The Alan Turing Institute

The Alan Turing Institute's response to the Ofcom Call for Evidence on the first phase of Online Safety Regulation

September 2022

Introduction

The Alan Turing Institute welcomes the opportunity to respond to Ofcom's Call for Evidence regarding the forthcoming Online Safety Regulation. We firmly believe that AI and data science can help solve online safety, although it is no 'silver bullet'. By submitting this response we hope to contribute nuanced evidence of how to better protect users and prevent harm without unduly restricting user activity. We also propose multiple routes available to improve online safety by increasing industry transparency and strengthening reporting mechanisms for users. Solving online safety truly requires strong multi-stakeholder collaboration; this response draws on our own expertise and research, as well as that of the wider academic community, civil society initiatives, industry best-practice and public policy papers.

The Alan Turing Institute

The Alan Turing Institute is the UK's national institute for data science and artificial intelligence. Our mission is to make great leaps in data science and artificial intelligence research in order to change the world for the better.

We have three ambitious goals:

- **Advance world-class research and apply it to real-world problems:** innovate and develop world-class research in data science and artificial intelligence that supports next generation theoretical developments and is applied to real-world problems, generating the creation of new businesses, services, and jobs.

- **Train the leaders of the future:** train new generations of data science and AI leaders with the necessary breadth and depth of technical and ethical skills to match the UK's growing industrial and societal needs.

- **Lead the public conversation:** through agenda-setting research, public engagement, and expert technical advice, drive new and innovative ideas which have a significant influence on industry, government, regulation, or societal views, or which have an impact on how data science and artificial intelligence research is undertaken.

The Alan Turing Institute is headquartered in the British Library, London. Since its inception in 2015 the Institute has been funded through grants from Research Councils, university partners and from strategic and other partnerships.

The Public Policy Programme

The Alan Turing Institute's Public Policy Programme[1] works alongside policy makers to explore how data-driven public service provision and policy innovation might solve long running policy problems and to develop the ethical foundations for the use of data science and artificial intelligence in policy-making. Our aim is to contribute to the Institute's mission – to make great leaps in data science and artificial intelligence research in order to change the world for the better – by developing research, tools, and techniques that have a positive impact on the lives of as many people as possible.

The Online Safety Team

Part of The Alan Turing Institute's Public Policy Programme, the Online Safety Team provides objective, evidence-driven insight into the technical, social, empirical and ethical aspects of online safety, supporting the work of policymakers and regulators, informing civic discourse and extending academic knowledge. We are working to tackle online hate, harassment, extremism and mis/disinformation. There are three core workstreams: (1) Data-centric machine learning, where we are building and critically examining cutting-edge technologies to flag and rate toxic content; (2) The Online Harms Observatory, mapping the scope, prevalence and impact of content and activity that could inflict harm on people online; and (3) Policymaking for Online Safety, where we are working to understand the challenges in ensuring online safety, and supporting the creation of ethical and innovative solutions.

[1] https://www.turing.ac.uk/research/research-programmes/public-policy

Q7: What can providers of online services do to enhance the transparency, accessibility, ease of use and users' awareness of their reporting and complaints mechanisms?

Enhancing user engagement with flagging mechanisms is critical for tackling online harms.

Given the scale of harmful content online, user engagement with reporting ('flagging') such content is considered crucial for helping to tackle online harms.[2] Most platforms have set procedures for users to flag content which they believe to be inappropriate, offensive, or as breaching community guidelines. However, these procedures may differ across platforms in terms of how accessible the reporting mechanisms are, how much information users are given about when and why content should be flagged, how much detail users are able to express regarding why they wish to flag something, and how much transparency there is about what happens to flagged content.[3]

It is difficult to measure how much potentially harmful online content is flagged by users as platforms do not always make this information available. However, as user engagement is increasingly central for tackling online harms, it is important to assess how platforms can improve their reporting and complaints mechanisms, both in encouraging users to flag potentially harmful content more routinely, and also in making sure the right kind of content is flagged. In the sections below, we outline key features which may enhance user engagement with flagging.

- **Providing more information about reporting mechanisms can increase flagging behaviours.**

  Platforms may differ regarding the amount of information they give to users about reporting mechanisms, such as how and when content should be flagged. Some research suggests that users are more likely to flag potentially harmful content if they are provided with detailed guidelines on reporting. Naab and colleagues[4] tested key factors underlying flagging of uncivil user-generated content in comments sections of news sites.

[2] See: Porten-Cheé, P., Kunst, M., & Emmer, M. (2020). Online civic intervention: A new form of political participation under conditions of a disruptive online discourse. *International Journal of Communication*, 14, 21.
[3] Outlined in: Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society,* 18(3), 410–428.
[4] Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, 20(2), 777–795.

