Written evidence submitted by Demos (OSB0159)

Who we are?

1. 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.

2. 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 social media, and other 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.

3. Our research has long included trying to understand the nature of online harms, which has often revealed overlooked complexities that are vital to recognise if we are to create effective online harms policy. We engage regularly with other civil society organizations on the Bill and wider digital regulation issues. Joint letters and reports we have participated in include *Free Speech for All: Why Legal but Harmful Content should continue to be included in the Online Safety Bill* and *Open Letter to EU Policy-Makers: How the Digital Services Act (DSA) can Tackle Disinformation*. We have also joined a joint submission to this Committee along with other civil society organisations working to defend democracy in the digital age.

4. Recent research relevant to the issues that arise in this Bill has included: *Online Harms: A Snapshot of public opinion*, investigating public experiences of and attitudes towards harmful content online; *A Room of One's Own: A guide to private spaces online,* examining how we can define private spaces online; *States, Corporations, Individuals and Machines*, explores proposed settlements on the balance of power and what they mean for the future of the web. *A Picture of Health: Measuring the comparative health of online spaces*, explores online behaviour, and finds that the design of online spaces fuels negative behaviour, including trolling. *Everything in Moderation: Platforms, communities and users in a healthy online environment*, arguing that the principle and practice underpinning most major platforms have failed to create healthy online spaces, and that current attempts by states to regulate these spaces will in all likelihood fall short of addressing the root causes of this failure; *Engendering Hate: The contours of state-aligned gendered disinformation online*, understanding how disinformation is being used online to exclude women from public life; *What's in a name? A forward view of anonymity online*, calling for a radical new approach to how we protect our identities online.

5. Our responses draw on this existing body of research, policy and advocacy work into online harms and our wider expertise on the subject. This submission has been prepared by Ellen Judson, Alex Krasodomski-Jones, Josh Smith, Ciaran Cummins and Akshaya Satheesh.

Summary

6. We welcome the proposals to hold tech companies to account for harms which people face that are exacerbated or facilitated by their services. Platforms can no longer be treated as 'neutral' entities, which passively host content that may be harmful. Platform design, platform systems and processes - from what users have the powers to do on the service, to how content is curated, scaled, amplified, and recommended, to what data is collected about them and how it is used to shape their online experience, to the ways users are encouraged or nudged to behave: these affect the risk of harms that arise from what occurs in online spaces.

7. We believe this Bill represents an unprecedented possibility to tackle these harms: from protecting children to protecting those targeted by abuse and hate to reducing the risks of dangerous misinformation - while also introducing oversight to ensure platforms are respecting people's fundamental rights and not further entrenching the marginalisation of certain groups.

8. However, we are concerned that the Bill as currently drafted runs the risk of inefficacy on the one hand, and serious overreach and infringement of user rights on the other. We describe these risks and how they can be mitigated in our submission.

9. In summary, the primary risks we see are:

    a. A focus on regulation and moderation of *content* rather than platform systems which affect the risk of harm arising from that content

    b. Inadequate protections for users' rights to privacy and freedom of expression online

    c. Over-reliance on platforms' own assessments and reporting rather than on independent audits of their systems and processes and their effects on harm and user rights

10. The ways we see these being mitigated, through the pre-legislative scrutiny process, amendments to the Bill, and development of secondary legislation and Codes of Practice, include:

    a. Clarifying and defining key concepts and expectations

    b. Ensuring a systems-focused approach rather than a content-focused one

    c. Embedding protections for users' rights more thoroughly

    d. Strengthening requirements for transparency and auditing

Will the proposed legislation effectively deliver the policy aim of making the UK the safest place to be online?

11. The policy aim, as stated, is one we do not think can be delivered. There is division as to what 'being the safest place to be online' means in practice, and this uncertainty is reflected in the Bill. Making the UK 'the safest place in the world to be online' implies a global approach to safety, implicitly setting itself the goal of protecting users online from all foreseeable harm. Not only is that an unachievable goal, it is also explicitly in opposition to the text of the Bill, which has limits on platforms' duties and excludes certain kinds of harm from scope. The Bill also frames rights such as privacy and freedom of expression 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 online abuse, hacking, scams, fraud or doxxing).

