Twitter UK
20 Air Street
London
W1B 5AN

twitter.com

Summary

We have welcomed the opportunity to participate in the Online Safety Bill consultation process over recent years and, in particular, the systems and processes approach aspired to in the Bill. We are also supportive of the designation of Ofcom as the regulator. However, there are areas where we do have concerns and feel further consideration must be given if the Bill is to be implementable and effective in prioritising the safety of those online. Below are five suggestions that we believe will help:

- **'Priority harms' should be detailed in the Bill, not in secondary legislation.** The lack of clarity on content that is harmful to children and to adults is making scrutiny near impossible on what speech will and will not be regulated. The lack of clarity is also highly likely to cause delays, as companies struggle to prepare in the absence of these harms being defined. From a freedom of expression perspective, establishing categories of legal speech on which companies will be expected to act must be considered in primary legislation.
- **The Secretary of State's powers to direct Ofcom on the detail of its work should be removed.** The specific power to modify the draft codes of practice for 'reasons of public policy' will undermine Ofcom's ability to credibly deliver independent, evidence-based regulation. This is among additional powers envisaged that will cause further political interference in the regulatory process.
- **Exemptions around journalistic content, news publisher content, and content of democratic importance should be removed.** We have concerns as to whether the definitions at present are workable in practice; and risk creating loopholes that bad faith actors could exploit to share harmful content online - which would be the antithesis of the Bill's online safety objective. While tighter definitions may help, this requirement in principle is likely to be incompatible with automated content moderation, and may ultimately lead to slower removal of content that breaks regulated companies' rules.
- **On user verification and empowerment, we believe that the Bill should focus on outcomes - that companies give users controls that enable them to filter out harmful content, without prescribing one specific solution of identity verification.** We would suggest that identity and verification could be looked at by an Advisory Committee of Ofcom, similar to as envisaged to review disinformation and misinformation (Clause 130). **If the current provisions, however, are maintained - both a technical feasibility assessment and an impact assessment of the user**

verification and empowerment requirements must take place immediately to enable Parliamentary scrutiny; as well as the government providing a detailed outline of how they see this duty working in practice. We support the government's objectives of giving users more choice and control. However, we have questions on how this specific provision would work in practice - and apply territorially. Failure to consider these challenges is likely, at the very least, to cause delays; and could see substantial resources invested to develop tools that fail to achieve the overall policy objective and/or substantially affect the online experience for social media users.

- **Categorisation needs to account for risk - and not just size - of a platform.** It is imperative that less established or well-known companies, who may be hosting profoundly harmful content - but may not receive public complaints or attention, or indeed make data available for research - are captured in the top tier of category. It is important to remember as the regulatory debate evolves, the internet is more than four companies and we need an approach to policymaking that reflects that.

We also remain concerned about the **criminal liability and business disruption powers given to Ofcom** in the Bill. If there is a question of how the UK leads the world in protecting the open internet and designing proportionate and effective regulation, we must carefully consider the precedent that will be set. People around the world have been blocked from accessing Twitter and other services by multiple governments under the guise of online safety, impeding peoples' rights to access information online. 10% of annual global turnover is a profound financial sanction; and, as demonstrated by the GDPR (which has a penalty of 4% of annual global turnover), financial sanctions alone are sufficient to drive compliance across the sector.

Earlier this month, Gavin Millar QC produced the first analysis of the implications of the Online Safety Bill on UK citizens' freedom of speech. He found the Bill will significantly curtail freedom of expression; **and does not comply with Article 10 of the European Convention on Human Rights.** We also note that no NGOs focused on freedom of expression were invited to give oral evidence to the Committee, despite Parliament's own Joint Committee on Human Rights writing to DCMS in May to raise a number of concerns.

Overall, we remain concerned that issues around technical feasibility, legal clarity and freedom of expression are being overlooked, which may lead to challenges for Ofcom as regulator and for companies who want to comply.

Position

While we are committed to ensure that there is healthy discourse on our platform, we continue to believe deeply in, and advocate for, freedom of expression and open dialogue - but that means little as an underlying philosophy if voices are silenced because people are afraid to speak up. With this in mind - we welcome the government's focus on online safety, and this Committee's work to consider the draft Online Safety Bill.

As debate around the world focuses on how to solve public policy challenges related to the technology industry, our approach to regulation and public policy issues is centred on protecting the Open Internet. We define the Open Internet as a global and singular internet that is open to all and promotes diversity, competition, and innovation.