Study details
In one experiment, participants viewed an article on a mock news site along with a comments section below, ostensibly written by other participants and containing an offensive comment about the individuals in the article. The researchers manipulated how much information participants were given about reporting content (high or low), along with targets of abuse (individuals or social groups) and whether other responses agreed or disagreed with the comment. Participants had the option to 'like', 'dislike', or 'flag' each comment.

Results showed that flagging an uncivil comment was overall more likely when participants were provided with increased information about how to use the reporting mechanism. The results offer preliminary evidence that giving clear information about community guidelines, how to use tools for flagging, and emphasising the importance of user engagement, can increase flagging behaviours. However, the effect was weaker if the target of the uncivil comment was an abstract group compared to individuals, suggesting that individualising targets of online abuse is also important for online civic intervention. Further, results should be interpreted with caution because they measure flagging behaviours in just one particular social context and online environment.

Despite this, the benefits of providing users with detailed information about community rules have been noted elsewhere.[5] We would therefore recommend that, to increase flagging behaviours, platforms should consider providing as much information as possible about usage policies and descriptions of how to intervene using flagging tools.

- **Increasing transparency about moderation processes may encourage flagging behaviours.**

  Reporting and complaints mechanisms for online platforms have been critiqued on the grounds that processes for flagged content can lack transparency.[6] Often, users are provided with little indication of how or whether a decision is made to remove the content that they flagged.[7] Usually, flagged content is not apparent to other users, and the reasons for removal or retention are not made public. Further, social media sites can be unclear about how and when a flag has an impact.

[5] Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, *116*(20), 9785–9789.
[6] For example see: Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[7] Though some platforms have made an effort to increase transparency in recent years, for example Facebook's Support Inbox allows users to monitor reports they have made, displaying what was flagged and when and its status in the review process.

Despite these critiques, empirical research is yet to directly test whether enhancing process transparency does increase and improve flagging behaviours online. One recent research study examined whether enhancing transparency of moderation systems led to an increase in users' level of trust in content moderation systems for hate speech and suicidal ideation.[8]

> Study details
> Participants were asked for feedback on a classification system ostensibly under development. Participants saw posts which they were told came directly from social media users and received information about how these posts had been classified. The researchers manipulated the level of transparency about how the system classified posts (transparency, interactive transparency or no transparency), the moderator type (human, AI or both) and whether the posts were flagged or not flagged by the system. In the transparency condition, participants were provided with details about keywords that the system used to classify the content. In the interactive transparency condition, participants were provided with these details and were also able to suggest words for inclusion or exclusion.

Results showed that trust in moderation systems was higher in both the transparency conditions compared to when no information about the process was given. This effect held irrespective of the source of moderation (AI, human, or both). The researchers suggest that knowledge about the rules of classification allows users to feel more agentic, resulting not only in better understanding of the system, but also in greater trust and agreement with its decisions. This is consistent with an analysis on process transparency by Suzor and colleagues.[9]

In relation to reporting mechanisms, it might be reasoned that users should be more willing to engage with a system that they have higher trust in. While further research is needed to examine the direct impact of process transparency on the quality and quantity of flagging behaviours, there are grounds to suggest that platforms would benefit from providing users with as much information as possible regarding how decisions are made to remove or retain flagged content.

[8] Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, *27*(4), zmac010.
[9] 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*, 18.

- **Platforms could benefit from a better understanding of how social norms and bystander effects can influence intentions to intervene.**

  Some research has conceptualised online flagging behaviours as a specific form of bystander intervention.[10] Obermaier and colleagues examined effects of perceived severity of harm along with the number of bystanders present on people's intentions to intervene in an instance of cyberbullying on a Facebook group. Results showed that a highly severe cyberbullying incident positively affected participants' feelings of responsibility and, in turn, their intentions to intervene. However, the presence of a large number of bystanders lowered participants' feelings of responsibility to act, reducing their intentions to intervene.

  Relatedly, another experiment[11] tested how responses to uncivil comments from other users influence flagging behaviours by altering feelings of self-responsibility. Results showed that flagging was lower when a response from another user disagreed with an offensive comment compared to when a response agreed with the comment, but only when the response was impolite. This implies that if strong disagreement with incivility has already been expressed, then users may feel that the situation has been dealt with and feelings of self-responsibility to intervene are diminished.

  The role of bystander behaviours and social norms in influencing flagging behaviours is likely to be complex and current knowledge on the matter is limited. Whilst difficult to address at the platform level, dynamics between users may be important to consider in optimising conditions under which users flag potentially harmful content. For example, cues about how many and which other people have viewed or responded to a piece of content may affect users' reporting of that content. Platforms could benefit from working towards a better understanding of the interplay between the presence and responses of other users and reporting behaviours.