12. 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. It also risks other methods alongside the Bill to reduce the risks of harm to users and improve user rights online being overlooked. 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.

13. We recommend:

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

    b. That consultations on the Codes of Practice should include consultation with different groups on what harms and risks they define 'safety' as being protected from, to avoid a narrow view of 'safety' being imposed on groups with different needs

Are children effectively protected from harmful activity and content under the measures proposed in the draft Bill?

14. The Bill rightly has a strong focus on child safety, and we welcome the expectation that platforms will take significant action to protect children from harmful content. However, we have concerns that the Bill will lead to children's rights being infringed in the pursuit of child safety, thereby putting children at risk unintentionally.

15. For instance, the requirement to have systems designed to prevent children from accessing harmful content risks that identity verification measures may be expected or implemented which would infringe children's own rights to privacy. This concern is exacerbated by the DCMS Safety guidance issued to help platforms meet their safety responsibilities, which suggests that 'you can also prevent end-to-end encryption for child accounts', explicitly recommending that children's accounts be made less secure.

16. We would also like to see systems in place for redress for overmoderation by platforms regarding their systems for protecting children. For instance, automated filters to stop children viewing sexual content that could be harmful have wrongly identified LGBT+ content as 'inappropriate' for children: a significant concern when the internet is so often a crucial resource and lifeline for LGBT+ youth.

17. We recommend:

    a. That the duty on all services to have 'due regard to the importance of' users' freedom of expression and privacy when deciding on their safety policies should explicitly include children within 'users'.

    b. That preservation of users' rights to freedom of expression and privacy are included in the online safety objectives, the pursuit of which Ofcom must ensure codes of practice are compatible with. Rights impact assessments should be carried out before codes of practices are implemented.

Does the draft Bill make adequate provisions for people who are more likely to experience harm online or who may be more vulnerable to exploitation?

18. Whether the provisions are adequate to protect all groups is currently difficult to assess, given that which harms will be priority harms, and what platforms will be expected to do to combat them, is deferred to later in the regulatory process.

19. However, it is a concern that currently, the primary vulnerable group focused on explicitly in the Bill is children, with little else about other groups who may be more likely to experience harm online.

20. We are concerned that the focus of duties to reduce the harms that users experience arising from legal but harmful content or activity is enforcement of a platforms' own terms and conditions, rather than there being an expectation of wider actions that could reduce the risks that vulnerable and marginalised groups face (such as increasing user powers, altering content curation systems, designing services to promote pro-sociality).

21. The focus on reducing harm caused to an individual also risks exempting platforms from taking action to tackle societal harms that disproportionately negatively impact marginalised groups: such as online disinformation campaigns, which can be used to exclude or increase hostility towards minoritised groups.

22. We are supportive of the mechanism of 'supercomplaints' through which groups may bring a complaint if they are being subjected to harm or rights infringements by online services. However, in the absence of clear criteria for 'eligible entities' which can bring supercomplaints and the power to define them given to the Secretary of State, there remains an unresolved risk of excluding certain groups from being able to exercise this power.

23. We recommend:

    a. The more specific ways the regulator will ensure that the rights of marginalised groups are being adequately protected.

b. That risk assessments carried out by platforms be subject to audit by the regulator (in consultation with affected communities) to ensure they adequately reflect the risks faced by marginalised groups on their services.

Is the "duty of care" approach in the draft Bill effective? Does the Bill deliver the intention to focus on systems and processes rather than content, and is this an effective approach for moderating content? What role do you see for e.g. safety by design, algorithmic recommendations, minimum standards, default settings?

24. We believe that a successful approach would be based on the following principle: that platforms have systems and processes in place which reduce the risk of harm to and protect the rights of their users. We are concerned that though a systems approach is the intention of the Bill, the current drafting as well as the public and political discussion around the Bill means there is a strong risk of instead delivering a content-focused regulatory approach.

25. 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. Given the public discussion around the Bill has also focused heavily on how far prevalence of content will be affected (rather than platform systems) 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.

26. Platforms are not 'neutral' spaces which have no effect on what is posted or shared in their space. From the number of characters you are allowed to type - what you are encouraged to engage with - who is allowed to read your content - who is shown your content - what content is taken down - this is all already determined by platforms. They are - and should be held - responsible for the systems and processes that they employ which *increase* the risk that users may be harmed on their services by content or behaviour.

27. This is of particular importance given that the unique difference between speech offline and speech online is the unprecedented speed and scale that speech online can achieve. The audience, reach and spread of speech online is determined not by an individual speaker but by platforms. Systems to reduce the risk of harm, therefore, should focus on preventing the incidence of harmful content and reducing its amplification and dissemination, rather than on its removal.

28. *Possible case 1: a content-based approach: An abusive post is sent to a politician during a televised debate and is reported to the platform for violating its terms of service. Due to the volume of posts being reported, it takes 24h to remove the post, during which time it is shared thousands of times. It also spawns copycat posts, which use the same language to target the public figure. These continue long after the original post is removed.*

29. *Possible case 2: a systems-based approach: An important political debate is upcoming. Platforms prioritise reviewing reports of abusive content or behaviour from the politicians involved, suspending users who have engaged in ongoing abusive*

behaviour against the politicians. On the night of the debate, they automatically implement measures meaning that no-one can tag the politician in a post unless she is already connected to them on the platform.

30. 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. In practice, this could incorporate:

    a. Design choices which increase the risk of harm

        i. *Example: Users creating a new account are allowed to post instantly, meaning throwaway accounts can be set up very quickly to facilitate harassment campaigns.*

    b. Content curation choices which increase the risk of harm

        i. *Example: Users being recommended extremist content*

        ii. *Example: Algorithms which demote harmful content being scrapped because they reduce user engagement*

    c. Business models which rely on prioritising user engagement over user safety

        i. *Example: Goal being to maximise engagement incentivising sensationalist, viral content*

31. Although references are included in the Bill to platforms including in their risk assessments design of services and systems and processes, these currently have less prominence than the content moderation requirements.

32. As well as being less effective, a content-focused approach runs greater risks of infringing upon freedom of expression: prioritising the takedown of legal content and incentivising overmoderation which may disproportionately affect marginalised groups. This model also risks being imitated by authoritarian regimes globally to justify the censorship of content they do not like.

33. Platforms' duties should be to minimise the risks associated with content and with user behaviour, based on evidence (which the service should provide to the regulator) of which measures are effective in reducing the harm that users experience. This should include content moderation, removal and demotion. However, the Bill should not *presuppose* that this will be the most effective way to protect users from harm: or, conversely, presuppose that preventing the removal of some types of speech on a platform will be the most effective way to protect freedom of expression.

34. For instance, an extremely efficient content removal system could easily be weaponised by malicious actors against marginalised groups to suppress their content. Assessment of the efficacy of systems, based on real-world effects - will be crucial to the success of the regulatory regime.

35. We would hope to see greater specificity in the Bill on the expectations for platforms' responsibilities as they relate to platform design, systems and processes. This should particularly focus on:

    a. how platforms will be expected to reduce risks of harm

    b. how these risks will be measured beyond presence or absence of harmful content

    c. how platforms will be expected to protect rights

36. 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.

How does the draft Bill differ to online safety legislation in other countries (e.g. Australia, Canada, Germany, Ireland, and the EU Digital Services Act) and what lessons can be learnt?

37. 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.

38. 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.

39. 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 only illegal content means that those affected by hate speech which is legal but harmful.

40. The current Bill includes provisions for business disruption measures. These, if retained, should be an absolute last resort and need 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 undermine the ability of citizens to access information and express themselves online, and interfere with citizens engaging in democratic processes.

Does the proposed legislation represent a threat to freedom of expression, or are the protections for freedom of expression provided in the draft Bill sufficient?

