The Alan Turing Institute, Public Policy Programme

Submission to the Joint Committee on the Draft Online Safety Bill

Dr. Bertie Vidgen & Professor Helen Margetts
The Alan Turing Institute

The Alan Turing Institute is the UK's national institute for data science and AI. Its mission is to make great leaps in data science and artificial intelligence research in order to change the world for the better. The Alan Turing Institute undertakes research which tackles some of the biggest challenges in science, society and the economy. For more information contact Dr. Bertie Vidgen.

The Online Safety Team is a new part of The Turing's Public Policy Programme. We provide objective, evidence-driven insight into the technical, social, empirical and ethical aspects of online safety, supporting the work of public servants (particularly policymakers and regulators), informing civic discourse and extending academic knowledge. High-priority issues include online hate, harassment, extremism, misinformation and disinformation. We have three main workstreams.

1. **Data-Centric AI for Online Safety**: building high-performing, robust, fair and explainable tools for detecting content and activity that could inflict harm on people online.
2. **Building an Online Harms Observatory**: a real-time monitor to understand the scope, prevalence, dynamics, impact and motivations behind content and activity that could inflict harm on people online.
3. **Policymaking for Online Safety**: understanding the challenges faced by policymakers, regulators and others in ensuring online safety, supporting the creation of ethical, responsible and innovative solutions.

Recommendations to the Joint Committee

1. The language of hazards, harms and risk factors should be adopted in the Online Safety Bill as the term "harms" has become unclear and is likely to cause confusion.
2. Platforms should report publicly what criteria they use to decide whether content and activity is likely to inflict harm. They should also report what interventions they apply once such content/activity has been identified.
3. Ofcom should be given powers to require platforms to explain how their technology for finding hazardous online content and activity is created, evaluated and deployed. This could include opening up platforms' technologies to external scrutiny by experts.

Evidence submission

1. What is an online harm?

We broadly welcome the shift from "online harms" to "online safety" in UK policymaking and regulation as it emphasises what we want to encourage (safety), rather than what we want to reduce (harm). However, it raises the longstanding problem of definitions and "what is a harm?"[1] Currently, the term "harm" is used to describe both (a) online activity and content and (b) its negative effects, such as inflicting emotional or physical suffering. This is unclear and could cause real confusion. We advise that the singular conflicted use of "harm" is replaced with three terms.

1. **Hazards:** Online content and activity which creates a risk of harm. This includes hate, harassment, extremism, grooming interactions, terrorist activity and others. In contrast to the Online Harms White Paper, the Draft OSB makes little mention of specific hazards (outside of terrorism, Child Sexual Exploitation and Abuse and misinformation). This omission makes it more difficult to fully understand which types of activity and content will fall under the OSB, outside of illegal online content and activity.
2. **Harm:** The negative impact on a person's wellbeing, which can be (a) internalised (affecting how a person thinks or feels) or externalised (affecting how a person acts) and (b) short-term or long-term in effect. Harm can take many forms, including physical harms, emotional harms, relation harms (e.g. reputational damage) and financial harms.[2] We recommend using the verb "inflicts" to describe how hazards affect victims (e.g. "bullying inflicts emotional harm on people who are targeted") because it makes clear both their blamelessness and its seriousness.
3. **Risk factors:** Aspects of how online content and activity are produced which increase the likelihood that harm will be inflicted, such as the socio-technical features of the host platform. In our previous work on online hate we examined: (1) the medium, (2) the perpetrator, (3) the actual or potential audience and (4) the communicative setting (e.g. private messaging services vs. broadcast-style social media platforms).[3] This is partly covered by the Draft OSB (OSB: Part 2, Chapter 2, Section 7, 8) but should be described more comprehensively.

In some cases hazard and harm are very closely connected. For instance, each time Child Sexual Abuse Material is shared or viewed it directly violates the rights, privacy and dignity of the abused child, recreating the exeperience of abuse. In this case, the hazard and the harm are exactly the same. In other cases, there is a far larger gap between the online activity and the harm. For instance, the mental health problems caused by bullying and harassment typically emerge from repeated interactions which take place over many weeks and months. The exact impact of each bullying interaction can be difficult to assess as they are insidious and cumulative. Unpicking the longer and more complex pathways from hazard to harm is crucial to fully understand the negative effects of hazardous content and activity.

---

[1] For a longer discussion of definitions, see The Turing's Response to the Online Harms White Paper.

[2] See Scheuerman et al.'s 2021 paper, A Framework of Severity for Harmful Content Online. Our points draw on forthcoming work by Vidgen, Cowls and Margetts (2021).

[3] See The Alan Turing Institute's Report on online hate regulation for Video Sharing Platforms, commissioned by Ofcom.

