DEMOS

EPISTEMIC SECURITY NETWORK
DEMOS

15 WHITEHALL,
LONDON, SW1A 2DD

020 3878 3955
hello@demos.co.uk
www.demos.co.uk

OFCOM CONSULTATION: ADDITIONAL SAFETY MEASURES

Demos Digital: Response
20 October 2025

About Demos

Demos is Britain's leading cross-party think tank.[1] Demos is an independent educational charity which conducts public benefit research on issues of politics, economics, technology, public deliberation, the environment and public policy.

Demos Digital is Demos' digital policy research hub.[2] We work to shape a future in which technological development and governance is aligned with the needs and values of the public.

At Demos Digital, we contend that our *epistemic security* – the resilience of the UK's information supply chains that our democracy depends on – is under threat.[3] By 'information supply chains', we mean the full lifecycle of the information that we use to understand the world and make decisions, from production to distribution to eventual action. Demos Digital co-ordinates the Epistemic Security Network to help address these challenges: a group of civil society organisations and individuals which collaborates to fortify our information supply chains.[4]

[1] https://demos.co.uk
[2] https://demos.co.uk/demos-digital/
[3] https://demos.co.uk/wp-content/uploads/2025/02/Epistemic-Security-2029_accessible.pdf
[4] https://demos.co.uk/epistemic-security-network/

Volume 14: Recommender systems

Question 31: Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

The Demos Digital team agrees with Ofcom's proposal to exclude illegal content from recommender systems until the content has been reviewed by content moderation teams.

We believe that this proposal will go some way towards mitigating the role recommender algorithms play in amplifying illegal content such as hate speech. However, we believe there are notable gaps to this proposal that risk undermining efforts to limit the impacts of recommender algorithms on the viral dissemination of illegal content. In order to tackle these gaps, we propose a number of additional measures which are expanded on below; specific guidance on the use of automated tools for identifying potentially illegal content to exclude from recommender algorithms; and greater transparency and user choice.

Point of agreement with the proposals

We strongly agree that additional safety measures regarding recommender algorithms are needed because of the existing flaws that have been identified in such systems. These algorithms currently play a significant role in the amplification of illegal content on online platforms – such as hate speech, incitement to violence and suicide content.

There is a growing body of research evidencing the role of recommender algorithms in contributing to offline violence, such as during the 2024 Southport riots. While the spread of illegal content on various platforms that contributed to the offline violence is well-documented, Amnesty International has identified that X's recommender algorithm does not currently assess a tweet's substance for potential harm before boosting a post. Instead, X's recommender algorithms boost content generating heated replies, which is often the case with harmful or divisive content.[5] For example, Lucy Connelly's posts in which she called for mass deportations and setting fire to asylum hotels were viewed 310,000 times, which is far beyond the reach of her immediate network, and for which she was later sentenced to 31 months in prison for Public Order Offences.[6]

While such examples stem from a period prior to the Online Safety Act coming into force, the Act continues to be insufficient to mitigate these outcomes because measures regulating recommender algorithms are absent from the Act. Recent research from the Molly Rose Foundation demonstrates that recommender algorithms on TikTok and Instagram are still promoting suicide-related content, even after the Online Safety Act came

[5] https://www.amnesty.org/en/documents/eur45/0618/2025/en/
[6] https://www.amnesty.org/en/documents/eur45/0618/2025/en/

into force.[7] After setting up a number of dummy accounts, they found: "Over half (55%) of recommended harmful posts on TikTok's For You page contained references to suicide and self-harm ideation and 16% referenced suicide methods, including some which the researchers had never encountered before."[8]

Therefore, we agree that additional safety measures relating to recommender algorithms and illegal content are urgently needed. We also agree that Ofcom's specific proposal to exclude illegal content from recommender algorithms until it has been manually reviewed is likely to reduce the risks of illegal content such as hate speech or suicide content being virally disseminated on online platforms.

Recommended improvements

There are three areas where the proposal to exclude illegal content from recommender algorithms until it has been reviewed could be improved detailed below:

(1) Recommendation 1: Specific guidance on the use of automated tools for identifying potentially illegal content:

Ofcom should provide platforms with specific guidance on the use of automated tools for identifying illegal content for excluding from recommender algorithms. Currently, Ofcom's proposal states: "We do not propose to be proscriptive about the information that providers should take into account to indicate content is potentially illegal." However, because the proposal both encourages platforms to use automated content identification tools and gives platforms a large degree of responsibility for the parameters of such tools, we believe that this lack of guidance may result in inadequate standards for automated content identification tools, carrying risks for the over-exclusion of legal content by online platforms, or 'shadow banning'.

Automated content identification tools are known to struggle with reliability and bias. This is especially the case for identifying illegal content such as hate speech or 'terrorist content', where a high degree of local and contextual knowledge is needed. A recent study by researchers from the University of Pennsylvania, OpenAI and DeepSeek found that, across seven AI moderation systems: "when it comes to hate speech, the AI driving these decisions is wildly inconsistent." The systems even had dramatic inconsistencies when evaluating identical hate speech content.[9] While this study evaluates automated

7 https://mollyrosefoundation.org/suicide-and-self-harm-content-still-recommended-at-industrial-scale-by-tiktok-and-instagram-eight-years-after-mollys-death/

8 https://mollyrosefoundation.org/suicide-and-self-harm-content-still-recommended-at-industrial-scale-by-tiktok-and-instagram-eight-years-after-mollys-death/

9 https://www.independent.co.uk/news/uk/home-news/ai-hate-speech-study-university-pennsylvania-b2826860.html

moderation systems, we can assume similarities for automated content identification systems, such as those included in Ofcom's proposals.

Because of these risks of inconsistency, Ofcom should provide specific guidance for platforms' responsible use of automated content identification tools, including: transparency reporting; quality control standards for automated identification systems, including bias, reliability and accuracy; impact assessments for evaluating the automated systems; and model parameters for identifying illegal content. We believe this would alleviate some of the risks of automated content identification systems – such as inconsistencies, inaccuracies, and bias – which could result in the over-exclusion of legal content, or under-exclusion of illegal content.

(2) Recommendation 2: Greater transparency of platforms' recommender algorithms

Ofcom should ensure that the proposals for mitigating the role of recommender systems in amplifying illegal content are activated in combination with transparency mechanisms. This is a crucial first step for ensuring greater accountability of recommender algorithms on online platforms. Greater transparency would allow researchers, regulators, and the public to assess the role of recommender algorithms in relation to illegal content, and the efficacy of Ofcom's measures, such as these new additional safety measures.

We understand that Ofcom is addressing transparency guidance in relation to the Online Safety Act separately and at a later date to the additional safety measures. However, we strongly believe that launching the additional safety measures for recommender algorithms without transparency mechanisms in place would undermine the efficacy of the measures because of the importance of being able to assess the impact of these changes both intended and unintended. As such, we offer proposals for transparency mechanisms for recommender algorithms below that should be included within the additional safety measures.

Currently, there is minimal transparency of recommender systems on online platforms. Recommender algorithms are widely described as a "black box" by members of the public, researchers, and even technical experts.[10] For example, in a 2024 qualitative study involving a public survey of perceptions on recommender algorithms, 583 respondents said there was "no sufficient information" on recommender algorithms, with only 58 saying there was "partially or fully sufficient information".[11] Despite the central role that recommender algorithms play in determining what content users receive on their feeds, it is extremely

[10] https://firstmonday.org/ojs/index.php/fm/article/view/13357/11634; https://www.independent.co.uk/tech/tiktok-black-algorithm-box-study-b2534606.htm; https://journals.law.harvard.edu/lpr/wp-content/uploads/sites/89/2024/08/18.1-Right-to-Know-Social-Media-Algorithms.pdf; https://www.politico.eu/article/facebook-whistleblower-frances-haugen-europe-parliament/

[11] https://firstmonday.org/ojs/index.php/fm/article/view/13357/11634

challenging for users, researchers, or regulators to understand how they work, such as through data inputs or model weightings.

This makes evaluating the impacts of recommender systems in relation to illegal content extremely difficult – including robustly considering the impacts of proposals such as Ofcom's additional safety measures. Indeed, Amnesty International's technical analysis of X's recommender system during the 2024 Southport riots was only possible because X made their code public, which is a rare exception.[12] The study undertaken by the Molly Rose Foundation evaluating the promotion of suicide content by TikTok's recommender algorithm relied on the organisation's creation and extended testing of dummy accounts to monitor content that is pushed.[13] This is a time-consuming and expensive method, placing significant barriers on independent evaluations of recommender algorithms.

Robust transparency measures would be a vital first step towards understanding and mitigating the impacts of recommender algorithms in relation to illegal content. Several academic studies have also shown that greater transparency of recommender algorithms on online platforms would: improve user satisfaction;[14] foster greater public trust;[15] and moderate privacy concerns.[16]

Specifically, we propose that greater transparency of recommender algorithms can be achieved through the following measures. These measures should be introduced alongside the additional safety measures:

a) Public disclosure of recommender algorithm parameters: As a crucial first step, Ofcom should encourage platforms to disclose the main parameters of their recommender algorithms, including input data and weightings.[17] Platforms should disclose the following input data: all sources of information used in rankings (eg. item content and metadata; engagement history; user survey data; quality feedback from users; annotations from raters; user settings; profile and social graph data; and context data (day, time, location).[18] Given that the parameters of recommender algorithms are regularly updated and tweaked by online platforms, these should be disclosed at regular intervals, such as monthly.

---

[12] https://www.amnesty.org/en/documents/eur45/0618/2025/en/

[13] https://mollyrosefoundation.org/suicide-and-self-harm-content-still-recommended-at-industrial-scale-by-tiktok-and-instagram-eight-years-after-mollys-death/

[14] https://www.sciencedirect.com/science/article/abs/pii/S1071581913002024

[15] https://doi.org/10.1108/INTR-02-2021-0087

[16] https://www.tandfonline.com/doi/full/10.1080/08838151.2022.2057984

[17] https://dsa-observatory.eu/2025/05/19/making-recommender-systems-work-for-people/

[18] https://kgi.georgetown.edu/wp-content/uploads/2025/03/Better-Feeds-EU-Policy-Brief-2025.pdf

b) Researcher access to data and insights about content flows: Ofcom should encourage platforms to publish key insights and data about their content flows, especially in relation to illegal content such as hate speech.[19] This should ensure that researchers are able to understand the prevalence, viral dissemination and engagement of illegal content on online platforms.[20] Specifically, platforms should publish: a sample of the public content that is most highly disseminated on the platform – such as the top, 1000 "most-viewed" posts per platform – and a sample of the public content that receives the highest engagement.[21] This is particularly important around major incidents.[22]

c) Researcher access to the latest recommender API: Finally, Ofcom should encourage platforms to provide researchers with access to the latest recommender API and/or create a safe data-sharing mechanism. This would enable experts to monitor trends in content flows of illegal content such as hate speech in real time.[23]

Taken together, these transparency measures would be particularly beneficial for enabling researchers and independent experts to evaluate the long-term impacts of recommender algorithms in relation to illegal content such as hate speech. The measures would be crucial for robust independent audits of recommender systems.

The EU's Digital Services Act (DSA) sets a valuable precedent for Ofcom in relation to greater transparency of recommender algorithms.[24] Several Articles in the DSA place mandatory obligations on platforms to provide transparency on their recommender algorithms: Article 27 requires online platforms to publish the main parameters used in their recommender algorithms;[25] while Article 40 requires providers of Very Large Online Platforms (VLOPs) to provide vetted researchers with access to non-public data for evaluating systemic risks such as illegal content.[26]

To mitigate the impacts of recommender algorithms on the viral dissemination of illegal content, Ofcom should follow the course of the EU DSA and adopt our proposals for transparency measures. We understand that Ofcom is considering transparency separately to the additional safety measures. We strongly encourage Ofcom to ensure transparency

---

19 https://counterhate.com/blog/show-us-whats-viral-a-request-to-platforms-to-share-the-most-viewed-posts-in-the-eu/

20 https://www.amnesty.org/en/documents/eur45/0618/2025/en/

21 https://counterhate.com/blog/show-us-whats-viral-a-request-to-platforms-to-share-the-most-viewed-posts-in-the-eu/; https://kgi.georgetown.edu/wp-content/uploads/2025/03/Better-Feeds-EU-Policy-Brief-2025.pdf

22 https://www.amnesty.org/en/documents/eur45/0618/2025/en/

23 https://www.amnesty.org/en/documents/eur45/0618/2025/en/

24 https://dsa-observatory.eu/2025/05/19/making-recommender-systems-work-for-people/

25 https://www.eu-digital-services-act.com/Digital_Services_Act_Article_27.html

26 https://www.eu-digital-services-act.com/Digital_Services_Act_Article_40.html

mechanisms for recommender algorithms are in place before launching the additional safety measures.

(3) Recommendation 3: Greater user choice over recommender algorithms

Ofcom should encourage platforms to give users greater choice over the recommender systems used to push content to them.[27] There is growing consensus among experts that platforms should provide users with more choice about the make-up of the algorithms on their social media feeds.[28] This would provide additional safety in relation to illegal content such as hate speech as users would be able to select alternative recommender systems if they felt that illegal content was particularly prominent on their feeds based on the system selected. Crucially this proposal is distinct from enabling users to understand their choices and ensuring they are aware of how to change the settings (as discussed in the 'How to promote media literacy consultation' - we are proposing here that Ofcom recommends platforms give users meaningful alternative choices in the recommender algorithm that is deployed on their service.

BlueSky sets a valuable precedent for how platforms could provide users with greater choice over recommender algorithms. In 2021, BlueSky opened up its data to allow developers to build custom algorithms. This established a 'marketplace of algorithms', giving users greater agency over what they see.[29] It was reported that already in 2021, 20 per cent of BlueSky's 265,000 users were using custom feeds.[30] Greater user choice could have a positive impact on users' exposure to illegal content as it would reduce reliance on profiling-based algorithms that are primarily optimised for user engagement over safety, and as such have been proven to amplify illegal content – as per research by Amnesty International and the Molly Rose Foundation. Additional benefits of algorithmic choice are reductions in market monopoly with positive impacts for economic growth.