41. We do not believe that the current protections for freedom of expression in the draft Bill are sufficient. It remains entirely unclear what a 'duty to have regard to the importance of' freedom of expression means, or what the criteria will be by which that duty will be judged to have been met. Similarly, though this is specified in more detail for Category 1 services, how the comprehensiveness, accuracy or efficacy of both the impact assessment and the statement of steps to protect rights will be assessed by the regulator is not clear, leading to a risk of this becoming a tick-box exercise rather than leading to meaningful change from platforms.

42. However, we do not think this is an insurmountable obstacle that can only be overcome by drastically changing the scope of the Bill. We do not agree with the concerns frequently expressed that this Bill is effectively censorship of legal speech: that by appeal to a vague notion of harms, platforms will be at best able to justify and at worst legally compelled to remove legal speech from their services, compromising freedom of expression.

43. As discussed above, we support a regulatory regime that seeks to regulate systems which manipulate content online, rather than one which seeks to remove content as its goal. However, increased content removal is a likely outcome of the regulation.

44. Our view is that this outcome is justified, subject to certain parameters and safeguards. Freedom of expression is not an absolute right, and particularly a) in the context of expression within a particular defined space and b) with no legal repercussions for an individual for their legal speech, we view that a significant risk of harm is a legitimate reason to restrict people's speech online, as long as any restrictions are transparent, predictable, consistent, and does in fact reduce the harm and is not merely performative.[1]

45. Moreover, we see far greater threats to freedom of expression arising from online services on which harmful speech is allowed to proliferate, driving people away from being able to safely express themselves online, while also algorithmically suppressing content (e.g. through 'shadowbans') by marginalised groups. It is therefore in the interests of promoting freedom of expression that platforms be expected to meet certain standards in their design, systems and processes.

46. The clarification from the Government in their evidence submission that companies will not be *required* to remove legal but harmful content, as long as they have clear and consistent policies, is welcome. We are concerned, however, by the Government's focus on the protection against 'arbitrary removal' of 'controversial' content. While aiming to ensure that speech and information freedoms are protected, this ambition could be co-opted by malicious actors online. Members of

[1] And do not, for instance, secretly exempt high-profile figures from moderation

marginalised groups who are calling out hateful speech online often face accusations that they are aiming to 'silence controversial opinions'. 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.

47. It is crucial to recognise that under the new regulation, overmoderation and undermoderation of content by platforms are inevitable: they are already commonplace. It is crucial that, instead of ensuring there is no removal of 'controversial' content, there are robust avenues of redress for both of these outcomes.

48. Redress should be available both at an individual and a systematic level: both for users to appeal decisions about content of theirs or targeting them, and for the regulator to be able to regularly review at a systematic level the effectiveness of the measures platforms are taking and if there are any disproportionate impacts arising on marginalised groups that require correction. Platforms should also as part of their transparency obligations be required to provide justifications for their moderation decision-making processes.

49. Platforms should also be free to set their own terms of service more broadly (for instance, a platform dedicated to a certain topic of discussion should be free to moderate over content that is irrelevant). We do not consider that a platform preventing certain forms of speech necessarily violates freedom of expression - freedom of expression does not guarantee the right to speak in any space one chooses. However, on equality grounds, it is appropriate that there should be recourse to the regulator if, for instance, the speech of a particular marginalised group is systematically suppressed by a platform (such as content moderation algorithms being employed which disproportionately remove content from BAME creators or suppresses or demonetise LGBT content).

50. Given the complexities of promoting and protecting freedom of expression, we recommend:

    a. That platforms should be required to submit more substantial rights impact assessments of their policies and procedures, which should be subject to independent audit by the regulator, rather than simply having to publish an assessment along with their specification of the steps they are taking to safeguard the rights protected in the Bill.

The draft Bill specifically places a duty on providers to protect democratic content, and content of journalistic importance. What is your view of these measures and their likely effectiveness?

51. We support the intention to protect political speech, as this is an important protection against government overreach. However, having additional protections for 'democratic content' specifically 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.

52. Moreover, we are concerned that 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:

53. Overmoderation of political speech

    a. 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.

54. Undermoderation of harmful speech

    a. 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.

55. 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.

56. We would recommend:

    a. 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.

Earlier proposals included content such as misinformation/disinformation that could lead to societal harm in scope of the Bill. These types of content have since been removed. What do you think of this decision?