- **It is important to recognise that flagging systems may be used inappropriately.**

  As outlined above, very few research studies have focused on enhancing user engagement with reporting and complaints mechanisms online. Work that has addressed this issue has tended to focus on how users might be encouraged to engage with flagging potential harms to a greater degree.[12] However, wider

[10] Obermaier, M., Fawzi, N., & Koch, T. (2016). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New Media & Society*, *18*(8), 1491–1507.
[11] Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, *20*(2), 777–795.
[12] E.g., Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, *20*(2), 777–795.

concerns with flagging mechanisms have been discussed. Perhaps of most concern surrounding the reporting mechanisms of online services is their ability to be 'gamed'. Crawford and Gillespie[13] describe the many instances in which flags are used inappropriately by users, from pranks between friends and sabotage attempts between rivals, to harassment and bullying attacks and highly coordinated hate campaigns. The researchers note the case of one conservative group coordinating members to flag LGBTQ+ groups on Facebook. Similarly, another article describes a case in which a group of bloggers coordinated their supporters to flag Muslim content on YouTube as promoting terrorism.[14] Research elsewhere suggests that users may flag content as misinformation more often when the content is in disagreement with their own ideology, shedding light on the potentially biased nature of flagging.[15] In attempting to improve the efficacy of reporting and complaints procedures online, it will be important for platforms to not only encourage flagging overall, but also to make sure the right kind of content is being flagged.

[13] Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428.
[14] Fiore-Silfvast, B. (2012). User-generated warfare: A case of converging wartime information networks and coproductive regulation on YouTube. *International Journal of Communication*, *6*, 24.
[15] Coscia, M., & Rossi, L. (2020). Distortions of political bias in crowdsourced misinformation flagging. *Journal of the Royal Society Interface*, *17*(167), 20200020.

Q11: Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?

Content moderation has been successful in reducing harmful content, but handling borderline cases remains a challenge - we suggest that 'second level' content should be considered.

Up until now the main focus of content moderation on social media has been on actions which restrict user activity in some way: for example, by blocking and deleting content, banning users or deleting whole areas of a site if they are found to contravene a particular policy. Efforts in this area have undoubtedly been successful in reducing (though not eliminating) the prevalence and visibility of harmful content online. However they also have potentially important consequences for free speech. In addition to this, content moderation efforts are inevitably troubled by 'grey areas': borderline content which is difficult to classify or will provoke significant debate as to whether it should be protected as free speech or not (content produced by high profile political figures that could be interpreted as an incitement to violence is a standout example of this).

Hence more recently there has been an increased focus on what might be called 'second level' content moderation options, which still restrict user exposure to harmful content in some way, but are arguably less restrictive on user activity and hence can be more comfortably applied to content in the 'grey area'. In this response we will provide an overview of the options in this domain and review what is known about their effectiveness (with a particular focus on the harm types listed as priority for this consultation).

- **Technologies which reduce the speed of interaction with social media ('FrictionTech') have been shown to increase user safety.**

  A first area to consider is what has sometimes been called 'FrictionTech' [16][17]: modifications to the ways users interact with social media platforms that reduce their speed of interaction and put up (small) barriers to the creation of new content. The aim here is to make users reflect on their actions, especially in cases where the content they are creating might appear to be potentially harmful. For example, both Instagram and Twitter have experimented with systems that flag when a user appears to be typing a harmful comment[18], whilst Twitter has also brought in a

[16] Polgal, RP. (2021, May 25). *Friction Tech Should Play a Bigger Role in Social Media.* Built In. https://builtin.com/software-engineering-perspectives/key-better-social-media-ecosystem-friction
[17] Hendricks, H. (2021, December 20) *Turning the Tables: Using Big Tech community standards as friction strategies.* OECD | The Forum Network. https://www.oecd-forum.org/posts/turning-the-tables-using-bigtech-community-standards-as-friction-strategies
[18] Statt, N. (2020, May 5) *Twitter tests a warning message that tells users to rethink offensive replies.* The Verge. https://www.theverge.com/2020/5/5/21248201/twitter-reply-warning-harmful-language-revise-tweet-moderation

feature which prompts people to read news articles before retweeting them[19], which internal research showed increased article readership by 40%. It is easy to imagine application domains for this type of technology in many of the application domains relevant to this call, for example 'threats to kill'. Such systems could also potentially prompt users to be aware of relevant legislation, for example around drugs and firearms, where they may not be aware they are committing an offence (sadly one common source of material which might ostensibly be classified as child pornography is from grandparents posting pictures of their grandchildren in the bathtub[20], unaware of the potential legal consequences).