Specifically, user choice architectures for recommender algorithms should include the following features:

a) Alternatives to profiling-based systems: Ofcom should encourage platforms to offer users alternatives to recommender algorithms that draw on surveillance-based methods to personalise user feeds, such as sensitive personal information. These alternatives could be provided to users by requiring them to opt-in to profiling-based algorithms. Articles 27 and 38 of the DSA offer valuable precedents

[27] https://kgi.georgetown.edu/research-and-commentary/better-feeds/

[28] https://cyber.fsi.stanford.edu/publication/report-working-group-platform-scale#:~:text=The%20Program%20on%20Democracy%20and,if%20appropriate%2C%20recommend%20remedial%20policies ; https://kgi.georgetown.edu/wp-content/uploads/2025/03/Better-Feeds-EU-Policy-Brief-2025.pdf

[29] https://bsky.social/about/blog/3-30-2023-algorithmic-choice

[30] https://www.nytimes.com/2023/08/17/opinion/social-media-algorithm-choice.html

for user choice architectures.[31] As per Article 27 of the DSA, these options should be direct and easily accessible to users.

b) Right to reset: Ofcom should encourage platforms to offer users the option to reset their recommender algorithms. This proposal is included in the Science, Innovation and Technology Committee's (SIT) recent report on 'Social media, misinformation, and harmful algorithms'.[32] Polling conducted by YouGov and commissioned by Demos in September 2025 also found that 65% of the public are "worried" about "social media algorithms using your background data to decide which content to show to you."

Question 32: Do you have evidence on what types of content are typically recommended to users as part of concerted foreign interference activity?

(N/A)

Question 33: Do you have evidence on whether services track the extent of algorithmic amplification, such as impressions and reach, of content that is later deemed illegal/violating. If so, do they (or does your service) use this information to enhance the safety of their systems?

(N/A)

Question 34: Do you agree with our assessment of the impacts (including costs) associated with this proposal? Please provide any relevant evidence which supports your position.
(N/A)

Question 35: Are there any impacts of the proposed measure that we have not identified? Please provide the rationale and any supporting evidence for your response.
(N/A)

[31] https://www.eu-digital-services-act.com/Digital_Services_Act_Article_27.html; https://www.eu-digital-services-act.com/Digital_Services_Act_Article_38.html
[32] https://committees.parliament.uk/publications/48745/documents/258221/default/

Volume 20: Crisis response

Question 49: Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

Demos Digital agrees with the principle behind the crisis response proposals: services should be required to create crisis protocols to respond to incidents that involve a high risk of the spread of illegal content.

We also recognise that the proposals are limited to addressing situations where services see an increased risk of priority illegal content and/or content harmful to children, based on service providers' existing duties under Section 10[33] of the Online Safety Act (OSA) as well as the illegal content definitions set out in OSA Schedules 5[34], 6[35], and 7.[36] As a result, Demos Digital understands why the proposals do not cover certain types of content which may be problematic during crises, such as false or misleading content, but are not an illegal priority offence as listed in Schedules 5-7 of the Act.

Overall, however, Demos Digital do not think the proposals, as they currently stand, go far enough to ensure that platforms' responses are sufficient to mitigate the kinds of outcomes witnessed during the Southport riots.

Below, we have detailed 4 points of agreement: avoiding a content-specific approach, avoiding prescribing a specific content moderation mechanism, maintaining and strengthening existing moderation systems, and minimising risks to freedom of expression. We have also made 14 recommendations for strengthening the proposals which are grouped into six themes: a need for more a more precise crisis definition, a need for clarity in who decides that a crisis is ongoing, a need for time scales for crisis responses, standards for designing crisis protocols, best practices for crisis responses and strong transparency and accountability measures. The following subsection addresses our points of agreement and is followed by a subsection which addresses each recommendation.

Points of agreement with the proposals

(1) Avoiding a content-specific approach

Demos Digital agrees with the decision to focus the proposals on procedures and teams that services should set up, rather than being proscriptive about what specific types of content should be moderated during a crisis. We think this facilitates flexibility across

[33] https://www.legislation.gov.uk/ukpga/2023/50/section/10
[34] https://www.legislation.gov.uk/ukpga/2023/50/schedule/5
[35] https://www.legislation.gov.uk/ukpga/2023/50/schedule/6
[36] https://www.legislation.gov.uk/ukpga/2023/50/schedule/7

different types of platform, content and provides a foundation for limiting restrictions on freedom of expression. This is because:

(a) Flexibility for platforms: Each service hosts a different mix of content, has a different user-base, operates different policies, and therefore has a different risk profile which they should identify through their risk assessments. This approach aligns with the principle that regulation should be technology and platform agnostic. In so doing, it aligns with the approach taken by Ofcom's Risk Assessment Guidance and Risk Profiles, which allow for services to tailor their risk assessments to their specific contexts and systems.[37]

(b) Flexibility for different types of content: The specific types of harmful content will vary depending on the crisis. For example, the Southport Riots saw a marked increase in incitements to violence and hate speech content[38], while election periods can see a rise in abuse and harassment content directed towards electoral candidates[39].

(c) Foundations for freedom of expression: A focus on improving resources for existing content moderation systems avoids the risk that Ofcom's codes lead directly to restrictions on Freedom of Expression, as Ofcom has recognised in its rights assessment (Paragraph 20.69).

(2) Avoiding proscribing a specific content moderation mechanism

Demos Digital supports the decision not to prescribe which methods or systems platforms should use for content moderation during a crisis. This approach recognises the variety of content moderation systems and audiences relevant to the different platforms:

(a) Flexibility for different content moderation systems. Guidance that specifies which technical approach services should take would risk being appropriate for some platforms, but not others. For example, TikTok appears to rely heavily on proprietary AI-based content moderation tools, with which remove up to 80% of content that is

---

37 https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/risk-assessment-guidance-and-risk-profiles.pdf?v=390984

38 https://www.isdglobal.org/digital_dispatches/evidencing-a-rise-in-anti-muslim-and-anti-migrant-online-hate-following-the-southport-attack/; https://demos.co.uk/research/researching-the-riots-an-evaluation-of-the-efficacy-of-community-notes-during-the-2024-southport-riots/

39 https://commonslibrary.parliament.uk/research-briefings/cbp-9192/; https://www.electoralcommission.org.uk/news-and-views/our-responses-consultations/evidence-speakers-conference-security-candidates-mps-and-elections

moderated on the site[40], forum-based services such as Reddit rely heavily on volunteer moderators[41].

(b) Flexibility for different contexts: Different services have different user-bases, mixes of illegal content, and operate at different scales. This approach aligns with the principle that regulation should be technology and platform agnostic. In so doing, it aligns with the approach taken by Ofcom's Risk Assessment Guidance and Risk Profiles, which allow for services to tailor their risk assessments to their specific contexts and systems.[42]

(3) Maintaining and strengthening existing moderation systems

Demos Digital agrees with the proposals' focus on boosting resources for existing moderation systems in the event of a crisis rather than requiring new moderation systems. This is because providing more resources for existing moderation systems ensures continuity and consistency with content moderation policies outside crisis periods. It also avoids prompting platforms to make qualitative changes in how moderation policies are applied during a crisis.

Demos Digital recognises that prompting different moderation policies during a crisis would have negative effects especially if not communicated to users transparently. Such inconsistency risks harming public trust and fuelling allegations of censorship, which could undermine the crisis response. Research has indicated that inconsistent applications of content moderation policies can fuel a perception of double standards, especially when these inconsistencies affect minoritised communities[43].

(4) Minimising risks to freedom of expression during a crisis

Demos Digital supports the decision not to "recommend that services have a higher tolerance for false positives in their content moderation processes during a crisis" (Paragraph 20.70).