57. We would support the re-inclusion of harms caused more widely by misinformation and disinformation into the Bill: particularly disinformation and misinformation which, at scale, is likely to exacerbate social harms such as racism and misogyny, but which might not fall under the 'individual harm' category currently outlined by the Bill.

58. For instance, disinformation campaigns often amplify true information but in a misleading way that creates the impression that a story is more significant than it is (for instance, amplifying news stories about violence committed by a person who was an immigrant in order to stoke fears about immigrants in general). Gendered disinformation campaigns, which weaponise gendered stereotypes for political, economic or social ends, seek to undermine women in public life, and can lead to women's exclusion from participation in public spaces due to the fear of being subject to such a campaign. These kinds of disinformation risk bolstering prejudice and identity-based violence more broadly within society. We believe these would thereby be justified coming within the remit of an online safety bill, even though the harm they cause is not directly tied to an individual targeted in or exposed to the content, as they still have a link to indirect collective harm to a group.

59. The Bill has not divorced itself entirely from concerns relating to political speech and democracy: it has introduced specific clauses intended to uphold the integrity of political speech and democracy online. It would hence be consistent to also include disinformation and misinformation within the scope of the Bill - and particularly necessary, given that without this scope inclusion, the Bill could end up over-protecting political disinformation as a form of 'protected speech' which platforms have no obligation to reduce the risk of.

60. There should be further consideration given in advance of legislation regarding how platform systems which increase the reach of political disinformation which could significantly affect democratic processes (such as disinformation which aims to achieve voter suppression) should be tackled under the Bill, rather than the issue being sidestepped altogether.

61. *Possible case: Before elections, posts begin going viral online that appear to be an official notice claiming that voting has been postponed by a day due to Covid or other restrictions. Platform systems in place are such that these posts are recommended and promoted to users, regularly shared thousands of times and reach hundreds of thousands of people, with no fact-checking or promotion of authoritative information to counteract the false information. Under the Bill as it stands, Ofcom would have no recourse to demand any action from online platforms.*

62. We would caution that inclusion of misinformation or disinformation in scope, however, should not elide into those forms of content being deemed harmful as a whole. For instance, misinformation, understood as false information, is not inherently harmful: to treat anything incorrect on the Internet as 'causing harm'

would lead to incredibly damaging overreach and lead to a regulator or a private platform being in the position of needing to determine the truth or falsity of content writ large. Rather, systems which are known to increase the risks associated with *harmful* misinformation should be regulated and systems introduced which are known to reduce these risks.

63. Many of the systems which platforms would likely need to implement to meet their safety duties under the Bill - more prosocial design, risk-assessed content curation, and more consistent moderation - for instance, would also help reduce the risks of harms such as disinformation. We do not consider it would be a significant extra burden on platforms to include risk assessments and responses to harms such as disinformation in their compliance.

64. We recommend:

    a. That harmful misinformation and disinformation should be in scope, even where the harm they cause is broader collective harm rather than targeted at a particular individual

    b. That further consideration should be given to how platforms should be expected to respond to online activity which significantly affects democratic processes.

Are the definitions in the draft Bill suitable for service providers to accurately identify and reduce the presence of legal but harmful content, whilst preserving the presence of legitimate content?

65. 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.

66. 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.' The impacts of this, with regard to overmoderation and undermoderation, we detail elsewhere in this submission.

67. We recommend that:

    a. Key concepts and definitions in the Bill should be defined so that the likely effects of the Bill can be better assessed in advance of legislation.

The draft Bill sets a threshold for services to be designated as 'Category 1' services. What threshold would be suitable for this? Are the distinctions between categories of services appropriate, and do they reliably reflect their ability to cause harm?

68. 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. This risk should be assessed by Ofcom in consultation with a wide range of stakeholders.

69. 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.

70. We recommend that: platforms should be assessed by Ofcom on the basis of propensity of risk of harm arising from a service before they are assigned a Category.