Most platforms do not have the means or motivation to assess whether hazardous content and activity actually lead to harm being inflicted on users. Instead, they make assessments based on their likely impact. This leads to what we have previously described as a "harms paradox": most online content which is taken down for being harmful may have not actually inflicted any harm, precisely because it has been taken down in anticipation of harm being inflicted.[4] There is no easy solution to this paradox; assessing the hazard of content is intrinsically difficult and often involves making subjective assessments but only moderating content that is proven to have inflicted harm is both undesirable and infeasible. It remains largely unclear how platforms make assessments about hazards, harms and risk factors (and even whether they use frameworks similar to this). The OSB should mandate platforms to make this information public, with heightened reporting expectations for Category 1 platforms.

A clear tension in the Draft OSB is that, in practice, implementing the adults' and children's risk assessments (OSB: Part 2, Chapter 2, Section 7, 9-10) will, given the resource constraints faced by platforms, involve prioritizing some hazards over others, materially affecting which hazards people are protected against. This reality is not fully acknowledged in the Draft Bill. This is a problem as harm prioritization is intrinsically complex, and platforms may face conflicting pressures when making decisions, such as time-sensitivity, public perception, resource constraints, and the likelihood of harm being inflicted. Deciding what hazards to tackle will involve making numerous normative decisions which are likely to be contested and controversial. How, for example, should a platform compare hate speech against bullying? Or suicide and self-harm ideation against doxxing? We advise that platforms are mandated to not only explain how they define hazards and harms (see above), but also to explain their rationales behind comparative judgements, i.e. hazard prioritization, as well as any protocols for making decisions in exceptional periods of high pressure.

The law is an appropriate starting point for hazard prioritisation but, given its inconsistency and narrow scope, will not fully resolve this issue.[5] This raises another point of tension in the Draft OSB: Codes of Practice are only provided for illegal terrorism and Child Sexual Exploitation and Abuse (OSB: Part 2, Chapter 4, Section 29, 2-3). We welcome both Codes but advise that similar Codes are developed for all hazards which can be punishable by law. The Carnegie Trust, working with a network of civil society and other organisations, has already developed an initial Code for hate crime.[6]

[4] See Turing, Report on online hate regulation for Video Sharing Platforms.
[5] See The Law Commission, Modernising Communication Offences: A final report
[6] See Carnegie Trust, Draft code of practice in respect of hate crime and wider legal harms

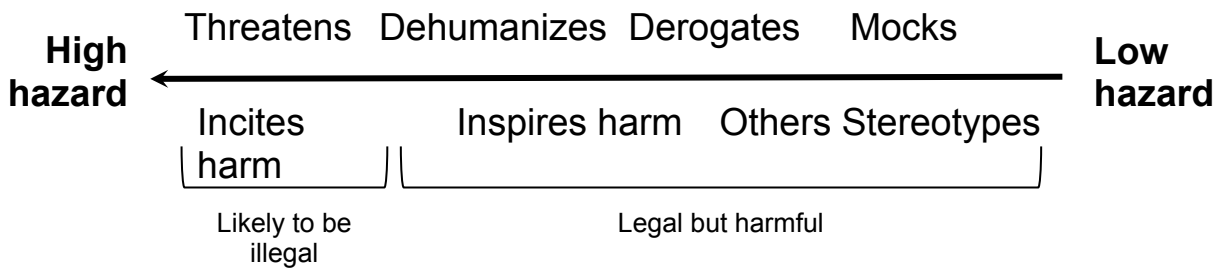
2. Policies and processes for online safety

Hazardous online content and activity is generally tackled by content moderation systems, implemented and managed by the host platform. The Draft OSB references a two-step process, stating that under the "duties to protect adults' online safety" platforms have to ensure that they have clear and accessible Terms of Service, and apply them consistently (OSB: Part 2, Chapter 2, Section 11, 3). This is a high-level approximation of the content moderation process, which we translate into three steps: (1) set the policy, (2) identify content/activity that contravenes the policy and (b) handle the contravening content/activity. A weakness of the Draft OSB is focusing too much on how policies are set, rather than how they are implemented. We illustrate the importance of all three steps, and how they are interrelated, with a short discussion of how a hypothetical platform might handle online hate.

2.1 Setting the policy

The platform first creates the policy. Outside of illegal content and activity, platforms set policies based on their social values and the proposition that they offer to users. For example, Gab, Bitchute and Parler emphasize free speech and, accordingly, implement minimal content moderation. In practice, all platforms have to balance the tradeoff between (i) protecting users from harm and (ii) protecting users' freedom of expression. This is essentially a decision about what degree of hazard is acceptable. For instance, prioritising protecting users from harm over protecting free expression means that even low hazard content will be considered unacceptable (see Figure 1). The spectrum of hazard tolerance should be reflected in the OSB, instead of treating online safety as a binary question of whether or not a platform is "protecting from harm".

