Call for evidence response form

Your response – Additional terms of service duties

Questions 1 – 5: Terms of service and policy statements

For all respondents

Question 1: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements?

Please submit evidence about what features make terms or policies clear and accessible.

Response:

First, platforms should offer a greater range of examples on what speech does and does not violate their terms of service. At present, platforms often present the abstract rationale for a given rule, gesturing toward the type of content that falls within it; but (with some important exceptions) they seldom specify examples. Inclusion of examples would empower users by making it easier for them to predict whether their content will or will not be caught by a particular content rule. Such prediction is essential for platforms to satisfy the principle of legality, which is increasingly recognised as central to rights-respecting content moderation. (It is sometimes suggested that too much detail will enable wrongdoers to "game" the system and circumvent the rules – but we would never accept this argument in ordinary legal contexts, and it's not clear why we should here.)

Second, while platforms have tended to provide extensive statements of their removal policies, they have failed to offer sufficiently detailed statements of their demotion policies. It is welcome that Meta has a "Types of content we demote" page, and TikTok has a "For You feed eligibility standards" page. But these are cursory compared with their elaborate statements of their removals policy, leaving users with less guidance than they should have about whether their content will be demoted. (Other platforms do much worse.) As we argue in a working paper (Howard, Kira, and Bartolo, "Remove or Reduce?" – available by email), demotion substantially impacts users' freedom of expression interests. For example, Meta has a policy of demoting some "borderline" content, i.e., speech that comes close to violating its rules but doesn't quite cross the line. But what exactly counts as borderline is left largely unspecified. Ditto for many other categories of demoted speech. Platforms should provide more information on what, exactly, they demote.

Finally, platforms should be more transparent about the ways in which their rules are sometimes crafted to be broader (or narrower) than they would otherwise ideally be, in response to the

feasibility constraints of at-scale mechanized content moderation. For example, if the accuracy rate of an "ideal" nuanced rule is very low, platforms may sometimes opt for a less discriminating rule with better rates of accurate enforcement. But it would be valuable for this reasoning to be spelled out publicly. Similarly, platforms would ideally be more transparent about the kinds of false negatives and false positives that their systems tend to produce (in which rules are misapplied). These errors are inevitable with at-scale mechanized content moderation (as Ofcom recognizes throughout its materials); it is important to level with the public on exactly this point so that people do not expect perfection, which is illusory in this space.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

Question 2: How do you think service providers can help users to understand whether action taken by the provider against content (including taking it down or restricting access to it) or action taken to ban or suspend a user would be justified under the terms of service?

In your response to this question please consider and provide any evidence related to the level of detail provided in the terms of service themselves, whether services should provide user support materials to help users understand the terms of service and, if so, what kinds of user support materials they can or should provide.

Response:

First, platforms should offer much more detail in its statements and justifications of removals and demotions policies (as we explain in our response to Q1).

Second, when a post is removed, platforms should seek to provide not merely an indication of the particular policy that the post violated, but an explanation of why the post was deemed a violation. This process will be imperfect, since it will have to be mechanized (e.g., large language models have promise in this area at producing the relevant explanations); but it is better than simply pointing to a general policy statement, which risks leaving the user with little sense of why exactly their speech was removed.

Finally, when platforms demote content (reducing its visibility) on the grounds that the content is harmful, users should ideally be notified that their posts may receive reduced reach. Just as users deserve notification in cases of deliberate content removal, they ideally deserve notification in cases of deliberate content demotion. It is not clear whether this is technically feasible, but if it is, platforms should be encouraged to provide it. (They should merely be encouraged, rather than required, since the costs of implementing such a notification system may be disproportionate.)

Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

Questions 9 – 15: Moderation

For all respondents

Question 9: Could improvements be made to content moderation to deliver more consistent enforcement of terms of service, without unduly restricting user activity? If so, what improvements could be made?

In your response to this question, please provide information relating to (a) –(c) where relevant.