<u>What role do algorithms currently play in influencing the presence of certain types of content online and how it is disseminated?</u>

71. Algorithms play a very significant role in influencing what content is online and what content is widely seen. The kinds that we consider to be most relevant to the Bill are:

72. Content moderation

    a. Algorithms which detect illegal content through hashing images and comparing them to known hashes of illegal images - e.g. of CSEA material

    b. Algorithms which detect content likely to be in breach of a platforms' terms of service - e.g. identifying Covid misinformation

73. Content curation

    a. Algorithms which determine what content should be surfaced to a user (e.g. used in YouTube's 'recommendations' or TikTok's 'For You' page, Google's search bar)

74. Content moderation algorithms determine what content is present or absent online, and as such are often the focus of discussion around online harms. However, we consider that content curation algorithms are just as an, if not more, important subject of regulation: the principles on which they operate are generally less clear, and measuring their effects is extremely difficult in current circumstances.

75. These algorithms have a significant role in determining what content is scaled - what is engaged with, shared, reproduced: and as such have a significant role in scaling the harm associated with harmful content - from supporting disinformation campaigns to recommending radicalising content, to facilitating pile-on abuse, we consider that there has been a lack of attention to content curation algorithms so far. This is of particular concern given reports of how platforms are failing to correct algorithms which promote and incentivise harmful content and behaviour online.

76. Much of the activity of algorithms online is unknown and difficult to assess: transparency about algorithmic systems used in the first place is limited, and access to data to try to assess the impact of different systems is severely restricted to researchers. There are some moves, such as transparency centers and publishing algorithms for debugging, towards more transparency, but these are few and far between and not systematised - while research is being done by platforms into the harms of their systems which is not made public.

77. A significant concern is also the lack of public understanding of algorithmic systems and how they work, compounded by the lack of transparency from companies on when algorithms are being used, for what, how they are being developed and tested, and reviewed. Users are sometimes allowed to 'feedback' on content they are shown, but what the actual effect on what content they are served is unclear.

78. The Bill could address these issues through a greater focus on algorithmic audit and accountability, as is currently being discussed in the EU and US. Algorithmic transparency is not straightforward. As the Ada Lovelace Institute has written: 'to develop transparency around ADM systems, we will need to gain insights into the decision-making processes that determine their goals, the impact of their outcomes, how they interact with social structures and what power relations they engender through continued use.'

79. We recommend that:

    a. The regulator is given the power and the responsibility to scrutinise (with support of independent researchers where necessary) how algorithms (particularly those used in content curation and moderation) are being developed, maintained, tested, adapted, and their efficacy against various safety and rights metrics.

What role might they play in reducing the presence of illegal and/or harmful content? Are there any foreseeable problems that could arise if service providers increased their use of algorithms to fulfil their safety duties? How might the draft Bill address them?

80. Algorithms enable curation and moderation at scale, and as such are essential to the functioning of online services. However, operating at scale means that a compromise will be made with accuracy. Different kinds of algorithm have different functions and different levels of accuracy that can be achieved.

81. As we saw during the height of the Covid-19 pandemic, during which time platforms were forced to rely significantly more on automated systems, the accuracy of moderation suffered. Automated systems are less able to respond to nuance and context than human moderators, and people are quickly able to adapt to better evade automated detection. For example, harassment campaigns can use images that are slightly edited or replacement words like 'b!tch' to avoid automatic detection. Users on Tiktok have adapted to using terms like 'unalive' in their content to avoid TikTok's automatic detection and removal of suicide-related content.

82. A significant concern is that if service providers are required or encouraged to use algorithmic systems to identify speech which is likely to cause harm, this could have

particularly significant impacts on the expression of marginalised groups, as algorithmic moderation systems can be often biased in ways that mean their speech is more often flagged as potentially harmful. This has been evidenced, for example, by a study from Sap et al (2019) for example with regard to racial discrimination, and by Gomes et al (2019) on how reclaimed speech by LGBTQ people may be disproportionately be identified by algorithms as 'toxic'.

83. In many cases, it may also not be that platforms need to *increase* their use of algorithms but change the use of existing algorithms where those have harmful consequences.

