WARNING: This consultation response contains language and/or material that may be distressing

Ofcom Consultation on Draft Guidance: A safer life online for women and girls

Response by the Minderoo Centre for Technology and Democracy at the University of Cambridge.

[redacted]

About MCTD

The Minderoo Centre for Technology and Democracy (MCTD) is an independent team of academic researchers at the University of Cambridge who are radically rethinking the power relationships between digital technologies, society, and the planet. Through our ambitious research agenda, we are enhancing public understanding of digital technologies and delivering positive changes to society's relationship with these technologies.

Question 1: Do you have any comments on our proposed approach to 'content and activity' which 'disproportionately affects women and girls'?

We broadly agree with the approach to 'content and activity' that 'disproportionately affects women and girls'. We welcome the framing of 'gender-based harms', acknowledging that the majority of individuals perpetuating abuse against women and girls are men, and we welcome that the draft guidance makes explicit that offline and online harms overlap.[1]

The guidance could go further in exploring misogynistic online cultures: the draft guidance cites research on incel culture, but it could benefit from a more explicit exploration of this in the main text, including the emerging threat of online misogyny and incel culture's connection with violence.[2] The guidance could also explore how men and boys experience, are drawn into, and perpetuate networks of harm, or examine the exploitation of the 'manosphere' by certain actors for political or economic gain.[3]

In addition, the guidance could explore the challenges of content moderation associated with incel culture. For example, the coded language of incel culture (e.g. 'high value males', 'Chads', '80-20 rule'), which may not be as easily detected in content moderation as explicitly misogynistic slurs, but still plays a role in networks of harm.[4]

---

[1] Online misogynistic content and domestic violence have been shown to correlate, though more research may be required for accurate predictions. See: Blake K. R., O'Dean S., Lian J., Denson T. F., 'Misogynistic tweets correlate with violence against women', *Psychological Science* 32.3, 315–325, (2021). https://doi.org/10.1177/0956797620968529.

[2] For the connection between incel culture and violence, see: Hoffman, B., Ware, J. & Shapiro, E., 'Assessing the threat of incel violence', *Studies in Conflict & Terrorism*, 43.7, 565–587 (2020). https://doi.org/10.1080/1057610X.2020.1751459.

[3] On the connection with political violence, see: Zimmerman, S., 'The Ideology of Incels: Misogyny and Victimhood as Justification for Political Violence', *Terrorism and Political Violence*, 36.2, 166−179 (2020). https://doi.org/10.1080/09546553.2022.2129014; and Barcellona, M., 'Incel violence as a new terrorism threat: A brief investigation between Alt-Right and Manosphere dimensions', *Sortuz: Oñati Journal of Emergent Socio-Legal Studies*, 11.2, pp. 170–186 (2020). Available at: https://opo.iisj.net/index.php/sortuz/article/view/1471 [Accessed: 15 May 2025].

[4] On the limits of content moderation focused only on specific terms or forums, see : Solea, A.I., and Sugiura, L., 'Mainstreaming the Blackpill: Understanding the Incel Community on TikTok', *Eur J Crim Policy Res* 29, 311–336 (2023). https://doi.org/10.1007/s10610-023-09559-5l; and Matter D., Schirmer M., Grinberg, N., and Pfeer, J., 'Investigating the increase of violent speech in Incel communities with human-guided GPT-4 prompt iteration', *Frontiers in Social Psychology,* 2:1383152 (2024) doi: 10.3389/frsps.2024.138315. On understanding

We welcome the foregrounding in the draft guidance of intersectionality and systemic harms, particularly as they relate to pile-ons and online harassment. The importance of intersectionality could be made even more explicit, for example with a brief case study exploring the experiences of survivors of gender-based violence whose experience is also shaped by, for instance, their racialised or queer backgrounds.

With intersectionality in mind, the guidance could also be more explicit in expressing the disproportionate harms experienced by transgender people. LGBTQ+ people are more vulnerable to online hate than their heterosexual and cisgender peers.[5] Trans people can experience particular hostility online.[6] The guidance could go further in outlining user-to-user abuse directed against these communities, and how platforms' lack of action facilitates it.

Question 2: Do you have any comments on the nine proposed actions? Please provide evidence to support your answer.

Comments on the specific actions:

Action 1: Ensure governance and accountability processes address online gender-based harms

It is unclear why setting policies against illegal activity including 'stalking, harassment, and intimate image abuse' is under the good practice steps, rather than the foundational steps, as the illegal content codes of practice include a duty on platforms to 'mitigate and manage the risk of offences taking place' through the service.[7]

Action 2: Conduct risk assessments that focus on harms to women and girls

We welcome that the good practice steps recommend the use of external assessors and consultation with survivors and victims. We also welcome the recognition that engaging with survivors and victims can be burdensome for them. This section could go further by recommending service providers pay such advisors for their time, to reduce the amount of free labour that is often required of them.