Response:

Our response focuses on the importance of improving content moderation "without unduly restricting user activity". We seek to draw attention to a potential point of tension between this commitment (protecting users' legitimate speech from undue burdens) and earlier guidance that Ofcom offered in its draft codes on illegal content duties.

In some forthcoming work, we examine the ways in which Ofcom's draft guidance on illegal content duties incentivises platforms to adopt what we call a "bypass strategy", whereby platforms create and enforce content moderation rules that are far broader than existing criminal laws, enabling them to bypass judgements of illegal content. This strategy aims to avoid complex legal interpretations of criminal intent and potential defences, which are typical of criminal adjudication but would be infeasible for at-scale automated content moderation systems. We argue, however, that the bypass strategy poses a significant risk to users' freedom of expression by substantially incentivising the over-removal of lawful speech—unless guidance from Ofcom is adjusted to provide countervailing incentives. (The paper offers insights that could help Ofcom to improve its guidance on how platforms should interpret such duties on moderating content and mitigate this risk within the constraints of the Act.)

Our main worry is that if platforms pursue the Bypass Strategy—enacting content rules far broader than the law to ensure they catch illegal content—there is a risk that they will indeed "unduly restrict user activity". It will be unduly restrictive for platforms to adopt content rules which prohibit far more speech than they are strictly required to restrict: both because it is unnecessary (given the possibility of more narrowly tailored rules) and disproportionate in the strict sense of involving excessive costs relative to the benefits attained. In sum, to comply with the illegal content duties (in the way Ofcom's current draft guidance incentivises them to do) platforms could be in breach of their terms-of-service duties.

To guard against this, Ofcom should specify that platforms adopting the Bypass Strategy should explain how they are nevertheless mitigating risks to freedom of expression (specifically, in their free speech impact assessments). For example, while it is reasonable for platforms to adopt rules that don't tightly track the contours of the criminal law—since it is infeasible for at-scale

mechanized content moderation to do so with anything approaching accuracy—that doesn't mean that *any* rules will suffice. Rules that sweep too widely, encompassing substantial amounts of lawful speech, do not take seriously users' freedom of expression interests. In other words, if Ofcom is going to encourage platforms to adopt the Bypass Strategy (which its draft codes for illegal content duties manifestly do), Ofcom must also supply a countervailing pressure: namely, instruct platforms to explain how they are mitigating risks to freedom of expression (as part of their impact assessments).

The work-in-progress ("The Bypass Strategy: Platforms, the Online Safety Act, and Future of Online Speech"), which has been solicited for a forthcoming volume of *The Journal of Media Law*, is available here: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4822405

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

Your response – News publisher content, journalistic content and content of democratic importance

For all respondents

Question 24: What, if any, measures can online service providers put in place to ensure that protection of content of democratic importance applies in the same way to a wide diversity of political opinion?

In your response to this question, please provide information relating to (a) where relevant.

Response:

First, platforms should clarify that where political content crosses a line (e.g., crossing into hate speech or threats or harassment or incitement), it will be removed. But such removals are done merely on the basis of the violation, not on the basis of the particular political/partisan identity of the speaker. So, speech promoting violence is intolerable, regardless who on the political spectrum is engaging in it. This is an obvious point to those working within this space, but it is not obvious to the public, and it needs to be emphasised.

Second (reiterating our response to Q1), we think clearer and more developed policies on demoting content would help reassure the public that a wide variety of political content is protected online. Recent complaints that platforms are not even-handed in their treatment of different political points of view loom largest in relation to demotion policy, whereby platforms demote certain

"problematic" (i.e., mildly harmful or undesirable) content – allowing it to be posted, but reducing the likelihood that audiences will actually encounter it. This leads to complaints that platforms are "rigging the debate" for a particular point of view, and "shadowbanning" disafavored views. There is no evidence, to our knowledge, that platforms have deliberately done this (e.g., applied their demotion policies selectively against certain politicians). But the point is about how platforms should communicate. Clearer and more detailed statements of demotion policies, emphasizing the fact that they are applied to points of view across the political spectrum, are imperative. (See Howard, Kira, and Bartolo, "Remove or Reduce?" – available by email).

Third (reiterating our response to Question 9), platforms should explain how exactly they have tailored their rules narrowly to take account of protecting users' legitimate political speech. We argued that the Bypass Strategy that Ofcom encourages (by which platforms adopt capacious content rules to enable them to bypass making illegality judgements) poses a risk to users' general interests in freedom of expression—including political speech. Ofcom notes that "Category 1 service providers must include provisions in the terms of service specifying the policies and processes by which the importance of the free expression of content of democratic importance is taken into account" (4.11, p. 25) as well as provide "an assessment of the impact of safety measuyres and policies" on "users' rights to freedom of expression" (4.12, p. 25). To satisfy these aims, platforms should explain—in their terms and conditions, and in their free-speech impact assessments—how precisely they have safeguarded legitimate political speech in the face of a clear incentive to over-remove speech. By doing so, these measures can provide the necessary countervailing pressure on platforms to steer them away from adopting excessively broad content prohibitions.

The work-in-progress ("The Bypass Strategy: Platforms, the Online Safety Act, and Future of Online Speech"), which has been solicited for a forthcoming volume of *The Journal of Media Law*, is available here: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4822405

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

Your response – User empowerment duties

Question 27 and 28: Experience of specific types of users

For all respondents

Question 27: Can you provide evidence around the types of adult users more likely to encounter relevant content, and the types of adult users more likely to be affected by such content?

Response:

First, given the Bypass Strategy (see Q9 and Q24 responses above), "relevant content" for the user-empowerment duties is highly likely already to be prohibited under their content rules. All the major platforms prohibit relevant self-harm content, speech abusive to those with protected characteristics, and speech inciting hatred to those with protected characteristics.

Second, even if some platforms technically allowed such speech, it is worth remembering that user-empowerment tools have limited harm-preventing efficacy. For some speech, the very people endangered by the content may be those who seek it out (as with speech promoting eating disorders or offering self-harm instructions). For other harmful speech, its ultimate victims may never even see it; speech inciting hatred risks harm by persuading audiences to engage in offline violence or discrimination, whose targets may not even be users of the platform. In these ways, user-empowerment duties are a poor substitute for the adult safety duties that appeared in earlier versions of the legislation. We mention this only because it is imperative to be clear-eyed about what these provisions can and cannot do.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

For all respondents

Question 28: How do you consider the experience of users who have a protected characteristic, or those considered to be vulnerable or likely to be particularly affected by certain types of content?

In your response to this request, please provide information relating to (a) – (c) where relevant.

Response:

See previous point (in reply to Q27).

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

DIGITAL SPEECH LAB

UNIVERSITY OF SUSSEX

SUSSEX LAW AND TECHNOLOGY GROUP

Submission Prepared By:

Jeffrey W. Howard is Director of the Digital Speech Lab and Associate Professor of Political Philosophy & Public Policy at University College London, and Senior Research Associate at the Institute for Ethics in AI at Oxford University

Beatriz Kira is Lecturer in Law at the University of Sussex, member of the Sussex's Law and Technology Research Group, and Fellow in Law & Regulation at the Digital Speech Lab at University College London.

About the UCL Digital Speech Lab

The Digital Speech Lab at University College London hosts a range of research projects on the proper governance of online communications. Its purpose is to identify the fundamental principles that should guide the private and public regulation of online speech, and to trace those principles' concrete implications in the face of difficult dilemmas about how best to respect free speech while preventing harm. The research team synthesizes expertise in political and moral philosophy, the philosophy of language, law, social science, and computer science.

About the University of Sussex's Law and Technology Research Group

An international hub for research, teaching, and engagement in law and technology, the University of Sussex's Law and Technology Research Group houses leading scholars with expertise in technology and information regulation, global governance of technology, intellectual property, and legal innovation. We conduct cutting-edge research through collaborative projects with policymakers, civil society organisations, and industry leaders.