DEMOS

EVIDENCE ON THE ONLINE SAFETY BILL

June 2022

1. *Demos is Britain's leading cross-party think tank, with a 25-year history of high quality research, policy innovation and thought leadership. CASM, Demos' dedicated digital research hub has unique insights and expertise across tech policy and its impact on our society, economy and democracy.*

2. *We are writing to the Committee to follow up from our oral evidence to the Committee on 24 May 2022. We would welcome enquiries from the Committee on any points of clarification or further details.*

3. Much of the discussion in our session centred around the merits of a systems-based approach and how this could help tackle the harms that users faced online better than the existing provisions in the Bill. The Bill is often said to represent a systems approach: we do not believe that is an accurate representation and that there are many ways it could be improved.

What is a systems approach?

4. A 'systems approach' to digital regulation means one which seeks to tackle the ways in which systems and processes contribute to the risks of harm. This means, that instead of seeking to reduce the incidence or prominence of certain forms of content on online services, regulation seeks to change *how platforms are designed,* and how technologies are being designed and deployed within online services. This can include a wide array of system and process considerations, including: how the design of a service encourages, facilitates or incentivises harmful behaviour; the functionalities that enable communication; how users' access to information is shaped by algorithmic systems; how decisions are made and systems tested within a company; what powers users are given over the services they use.

5. This is not entirely separate from content considerations: design choices affect how and what content users are exposed to. However, they are importantly different approaches - a systems approach:
    - Identifies the appropriate point of intervention as *upstream* and preventing harm where possible
    - Locates harm to users as arising from how systems, users and content interact

6. While a content approach accepts the existence and circulation of harmful content, and so:
    - Identifies the appropriate point of intervention as *downstream* and mitigating harm once caused

- Locates harm to users as within pieces of content

7. The Bill intends to take a systems approach: this is why it has been designed, in a way we support, as being based on duties to tackle risks of harm identified in risk assessments of platform systems and processes. The language of systems and processes is used throughout, and we welcome this ambition, and the inclusion of functionalities and the design and operation of services in risk assessments is crucial.

8. However, the details of the Bill and how it is likely to be implemented **will fail to deliver a true systems approach**. This has led to significant tensions within the Bill between making sure safety duties are robust enough and ensuring user rights are protected, when at a system level, the two can be mutually reinforcing rather than in conflict.

9. The Bill:

   - Links risk assessments and safety duties to *categories of content* and not, (for instance), types of *harm* as they arise from systems in the first place
   - Prioritises, explicitly and implicitly, content moderation and content curation above other forms of process and systemic change by platforms, by focusing on content-based measures (such as identifying and removing content, preventing users from accessing content, and setting out in terms and conditions how content will be treated)
   - Includes exemptions for certain forms of content

Why is a systems approach better at protecting users?

10. Examples of a systems approach (from our Joint Briefing)

Risk of harm to user	Content approach: the regulator might ask	Systems approach: the regulator might ask
Exposure to promotion of suicide	Is all content promoting suicide taken down?	If users search for, post, share or are exposed to content promoting suicide, is there a system through which they can be directed to/access emergency and longer-term support? What does your recommender algorithm serve a user who has searched for suicide content more than once?
Exposure to vaccine disinformation	Is vaccine disinformation content demoted?	How do platforms identify when vaccine disinformation is reaching wide audiences and are mitigations (such as promoting authoritative information or fact-checking vaccine content) able to be quickly put in place?
Subjected to racist pile-on harassment	Can users report content which is harassing them?	What functionalities or design choices encourage or incentivise pile-on harassment - do pile-ons feed into 'trends', can they be easily monetised?

Better protection for freedom of expression and privacy of users

11. Many of the concerns about freedom of expression and privacy protections arise from requirements in the Bill which either require or strongly incentivise the identification and takedown of content, much of which may be legal speech. These concerns are compounded by the fact that the approach of designating certain forms of harmful content as 'priority' requires those forms of content to be defined in a way that is both wide enough to ensure risks are acted upon but specific enough to be actually implementable.

12. Requiring platforms to act on content that is 'disinformation', for instance, requires platforms to have a proportionate and effective way of defining 'disinformation' - a complex and very context-based phenomenon (platforms cannot judge 'falsity' at scale). Not requiring them to act on disinformation or the associated harms at all (within the current content-based approach) would mean that the harm would go completely untackled. Requiring them to risk-assess and act to improve systems, however, which facilitate and incentivise coordinated inauthentic behaviour, would

reduce the risks associated with disinformation, without having to enter into precise debates about what content qualifies as 'disinformation' or not.

13. This is a similar paradox as the 'legal but harmful' clauses are subject to. Either they will be interpreted to mean that platforms are free to set any terms and conditions at all they wish - in which case harms to users arising from legal content and activity could go completely unaddressed, which is not the apparent intention of the Bill. Or they will be interpreted to mean that platforms are expected to act to take down or demote certain forms of content, which runs the risk of the Government setting out forms of speech which are legal but in effect not permitted online - a clear threat to freedom of expression.