84. Algorithmic systems will also not be able to replace human moderation and review. We are deeply concerned that the outsourcing of moderation, the poor conditions and lack of psychological support that content moderators are provided with, are undermining both standards of moderation and the health and wellbeing of content moderators.

85. We recommend:

    a. That, as above, the regulator is given the power and the responsibility to scrutinise (with support of independent researchers where necessary) how algorithms (particularly those used in content curation and moderation) are being developed, maintained, tested, adapted, and their efficacy against various safety and rights metrics.

    b. That Ofcom consider ways in which platforms can be audited not only for the design of their content moderation systems but how those systems are being enacted, with a focus on appropriate employment of and support for human moderators.

Does the draft Bill give sufficient consideration to the role of user agency in promoting online safety?

86. The consideration of user agency in ensuring terms and conditions are clear and accessible, user powers to report and redress, and the promotion of media literacy is welcome, but we would support a more holistic approach to user agency.

87. As we have set out, 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. To achieve the former, a focus should be given to a range of methods beyond the present focus on content removal risks, which protect, support and empower 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. These and other approaches shift agency to users to promote safety, by empowering them to be part of a safer online environment.

88. However, it is crucial that the Bill does not shift *responsibility* for preventing harm to users, while not granting them the power to do so. For instance, a platform with a

persistent problem of harassment campaigns having an excellent user reporting system should not count as sufficiently compliant: given that the burden of identifying and flagging campaigns is being taken on by users, while the platform is still determining and controlling the rest of the ecosystem in which these campaigns flourish.

89. We recommend:

    a. That the Bill and Codes of Practice consider more ways in which user agency may be promoted: focusing on how fundamental platform design can support greater user agency in addition to duties to increase users' understanding of and access to information.

Are Ofcom's powers under the Bill proportionate, whilst remaining sufficient to allow it to carry out its regulatory role? Does Ofcom have sufficient resources to support these powers?

90. We consider that Ofcom's powers may not be sufficient to allow it to carry out its regulatory role, and that support from other stakeholders may be required to support the exercise of its powers.

91. 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: likewise for Category 1 rights impact assessments.

92. How these will be audited, however, is not yet clear. Platforms simply producing reports cannot be taken as compliance without an audit of the efficacy and accuracy of these reports. 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.

93. Moreover, 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. Independent researchers and civil society have an important role to play in helping fill this gap.

94. More clarity overall is needed on how precisely Ofcom will seek, act and report on advice it has sought on its regulation in practice. For example, in preparing codes of practice the Bill (Section 5, Subsection (5) - (6)) requires Ofcom to consult a variety of actors "who appear to Ofcom" to have relevant expertise or be representative of certain groups. It is unclear, however, what exact mechanism for consultation will be used to do this, how Ofcom will decide who is a representative or has relevant expertise, and whether and how they will publish records of who they have consulted.

95. For instance, little detail is given on how Ofcom will institute public engagement with its regulation in practice. In Section 99, Subsection 2 references are made to how the regulator ought to "make arrangements for ascertaining" public opinion and

experience of the regulated services from “time to time”, though without further specification. Given the centrality of public opinion and experience to so many of the core concerns of the Bill - from freedom of expression, to psychological harm and perception of safety - more detail here is necessary to ensure public trust that the regulator will be adequately responsive to the lived experience of users. Given Ofcom’s existing programme of regular internet research, capacity for this appears to be there for the regulator, however this must be detailed explicitly in the Bill.

96. We would recommend:

    a. that Ofcom have greater powers to audit platforms’ systems and information themselves, including auditing algorithmic systems used by platforms to determine the risk of harms posed to users.

    b. 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.

    c. That the Bill should detail explicitly how Ofcom will seek, act and report on advice it has sought on its regulation and institute public engagement in practice.

<u>How much influence will a) Parliament and b) The Secretary of State have on Ofcom, and is this appropriate?</u>

97. The powers granted to the Secretary of State are significant, and in our view, (a <u>view widely shared across civil society</u>) 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.

98. We recommend that:

    a. 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 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.

    b. 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.

