DEMOS

EPISTEMIC SECURITY NETWORK
DEMOS

EPISTEMIC SECURITY BRIEFING

ADDRESSING PLATFORM DESIGN TO MITIGATE ONLINE HARMS

DEMOS' RESPONSE TO THE SCIENCE, INNOVATION AND TECHNOLOGY COMMITTEE'S REPORT ON SOCIAL MEDIA, MISINFORMATION AND HARMFUL ALGORITHMS

JAMIE HANCOCK
HANNAH PERRY

AUGUST 2025

Demos welcomes the recent findings from the **House of Commons Science, Innovation and Technology (SIT) Committee's** ***Inquiry on Social media, misinformation and harmful algorithms.*** The report is an in-depth examination of a serious vulnerability facing the UK: digital platforms routinely amplify unreliable and hateful content in ways which can spur offline violence, as seen during 2024's Southport riots. The Committee's findings represent an urgent call to action for the UK government, Ofcom, and platforms.

At Demos, we have argued that such risks to the UK's information supply chains pose a fundamental challenge to our epistemic security – the security of our knowledge and information – and therefore for our democracy. We have also published a detailed briefing, Community Disorder, setting out how the UK government's response to information crisis scenarios can be improved.

In this briefing, we have selected four cross-cutting recommendations that are especially relevant to the Epistemic Security agenda. We highlight how these recommendations could be implemented by the UK government and where they could be strengthened.

This briefing is designed to inform the government's response to the committee's report in September.

RECOMMENDATIONS

1. REQUIRE PLATFORMS TO CONDUCT AND ACT ON RISK ASSESSMENTS

Several SIT recommendations call for the government to require digital platforms to conduct and act on risk assessments for 'legal but harmful' content:

- Platforms should have a duty to assess the risk of legal but harmful content (Paragraph 48)
- Platforms should be required to report on risk assessment results (Paragraph 48)
- Platforms should be "required to act on all risks identified in risk assessments, regardless of whether they are included in Ofcom's Codes of Practice" (Paragraph 49)
- Similar requirements should be placed on generative AI services (Paragraph 76)

Services regulated by the OSA are already required to conduct risk assessments into potential harms from illegal content on their platforms – such as hate speech, child abuse material and terrorism content – to set out mitigation measures and, to varying extents depending on the service's classification, report the results. These assessments are required to identify the likelihood of such illegal content appearing and causing harm on their platform. The SIT Committee's recommendation would extend this existing system to include legal, but harmful content too, such as "potentially harmful misinformation".

At present, services are only required to act on risks they identify in their risk assessments if these fall within the scope of OSA's illegal content duties or child protection duties. Therefore, the Committee's recommendation for services to be "required to act on all risks identified in risk assessments, regardless of whether they are included in Ofcom's Codes of Practice," would be a major expansion of the duties placed on platforms to respond to risks they identify. It would require them to address all potentially harmful and risky content they identify in their assessments, even if this is not illegal under the OSA.

We support the overall aim behind these proposals: the current arrangements under the OSA do little to address the impacts of legal but harmful content, and risk and impact assessments are a potentially useful way of addressing this gap. However, there are some issues with the SIT report's approach that require further clarification.

(A) We need a precise definition of legal but harmful content:

Whereas illegal content offences are defined in the OSA and elaborated on in Ofcom's guidance, more clarity is needed on what would count as 'legal but harmful' for the purpose of these risk assessments. Without an official definition, platforms could benefit from legal uncertainty and have more say over what counts as a 'risk' on their platforms. This definition should be set out in any regulations which give platforms additional duties to assess or act on the risk of such content. For example, it could specify priority types of risky content, such as misleading medical claims. Freedom of expression is a critical element of any free and fair democracy, and supporting these core values that underpin liberal democracy are at the heart of Demos's mission. Any definition should be as precise as possible to avoid the over-moderation of legal speech and should incorporate safeguards for freedom of expression. These could include a requirement to conduct human rights impact assessments, as we suggest below.

(B) A "duty to act" on all risks in risk assessments could introduce threats to freedom of expression:

The Committee's call for platforms to "act" on all risks identified in their risk assessments is currently very open-ended. Requiring platforms to act upon all risks they identify in their risk assessments – regardless of whether the risk is identified by Ofcom's Code of Practice – could give platforms stronger grounds to justify removing lawful speech. As it stands, the OSA requires services to implement measures to protect freedom of expression and privacy rights. Ofcom's Codes of Practice reflect this by including measures designed to mitigate potentially negative impacts on these rights. Services which implement mitigations that are not in the Codes of Practice are required to conduct impact assessments of their potential effects on freedom of expression and privacy. Ofcom should set out guidance on what these mitigations could look like.