This is because any measures which lead platforms to have a higher tolerance for false positives in content moderation – i.e., a higher rate of non-illegal content being moderated

---

40 https://www.reuters.com/technology/bytedance-cuts-over-700-jobs-malaysia-shift-towards-ai-moderation-sources-say-2024-10-11/

41 https://redditinc.com/policies/moderator-code-of-conduct

42 https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/risk-assessment-guidance-and-risk-profiles.pdf?v=390984

43 https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation, https://www.article19.org/wp-content/uploads/2022/06/Summary-report-social-media-for-peace.pdf

as if it were illegal – would disproportionately undermine users' right to freedom of expression. Ofcom have correctly identified in the Rights Assessment that crisis response measures that would "recommend that [services] take down more content than they otherwise would have" would pose an unjustified risk to users' freedom of expression (Paragraph 20.70).

However, we recommend that Ofcom should set out further requirements to ensure that platforms do not choose to implement crisis response policies of their own accord which lead to a higher rate of non-illegal content being moderated. Ofcom has noted this risk in its Rights Assessment (Paragraphs 20.71-20.72), but does not identify proactive measures that would mitigate it. Recommendation 14 outlines measures that could be taken to address this.

Recommended improvements

We have identified 14 gaps and concerns for which we have set out recommendations. We have grouped these into six themes: adding precision to the crisis definition; determining who decides that a crisis is ongoing; time scales for crisis responses; standards for designing crisis protocols; best practices for crisis responses; and strong transparency and accountability measures. The following paragraphs address each theme to summarise each recommendation, explain our reasoning, and provide evidence.

Adding precision to the crisis definition

Recommendation (1): Resolve ambiguities in how a crisis is defined and provide more examples for what would be considered a crisis

Demos Digital suggests the current definition of crisis requires more specificity and detail. This is important to ensure that the definition is applied consistently and proportionately. A vague definition is not appropriate for high-risk crises with a narrow margin of error: public trust is on the line and the consequences of misidentifying the situation could be drastic.

It has been shown that inconsistent and poorly explained platform moderation decisions can fuel mistrust during crisis situations such as the Covid pandemic[44]. Polling in the immediate aftermath of the Southport riots suggested that "two thirds of Britons (66%) said that social media companies should be held responsible for posts inciting criminal behaviour" during the unrest[45]. However, a recent poll we ran indicated that 60% of the public are somewhat or very worried about free speech being limited by regulations put on

44 https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf

45 https://yougov.co.uk/politics/articles/50288-two-thirds-of-britons-say-social-media-companies-should-be-held-responsible-for-posts-inciting-riots

social media companies and their algorithms. This discrepancy – between a desire for more accountability when social media respond to illegal content and concerns about freedom of expression – suggests that the public's expectations and fears are both high. Getting the balance wrong in a future crisis could be disastrous for public trust. Therefore, a clear, detailed, and strongly delineated definition is a crucial means to ensure compliance, avoid misuses and prevent overreach that undermines users' rights as well as to shore up trust in their justified application.

Here, we suggest referring to the Civil Contingencies Act 2004 (CCA)'s definition of an emergency as a useful example of how crises may be defined for legal purposes.[46] The CCA provides a definition of 'emergency' which is both detailed and covers a range of situations. To do so, it describes multiple kinds of emergencies: either

(a) "an event or situation which threatens serious damage to human welfare in a place in the United Kingdom"

(b) Or "an event or situation which threatens serious damage to the environment of a place in the United Kingdom"

(c) Or "war, or terrorism, which threatens serious damage to the security of the United Kingdom.

The CCA provides further details on the risks that these types of emergencies must pose. These risks include "loss of human life" and "disruption of a system of communication". Additionally, it sets out an extensive list of examples of emergencies that it applies to. These include the "disruption of a supply of money, food, water, energy or fuel", "disruption of a system of communication", "disruption of facilities for transport," or the "disruption of services relating to health."

We also recommend providing additional examples of situations for how the definition of crisis could be applied. For context, Ofcom's current proposal provides a short list of examples of situations which "may, depending on the circumstances, satisfy [the] definition of a crisis". These are "nationwide riots, large scale terrorist attacks and/or inter-religious or inter-ethnic violence". We recommend that Ofcom should set out a much more expansive – but non-exhaustive – list to guide decision-making. The list of examples would not need to be exhaustive, but should be long and detailed enough to provide a better indication of the scope of the definition.

Not all risks to public safety involve mass violence or hate. Non-violent crisis situations that pose risks to public safety include environmental disasters and public health emergencies,

[46] HM Government (2025). Civil Contingencies Act 2004 Section 1. https://www.legislation.gov.uk/ukpga/2004/36/section/1

such as pandemics. Therefore, the list should include examples of such non-violent crisis situations.

As an illustration of best practice, the Civil Contingencies Act 2004 sets out an extensive list of examples of emergencies that it applies to.[47] These include the "disruption of a supply of money, food, water, energy or fuel", "disruption of a system of communication", "disruption of facilities for transport," or the "disruption of services relating to health." Similarly, the European Union's Digital Services Act (DSA) suggests that relevant crises may include "armed conflicts or acts of terrorism, including emerging conflicts or acts of terrorism, natural disasters such as earthquakes and hurricanes, as well as from pandemics and other serious cross-border threats to public health."[48]

We have provided a more comprehensive set of recommendations on how to make the crisis definition more precise in our response to Question 50.

Our recommendation is based on the need to provide services, users, civil society, and policymakers beyond Ofcom with clarity about what the definition will apply to. This is crucial for providing regulatory certainty and ensuring that there is consistency in the application of the definition.

Recommendation (2): Use a graduated definition of crisis severity to set minimum standards for responses

The current proposal does not reflect the fact that crisis situations may be of differing levels of severity, or that some crises may require more intensive responses than others. Instead, it presents a binary view of crises: either a situation is a crisis or it is not.

Demos Digital recommends that Ofcom should reconsider this approach by introducing a graduated understanding of crises based on levels of severity. One option for defining such severity levels would be to tie them to the risk of harm or degree of threat to public safety. These levels would then be used to set higher minimum standards for responses to higher severity crises. For example, the definition could outline what would be considered to be crises with medium, high, and very high risks of harm to users and public safety.

Taking this approach would allow for greater flexibility in how services apply their crisis protocols, depending on needs of the situation, while also potentially raising minimum standards. Some crises are not as severe as others and will not require the same level of response. Rather than giving services total discretion about the intensity of response that is warranted, it would be best if Ofcom can provide a framework to guide and regulate such decisions.

---

[47] HM Government (2025). Civil Contingencies Act 2004 Section 1. https://www.legislation.gov.uk/ukpga/2004/36/section/1

[48] https://www.legislation.gov.uk/ukpga/2004/36/section/1

It is common practice to use graduated definitions of crisis severity as a way of making crisis responses more effective. For example, the UK Health Security Agency uses four severity levels for its incident response plan: routine, standard, enhanced, and severe.[49] The National Cyber Security Centre (NCSC) uses a six-tier grading system to categorise the severity of cybersecurity incidents.[50] Likewise, it is our understanding that the government's Defending Democracy Taskforce (DDTF) and National Security Online Information Team (NSOIT) use a three-tier grading system for its information incident response. Finally, Full Fact's *Framework for Information Incidents* uses five tiers of severity[51]. All of these examples could provide inspiration for how Ofcom could approach its severity levels.

Determining who decides that a crisis is ongoing

Recommendation (3): Specify a mechanism whereby a publicly accountable body may identify a crisis and trigger protocols

The current proposals imply that services will have the responsibility for identifying crisis situations, based on internally-devised indicators, and triggering their internal crisis protocols (Paragraphs 20.30-20.31). The proposals also imply that for the purpose of holding services to account for compliance, Ofcom will be able to retrospectively assess whether a service has failed to trigger its protocols during a crisis.