We believe that the Open Internet has driven unprecedented economic, social and technological progress, and while not without significant challenges, it has also led to greater access to information and greater opportunities to speak that are now core to an open society.

We support smart regulation that is forward thinking, understanding that a one-size-fits all approach fails to consider the diversity of the online environment, and poses a threat to innovation. We have welcomed the opportunity to participate in the Online Safety Bill consultation process over recent years.

Our view is that regulatory frameworks that look at system-wide processes with clear parameters, as opposed to individual pieces of content, will be able to better reflect the diversity of our online environment and the challenges of scale that modern communications services involve - and we are therefore supportive of the government's original stated commitment to this approach.

Similarly, we welcome Ofcom's designation as the regulator for Online Harms. As we stated in our submission to the White Paper back in 2019, we think that Ofcom is the most appropriate and qualified body to be designated as the independent regulatory authority.

While our submission is focused on areas of the Bill where we believe change is necessary to achieve the stated objectives, we remain supportive overall of the government's ambition to improve online safety.

Suggested changes

Priority harms

The Bill in its present form fails to achieve a key objective: providing clarity to UK internet users - and providers - on what online speech should and should not be considered harmful. What's more, as Ellen Judson, senior researcher at Demos has stated: "*This Bill is a jigsaw: not only internally complex, but so much of what it means for the world relies on things that don't exist yet - secondary legislation, Codes of Practice - that can't even exist until the Bill is finalised, which makes scrutiny difficult.*"

Fundamentally, the consequence of this approach is confusion for internet users on what speech will and will not be permitted, a significant lack of clarity for service providers, the potential for freedom of expression to be curtailed - and delays on implementation as we await Ofcom or secondary legislation filling in these critical details.

In other areas of law, the challenges posed by overly vague definitions are well-documented (see, for example, the 2018 Law Commission report). Clear definitions are critical to avoid ambiguity, help those within scope fully understand what is required to comply with the law and, crucially, ensure that UK internet users know exactly what is and is not permissible - while trying to protect freedom of expression and ensure that the reaction from service providers is not just to remove large amounts of content for fear of being in breach (or alternatively, face penalties or litigation as a result of incorrect actioning).

We believe these problems would be alleviated by detailing in the Bill - not secondary legislation - the priority harms for children and adults, perhaps as a new Schedule 7(a) and 7(b) (as suggested by the Carnegie Trust). This would enable full scrutiny of the regulatory proposals, provide legal clarity on fundamental issues of speech, and reduce business uncertainty while aiding preparation.

It is, however, also worth highlighting the additional challenge of how platforms can comply with the illegal content duties (Schedule 7). It is hard to see how providers, and particularly automated moderation tools, will be able to determine whether content on their services falls on the legal or illegal side of some of the categories outlined. Even the UK Independent Reviewer of Terrorism Legislation Jonathan Hall QC recently publicly criticised the Bill as being "ineffective on terrorism". He said "It's hard to see how it provides a workable framework for regulation." **Given illegal content duties represent some of the most harmful content online, we would suggest greater detail of how to comply with these duties should be included in the Bill.**

Secretary of State's powers

The issues above are further complicated by the discretion given to the Secretary of State in the Bill to modify codes of practice for 'reasons of public policy' (Clause 40). One can imagine a code of practice is developed by Ofcom via an evidence-based, consultative process; if this

code is overridden by the Secretary of State, this both undermines Ofcom's credibility and creates the risk of regulated companies having to comply with requirements that are either less effective or are representative of political interference. As Ofcom stated in oral evidence on 24 May:

"We feel it is important that the independence of a regulator can be seen to be there and is there in practice … We must be able to show why and how we have created those codes of practice, so that we can be accountable and there is absolute clarity between regulator and Government."

The Secretary of State has a range of additional powers envisaged in the Bill. In particular, the ability to set out what Ofcom's strategic priorities should be "relating to online safety matters" (in theory, potentially overriding Ofcom's evidence-based process on the harms of most concern). In addition, the Secretary of State has powers to direct Ofcom under specific circumstances around ill-defined threats to 'the health or safety of the public' or 'national security' (Clause 146); and to issue guidance to Ofcom about its exercise of their functions and certain of their powers under the Act (Clause 147).

There is a clear consensus across stakeholders that the powers of the Secretary of State are problematic; this was also highlighted by the Draft Online Safety Bill Committee in 2021. **The Secretary of State's powers to direct Ofcom on the detail of its work should be removed.**

Exemptions

We, alongside several civil society organisations and other technology companies, have raised both technical and philosophical concerns with the notion of exemptions in the Bill (Clauses 15 and 16) since they were first put forward. Unfortunately, these concerns have not been addressed.

At the very least, the proposed exemptions require far greater clarity. From a philosophical point of view, though well-intentioned, they may have unintended consequences. For instance, would the 'content of democratic importance' exemption create a loophole that people suspended from Twitter would be able to challenge their suspension if they ran for election or established a political party?

'Journalistic content' is equally ill-defined. Every day we see Tweets with screenshots of newspaper front pages, links to blogs, updates from journalists and firsthand accounts of developing events. Crucially, there are accounts we have suspended for breaking our rules who have described themselves as 'journalists.'[1] Similarly, we have previously seen examples where journalistic content has included visible links to terrorist material, such as that produced by ISIS. Indeed, after the Christchurch mosque shootings, a number of news organisations broadcast the attacker's videos in full. Is the expectation that services should not remove this content? The

[1] The term 'journalism' has been held to have a wide ambit (see, for example the Supreme Court decision in Sugar v. BBC [2012] UKSC 4

lack of detail around these provisions risks significant confusion and potentially undermines the overall objectives of the Bill.

Of particular concern is the recent proposal from the Secretary of State to require companies to leave 'news publisher' content online - even if it clearly breaks our rules - while an appeals process is ongoing. The challenge is the criteria outlined to qualify as a 'news publisher' under the Bill - a threshold that is not difficult to meet. Again, this could mean not only that harmful content is left online for periods of time as an appeals process is ongoing, but that bad faith actors exploit this loophole to meet the definition of a news publisher in order to post clearly violative or potentially illegal content online.

Beyond the philosophical concerns, there are serious workability questions. The Bill requires us to be able to a) identify content that fits into these categories of exemption, and b) treat it differently to other types of content. As noted above, with the definitions as vague as they are, it would take a great deal of nuance to try and separate out such content. It would also take investigation into the poster, including whether they have "their principle purpose the publication of news-related material", whether they had a code of practice, complaints mechanism, or a name on their website (Clause 50); and for journalistic content whether "the content is generated for the purposes of journalism" (Clause 16(8)(b)). In other words, one is having to look at the nature, intentions and circumstances of the poster/entity behind the post rather than just the content itself, which will create very substantial difficulties for platforms.

Even then, companies may still have limited confidence on whether we have done so correctly, and potentially face sanctions or penalties, and/or litigation, as a result. There is no algorithm that can do this. Once identified, the extra step of ensuring "that the importance of the free expression of journalistic content is taken into account" is equally ill-defined.

In sum, this may lead to subjectivity and, ultimately, delays in taking down harmful content. We see about half a billion Tweets posted every single day - in order to scale enforcement, therefore, technology and automation is essential. Now, a majority of abusive content we take down is detected proactively. When it comes to terrorism and Child Sexual Expoitation, 93% and 89% respectively of content we removed was detected proactively. It is challenging to envisage at this point how we can effectively reconcile our automated tools with such vague, contextual and subjective exemptions.

Clause 19 already details duties about freedom of expression and privacy. **Given the potentially counterproductive nature of these exemptions - and the challenges of how this would work in practice - they should be removed. Instead, companies should continue to consider public interest and freedom of expression in their policies and processes overall, as set out in Clause 19.**

User verification requirements

The question of how to manage and verify your identity online, whether that means using your real name, checking an ID or using emerging technologies like blockchain is a debate that's happening right now - there are no easy solutions, but it is clearly an issue that is at the heart of the future of a range of industries. Consideration in this context must also be given to alignment with existing regulation, such as the ICO's Age Appropriate Design Code.

Our view is that users should have the ability to choose which Twitter accounts can interact with them. For example, at present, the following options are available on Twitter:

- *Protected account* - When you sign up for Twitter, you can choose to keep your Tweets public or protect your Tweets. If you protect your Tweets, you will receive a request when new people want to follow you, which you can approve or deny. Your Tweets will only be visible to approved followers.
- *Control who can reply to your Tweets - or turn replies off* - When you compose a new Tweet, you can now choose who will be able to reply to it. You'll see a default setting of 'Everyone can reply' next to a globe icon in the 'Compose Tweet' box. Clicking or tapping this prior to posting your Tweet allows you to choose who can reply to you - if you select 'only people you mention' and choose not to mention any other Twitter user, then nobody can reply to your Tweets (although Quote Tweets will still be possible). After you Tweet, you can change who replies by navigating to the top right of the Tweet and tapping the 'More' three dot icon. Tap 'Change who can reply' from the list of options and select who you'd like to reply to your Tweet.
- *Filter accounts* - Filter the types of accounts you see in your notifications timeline. You can access filter settings by navigating to 'Settings and Privacy' > 'Notifications' > 'Filters' > 'Muted notifications.' Here you can mute notifications from a range of people, such as those with accounts who haven't confirmed their phone number or email address, new accounts, accounts who have a default profile or accounts you don't follow.
- *Safety mode* - Safety Mode is a feature that temporarily blocks accounts for seven days for using potentially harmful language - such as insults or hateful remarks - or sending repetitive and uninvited replies or mentions. When the feature is turned on in your Settings, our systems will assess the likelihood of a negative engagement by considering both the Tweet's content and the relationship between the Tweet author and replier. It is currently being trialled.

The Bill envisages that all Category 1 companies must give users the option to verify their identity (Clause 57) - and, if they do so, there is a "duty to include in a service features which adult users may use or apply if they wish to filter out non-verified users" (Clause 14). This requirement was a late addition to the Bill and, as a result, has not received as much consideration as other components.

There are several issues with this proposal at present. First, how this applies territorially. It is unclear whether, for example, if a verified UK user turns such a feature on, the expectation is that 'unverified' users across the rest of the world (as they would be, because this duty only

applies to UK users) would be able to interact with the individual or not. This creates one of two situations. Either, a digital separation from the rest of the world - as only UK-based verified users could interact with such an account; or, perversely, everyone around the world is still able to interact with such an account, but only unverified accounts in the UK cannot. This is further complicated by the industry-wide challenge of technically faking your location (e.g. virtual private networks, VPNs) - it does not take much technical savvy to appear to be in a location that you are not. This means the feature itself could be easy to bypass, despite companies' best efforts.

Second, further consultation could help test how many people would be likely to opt into this feature, and an impact assessment would help understand if any particular communities would be adversely affected. According to a survey from YouGov in April 2022, 78% of the 2,000 adults surveyed would not be willing to verify their age to access adult websites by uploading a document linked to their identity such as a driver's license, passport or other ID card. While this survey was focused on adult websites, it may indicate a general reluctance to share personal data with large service providers more broadly. In addition, some of the communities who may lack access to government IDs are exactly those who we strive to give a voice to on Twitter. Estimates have suggested there are 3.5 million people in the UK who don't have access to official forms of photo ID; it is unclear what form of verification companies would be required to put in place.

If a majority of users do not take up the opportunity to verify themselves, this may be a disincentive for verified users to turn the 'filter' feature on. This is because they could see a substantial decline in engagement, such as the number of Likes, Retweets or Replies to such posts - including from exactly those communities they most want to be able to interact with. Alternatively, if the feature does prove popular, but disproportionately affects certain communities being able to interact with verified accounts, this could reverse one of the most important benefits of social media. Many people who lacked equal access to public platforms 15 years ago - especially the young and members of marginalised groups in particular - use social media every day; and this proposal could limit the accounts they are able to interact with.

One risk, therefore, is that companies invest resources in creating this system as required - but that it substantially affects the experience of many users on social media, and/or such resources could have been focused on other solutions to online harms.

While we share the government's objectives of giving users more choice and control, this particular requirement poses a number of risks at present. **We would suggest instead that the Bill focuses on outcomes - that companies give users controls that enable them to filter out harmful content, without prescribing one specific solution. Identity and verification could be looked at by an Advisory Committee of Ofcom, similar to as envisaged to review disinformation and misinformation (Clause 130).** If the government wishes to maintain this provision, at the very least we would suggest **undertaking immediately both a technical feasibility assessment and impact assessment of the user verification requirements, which may help alleviate these issues and aid Parliamentary scrutiny; as well as the government providing a detailed outline of how they see this duty working in practice.**

Categorisation

In our submission to the Draft Online Safety Bill Pre-Legislative Scrutiny Committee, we argued that it is imperative that Ofcom ensure less established or well-known companies, who may be hosting profoundly harmful content - but may not receive public complaints or attention, or indeed make data available for research - are captured. This is something we also raised in our original White Paper submission in 2019, perhaps suggesting a greater role for technology research organisations.

With this in mind, **categorisation (Schedule 10) needs to account for risk - and not just size - of a platform.** Given the evidence provided by the Antisemitism Policy Trust on this issue during the oral evidence sessions, we have nothing further to add and simply endorse the concerns they and others have continually raised.