Other areas of concern

Privacy protections

99. We have serious concerns about the lack of protections for privacy in the Online Safety Bill. Giving platforms a duty to 'have regard to the importance of protecting users from unwarranted infringements of privacy, when deciding on, and implementing, safety policies and procedures' with impact assessments and a public statement of steps taken for Category 1 services risks becoming a tick-box exercise that will not offer users significant protection of privacy.

100. 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. 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 concern is exacerbated by the potential conflict of privacy protection with other clauses of the Bill).

d. It is unclear how the impact assessments and public statement of steps taken will be assessed, audited, or corrected if it is not up to scratch.

101. 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.

102. The OSB does distinguish between obligations in public and private channels in the case where it can require the use of accredited technology to identify and take down CSEA content present on any part of the service (public or private), but to do so only for 'public terrorism content'. This indicates a call has been made that there are different expectations for services to act on different harms in public or private channels, but these are not made explicit.

103. We agree that platforms should still have a duty of care in private channels. However, the nature of private channels means that what platforms can proportionately be required to do will be significantly different to what they can do on public channels.

104. Steps they could be required to take to reduce the risk of harm would include: having the option to forward messages to moderators, be able to block or report users, be told when messages have been forwarded, analysing metadata to assess high risk users.

105. We do not support the mandating of measures incompatible with the use of end-to-end encryption to protect private channels. We are deeply concerned that this Bill could facilitate a requirement that platforms remove end-to-end encryption, with the intention of improving child safety, but leading to significant risks of harms for all users (including children) and meaning overreach of a regulatory regime is that much more likely. It would also set a dangerous international precedent, whereby companies which were required in the UK not to preserve the integrity of end-to-end encrypted channels would likely face significant similar pressure by authoritarian regimes to follow suit, with significant ramifications for human rights globally.

106. We are also concerned that the Bill's requirements on designing systems to prevent certain forms of harm (in particular, protecting children from age-inappropriate content and protecting adults from online abuse and harassment) may lead to platforms being required or strongly incentivised to require identity verification from users before they are allowed to use their services. We are supportive of measures which would tackle these issues: as long as those measures are effective in reducing harm and, crucially, do not infringe on users' wider digital rights, in particular, anonymity.

107. Being able to be anonymous from other users online is a necessary but not sufficient protection: being able to be anonymous from the services that you use online is also crucial: and the intrusion into privacy caused by data collection and tracking people across the websites they use is widely acknowledged. This kind of anonymity does not preclude service providers providing information on criminal activity to law enforcement in response to a legitimate enquiry with judicial oversight.

108. The concern with introducing identity verification requirements or restricting essential functionalities to unverified accounts, are as follows:

a. Being able to access online services anonymously is a crucial protection for the rights to freedom of information and expression. It allows people to access information, seek support, develop their understanding, allowing them to disclose private or sensitive information (such as about their sexuality, gender, health, immigration status) without the risk that it will be connected to their identity and compromise their privacy.

b. Moreover, there are groups for whom providing identity details may be prohibitive to their being able to safely engage in a space, as having to share details of their identity (even with a platform) would put them at significant risk for their personal safety: including, but by no means limited to: journalists, sex workers, whistleblowers, LGBT+ people, or undocumented migrants.

c. Although there are many third-party identity providers, it is likely that this market would be instantly captured by the large tech companies who already facilitate identity provision across platforms, such as Facebook and Google.

This would further consolidate their market power and their control of and ability to use and monetise people's personal data.

109. Our concerns are amplified by the fact that in the DCMS Safety Guidance published in June 2021, it is recommended to platforms that if they have private channels or end-to-end encryption, that they should restrict the use of these features: for instance, that they could 'prevent[] unverified users from using features such as: using encrypted messaging': when marginalised groups and people facing persecution and oppression are both the groups who most need anonymity online and to have access to secure private communications channels.

110. We recommend:

a. That platforms should be required not to infringe privacy across their policies and procedures not simply those related to 'safety'.

b. That the Bill include the preservation of users' rights to privacy and anonymity online within the online safety objectives which the Codes of Practice must further.

c. That any recommended platform actions within the Codes of Practice should be evidence-based and subject to a rights impact assessment before inclusion.

28 September 2021