One useful mechanism to ensure that responses to the risk of legal but harmful content comply with human rights standards could be to require platforms to incorporate human rights impact assessments (HRIAs) into their risk assessment processes. HRIAs are a tool organisations use to prove their compliance with human rights law and principles. Standards for conducting HRIAs have been set out by the OECD, the International Business Leaders Forum (IBLF), and others. In the UK, HRIAs are used by companies like the John Lewis Partnership and by some Scottish government bodies. Useful guidance on applying HRIAs to content moderation has been set out by organisations such as BSR.

2. ESTABLISH THE 'RIGHT TO RESET' AND OTHER USER CONTROLS

The report includes calls for users to be given more control over how platforms operate. This includes a suggestion that platforms should be required to give users the 'right to reset' their recommendation algorithms (Paragraph 32).

The proposal is an interesting way to provide users with more agency and there is precedent for its use. Platforms like TikTok and Instagram are already offering these features. However, more detail is needed on how it should be implemented to ensure these changes are meaningful for users.

A poor way to implement this proposal would be to require services to offer users the option to reset their recommendation algorithm without requiring that they make this feature easily accessible or to explain it in plain language. One consequence of this approach could be that platforms could use deceptive design techniques to limit how often the tool is used. For example, while TikTok and Instagram already have this feature, they bury it within settings in a way which makes it harder to access. A requirement to give users the feature as a prominent option in accessible language would be similar to existing requirements to provide users with clear opt-outs for cookies and data processing.

To prevent a situation such as this, platforms could be required to:

- Make the feature clearly visible, easily accessible, and explained to users in a way that is easy to understand.
- Include greater algorithmic transparency requirements, including a duty to provide meaningful information about how the recommendation algorithm works so that users can know if anything has actually changed.

Another option for enhanced user controls, which would further the goal of SIT's original 'right to reset' recommendation, could be to give users the '*right to decide*': the ability to select between the recommendation algorithms used to prioritise content. This idea aligns with

findings from researchers from the Knight-Georgetown Institute and other organisations, who have called for users to be given "choices and defaults that allow individuals to tailor their platform experiences and switch between different recommendation systems". Platforms like BlueSky and Reddit already offer this feature.

These more expansive user controls and rights could complement measures like Ofcom's voluntary 'media literacy by design' principles, which are intended to ensure users are given better information to allow them to make informed decisions. Where 'media literacy by design' focuses largely on changes to user interfaces to provide users with relevant information and prompt critical thinking, the kinds of user controls that the Committee calls for would give users more power over how a platform works for them.

3. BETTER DATA ACCESS FOR RESEARCHERS

The Committee identifies that a lack of access to high-quality data on platforms' inner workings has made it extremely difficult for researchers to conduct online safety research. The studies which do exist are either restricted to analysing the small amount of publicly-facing information that platforms provide to users – which can limit them to making educated guesses about how systems work – or they are contingent on platforms choosing to give them access. For example, studies that examine X's Community Notes system, such as the Center for Countering Digital Hate (CCDH)'s report on the system's effectiveness during elections and Demos' own research on its performance during the Southport riots, are dependent on X continuing to make this data available to researchers. To address this challenge, the Committee calls for social media and generative AI platforms to be required to provide online safety researchers commissioned by the government with "full" data access, including all data used to train recommendation systems and the weights such algorithms use when they tailor content to different users (Paragraphs 29 and 77).

This recommendation could be addressed through regulation soon. The recently-passed Data Use and Access Act (DUAA) gives the government powers to set out regulations to require platforms to provide independent researchers with data access for online safety research. The DUAA does not itself set out what these regulations will be or how they would be enforced.

Furthermore, Ofcom has recently proposed three possible measures to give researchers access to information about digital platforms, which could address the Committee's recommendation:

A. Clarify existing legal rules on what researcher access is already allowed.

B. Create new duties for platforms to give researchers access, such as requirements to provide data directly or via an interface, "enforced by a backstop regulator".

C. Grant legal powers to a "trusted third party" which would act as an "independent intermediary" to "facilitate and manage researchers' access to data".

We agree with the Committee that researchers must be given greater access to platforms' data. However, we would build on SIT and Ofcom's proposals by mandating that platforms should be mandated to make such data available much more widely. Rather than just providing data to researchers as Ofcom suggests, platforms should give access to any pre-vetted public interest researchers at no additional cost. This access should be regulated by an independently facilitated request system and Code of Practice – similar to the system set out by the EU's Digital Services Act Article 40.4 and 40.12 – and could be facilitated through a secure online access environment. By 'vetted public interest researchers', we include civil society organisations who play a crucial role in identifying harmful content as well as researchers based at academic institutions.

4. FUND AND CONDUCT MORE RESEARCH

The Committee evidences how digital platforms employ "algorithms [which] can amplify content regardless of accuracy or potential for harm". Crucially, the Committee highlights gaps in the knowledge-base regarding how such recommendation algorithms work and the comparative efficacy of different content moderation systems.