14. The inclusion of content exemptions, such as the media exemption, confirm that this is a worry - that content may end up being regulated by proxy. Content exemptions in a systems approach, however, are not needed - in fact they are likely counterproductive. If a platform has conducted an evidence-based risk assessment, and amended its systems and processes in a proportionate and effective way to reduce the risks of harm to users while preserving rights, to then *not* implement those amendments on the basis that it could affect a particular form of content, is to ask platforms to use a *less safe and effective* system.

15. Moreover, it is fundamentally assigning responsibility to platforms *for those decisions platforms are in control of*, rather than seeking to regulate user activity by proxy. Platforms should not be held liable for every individual piece of content that a user posts on their site - that level of liability would pose an unacceptable risk of chilling freedom of expression. But where platforms are making decisions and designing and deploying systems which can be reasonably expected to cause significant harm to users, they should be held accountable.

More effective at tackling real drivers of harm

16. Locating the harms that users face online *in individual pieces of content* is to misunderstand how harms occur online. Certainly, users can be harmed by individual pieces of content - in which case, where this harm meets an illegal threshold, the recourse is through the judicial system, with its appropriate checks and balances. The harms which the Online Safety Bill is designed to tackle, however, are frequently those which arise *at scale.* Pile-on harassment campaigns, health misinformation, conspiracy theories, hateful speech, extremism, exposure to pro-self harm content - these pose the greatest risk when they are scaled and amplified. One post promoting conspiracy theories about an election being rigged may make a few people question an election result - thousands of these posts amplified, can help incite an insurrection.

17. Tackling these harms on an individual, post-hoc basis fails to engage with the holistic environment which encourages, facilitates and incentivises harm to users. Seeing the purpose of the Bill as identifying individual illegal acts online fails to preventatively

tackle harms at scale, while dangerously outsourcing the processes of law enforcement to private companies.

More futureproof for evolving technologies

18. The Bill is intended to establish a futureproof regulatory framework which allows new technologies and new platforms to be regulated - such as new metaverse platforms which are the subject of growing investment.

19. However, a content-based approach fails to engage with the reality of how new platforms may operate. In the virtual reality of the metaverse, where harmful behaviour can occur much more directly rather than via the medium of content posted or shared, it is unclear how the safety duties will be applied in an effective and coherent manner.

20. A systems-based approach is more futureproof, and enables platforms to be required to risk assess the systems and processes they have in place which pose risks of any kind to users, not just those which are linked to a form of content.

More technically achievable

21. A systems-based approach is more likely to be technically achievable than a content-based approach. The current approach outlined in the Bill requires platforms to be able to confidently identify many different kinds of harmful content, reliably and at scale, where the technology available does not match the ambition. For instance - there exists effective and proportionate technologies, such as PhotoDNA, which enables CSEA imagery to be detected at scale with a high degree of accuracy. For other forms of harmful content, such as those which rely much more on contextual information or intent to accurately identify, there is likely to be significant error in platforms trying to identify these at scale. Allowing users to specify what forms of harmful content they are happy to be exposed to, for instance, as in the user empowerment duties, appears to promise a degree of control to a user that will technically be impossible to deliver.

22. A systems-based approach, by contrast, tackles the ways in which platforms are designed and how decisions are made. Much information about systems and processes will already be held by platforms - where it requires further scrutiny and audit (e.g. conducting testing on the effect of different design changes) this can be tested independently, assuming that there is sufficient data access for the regulator and independent researchers to engage in this kind of research.

More holistic

23. A content-based approach naturally pits ‘rights’ against ‘safety’ - the right to express certain views online against the need to protect users from the harmful effects of what is being said: hence why thorny questions arise around freedom of expression.

A system-based approach, however, assesses how systems are already shaping what information and content users can see and share, the risks - both in terms of physical and psychological harm, and in terms of threats to rights - of those systems, and how it can be improved so that speech is being facilitated in ways that are more consistent with democratic principles rather than with commercial imperatives.

How could the Bill be amended to make it more of a systems approach?

24. This could be achieved by:

- Decoupling the risk assessments from categories of harmful content, and require platforms to assess against risk of harms, and to publish these risk assessments.
- Requiring OFCOM's report on independent researcher data access to lead to the development of a code of practice on ensuring data access
- Removing the content exemptions (for media, democratic and journalistic content) and strengthen the systemic protections instead, by including the protection and promotion of human rights within the online safety objectives
- Specifying that in transparency reporting, reports must include details of how systems and processes are operating, how any design changes have been tested, and who has been involved in decision-making about design changes

Further reading and examples of a systems-based approach:

Joint Briefing on Exemptions, Exclusions and Exceptions in the Bill
System Change for System Change's Sake
The Online Safety Bill Position Paper
Joint Briefing on Privacy and Anonymity Online