This type of system is especially useful because it may allow users themselves to navigate tricky questions of whether contextual factors may mean that content which might be impermissible in one setting is nevertheless justified in another (for example, the horrific photo of a child covered in Napalm in Vietnam is at once an example of graphic violence but at the same time a vitally important piece of war reporting, and as such has frequently troubled content moderation systems[21]). However, of course it is worth stating that people who are intent on creating harmful content will not be deterred by such technology. Furthermore, FrictionTech also opens the possibility that users can try and game the system by experimenting with different phrasings and seeing which ones a platform will classify as potentially illicit.

Once content has been created, a variety of types of secondary protection can be applied. One example of this type of protection is further friction being introduced before the content can be consumed. For example, explicit content warning labels may be added to content which appears to be pornographic or graphically violent. During the most recent US presidential elections, a Facebook employee was quoted as saying that such filters reduced sharing of misinformation posts by Donald Trump by around 8%.[22] However of course it is difficult to know if that number would generalise to other contexts. In 2021 Twitter said that their redesigned labels resulted in a 17% increase in users clicking on the information to find out more, though it is not clear what the baseline is. [23] Some exploratory work

[19] Porter, J. (2019, December 16) *Instagram to start warning users before they post 'potentially offensive' captions.* The Verge. https://www.theverge.com/2020/5/5/21248201/twitter-reply-warning-harmful-language-revise-tweet-moderation

[20] Constitutional Fights. (2009, May 5) *Grandma Arrested For Photos of Grandchild in Tub.* https://constitutionalfights.wordpress.com/2009/05/05/grandma-arrested-for-photos-of-grandchild-in-tub/

[21] Kleinmann, Z. (2016, September 9) *Fury over Facebook 'Napalm girl'.* BBC. https://www.bbc.com/news/technology-37318031

[22] Kraus, R. (2020, November 19) *Facebook labeled 180 million posts as 'false' since March. Election misinformation spread anyway.* Mashable. https://mashable.com/article/facebook-labels-180-million-posts-false

[23] Business Standard (2021, November 17) *Twitter launches more effective, redesigned misinformation warning lables.* https://www.business-standard.com/article/technology/twitter-launches-more-effective-redesigned-misinformation-warning-labels-121111700056_1.html

has said warning labels are effective in reducing beliefs in Covid-19 misinformation.[24]

This type of protection does not place a great deal of limits on user freedom as it is of course still possible to 'click through' the warning but can be effective in meaning that users are not exposed to content that they know that they do not want to see. In addition to this, social platforms can also provide options to users to explicitly set limits on the type of content that they might be exposed to on their platform. Google's SafeSearch is an obvious example of this, as well as Twitter's options for limiting who can reply to content posted by a user. To our knowledge however research about usage rates of such technology is limited.

- **Limiting the spread and visibility of content is an option which may reduce harms whilst only having a moderate impact on free speech.**

  Another type of example are general actions made by a platform to limit the spread of a given piece of content (even whilst allowing its creation). All social media companies have an automated mechanism for ranking the 'feed' of content to which a user is exposed, which is based on a variety of factors specific to each platform. Including potential harmfulness as a 'penalty' factor in these algorithms can be one way to limit the automatic distribution of some content, without deleting it entirely. [25] Such actions do restrict user freedom to an extent, as users may miss pieces of content that they otherwise would have seen. But they also mean that content is still discoverable to those who want to see it, whilst removing the incentives for content creators to create borderline content as a way of generating exposure.

- **Robust appeals processes are critical in online content moderation systems.**

  A final type of improvement worth mentioning in this context is in the domain of appeals processes: i.e. routes that a user can make use of if their content was removed or their account / channel was deleted. It is important to recognise that, for some people, social media represents the key place where they carry out their business activities or perhaps the only way they have of keeping in touch with certain friends and family: hence decisions to delete content or accounts can sometimes carry serious consequences for an individual. This makes appeals processes with the ability to reverse incorrect decisions very important. Appeals processes are multi-faceted: people first need to be made clearly aware that some of their content was deleted, and also the rationale behind the deletion.[26] They then

[24] Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2022). Misinformation warnings: Twitter's soft moderation effects on covid-19 vaccine belief echoes. *Computers & Security*, *114*, 102577. https://doi.org/10.1016/j.cose.2021.102577
[25] Twitter. (n.d.) *Abusive behaviour.* https://help.twitter.com/en/rules-and-policies/abusive-behavior
[26] Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter? *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–27. https://doi.org/10.1145/3359252

need to be provided with appeals options. Knowing exactly how to structure these appeals is difficult: with millions of pieces of content removed every day, having a detailed human appeal process for all of them would be impossible. Nevertheless, routes to correct mistakes are key in not unduly restricting user activity.

