Written evidence submitted by Demos

Who we are

Demos is Britain's leading cross-party think tank, with a 25-year history of high-quality research, policy innovation and thought leadership. Our priority is to bring ordinary citizens' voices into policy making.

CASM, Demos' dedicated digital research hub, has unique insights and expertise across tech policy and its impact on our society, economy and democracy. CASM has spent the last seven years developing methods and technology to undertake policy-focussed research on online platforms on which public conversation is taking place.

CASM is also the home of the Good Web Project, a joint project with the Institute for Strategic Dialogue, the Alliance for Securing Democracy and Arena at Johns Hopkins University, to empower the UK and international governments to ensure the future of the internet is compatible with liberal democracy. It seeks to measure and build public support for an internet that resists the authoritarian alternative and empower policymakers to fight for this cause.

Our recent research relating to online harms and digital regulation includes:

- *Online Harms: A Snapshot of public opinion*
- *A Room of One's Own* (a guide to defining and regulating private spaces online)
- *States, Corporations, Individuals and Machines* (on the balance of power online)
- *A Picture of Health: Measuring the comparative health of online spaces* (on how the design of online spaces fuels negative behaviour)
- *Everything in Moderation: Platforms, communities and users in a healthy online environment* (on how current attempts by states to regulate online spaces will in all likelihood fall short)
- *Engendering Hate: The contours of state-aligned gendered disinformation online* (on how disinformation is being used online to exclude women from public life)
- *What's in a name? A forward view of anonymity online* (an approach to how we can protect our identities online)

Our responses draw on this existing body of research, policy and advocacy work into online harms and our wider expertise on the subject.

This briefing has been prepared by Ellen Judson, Alex Krasodomski-Jones, Josh Smith, Ciaran Cummins and Akshaya Satheesh.

How has the shifting focus between 'online harms' and 'online safety' influenced the development of the new regime and draft Bill?

The shift from 'online harms' to 'online safety' has led the draft Bill and regime to have both a peculiarly wide and narrow scope in ways which make successful implementation challenging.

The Bill is being described as making the UK 'the safest place in the world to be online', promising to ' keep children safe, stop racial hate and protect democracy online'. As such, it implies a global

approach to safety, implicitly setting itself the goal of protecting users online from all conceivable harm.

Not only is that an unachievable goal, it is also explicitly in opposition to the text of the Bill itself, which has limits on platforms' duties and excludes certain kinds of harm from scope altogether (such as some kinds of financial fraud, or social/democratic harms arising from disinformation).

Similarly, the text of the Bill poses 'safety' in opposition to rights such as privacy and freedom of expression, which are described as constraints on the pursuit of safety. This fails to consider how measures taken to protect these rights may in fact support greater user safety (privacy measures, for instance, often make users less vulnerable to threats such as hacking or doxxing).

The risk of these conflicting accounts of the purpose and scope of the Bill mean that public expectations of what the Bill can and will achieve may be out of step. Other methods to reduce the risks of harm to users and improve user rights online may also be wrongly deemed superfluous if the Bill is understood as a panacea. This Bill should be part of a suite of measures taken to build a better and more democratic Internet, and not treated as the whole solution in itself.

The Bill should define its objectives much more clearly and be clear on how 'safety' is being understood, in order to better inform implementation and enable the success or failure of the Bill to be measured.

Does the draft Bill focus enough on the ways tech companies could be encouraged to consider safety and/or the risk of harm in platform design and the systems and processes that they put in place?

We are supportive of the intention of the Bill, to introduce a duty of care on platforms to identify and reduce risks of harm on their services by improving their systems and processes. However, we are concerned that the Bill as it currently stands may not achieve this aim.

Firstly, many of the key concepts in the Bill remain undefined, or deliberately left to be defined at a later point in the regulatory process (such as the priority harms, the platforms in scope, what systems platforms will be expected to have in place, and against what metrics compliance will be assessed). This means that scrutiny of the Bill at this stage is essentially an exercise in clarification and assessing how far specific outcomes will or will not result is very difficult.

The protections for freedom of expression and privacy are crucial to be included: but as they stand, they are also at a level of generality that has the potential to allow significant infringements - as platforms are required only to 'have regard to the importance of [these rights] when deciding on, and implementing, safety policies and procedures.'

Secondly, although the Bill sets out a systems-based approach, there is a focus on reducing harm through content takedown measures, measuring the incidence of harms online and a focus on enforcing terms and conditions. Amid the promises that the Bill will 'end' harmful content online, we are concerned that in implementation this will turn into a 'content-based approach' by proxy, by prioritising the regulation of content moderation systems above other systems and design changes.

It is well understood that while certain kinds of illegal content can be policed at a content level, reducing the risk of most harms requires systemic change. Concrete examples in the offline world

include our approach to food and public health.[1] To reduce the harm to people's health from what they eat, we would fail if we just relied on banning toxic substances from foodstuffs: we also have restrictions on food advertising, taxes on certain foods, requirements about labelling food accurately, community initiatives to support people to access healthy food, and so on. Likewise, there is no simple policing solution and no individual content-level solution, to harm arising from hugely diverse and scaled forms of online content.

Moreover, for some forms of illegal content (such as abuse, hate speech &c), it is very difficult to define easily which pieces of content definitely qualify as legal or illegal. Having a content-based approach, therefore, which requires that illegal speech be taken down and legal speech remain up, risks both over-moderating legal speech that gets accidentally taken down, and *undermoderating* extremely harmful speech that are against platforms' terms of service and which users expect to be dealt with, and which cause significant harm to its targets. Both of these risks can be mediated (if not removed entirely) by a systems-based approach that seeks to regulate systems that increase or decrease the risk of users suffering harm, rather than seeking to regulate content by proxy.

The aim of the Bill should be to reduce the risk of harm occurring in the online environments in its scope, rather than simply to reduce a given incidence of harmful content. Focusing only on content removal risks overlooking other systems platforms should be using in a way that protects and supports users, including: reporting processes and resources offered, behavioural nudges, user powers to shape their online experience, support and incentivisation for communities setting their own standards, content interaction and labelling systems, content curation systems and promotion systems, and data collection and tracking systems. A focus on reducing risk also encourages more proactive measures to reduce harm, rather than only retroactive content takedown.

We would hope to see greater specificity in the Bill on the expectations for platforms. This should particularly focus on:

- a) how they will be expected to reduce *risks of harm*
- b) how this risk will be measured beyond incidence of harmful content
- c) how they will be expected to protect rights

This should be set out ahead of legislation being passed so that these expectations can be subject to scrutiny and create greater certainty for users and platforms.

The algorithms used by the service and the design and operation of the service (including the business model, governance and other systems and processes) are already included in the Bill. These could be given much greater prominence and substantiation throughout the Bill.

Are there any contested inclusions, tensions or contradictions in the draft Bill that need to be more carefully considered before the final Bill is put to Parliament?

Platforms in scope

[1] For discussion of a public health approach to online abuse, see Glitch's work on online abuse).

The Secretary of State having the power to set thresholds based on general concepts such as number of users and functionalities, and 'any other factors', means there is significant uncertainty for services about which Category they will be in, and what the expectations of them will be accordingly.

We submit that platforms should be assessed by Ofcom on the basis of propensity of risk of harm arising from a service *rather* than number of users and functionalities (though both of those factors could inform a judgement about risk): both to ensure that low-risk platforms are not overburdened with compliance requirements and that high-risk but low-user, low-functionality platforms do not escape requirements to reduce the risk of harms on their services.

Privacy protections

We are concerned that the privacy protections currently in the Bill are inadequate.

The current duty has several limitations:

- a) it is a duty to 'have regard to' the importance of, rather than a duty *to* protect users from unwarranted infringements of privacy
- b) it is restricted to the implementation of safety policies rather than policies more generally
- c) the lack of definition about the process or framework for deciding when an infringement of privacy is 'warranted' risks privacy infringements being too easily justified.

This last concern is exacerbated by the potential conflict of privacy protection with other clauses of the Bill.

Without further clarification, clauses which require systems designed to present children from accessing certain content, that allow OFCOM to require the use of certain technologies be used to identify illegal content, including in private channels, may lead to platforms removing essential privacy protections under the guise of it being 'not unwarranted', despite significantly undermining user privacy. This is an opportunity for the government to stand up for users' privacy in the face of enormous overreach by corporations.

Freedom of expression protections and democratically important or journalistic content

Having additional protections for democratic content leads to the question of why the freedom of expression protections in the Bill are not already sufficient to adequately protect freedom of expression for political/democratic speech. If they are not, they should be strengthened to better protect expression more broadly.

As currently written, 'freedom of expression' and 'democratic content' are extremely open to interpretation through the Codes of Practice. Currently, the process of content moderation must be 'designed to ensure the importance of the free expression of content of democratic importance is taken into account when making decisions'. Combined with the definition of democratic content as being/appearing as 'intended to contribute to democratic political debate in the United Kingdom', this risks:

Overmoderation of political speech

If this requirement is interpreted merely to mean that there should be some consideration of free expression of democratic speech, then this requirement would likely become a tick-box exercise. For instance, a platform could have an appeal process which allowed political expression as grounds for appeal, which in practice made no difference to permitted expressions online. As a House of Lords Committee has warned, it would also risk privileging speech which is more easily defined as 'political': such as political contributions by politicians or those involved in policy debates; or prioritising freedom of expression about political issues which, for instance, are actively being debated by the Government. This could see political speech by members of the public, political discussion about countries outside of the UK, or speech about wider political and social issues disadvantaged.

Undermoderation of harmful speech

However, if the requirement is interpreted to mean that no content which can be argued to be democratically important may be removed, demoted, or a user banned on account of it, this also has risks for allowing widespread harm to be perpetuated. Since much abuse and disinformation is or can appear to be linked to political issues or political figures (for instance, abuse of women in public life, racist comments about immigrants, transphobic abuse in discussions of e.g., gender recognition), having general exemptions from enforcement of terms and conditions would risk allowing significant harm to be perpetrated. In the US, for instance, we have seen false claims that platforms are 'censoring' certain political viewpoints being used to undermine platforms taking action on extremist or harmful speech online.

The same worries apply to the protections for 'journalistic content'. Without a more precise definition, or expectations clearly spelt out, overmoderation of legitimate journalistic content and undermoderation of extremist or hateful content under the guise of journalism are both significant risks.

We would recommend that the Bill include further details of how platforms should protect freedom of expression in general, and include within that more specific expectations for how platforms should approach the protection of political speech and journalistic speech.

Secretary of State powers

The powers granted to the Secretary of State are significant, and in our view, excessive: the purpose of having an independent regulator, with Parliamentary oversight, is to ensure an independent and democratically legitimate process for regulating platforms. The Secretary of State having extraordinary powers, including being able to exempt certain kinds of services, vary the online safety objectives to be pursued, direct OFCOM to modify a code of practice, specify offences and priority harms, is not consistent with the pursuit of this aim.

In particular, the power to direct a modification of a code of practice to ensure that the code of practice reflects government policy should be removed. This runs a high risk of allowing the government to demand platforms change their policies to benefit the government or to further other government policies which are not effective in reducing the risk of harm to users.

The power to modify a code of practice for reasons of national security or public safety, if retained, should have additional safeguards included (such as requiring judicial oversight) given that in certain

circumstances the Secretary of State is not required to submit reasons for these modifications, reducing the possibility of external scrutiny.

Risk assessments and transparency

The Bill relies on platforms to a significant degree to produce their own credible risk assessments, procedures to deal with those risks, and information through transparency reporting about the success of those measures. How these will be audited, however, is not yet clear. Genuinely reducing risks of harm to users, as opposed to reducing the incidence of harmful content, requires much more analysis than simply numbers of reports, takedowns and appeals, which transparency reports currently often focus on. Though OFCOM has significant information powers in the Bill, to be able to scrutinise compliance from all services in scope will most likely be beyond its resourcing capabilities.

We would strongly recommend that greater priority be given than is in the current Bill to facilitating independent researcher access to platform data, with appropriate privacy safeguards, so that platform action can be better scrutinised and improve accountability for any failures to take meaningful measures to reduce risks of harm.

What are the lessons that the Government should learn when directly comparing the draft Bill to existing and proposed legislation around the world?

We believe that by framing the Bill in terms of what the government wants to see more of, rather than content it wants stamped out, the UK can pave the way in a progressive defence of the open web in the face of its corporate and state opponents, rather than joining the long list of countries whose lexicon is limited to reactive policing of whatever harmful content is currently in vogue.

Proposed Online Harms legislation in Canada bears similarity to the UK proposals, but is a clear example of the dangers of seeking to legislate harmful content rather than harmful systems: and has as such led to criticism that it will unacceptably chill freedom of expression and lead to significant overmoderation. It proposes mandating proactive monitoring and 24h takedown of certain kinds of harmful content, along with requirements for platforms to report users who share it to law enforcement. There is no substantial discussion in the proposals of how it would protect freedom of expression in a context where automated tools for monitoring and takedown are imperfect and often biased against marginalised groups, and the incentives for overmoderation are strong.

Similarly, a recent update to NetzDG in Germany that would require platforms to report users who post what the platform judges to be illegal hate speech has been criticised on the grounds of privacy and the likelihood of mistakes being made by content moderators trying to ascertain illegality of posts. This has led to Google taking legal action. At the same time, the restriction of platform action to illegal content *only* means that those affected by hate speech which is just on the legal side of the law are denied recourse or justice.

The current Bill includes provisions for business disruption measures. These, if retained, should be an absolute last resort and needs significant safeguards. Internet shutdowns are regularly used around the world (such as the regular internet shutdowns in India), often under the guise of reducing the risk of violence being fuelled on social media. These shutdowns violate human rights, significantly undermines the ability of citizens to access information and express themselves online and interfere with citizens engaging in democratic processes.