Demos Digital strongly recommends that Ofcom should reconsider this approach. We propose that Ofcom should outline a mechanism whereby a democratically accountable body – such as Ofcom or a Secretary of State – is able to identify that a crisis is ongoing and inform platforms of this in a way which provides legitimacy for services to trigger their protocols. The body responsible for this should be one that is subject to Parliamentary oversight. By providing a consistent process to flag a crisis situation to all services, such a mechanism would help promote a consistent approach across services in a way that addresses how illegal content often spreads between services during a crisis. The details of the proposed mechanism should be made public and the public should be informed whenever the mechanism is triggered.

Service providers would remain able to identify that a crisis is underway, using their internal data, and would communicate this to the relevant responsible body. If a service were to trigger its internal crisis protocol on its own, the service would then be accountable to Ofcom to justify its decision.

Our reasoning for this proposal is grounded in the principle that significant decisions which could affect responses to public emergencies and which have implications for freedom of

49 https://www.gov.uk/government/publications/emergency-preparedness-resilience-and-response-concept-of-operations/incident-response-plan

50 https://www.ncsc.gov.uk/information/categorising-uk-cyber-incidents

51 https://fullfact.org/policy/incidentframework/report/

expression should be made by a body which is subject to democratic oversight, rather than private corporations. Our proposed measure would create an avenue for public accountability regarding the implementation of measures which could, in theory, lead to users' civil liberties being curtailed. It would also mean that there is a timely mechanism that indicates that services should trigger their crisis protocols when they are needed, rather than relying solely on after-the-fact accountability procedures. We acknowledge that some services may not see the same crisis situation and may not need to initiate their protocols. Hence, our proposal allows for flexibility in how protocols may be triggered and which services the measure applies to.

This change is intended to address the risk that services could be inconsistent in applying their protocols or could fail to trigger them altogether. In an October 2024 letter sent by Ofcom's Chief Executive Dame Melanie Dawes to the Secretary of State for Science, Innovation and Technology, it was noted that Ofcom had received evidence that online services' responses to the Southport riots were "uneven"[52]. Dawes stated that "some services" had implemented incident response protocols – but not all. Likewise, Meta, TikTok, Google and X told the House of Commons Science, Innovation, and Technology (SIT) Committee that they had implemented crisis protocols during the riots[53]. Yet the details given to the SIT Committee varied considerably, and "neither X nor TikTok provided [the Committee with] a date for when their protocols were triggered". This underlines the importance of providing an external, consistent and democratically legitimate indicator for when crisis protocols should be triggered.

Recommendation (4): Clarify Ofcom's role in holding platforms accountable for triggering their protocols

As mentioned above, the current proposals indicate that Ofcom will not determine when and how services should trigger their protocols. However, it is our understanding that Ofcom intends to assess after the fact whether services have failed to trigger their protocols during crises that would have warranted them.

Demos Digital recommends that Ofcom should clarify when and how it will assess whether services should have triggered their protocols. In doing so, Ofcom should specify:

(a) When it will conduct these assessments.

(b) The procedures it will follow to evaluate whether a situation constituted a crisis that warranted action.

---

52 https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/public-correspondence/2024/letter-from-dame-melanie-dawes-to-the-secretary-of-state-22-october-2024.pdf?v=383693

53 https://publications.parliament.uk/pa/cm5901/cmselect/cmsctech/441/report.html

(c) The data and indicators it will use as part of these evaluations.

(d) The actions it will take if it finds that a service has failed to trigger its protocol.

This recommendation is intended to address an ambiguity in the current proposals. Without further clarification, there is a risk that services and the public may be left unsure about Ofcom's procedures for ensuring that the proposed Code amendments are implemented.

Time scales for crisis responses

Recommendation (5): Set indicative time scales and time limits for reviewing crises protocol application

At present, the proposals for services to prepare and apply an internal protocol for responding to a crisis (Paragraphs 20.31-20.37) do not specify what time scales the protocol should operate within. Services' crisis protocols should set out an indicative time limit for the crisis protocol to apply and specify a timeline in which actions should be taken including at what stage it should be reviewed for extension. In the event that a crisis extends beyond the time limit, there should be a clearly defined and transparent procedure to extend the protocol.

These measures are needed because crises are, by definition, time limited. Without a time limit within which to review and assess if the protocol is still needed, there is a risk that a service could declare an indefinite crisis situation. Likewise, without requirements for services to set out a timeline for action, there is a risk that services' responses could be subject to delays and inconsistencies. We are not recommending that Ofcom should specify the exact length of time that crisis protocols should be in place, as we understand that this may need to vary depending on the crisis situation and service.

Implementing time limits is a standard best practice for crisis protocols. For comparison, the European Union's Digital Services Act 2024's provisions for platform crisis protocols states that: "crisis protocols should be activated only for a limited period of time and the measures adopted should also be limited to what is strictly necessary to address the extraordinary circumstance" (EU DSA 2024, paragraph 108). Meanwhile, during the COVID-19 pandemic, the Coronavirus Act 2020 included temporary provisions which were subject to a two-year time limit and required further legislation to extend[54].

[54] https://commonslibrary.parliament.uk/expiry-of-the-coronavirus-acts-temporary-provisions/

Recommendation (6): Set a minimum response time for platforms to respond to a crisis once it has been identified.

The current proposals do not set a **minimum** time after a crisis has been identified for services to trigger their crisis protocols. We recommend that services should be required to respond within at least 8 hours of identifying that a crisis is ongoing.

This measure is needed because a timely response is crucial during crisis situations. During previous crisis situations, research indicates that the most high-risk and harmful illegal content receives the most attention during the initial hours of a crisis. For example, during the Southport riots, our research indicated that hateful posts on X which were flagged using the platform's 'Community Notes' system received their highest engagement within the first 36-hours of the attack[55]. Yet the time it took between when a post was first created and when a Community Note was made public was 1,193 minutes (19.8 hours) on the day the riots began (30th July). This meant many harmful posts lacked X's desired contextual labels, a core function of its content moderation system, for a significant portion of time during which members of the public began rioting and platforms were subsequently blamed for contributing to such violence. Similarly, research by the Institute for Strategic Dialogue found that the most hateful false narratives that spread about Southport went viral soon after the attack.[56] For example, posts which shared a false name for the attacker that implied he was Muslim received over 30,000 mentions on X by 3pm on the day after the attack. The false name was recommended to users on X as a 'Trending in the UK' topic in the platform's 'What's happening' sidebar. These studies demonstrate that the speed at which a crisis protocol is implemented is crucial to its effectiveness in mitigating negative offline outcomes.

If the crisis response proposals are intended to mitigate future occurrences like the Southport riots, then they must include a time limit for platforms' responses to not only be triggered, but clearly taking effect.

Standards for designing crisis protocols

Recommendation (7): Recommend civil society involvement in developing and implementing platforms' crisis protocols

The current proposals do not set out recommendations or best practice examples for services to involve civil society in their crisis protocols. Demos Digital proposes that the guidance should recommend that Category 1 services incorporate civil society involvement

[55] https://demos.co.uk/wp-content/uploads/2025/07/Researching-the-riots_2025_July.ac_.pdf

[56] https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/

at both the development and implementation stages. These arrangements should be made public (see Recommendation 13).

Civil society organisations have valuable insights that should be used to shape crisis protocols to be effective. Community organisations, faith groups, fact checkers, research organisations, and other civil society organisations can sense-check protocols as they are developed and can help ensure their implementation is successful. Moreover, civil society engagement would help establish public trust and build up legitimacy of protocols. One option could be for services to create oversight boards for their protocols which feature civil society involvement.

We understand that large services do already engage with civil society on an informal basis. Likewise, civil society organisations already submit information and flag content to platforms, even if they do not have special status. As stated in the October 2024 letter from Ofcom's Chief Executive Dame Melanie Dawes to the Secretary of State for Science, Innovation and Technology:

> "In one instance, a service proactively reached out to a civil society organisation focused on anti-Muslim hatred requesting training to improve their moderation systems and took down potentially illegal content based on a referral from law enforcement agencies."[57]

Meanwhile, Article 48 of the European Union's Digital Services Act states that "the Commission may, where necessary and appropriate, also involve civil society organisations or other relevant organisations in drawing up [voluntary] crisis protocols" for services.[58]

Such community engagement is an example of best practice which other services should follow if they have the resources to do so. We recommend that these arrangements should be formalised through the crisis protocol guidance.

Recommendation (8): Specify integration with existing civil contingencies arrangements

At present, the proposals suggest that large service providers should set up communication channels with law enforcement, but do not mention other bodies that perform crisis response functions, such as the NHS, Fire Service, or local governments. As a result, this leaves out significant bodies involved in the UK's responses to public emergencies.

Demos Digital recommends that services should communicate with non-law enforcement bodies that hold primary responsibility for responding to civil contingencies, as set out by the Civil Contingencies Act 2004 and related policies. Such bodies include local

---

57 https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/public-correspondence/2024/letter-from-dame-melanie-dawes-to-the-secretary-of-state-22-october-2024.pdf?v=383693

58 https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

government and the NHS at a local level and the Civil Contingencies Secretariat at a national level.

'Civil contingency' is a legal term for a public emergency or major incident, such as a pandemic or flood.[59] In UK law, the Civil Contingencies Act 2004 (CCA) sets out arrangements for local public protection during emergencies.[60] The CCA regulates responses to three categories of emergencies[61]:

(1) Events that seriously threaten "human welfare", such as mass injury, illness or loss of life.

(2) Events that seriously threaten the environment.

(3) War or terrorism which threatens the UK's security.

Implementation of the CCA is shaped by a mixture of statutory and non-statutory guidance [62]. The CCA places significant emphasis on local-level emergency planning and response, according to what is known as the principle of 'subsidiarity'[63]. It establishes local authorities, the NHS, Fire and Rescue, police forces, and The Environment Agency as the bodies with principle responsibility for responding to public emergencies at a local level (termed 'Category 1' services). We recommend that crisis protocols should include measures to create communication channels with all of these Category 1 services, where appropriate.

National emergency planning and response is addressed by the UK government's Resilience Capabilities Programme (RCP)[64], which is managed by the Civil Contingencies Secretariat (CCS). The RCP's activities include coordinating emergency planning across government departments, conducting risk evaluations, and running public information campaigns.

---

[59] College of Policing (2024). 'Civil contingencies'. https://www.college.police.uk/app/civil-emergencies/civil-contingencies

[60] HM Government (2025). *Civil Contingencies Act 2004.* https://www.legislation.gov.uk/ukpga/2004/36/contents

[61] HM Government (2025). Civil Contingencies Act 2004, Section 1. https://www.legislation.gov.uk/ukpga/2004/36/section/1

[62] Cabinet Office (2013). 'Guidance: Preparation and planning for emergencies: responsibilities of responder agencies and others'. HM Government. https://www.gov.uk/guidance/preparation-and-planning-for-emergencies-responsibilities-of-responder-agencies-and-others, https://www.gov.uk/government/publications/emergency-response-and-recovery, Cabinet Office (2013). 'Guidance: Emergency preparedness'. HM Government. https://www.gov.uk/government/publications/emergency-preparedness

[63] Cabinet Office (2012). 'Chapter 16: Collaboration and Co-operation between Local Resilience Forums in England: Revision to Emergency Preparedness'. HM Government. https://assets.publishing.service.gov.uk/media/5a798cc2ed915d04220694f4/Chapter-16-final-post-consultCCS_amends_16042012.pdf

[64] Cabinet Office (2018). 'Guidance: Preparation and planning for emergencies'. HM Government. 'https://www.gov.uk/guidance/preparation-and-planning-for-emergencies-the-capabilities-programme

Coordination is also conducted at a sub-national level, which in turn feeds into local-level emergency preparations.This is led by the Ministry of Housing, Communities & Local Government's Resilience and Emergencies Division (RED)[65].

Because the proposed definition of a crisis would overlap significantly with the categories of civil contingencies defined by the CCA, we recommend that digital services' crisis protocols should include measures to create communication channels with all Category 1 services mentioned in the CCA – rather than just the police – as well as the Civil Contingencies Secretariat and the Resilience and Emergencies Division. This would ensure that services are in close communication with crisis responses at local, regional, and national levels. Otherwise, there is a risk that services only correspond with centrally-run national bodies and therefore are not in communication with the broader crisis response infrastructure.

Best practices for crisis responses

Recommendation (9): Require platforms to notify the public that a crisis protocol has been triggered

The current proposal for what a crisis response protocol should include does not include a requirement for services to notify the public that the protocol has been triggered. As a result, services would trigger their protocols without users knowing.

Demos Digital strongly recommends that services should be required to inform the public when a crisis protocol has been triggered. The public notification should follow Ofcom's 'media literacy by design' principles.[66] It should be provided immediately, be clearly visible to users via the service's user interface, and be accompanied by easily accessible and understandable information for what this means for the service's operation. It should also include reference to the specific locations, people, systems, or services which are affected. This information should be accompanied by a summary which discloses the evidence upon which the service decided to trigger the protocol. Additionally, any notification that a crisis protocol has been triggered must be accompanied by a publicly-available post-crisis publication which reviews the crisis response (see Recommendation 12).

Our recommendation is intended to embed transparency throughout the lifecycle of a service's crisis protocol. Without transparency around the implementation of a crisis protocol, there is a risk that public trust would be undermined. If a protocol is triggered but this is not disclosed until a later date - whether through a leak, a public evidence

[65] Cabinet Office (2018). 'Guidance: Preparation and planning for emergencies'. HM Government. 'https://www.gov.uk/guidance/preparation-and-planning-for-emergencies-the-capabilities-programme

[66] https://www.ofcom.org.uk/media-use-and-attitudes/media-literacy/principles-for-media-literacy-by-design

submission, or a post-crisis report – this could trigger concerns about shadowy forces, censorship, and double standards.

Concerns already abound regarding a lack of transparency in digital services' content moderation policies. Such opacity has been found to "lead to mistrust, backlash and uncertainties surrounding platforms' policies and operations"[67]. This has frequently led to users speculating about the possibility of censorship by nefarious forces[68]. We have seen how unexplained platform moderation decisions can fuel mistrust during crisis situations such as the Covid pandemic[69]. Indeed, the only indication that social media services implemented crisis response protocols during the Southport riots came from evidence revealed by Ofcom[70] and the SIT Committee[71] - and was disclosed months after the riots took place. Mistrust has been shown to trigger accusations of censorship and double standards, fuelling conspiracy theories. The result is that accusations can inflame situations and may make crisis responses even harder for emergency services and other responders offline.