Q18. Are there any functionalities or design features which evidence suggests can effectively prevent harm, and could or should be deployed more widely by industry?

- **Identifying illegal content automatically is difficult and therefore any measures built on top of such tools should be used with caution.**

  Deploying automated measures to tackle illegal content requires the capability to identify this content accurately, and automatically. When discussing the efficacy of these measures it is important to remember the many challenges automated solutions still face, despite the great leaps in technology being made – many models still struggle to correctly illegal content, and a large grayzone remains.

  Therefore, preventative measures should focus on dealing with "potentially illegal" content. Furthermore, evaluating the effectiveness of preventative measures is difficult since it requires (1) clarifications on what qualifies effectiveness, and (2) properly controlled setups and sufficient time to monitor how measures perform.

- **Counterspeech could potentially induce behaviour change among perpetrators or bystanders**

  Counterspeech measures (referring to a direct response challenging the narrative or viewpoint in a post[27]), either in response to harmful content or used more generally in conversations tangential to, or preceding harmful content, can potentially induce positive behaviour and attitudinal changes of perpetrators or bystanders who may otherwise go on to become purveyors of illegal content in the future.

  Most studies on the effectiveness of counterspeech rely on predefined messages (e.g., manually written responses) rather than new messages generated by language models on-the-fly, partially due to the limitation and controllability of existing technology. We summarise the key findings regarding the effectiveness of counterspeech in real-life scenarios from academic research below.

  A common speculation about counterspeech is that exposure to counter messages would make haters more hateful or provoke abuse. However, there is no evidence showing that exposure to counterspeech leads to radical behaviour by perpetrators or audiences[28].

[27] Benesch, S. (2014). Countering dangerous speech: New ideas for genocide prevention. *SSRN Electronic Journal*.
[28] Saltman, E., Kooti, F., & Vockery, K. (2021). New models for deploying counterspeech: Measuring behavioral change and sentiment analysis. *Studies in Conflict & Terrorism*, 1-24.

Among three studied strategies (empathy, warning of consequences, and humour), empathy-based counterspeech is shown to be most effective in encouraging Twitter

users to (1) delete the original harmful tweets and (2) reduce the volume of racist posts over a 4-week follow-up study.[29] In response to a hate incident, an example of empathetic response might be: "This is heart-breaking news. As a woman, any type of abuse is always disturbing. Let's stop similar events from happening to anyone." Furthermore, according to an intervention study[30] conducted on Reddit, exposure to messages containing normative induction (e.g., suggest users to follow socially appropriate behaviour) is effective in reducing verbal aggression.

The identity of counterspeakers can be an indicator for the effectiveness of counterspeech. For instance, based on a 2-month study counter messages posted by high-follower and in-group users are shown to significantly reduce the use of racist slurs.[31]

The encouraging results of counterspeech interventions lay the groundwork for the potential of automatising hate prevention via, e.g., conversational agents at scale. See more details on counterspeech automation from sectors such as research institutions[32], civil organisations[33], and social network companies[34].

- **Anticipatory de-escalation may be effective to intervene prior to hateful content being posted**

  In a similar vein to our response to Question 11, de-escalatory measures can be used to intervene in the lead up to illegal content being published, either by disrupting conversations (as with counterspeech), or by interrupting the train-of-thought or emotional drivers behind a user intending to post harmful content.

  For example, to combat cyberbullying, Rethink application takes this initiative and offers an in-the-moment nudge to pause, review and rethink before posting illegal content using algorithms.[35] This approach encourages users to react responsibly

[29] Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., ... & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50).
[30] Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive behavior*, *47*(3), 260-266.
[31] Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629-649.
[32] Chung, Y.-L. & Vidgen, B. (2022, July 1). Counterspeech: a better way of tackling online hate? *The Alan Turing Institute*. https://www.turing.ac.uk/blog/counterspeech-better-way-tackling-online-hate
[33] How to counter hate speech on Twitter? *Get the trolls out*. https://getthetrollsout.org/stoppinghate
[34] *Counterspeech*. (n.d.). Retrieved September 12, 2022, from https://counterspeech.fb.com/en/
[35] ReThink, Inc. (n.d.). *ReThink - Before the Damage is Done*. Retrieved September 12, 2022, from https://www.rethinkwords.com/

by spontaneously measuring the potential harms and consequences their actions can cause. Rethink is a student-led movement, and, according to the website, the application reaches at least 1500 schools and 500K downloads, and successfully convinces users not to post harmful content 93% of the time. [36]

- **It may be effective to suppress content or in other ways combat potential harms without removing it**

  Suppressive measures involve reducing the degree to which potentially illegal content is shared and is able to cause harm, such as by advertising positive information or replacing illegal content with predefined pages or helpline information. For instance, the Redirect Method[37], designed by Jigsaw and Moonshot, is a module that attempts to show alternative counterspeech or counter videos in the search results, when users input queries that can imply intent for extremist content or groups.

  With the help of algorithmic ranking design, internet users have the opportunity to control the online experience they prefer. Opt Out[38] is a Firefox extension that blocks misogynist posts from a user's Twitter feed. Tune[39] is a Chrome add-on that lets users decide the amount of potential toxic content allowed in the online content they consume.

- **More approaches are needed to tackle multimodal content**

  In addition to these measures, we identify major gaps in industry to be addressed in tackling illegal content. The preventative measures discussed above are focused on tackling textual content. As video streaming services are becoming popular and accessible for content dissemination, perpetrators can take advantage of such platforms to both share harmful content and engage in abuse in real-time. To create a safe online sphere and avoid live-streamed abuse, measures for limiting the sharing of illegal videos and keeping users from being subject to potential threats are urgently needed.

[36] ReThink, Inc. (n.d.). *ReThink - Before the Damage is Done*. Retrieved September 12, 2022, from https://www.rethinkwords.com/
[37] The Redirect Method. *Moonshot*. https://moonshotteam.com/the-redirect-method/
[38] Opt Out application. *Mozilla*. https://addons.mozilla.org/en-GB/firefox/addon/opt-out-tools/
[39] Tune (experimental). (n.d.). Chrome Web Store. Retrieved September 12, 2022, from https://chrome.google.com/webstore/detail/tune-experimental/gdfknffdmmjakmlikbpdngpcpbbfhbnp?hl=en

Q27: For purposes of transparency, what type of information is useful/not useful? Why?

In the following points, we highlight some key considerations we would recommend using in order to choose and define appropriate transparency metrics: (1) what is being actioned, (2) how it is being actioned and (3) when it is being actioned. Lastly, we consider the "what then" of transparency metrics in the risk of unintended consequences and misaligned incentives, and how the process can be continually improved and iterated upon.

What is being actioned:

- **Covering different levels of reporting is important for transparency.**

  Actions taken by user-to-users services can be analysed at the content-level or user-level. Content-level metrics cover actions taken for specific pieces of online content such as tweets, videos, comments, posts or conversation threads. Appropriate metrics include both those already reported by various in-scope services[40], such as the number of flagged pieces of content, number of removed pieces of content (takedowns) as well as the addition of metrics relating to other technical solutions (discussed in our response to Q11 and Q18) to decrease the engagement with a specific piece of content without removing it. User-level metrics cover actions taken for specific accounts, usernames or groups of users. Appropriate metrics include number of temporary suspensions, number of bans, the number of followers offending accounts have, or percentage of repeat offenders after a first account action. These metrics are not to be understood as policing individual pieces of content, but rather as painting a picture of, if in the aggregate, there are sufficient systems and processes in place to protect users from illegal content.

- **Any metric should be broken down by harm category and medium.**

  There are already examples of major user-to-user platforms including a breakdown between various harms across metrics such as removals.[41] We encourage the mandatory inclusion of these under the Online Safety Regime' transparency reports. In addition, we suggest including a breakdown by medium, as different mediums pose different severity and immediacy of risk - for example, live-streaming versus asyncronous sharing in child abuse imagery.[42] Mediums of

[40] *Rules Enforcement—Twitter Transparency Center*. (2022, July 28). Twitter. https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec
[41] *Rules Enforcement—Twitter Transparency Center*. (2022, July 28). Twitter. https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec.
[42] *E.2: Challenges posed by live streaming*. (2020, March 3). IICSA. https://www.iicsa.org.uk/reports-recommendations/publications/investigation/internet/part-e-live-streaming/e2-challenges-posed-live-streaming

exchange include text, images, video, livestream and audio, and some multimodal mediums like text-image content in memes.

How actions are taken:

- **Metrics should indicate the balance between proactive and reactive content moderation.**

  The scope for user harm depends in part on the public visibility of the content. Platforms should report how much content was automatically actioned before posting (proactive) versus after being posted and flagged by users (reactive).

- **Platforms should evidence the functioning of reporting systems.**

  Platforms often heavily rely on user reporting or user flags as a signal.[43] Thus, assessing the successes and failures of the user reporting mechanism is vital. For example, the amount of content which is reported by users and how much of this content was then actioned. Evidence has shown the majority of Twitter user flags are not upheld due to trolling or misuse of the flagging mechanism. [44]

  Platforms could also report metrics proxying the efficiency and usability of the flagging mechanism e.g. the number of clicks, how many subcategories of flags are available and how often each of these are used, as well as the average time spent from creating to submitting a report. Evidence has shown platforms often rely on a small number of super-flaggers, resembling the 90:10 power law in online platform participation.[45] So, platforms should report what percentage of reports come from what percentage of users.

- **Platforms should report the functioning of decision-making systems to Ofcom.**

  It is important to assess the successes and failures of content and user-based decisions. For example, platforms should report (1) the amount of content or number of accounts which were actioned erroneously and "restored" after an appeal or review process[46] (false positives) and (2) the amount of content or number of accounts which were not actioned erroneously and instead subsequently flagged (false negatives). We also urge that these metrics are broken

[43] Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[44] Matias, J., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). *Reporting, Reviewing, and Responding to Harassment on Twitter* (SSRN Scholarly Paper No. 2602018). https://papers.ssrn.com/abstract=2602018
[45] Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[46] *Restored content | Transparency Centre.* (2022, January 19). Facebook. https://transparency.fb.com/en-gb/policies/improving/restored-content-metric/

down by language, as automated content moderation is known to perform significantly worse on non-English non-Roman or low-resource languages.[47]

Research conducted at The Turing has also illustrated the importance of functional model testing and error analysis.[48] [49] Functional testing enables more targeted model diagnostics, such a reporting requirement would therefore entail that user-to-user services share with Ofcom known weaknesses in their decision-making systems. In-scope services should address how these system-flaws are being mitigated, and Ofcom track the progress of the mitigation efforts. Note that we recommend that this information only be shared with Ofcom to avoid malicious exploitation of known weaknesses.

When actions are taken:

- **We suggest the collection of metrics on time taken before harmful content is actioned.**

  The more users that see illegal content or the longer the potential exposure time, the greater the potential scope for harm. The need for speed in removing "evidently unlawful" material is acknowledged in German Law, where such content must be removed within 24 hours.[50] Thus, it is important to collect metrics on the "time decay" of actions. Similar to a half-life graph for nuclear decay, we suggest reporting the % of content taken down (y-axis) versus the time after it was published or flagged (x-axis). Other metrics such as time from submission of a flag report to a decision, or time from appeal of a decision to a resolution could also be valuable information for how sharp platform's reactions are.

- **Transparency reports need a long enough time horizon to smooth across events.**

  Our own research on abuse towards footballers on Twitter has demonstrated that offline events correlate to peaks in online abuse.[51] Thus, any metrics should be

[47] Röttger, P., Seelawi, H., Nozza, D., Talat, Z., & Vidgen, B. (2022). Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models. *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 154–169. https://doi.org/10.18653/v1/2022.woah-1.15

[48] Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 41–58. https://doi.org/10.18653/v1/2021.acl-long.4;

[49] Kirk, H. R, Vidgen, B., Rottger, P., Thrush, T., & Hale, S. (2022). Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-Based Hate. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1352–1368. https://doi.org/10.18653/v1/2022.naacl-main.97

[50] Schulz, W. (2018). Regulating Intermediaries to Protect Privacy Online – The Case of the German NetzDG. In M. Albers & I. Sarlet (Eds.), Personality and Data Protection Rights on the Internet. https://papers.ssrn.com/abstract=3216572

[51] Vidgen, B., Chung, Y.-L., Johansson, P., Kirk, H. R., Williams, A., Hale, S. A., Margetts, H., Röttger, P., & Sprejer, L. (2022). *Tracking abuse on Twitter against football players in the 2021-22 Premier League*

reported in various time windows e.g. quarterly, biannually, annually to ensure smoothing over short-periods of more or less violating content.

- **External events can require changes to the platform's internal regulation.**

  The online safety regime is one which relies heavily on platforms upholding and consistently enforcing their Terms of Service. Changes and edits made to these documents have great implications for how the services are run, how they comply with the overarching regime, and under which terms users agree to use their services.[52] Therefore, we suggest that transparency measures should be added which track the edits made to the Terms of Service. In addition to in-scope services accounting for significant changes made to their Terms of Service, we also suggest collecting version tracking metrics similar to software like .git, metrics such as how many character edits, additions and/or removals have been made to these documents during the reporting window.

Unintended consequences of transparency reporting:

- **Specifying a set of metrics may lead to platforms over-focusing or gamifying their actions.**

  We are cautious of the unintended consequences that transparency reporting might bring where the metrics in turn endogenously change platform behaviour - akin to "Goodhart's law" which states that "when a measure becomes a target, it ceases to be a good measure". Having a standardised and fixed reporting criteria may create a problem of misaligned incentives, whereby platforms focus only on optimising these metrics and design manual or automated systems accordingly, at the expense of non-reported metrics.[53] For example, asking for specific metrics such as number of users banned or pieces of content removed may incentivise excessive banning or removal and encourage false positives at the expense of user freedom of speech.[54] Having a set of transparency metrics could also lead to 'trust and safety washing'[55]; where in this case, an over-emphasis of success on the reported metrics may result in an underspecification of errors and failures.

---

season (p. 37). The Alan Turing Institute. https://www.ofcom.org.uk/__data/assets/pdf_file/0019/242218/2021-22-tracking-twitter-abuse-against-premier-league-players.pdf

[52] Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

[53] Thomas, R., & Uminsky, D. (2020). The Problem with Metrics is a Fundamental Problem for AI. Ethics of Data Science Conference. https://doi.org/10.48550/arXiv.2002.08512

[54] Llansó, E. J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society*, *7*(1), 2053951720920686. https://doi.org/10.1177/2053951720920686

[55] Zalnieriute, M. (2021). "Transparency-Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism. University of Illinois Journal of Law, Technology & Policy, 2. https://papers.ssrn.com/abstract=3805492

- **Public-facing transparency reporting may lead to unintended user behaviours.**

  A focus on transparency reporting and content or user actions may lead to users self-censoring their online activity, known as the "chilling effect" on freedom of speech.[56] Furthermore, we are wary that public transparency reporting might unintentionally inspire malicious groups and users to flood underground or niche

  platforms that hold less transparent or less stringent policies.[57] The attention given to the exsistence of illegal content might also inspire actors to flood platforms with more of it, similar to the media incentive related to public acts of violence, for example terrorist attacks.[58]

Suggestions to improve the transparency reporting process:

- **All metrics should be reported in a standardised and machine-readable format.**

  It is essential that reports produced by the user-to-users services in scope of the regime, as well as subsequent Ofcom reports, are machine readable and published in a standardised format. This is essential so they can be meaningfully compared and analysed in the aggregate or disaggregate.[59] We recommend that the data is shared in a common data format such as CSV or JSON files. The top-level metrics would be required by all responses, with additional logic flow for further reporting - for example, in a nested JSON, the top-level dictionary should be filled out by all platforms, with subsequent nesting used for optional or specific categories.

- **Continual iterative, critical and reflexive evaluation is needed from Ofcom.**

  It is likely that it will take multiple rounds of transparency reporting to assess which metrics are most suitable and efficient to track how platforms deal with specific harmful content. We therefore strongly urge Ofcom to assess the metrics regularly, especially as platforms and content evolves.

- **Opening up platforms to researcher access encourages greater external scrutiny.**

  One way to improve transparency is to allow third-party analysis and audit of content and actions from in-scope services; we therefore strongly support the

[56] Brown, A. (2017). What is hate speech? Part 1: The Myth of Hate. *Law and Philosophy*, *36*(4), 419–468. https://doi.org/10.1007/s10982-017-9297-1

[57] Plucinska, J. (2018, February 7). Hate speech thrives underground. POLITICO. https://www.politico.eu/article/hate-speech-and-terrorist-content-proliferate-on-web-beyond-eu-reach-experts/

[58] Jetter, M. (2017). The effect of media attention on terrorism. *Journal of Public Economics*, *153*, 32–48. https://doi.org/10.1016/j.jpubeco.2017.07.008

[59] Wagner, B., Rozgonyi, K., Sekwenz, M.-T., Cobbe, J., & Singh, J. (2020). Regulating transparency? Facebook, Twitter and the German Network Enforcement Act. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 261–271. https://doi.org/10.1145/3351095.3372856

recommendation made by the Joint Committee on the Draft Online Safety Bill on independent researcher access.[60] Improved researcher access has the opportunity to create a healthier online environment by allowing expert voices to conduct rigorous research on platforms systems and processes. The research community could play a supportive role to Ofcom and their duties as the online harms regulator through early detection trend analysis, and scoping the prevalence and dynamics of potential online harms. However, we suggest that guidance is issued to ensure ethical and safe research into online harms while protecting the wellbeing of researchers and data subjects.[61] We recommend services to release API endpoints, similar to those made available by Twitter and YouTube, which Ofcom in turn could monitor for metrics such as number of API calls, number of API endpoints or the number of unique API users.

[60] *Joint Committee on the Draft Online Safety Bill | Draft Online Safety Bill | Report of Session 2021–22* (2021, December 14). UK Parliament. https://committees.parliament.uk/publications/8206/documents/84092/default/

[61] Derczynski, L., Kirk, H. R., Birhane, A., & Vidgen, B. (2022, April 29). Handling and Presenting Harmful Text. http://arxiv.org/abs/2204.14256

turing.ac.uk
@turinginst