Figure 1, The hazard of different types of online hate



2.2 Identifying contravening content

Hazardous content and activity is typically identified by platforms through a mix of human- and AI-based approaches (see below). If the systems for identifying hazardous content are inadequate then the platform's desired tradeoff between protecting from harm and protecting free expression will not be achieved, leading to one of two outcomes[7]:

[7] For a longer discussion of challenges which may lead to lower precision or recall in hazardous content detection, see our previous paper on Challenges and Frontiers in Abusive Content Detection.

1. Over moderation from low precision: if the moderation system flags too much content as hateful then users' free expression will be unduly restricted. This is a particular concern with monitoring systems that have time-based targets, such as the EU's Code of Conduct on Online Hate which creates a 24-hour target for handling hate. There is a fear that such targets incentivise over-moderation to avoid legal penalties.
2. Under moderation from low recall: if the moderation system cannot identify enough of the hate then it will fail to adequately protect from harm. With high priority hazardous content, such as illegal forms of hate, failure to identify presents a serious risk of harm. It is crucial that moderation systems prioritise the most damaging forms of content and activity.

The Draft OSB gives comparatively little attention to the processes by which contravening content is identified. Ofcom should be given powers to interrogate the processes used by platforms, as we discuss below in relation to the use of technologies such as Artificial Intelligence.

2.3 Handling contravening content

Once contravening content has been identified it needs to be handled. In our previous work we identified 14 interventions which are known to be applied by platforms to hazardous content, including suspending content, banning the creators, limiting how many times content can be engaged with, and limiting how many times it can be shared.[8] Each of these interventions apply different amounts of *friction* to constrain the content's spread and impact.

The Draft OSB requires that platforms explain "how priority content that is harmful to adults is to be dealt with" (OSB: Part 2, Chapter 2, Section 11, 2). This needs to be better specified, and we advise that platforms are required to explain (a) which interventions they use to apply friction and (b) what criteria is used to justify different types of friction being applied. In principle, we advise that the degree of friction is based on the assessment of hazard, with content and activity assessed to be more hazardous being subjected to more friction -- although platforms may have other well-justified rationales. We also advise that Category 1 platforms should be mandated to assess how different interventions affect users' freedom of expression and privacy, as part of their requirements under Part 2, Chapter 2, Section 12 of the Draft OSB.

[8] See Turing, Report on online hate regulation for Video Sharing Platforms.

3. Artificial Intelligence for online safety

Computational technologies, particularly Artificial Intelligence (AI), have been lauded as a way of achieving online safety because they are scalable, reproducible, improvable and fast. AI reduces the number of human touchpoints in the moderation process, minimising how many hazards trust and safety professionals are exposed to -- a growing concern given they routinely suffer from mental health problems and are subjected to poor work practices.[9] AI is most often used to find hazardous content, and can also be used to automatically decide how content should be handled.

Our research identifies numerous problems when using simplistic or badly-designed AI to detect hazardous content, three of which we outline below.[10]

1. AI struggles to understand context, such as the content of previous posts in conversational threads, the broader social or historical setting and the norms of different communities.[11]
2. AI can be easily tricked. In our recent work, in collaboration with the University of Oxford, we showed that AI moderation tools for abusive content are easily tricked when emojis are used.[12] More broadly, a range of complex and unusual forms of content are likely to avoid detection.
3. AI can be biased, performing unevenly across different groups and different types of content. This can lead to serious problems when models are deployed in the real-world, reducing performance and introducing new sources of societal unfairness.[13]

As far as we can identify, the Draft OSB makes no mention of Artificial Intelligence, which we understand is so that the Bill does not go out-of-date and does not implicitly encourage use of a single technology. This is sensible but raises a clear problem given that *how* content is moderated matters, and human- and AI- based moderation systems are fundamentally different. To better understand the challenges of using AI (a very fast-moving area of research) Ofcom should be given specific powers to investigate and scrutinise the technologies that platforms are using for content moderation, with a mandate for AI.

Finally, we note that assessments of how AI is used in content moderation are by nature partial and incomplete as almost all platforms do not open up their models for external evaluation. Most research focuses on proxies; either academic models, which are trained on open source datasets, or the Perspective API from Google's Jigsaw. There are powerful commercial and logistical reasons for not fully opening up all AI for scrutiny. It constitutes IP, reflecting tens of millions of dollars of investment, and often relies on some signals which cannot be publicly shared, such as users' social network graphs. Nonetheless, if we do not identify ways of scrutinising the AI used by platforms then we will continue to not understand a key part of the content moderation process. Potential solutions which could be explored include use of TREs, releasing models in a sandbox, and granting permissioned access to simplified versions of models.

---

[9] In 2020 it was reported that Facebook agreed to a $52 million settlement with its moderators.

[10] For more discussion of challenges of abusive content detection see Learning From the Worst and Challenges and Frontiers.

[11] See CAD: Introducing the Contextual Abuse Dataset.

[12] See Hatemoji, lead authored by Hannah Rose Kirk.

[13] See HateCheck, lead authored by Paul Röttger.