Recommendation (10): Require services to implement crisis communication policies

The current proposals do not address the need for services to communicate effectively with the public regarding their crisis response. Well-managed public communications are a best practice for any crisis response: by sharing key information, selecting language carefully, and helping to set expectations, crisis communications policies can mitigate threats to trust. Without well-managed crisis communication, there is a risk that a service's crisis response could create distrust. For example, during the Southport riots, services such as TikTok[72] and X[73] appear to have implemented crisis protocols without clear communication to their users about the measures.

Therefore, Demos Digital recommends that Ofcom should require services to implement crisis communication policies as part of their crisis protocols. These policies should cover the service's communications objectives, key audiences, and guiding principles. Where possible, they should identify key messages and communications channels. The policies should also identify the teams responsible and ensure these will receive sufficient support.

---

[67] https://policyreview.info/articles/analysis/transparency-content-moderation
[68] https://journals.sagepub.com/doi/10.1177/1461444818773059
[69] https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf
[70] https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/public-correspondence/2024/letter-from-dame-melanie-dawes-to-the-secretary-of-state-22-october-2024.pdf?v=383693
[71] https://publications.parliament.uk/pa/cm5901/cmselect/cmsctech/441/report.html
[72] https://committees.parliament.uk/writtenevidence/134454/html/
[73] https://committees.parliament.uk/writtenevidence/133665/html/

As an example of best practice, the UK Government Communication Service has published a Crisis Communications Planning Guide.[74] The Guide uses a STOP framework (Strategy, Tactics, Organisation, and People) to give government departments a template for their crisis communications policies.

Recommendation (11): Require services to promote public interest information during crises

The current proposals do not address the need for public bodies to reach the public with vital public interest information during crises. As a result, the proposals do not address the risk that critical communication from public bodies such as the emergency services may be drowned out by more engaging posts from users with less authoritative access to information.

Demos Digital recommends that services' crisis protocols should include provisions to ensure that important public interest information is displayed prominently during a crisis situation. This should apply to a select group of public bodies, such as the NHS, Fire Services, police, and local government. For example, services which serve content using recommendation algorithms could boost the ranking of content from these bodies when their crisis protocol is in effect.

This measure is important as a way to ensure that the public is able to access critical information during crisis situations. Public services can struggle to communicate effectively with the public over social media[75], where their messages may be drowned out by more attention-grabbing sources. In public health communications, for example, health misinformation may easily win out in users' feeds[76]. During the Covid-19 crisis, this information environment presented significant challenges for the NHS[77]. In response, the Government Communications Service has stated that it worked closely with Google, Facebook and Twitter to "ensur[e] that public health campaigns were promoted through reliable sources" on their services during the pandemic.[78] As this case demonstrates, public services may need additional resources and support from services to ensure that critical communications reach the public.

---

[74] https://www.communications.gov.uk/publications/crisis-comms-planning-guide/

[75] https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/12/public-communication-scan-of-the-united-kingdom_6c3acae1/bc4a57b3-en.pdf

[76] https://www.tandfonline.com/doi/full/10.1080/13648470.2024.2386887#abstract

[77] https://committees.parliament.uk/writtenevidence/132776/html/, https://www.communications.gov.uk/wp-content/uploads/2020/10/COVID-19_Communications_Advisory_Panel_Report.pdf

[78] https://www.communications.gov.uk/wp-content/uploads/2020/10/COVID-19_Communications_Advisory_Panel_Report.pdf

For comparison, the European Union's Digital Service Act's requirements for crisis protocols for digital services specifies that such protocols should include measures for "prominently displaying information on the crisis situation provided by Member States' authorities or at Union level, or, depending on the context of the crisis, by other relevant reliable bodies" (Article 48, Paragraph 2(a))[79].

Strong transparency and accountability measures

Recommendation (12): Establish strong transparency and accountability measures, including strong requirements for access to data for independent public interest researchers

The current proposals lack strong transparency reporting requirements or measures which would facilitate accountability, such as mandatory post-crisis reporting or real-time data access for independent public interest researchers. We note that the proposals explicitly state that Ofcom is "not proposing to recommend that services submit the post-crisis analysis to Ofcom or publish it" (Paragraph 20.40).

We acknowledge that Ofcom's proposals for transparency requirements under the OSA are forthcoming and await their publication. We also acknowledge that Ofcom has published proposals for providing researchers with access to services elsewhere.[80]

Demos Digital recommends that the guidance should include much stricter transparency and accountability requirements for services' crisis protocols. We propose that:

(a) All relevant services should provide details of their crisis protocols to Ofcom.

(b) All relevant services should notify the public and Ofcom that a crisis protocol has been triggered without delay. Ofcom may then notify other platforms as it sees fit, to help address the risk of cross-platform virality. Platforms should display their public notice prominently in their user interfaces alongside an explanation of what this means (see Recommendation 9).

(c) Relevant Category 1 services should publish details of their crisis protocols publicly.

(d) Relevant Category 1 services should publish post-crisis analyses of their crisis responses in a timely manner. Ofcom may then respond to these public analyses as it sees fit.

[79] https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

[80] https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/call-for-evidence-researchers-access-to-information-from-regulated-online-services

(e) These post-crisis publications should include a summary of steps that were taken, including:

(i) What training and support content moderation staff are typically given for this kind of response.

(ii) What additional training and support staff were actually given during the crisis response.

(iii) How staffing was different during the crisis response.

(iv) How content moderation response times were different on each day of and on average during the crisis response.

(f) Relevant services should archive versions of all content they remove. This is to assist transparency and accountability processes, and may also assist law enforcement

(g) All relevant services should have their crisis protocols undergo regular independent audits by a regulated auditing service to ensure their compliance. This audit should be made public.

(h) Post crisis, relevant Category 1 and 2B services should provide independent public interest researchers with access to data on their crisis responses so they can assess the service's performance. This should include information on:

(i) The rate of content moderation decisions

(ii) Which content has been moderated and why

(iii) The speed of moderation decisions

(iv) How content moderation staff teams were increased in size

(v) What was communicated to users

These measures are necessary to ensure that the implementation of crisis protocols does not harm public trust, which can undermine the crisis response in turn. If a protocol is triggered without sufficient transparency, this could trigger concerns about shadowy forces, censorship, and double standards. As we have discussed above, concerns already abound regarding a lack of transparency in digital services' content moderation policies. Such opacity has been found to "lead to mistrust, backlash and uncertainties surrounding platforms' policies and operations"[81]. During the Southport Riots, for example, allegations of inconsistent content moderation and censorship arose which threatened to undermine

[81] https://policyreview.info/articles/analysis/transparency-content-moderation

the credibility of the government's response[82]. As Demos found in our research on X's Community Notes system during the riots[83], a lack of data access can hamper accountability. We were only able to conduct our research because X provided access to data on this system, yet no data is available on how X's other content moderation systems performed in the period.

For comparison, the EU DSA requires that services with crisis protocols should "report to the Commission by a certain date or at regular intervals specified in the decision, on the [crisis risk] assessments referred to in point (a), on the precise content, implementation and qualitative and quantitative impact of the specific measures taken [...] and on any other issue related to those assessments or those measures, as specified in the decision" (Article 36).[84] We think this offers an example of best practice that Ofcom should also replicate.

Recommendation (13): Require platforms to publicly disclose which civil society and governmental organisations they work with during crises, including which have been given 'trusted flagger' status

Demos Digital recommends that platforms be required to publicly disclose which civil society and governmental organisations they work with during crises, including which have been given 'trusted flagger' status.