Action 3: Be transparent about women and girls' online safety

We welcome the expression that transparency is an important source of online safety information for users. However, in our response to Ofcom's call for evidence on researchers' access to information from regulated online services, we outlined how platforms have increasingly closed off

---

the toxic language on incel communities, see: Pelzer, B., Kaati, L., Cohen, K., et al., 'Toxic language in online incel communities', *SN Social Sciences,* 1.213 (2021). https://doi.org/10.1007/s43545-021-00220-8.

[5] Powell, A., Scott, A. J., & Henry, N., 'Digital harassment and abuse: Experiences of sexuality and gender minority adults', *European Journal of Criminology*, 17.2, 199-223 (2018). https://doi.org/10.1177/1477370818788006.

[6] Haimson, Oliver L., Buss, J., Weinger, Z., Starks, D. L., Gorrell, D., and Sweetbriar Baorn, B., 'Trans Time: Safety, Privacy, and Content Warnings on a Transgender-Specific Social Media Site', *Proceedings of the ACM on Human-Computer Interaction*, 4, CSCW2, Article 124, 1-12 (October 2020). https://doi.org/10.1145/3415195.

[7] Ofcom, *Quick guide to illegal content codes of practice*, (9 November 20254). https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/codes-of-practice [accessed 16 May 2024].

routes for researchers to access data for reliable and verifiable research, making meaningful transparency practically impossible.[8]

We recognise that categorised services will be subject to annual transparency notices, but we believe a broader framework for researcher access to data is required to address imbalances of power between the public interest and those of platform companies. For example, the draft guidance says non-categorised services 'could' engage in transparency reporting, but we have seen no evidence that platforms are willing to do that consistently, nor evidence that they would take such a suggestion seriously without stronger guidance from the regulator. A broader transparency framework would be more effective in assessing all nine actions in the draft guidance, ensuring that service-side measures are independently verified and assessed as part of the online safety regime.

Action 4: Conduct abusability evaluations and product testing

No specific comments.

Action 5: Set safer defaults

No specific comments.

Action 6: Reduce the circulation of online gender-based harms

We have previously argued that content moderation that is focused on assessing individual instances of content is far less effective in preventing harm than approaches aimed at addressing networks of harmful content.[9] This action aimed at reducing the circulation of gender-based harms is therefore welcome.

The diffusion of online information, harmful and otherwise, responds to relational dynamics and follows predictable 'network laws'.[10] Gender-based violence against women and girls follows similar patterns as other information disorders, such as disinformation campaigns directed against migrants and minorities. These 'networked harms' cannot be efficiently monitored at the individual scale, nor are mitigations effective at the individual level. This means that while users' individual actions to stay safe online are important, they are not sufficient to address the problems of networked harms and, more specifically, networked harassment to which women and girls are more likely to be subjected.[11]

[8] Minderoo Centre for Technology and Democracy, *Written Evidence: Data-Driven Research for Evidence-Based Interventions,* (24 January 2025), https://doi.org/10.17863/CAM.115665. https://www.mctd.ac.uk/written-evidence-ofcom-data-driven-research-evidence-based-interventions/

[9] Minderoo Centre for Technology and Democracy, *Written Evidence: Harmful by Design: Current approaches to content moderation and how to improve harm mitigation* (18 December 2024), https://doi.org/10.17863/CAM.115614. https://www.mctd.ac.uk/written-evidence-science-innovation-tech-select-committee-inquiry-social-media-misinformation-harmful-algorithms/

[10] See, for example: Centola, D., and Macy, M., 'Complex Contagions and the Weakness of Long Ties', *American Journal of Sociology,* 113.3, 702–34, (1 November 2007); Vosoughi, S. Roy, D., and Aral, S., 'The Spread of True and False News Online', *Science,* 359.6380, 1146–51 (2018); or Leal, H., 'Networked Disinformation and the Lifecycle of Online Conspiracy Theories', in R*outledge Handbook of Conspiracy Theories*, ed. Butter, M. and Knight, P. (Routledge, 2020).

[11] Marwick, A. E., and Caplan, R., 'Drinking Male Tears: Language, the Manosphere, and Networked Harassment', *Feminist Media Studies,* 18.4, 543–59 (4 July 2018). https://doi.org/10.1080/14680777.2018.1450568

We caution against overreliance on purely automated content moderation, as it often does not account for these network factors, and several platforms lack tools to allow people to address the effects of networked harms.[12] For instance, automated sifting of billions of individual posts cannot replace informed analysis of the networked circulation of information. Informed analysis relies on access to data to assess networked harms at the network level.

Except for a limited number of cases, the guidance should avoid recommending specific technical patches. While they will work well in some contexts, they have limited capacity to prevent the diffusion of emerging practices and new technologies across networks.

The guidance should also more clearly acknowledge that algorithms prioritising sensational or emotional content increase the risk of misogynist or gendered harms. The business models of many large platforms perpetuate this factor by design by prioritising user engagement (and, therefore, profitability) over user wellbeing.[13]

Action 7: Give users better control over their experiences

No specific comments.

Action 8: Enable users who experience online gender-based harms to make reports

We broadly agree with the principles under this action. However, we present two important caveats.

First, company guidelines on what is tolerated will determine the effectiveness of such reports. Much content may be reported but not taken down because it does not violate internal policies. Societal assumptions about what is appropriate already complicate content policies in ways which can foster sexist and misogynistic online cultures. For example, semi-nudity and sexually suggestive content may be allowed, while health information such as reproductive or menstrual content is removed.[14] Again, this is a heightened issue for trans people who often experience censorship of their bodies and experiences online. For example, recent Meta online content policy changes suggest that hate targeted at trans people is specifically permitted under the guise of free speech.[15]

Second, while improving the experience of individuals who experience online abuse is a commendable goal, individual action is not the solution to collective online harms. The right of one person to

[12] Neff, G., and Chowdhury, R., 'Platforms Are Fighting Online Abuse—but Not the Right Kind', *Wired* (28 Feb 2023). https://www.wired.com/story/platforms-combat-harassment-but-theyre-focusing-on-the-wrong-kind/ [accessed 23 May 2025].

[13] Ibid.

[14] Centre for Intimacy Justice, *The Digital Gag: The suppression of sexual and reproductive health on meta, tiktok, amazon, and google* (2025). https://www.intimacyjustice.org/report2025 [accessed 23 May 2025].; *Feminist Insider* 'Six Women's Health Startups File EU Complaint Against Tech Giants Over Content Censorship' (11 March 2025). https://femtechinsider.com/womens-health-startups-file-eu-complaints-against-tech-giants-over-content-censorship/ [accessed 23 May 2025].

[15] *The Independent* , 'Facebook lifts restrictions on calling women 'property' and transgender people 'freaks' (08 January 2025). https://www.independent.co.uk/tech/facebook-meta-announcement-fact-checking-hate-speech-b2675594.html [accessed 23 May 2025]; *Human Rights Campaign*, 'Meta's New Policies: How They Endanger LGBTQ+ Communities and Our Tips for Staying Safe Online' (15 January 2025). https://www.hrc.org/news/metas-new-policies-how-they-endanger-lgbtq-communities-and-our-tips-for-staying-safe-online [accessed 23 May 2025].

file a complaint against an abusive post by a male supremacist will not significantly change the dynamics of online diffusion of gender-based violence and the increasing prominence of the male supremacist networks.

Moreover, the individualistic approach to online harms risks transferring the responsibility from the provider to the user and runs counter to the spirit, in our assessment, of the gendered harms provision of the Online Safety Act. It should not be incumbent on victims to solve either the single episodes or the networked abuse. Regulatory and moderation efforts should concentrate on the proliferation of abusive discourses and behaviours against women and girls. This means identifying, mapping and countering misogynistic cultures, narratives and behaviours at the network scale.

Action 9: Take appropriate action when online gender-based harms occur

See our comments on network harms under Action 6, and on the limits of reporting content under Action 8.

Question 3: Do you have any comments about the effectiveness, applicability or risks of the good practice steps or associated case studies we have highlighted in Chapter 3, 4 and 5? Are there any additional examples of good practices we should consider? Please provide evidence to support your comment.

We have grouped our answers to questions 3 and 4 together. See below.

Question 4: Do you have any feedback on our approach to encouraging providers to follow this guidance, including our proposal to publishing an assessment of how providers are addressing women and girls' safety? Do you have any examples or suggestions of other ways we could encourage providers to take up the 'good practice' recommendations?

The nine proposed actions would be even more effective if they could be used to further develop a framework for legal requirements and other incentives for platforms to implement them. The guidance needs to rest on sufficiently robust legal codes of practice and risk assessment duties, with adequate enforcement. Many civil society organisations, including the Online Safety Act Network, have expressed legitimate concerns that current frameworks do not yet go far enough.[16]

Ofcom has multiple tools to encourage better practice, for example through the risk assessment process. Paragraph 2.20 of the consultation document states that Ofcom will 'strongly encourage providers to implement relevant good practice steps'. The guidance could more strongly express Ofcom's, and society's, expectations in this matter, encouraging good practice and opening the door for new, robust frameworks.

Another way to support all nine actions would be through independent verification or validation by independent researchers. Platforms have increasingly closed off routes for researchers to access data.[17] Encouraging non-categorised providers to be transparent as to the effectiveness of their im-

[16] See for example: Online Safety Act Network, *Statement on Ofcom's Illegal Harms Code of Practice*, (15 January 2025). https://www.onlinesafetyact.net/analysis/statement-on-ofcom-s-illegal-harms-code-of-practice/ [accessed 23 May 2025]; or Online Safety Act Network, *Statement on Ofcom's protection of children codes*, (06 May 2025). https://www.onlinesafetyact.net/analysis/statement-on-the-ofcoms-protection-of-children-codes/ [accessed 23 May 2025].

[17] Minderoo Centre for Technology and Democracy, *Written Evidence: Data-Driven Research for Evidence-Based Interventions* (24 January 2025). https://doi.org/10.17863/CAM.115665.

plementation of the guidance is unlikely to be sufficient to address the problem. Including mechanisms for independent measurement and verification of good practices would be a powerful incentive to show that women and girls have positive experiences on their platforms.

Question 5: Do you have any comments on our impact assessment, rights assessment, or equality impact assessment? Please provide any information or evidence in support of your views.

We agree with Ofcom's position in the rights assessment that the draft guidance represents 'a fair balance between securing adequate protections for women and girls from harm' and the rights of users, for example of freedom of speech.

Ensuring a safer life for women and girls is fundamentally a question of human rights because online violence against women and girls is a threat to their rights to dignity, security, and safety.

Moreover, the Council of Europe Convention on preventing and combating violence against women and domestic violence, and UN Women, identify online violence against women and girls as crucial factors that can limit women's public participation and their right to express themselves.[18] Online harms can therefore also represent a threat to women's and girls' freedom of speech.

Question 6: Do you agree that our draft Guidance is likely to have positive effects on opportunities to use Welsh and treating Welsh no less favourably than English? If you disagree, please explain why, including how you consider the draft Guidance could be revised to have positive effects or more positive effects, or no adverse effects or fewer adverse effects on opportunities to use Welsh and treating Welsh no less favourably than English.

The duty to treat the Welsh language no less favourably than English raises important considerations about content moderation across language boundaries, which are absent in the draft guidance in relation to content in Welsh or any other language.

Content moderation in non-English languages receives less attention across the board.[19] Statistics suggest that Welsh is spoken by around 843,000 people in the UK.[20] Census data from 2021 suggests only 91% of people in the UK use either English or Welsh as their main language.[21] Combined with

[18] UN Women Headquarters Office, *Accelerating Efforts to Tackle Online and Technology-Facilitated Violence Against Women and Girls* (2022), https://www.unwomen.org/sites/default/files/2022-10/Accelerating-efforts-to-tackle-online-and-technology-facilitated-violence-against-women-and-girls-en_0.pdf; Council of Europe, *Convention on Preventing and Combating Violence against Women and Domestic Violence* (Istanbul Convention), CETS No. 210, opened for signature May 11, 2011, entered into force August 1, 2014, https://www.coe.int/en/web/istanbul-convention.

[19] See for example: Global Witness *How Big Tech platforms are neglecting their non-English language users*, (30 November 2023). https://globalwitness.org/en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users/ [accessed 23 May 2025]; or on automated moderation, see: Centre for Democracy & Technology, *Languages Left Behind: Automated Content Analysis in Non-English Languages*, (18 August 2022). https://cdt.org/insights/languages-left-behind-automated-content-analysis-in-non-english-languages/ [accessed 23 may 2025].

[20] Welsh Government, *Welsh language data from the Annual Population Survey: 2024* (16 April 2025), https://www.gov.wales/welsh-language-data-annual-population-survey-2024-html [accessed 17 May 2025].

[21] Office for National Statistics, *Language, England and Wales: Census 2021* (29 November 2022),

https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/bulletins/languageenglandandwales/census2021 [accessed 17 May 2025].

the fact that users in the UK will routinely access content originating outside the UK, this means harms in languages other than English represent a not insignificant risk.

Different languages carry different biases, including biases related to gender (for example relating to social misconceptions of grammatical gender), and sub-cultures in different languages can be hard to moderate. Platforms may also neglect moderation in other languages, which has enabled harm and offline violence.[22] The guidance would benefit from an exploration of these risks.

[End].

[22] *BBC*, 'Facebook admits it was used to "incite offline violence" in Myanmar', (6 November 2018). https://www.bbc.co.uk/news/world-asia-46105934 [accessed 23 May 2025]; or Amnesty International, *Myanmar: Facebook's Systems Promoted Violence Against Rohingya; Meta Owes Reparations*, (29 September 2022). https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/ [accessed 23 May 2025].