The report recommends more government-funded research to address these gaps. It calls for:

- The government to commission independent researchers to examine how social media recommendation systems "spread, amplify or prioritise harmful content". The researchers should have " full access to the inner functions of the systems that major platforms use to algorithmically recommend content" (Paragraph 29).
- The government to commission independent research on the relative benefits of different content moderation methods for preventing the spread of harmful content, such as "independent third-party fact-checkers, crowd-sourced context provision, and AI driven detection of misinformation" (Paragraph 41).

We agree with the aim of the SIT Committee's recommendations and have previously called for more research funding. However, we disagree that the UK government should be funders of this research. Instead, we recommend that funding is granted by bodies who already regularly fund or publish independent research, such as UK Research and Innovation (UKRI) or the online safety regulator, Ofcom.

If funding is granted by the UK government directly, there is a risk the research's independence could be undermined. For example, if the government commissions research that results in findings which are critical of platforms' content moderation strategies, these could be viewed as in conflict with other policy goals. Take trade policy: there is a chance that such research would aggravate the US government, an important trading partner, given that key figures in the Trump administration – including Vice President JD Vance – have already made critical comments about the rollout of the OSA. The US reportedly pressured the UK to make changes to the online safety regime as part of trade talks and also criticised the application of the UK's Digital Services Tax to American tech platforms. While the UK appears to have finished its trade negotiations without making concessions on the OSA or other digital policies, this does not rule out similar pressures in the future.

Commissioning the research through external institutions that do not have this additional layer of complexity, such as UKRI or Ofcom, could help to safeguard its independence even if the findings challenge the government's existing policy positions or trade policy objectives.

CONCLUSION

One year on from the Southport riots, it is clear that much more must be done to improve the digital environment and avoid another crisis situation. The Committee's report is a powerful summary of many of the challenges facing our epistemic security and identifies several important first steps towards addressing them. The government and regulators should take heed of the recommendations, as well as the calls for action that Demos and other civil society organisations have made. We look forward to September 11th, when we expect to hear from how the UK government responds to the Committee's findings.

Licence to publish

Demos – Licence to Publish

The work (as defined below) is provided under the terms of this licence ('licence'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this licence is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this licence. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions

a 'Collective Work' means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this Licence.

b 'Derivative Work' means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this Licence.

c 'Licensor' means the individual or entity that offers the Work under the terms of this Licence.

d 'Original Author' means the individual or entity who created the Work.

e 'Work' means the copyrightable work of authorship offered under the terms of this Licence.

f 'You' means an individual or entity exercising rights under this Licence who has not previously violated the terms of this Licence with respect to the Work, or who has received express permission from Demos to exercise rights under this Licence despite a previous violation.

2 Fair Use Rights

Nothing in this licence is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 Licence Grant

Subject to the terms and conditions of this Licence, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) licence to exercise the rights in the Work as stated below:

a to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

b to distribute copies or phono-records of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions

The licence granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this Licence, and You must include a copy of, or the Uniform Resource Identifier for, this Licence with every copy or phono-record of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this Licence or the recipients' exercise of the rights granted hereunder. You may not sublicence the Work. You must keep intact all notices that refer to this Licence and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this Licence Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this Licence. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

b You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended

for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

c If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, you must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer

a By offering the Work for public release under this Licence, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:

i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;

ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.

b Except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability

Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination

a This Licence and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this Licence. Individuals or entities who have received Collective Works from You under this Licence, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this Licence.

b Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this Licence (or any other licence that has been, or is required to be, granted under the terms of this Licence), and this Licence will continue in full force and effect unless terminated as stated above.

8 Miscellaneous

a Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this Licence.

b If any provision of this Licence is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Licence, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

c No term or provision of this Licence shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

d This Licence constitutes the entire agreement between the parties with respect to the Work licenced here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This Licence may not be modified without the mutual written agreement of Demos and You.

DEMOS

Demos is a champion of people, ideas and democracy. We bring people together. We bridge divides. We listen and we understand. We are practical about the problems we face, but endlessly optimistic and ambitious about our capacity, together, to overcome them.

At a crossroads in Britain's history, we need ideas for renewal, reconnection and the restoration of hope. Challenges from populism to climate change remain unsolved, and a technological revolution dawns, but the centre of politics has been intellectually paralysed. Demos will change that. We can counter the impossible promises of the political extremes, and challenge despair – by bringing to life an aspirational narrative about the future of Britain that is rooted in the hopes and ambitions of people from across our country.

Demos is an independent, educational charity, registered in England and Wales. (Charity Registration no. 1042046)

Find out more at **www.demos.co.uk**

PUBLISHED BY DEMOS AUGUST 2025

15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK