).

Mohan, N. (2022) 'Inside Responsibility: What's next on our misinfo efforts', YouTube Official Blog. Available at: blog.youtube/inside-youtube/inside-responsibility-whats-next-on-our-misinfo-efforts/ (Accessed: 30 November 2022).

OHCHR (2011). Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, HR/PUB/11/04. United Nations. Available at: ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf (Accessed: 25 April 2023).

Saltman, E. (2022). GIFCT Executive Summary and Discussion of Dr Jazz Rowa's Algorithms Research. Available at: gifct.org/wp-content/uploads/2022/09/GIFCT-22WG-ContextualityIntros-1.1.pdf (Accessed: 30 November 2022).

Saltman, E. and Hunt, M. (2023) Borderline Content: Understanding the Gray Zone (GIFCT). Available at: gifct.org/wp-content/uploads/2023/06/GIFCT-23WG-Borderline-1.1.pdf (Accessed: 29 September 2023).

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), pp. 939-1006.

Tech Against Terrorism (2022). State of Play: Trends in Terrorist and Violent Extremist Use of the Internet 2022. Available at: techagainstterrorism.org/wp-content/uploads/2023/01/FINAL-State-of-Play-2022-TAT.pdf. (Accessed 29 September 2023).

Tech Against Terrorism (2023). 'Membership Criteria'. Available at: techagainstterrorism.org/membership/trustmark/ (Accessed 29 September 2023).

Twitter (2022). Twitter Transparency Center. Available at: transparency.twitter.com/ (Accessed: 20 December 2022).

UK Government (2020). Online Harms White Paper: Full Government Response to the consultation. CP 354. London: The Stationery Office.

YouTube (2019). 'The Four Rs of Responsibility, Part 1: Removing harmful content', Inside YouTube, 3rd September. Available at: blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/ (Accessed: 21 December 2022).

Professor Stuart Macdonald (Swansea University)

Dr. Katy Vaughan (Swansea University)