The current proposals do not mention whether services should make public disclosures about their co-ordination with civil society and governmental organisations during crises. This creates a risk for public trust: without transparency, members of the public may suspect collaboration is occurring and inaccurate narratives may arise about interventions with nefarious intentions. The risk of mistrust becomes acute when collaborations are revealed through media reporting or leaks, which may not provide a full picture of events.

For example, recent news stories about the role of DSIT's National Security and Online Information Team (NSOIT) in identifying and flagging hateful content directly to platforms during the Southport riots have included allegations that NSOIT is engaging in censorship[85] . One of the key concerns in the NSOIT case has been the team's 'trusted flagger' status with social media platforms, which we understand means that the team receives priority attention from these platforms when it provides them with notice of content that might violate their terms of service[86]. Partly because trusted flagger status in the UK is an informal

[82] https://www.bbc.co.uk/news/articles/cr548zdmz3jo

[83] https://demos.co.uk/research/researching-the-riots-an-evaluation-of-the-efficacy-of-community-notes-during-the-2024-southport-riots/

[84] https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

[85] https://www.telegraph.co.uk/news/2025/07/31/exposed-labour-plot-silence-migrant-hotel-critics/

[86] https://bigbrotherwatch.org.uk/wp-content/uploads/2024/11/BigBrotherWatch-Briefing-on-the-National-Security-Online-Information-Team.pdf

role given by platforms – unlike the EU where it is a legal role defined by the DSA[87] – there is not sufficient transparency regarding what NSOIT's status entails or how platforms have responded to its takedown requests.

Moreover, we understand that services do engage with government, police, and civil society organisations. For example, in the October 2024 letter from Ofcom's Chief Executive Dame Melanie Dawes to the Secretary of State for Science, Innovation and Technology, it is disclosed that "firms told [Ofcom] they took a range of actions in response" to Southport" which included "proactive engagement with civil society and/or law enforcement partners to seek guidance."[88] However, these interactions are also not transparent to the public and tend to only come to light through *ad hoc* disclosures.

Therefore, Demos Digital recommends the services' crisis protocols should come alongside transparency requirements which set out which civil society and governmental organisations services work with during crises. These disclosures should detail which partners the services engage with, how they do so, and the outcomes of these engagements. They should incorporate details on any 'trusted flagger' systems that are in place, including which organisations have been given this status.

Recommendation (14): Set out explicit requirements to protect human rights

The current proposals do not require services to set out how they will avoid the risk that their crisis responses will disproportionately limit users' right to freedom of expression, privacy, and other rights. While Ofcom's rights assessment does identify the risk that services might respond to the proposals by infringing on freedom of expression (Paragraphs 20.71-20.72), the implications of this assessment do not appear to be reflected in the content of the proposal.

To mitigate the risk that services' respond to crises by implementing measures which disproportionately limit users' right to freedom of expression, privacy, and other rights, Demos Digital recommends that services should be required to conduct human rights impact assessments as part of their crisis protocols. Such assessments should identify how their crisis response policies would affect users' rights and set out appropriate mitigations. These assessments should be made public to ensure transparency (see Recommendation 12). Ofcom could set out examples of mitigation measures designed to safeguard users' rights, as they have in the existing Illegal Content Codes of Practice[89]. This measure would help to ensure that services adhere to their duty to uphold these rights, as set out in the OSA.

[87] https://digital-strategy.ec.europa.eu/en/policies/trusted-flaggers-under-dsa

[88] https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/public-correspondence/2024/letter-from-dame-melanie-dawes-to-the-secretary-of-state-22-october-2024.pdf?v=383693

[89] https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/codes-of-practice

As an example of good practice, the EU DSA's measures for platforms to implement crisis protocols include a requirement for services to "take due account [...] of the actual or potential implications for the rights and legitimate interests of all parties concerned, including the possible failure of the measures to respect [users'] fundamental rights" (Article 36). The DSA requires the EU Commission to set out how its voluntary crisis protocols will include "safeguards to address any negative effects on the exercise of the fundamental rights enshrined in the Charter, in particular the freedom of expression and information and the right to non-discrimination" (Article 48).[90]

Question 50: Do you agree with our proposed definition of 'crisis'? Please explain your reasoning, and if possible, provide supporting evidence.

Demos Digital agrees with the decision for the definition to focus on extraordinary situations that pose risks to public safety. However, we suggest the definition requires much more specificity and detail. This is important to ensure that the definition is applied consistently and proportionately without undue impacts on users' rights.

A vague definition is not appropriate for high-risk crises with a narrow margin of error: public trust is on the line and the consequences of misidentifying the situation could be drastic. A clear, detailed, and strongly delineated definition is a crucial means to ensure compliance, avoid misuses and prevent overreach that undermines users' rights as well as to shore up trust in their justified application. Here, we suggest referring to the Civil Contingencies Act 2004 (CCA)'s definition of an emergency as a useful example of how crises may be defined for legal purposes[91].

Below, we set out seven recommendations for how the definition could be made more detailed and specific:

- Recommendation (1): Include a reference to specific affected locations, people, systems, and/or services
- Recommendation (2): Provide more examples of crisis situations to guide identification
- Recommendation (3): Define 'public safety' clearly and narrowly
- Recommendation (4): Provide more detail regarding how an actor is expected to determine if illegal content has caused "a serious threat to public safety"
- Recommendation (5): Provide more detail on the threshold criteria to be used to identify when situations become crises

[90] https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng
[91] HM Government (2025). Civil Contingencies Act 2004 Section 1. https://www.legislation.gov.uk/ukpga/2004/36/section/1

- Recommendation (6): Set out a a graduated definition of crisis severity
- Recommendation (7): Publish detailed information on how it will use its definition to identify a crisis for the purpose of assessing whether platforms should have initiated their protocols after an incident.

Recommendation (1): Include a reference to specific affected locations, people, systems, and/or services

The current definition does not provide specificity about who, what, or where is affected by the extraordinary situation that poses a threat to public safety. Instead, it refers generally to the United Kingdom.

While we appreciate that this wording may be intended to afford maximum coverage and flexibility, we are concerned that it fails to acknowledge the subjects of threats to public safety as a result. Crises always impact specific people, systems, and/or places. They may, for example, impact a specific area or region in the UK rather than the country at a national level. Therefore, we suggest that the definition should include wording to the effect of: ' A crisis is an extraordinary situation in which there is a serious threat to public safety in relation to a location, community, system or service in the United Kingdom '.

This wording would draw attention to the communities, places, and systems that are at risk. It would direct the service responsible for implementing its crisis protocol to identify and justify the risk to these specific communities, places, and/or systems. Some crises may not take place at a national level and may only impact a slice of a service's userbase. Moreover, adding this specificity would reduce the risk that a 'general crisis' could be declared.

There are several useful examples of crisis definitions which include wording to this effect. The Civil Contingencies Act 2004 specifies that an emergency must pose a risk to "human welfare" or "the environment" of "a place in the United Kingdom".[92] This may involve risks including "loss of human life" and "disruption of a system of communication". Likewise, Full Fact's proposed *Framework for Information Incidents* defines an information incident as a "cluster or proliferation of inaccurate or misleading claims or narratives" which "have a substantial and material impact on the people, organisations, and systems that consume, process, share or act on information."[93]

Recommendation (2): Provide more examples of crisis situations to guide identification

The current proposal provides a short list of examples of situations which "may, depending on the circumstances, satisfy [the] definition of a crisis". These are "nationwide riots, large scale terrorist attacks and/or inter-religious or inter-ethnic violence".

[92] https://www.legislation.gov.uk/ukpga/2004/36/section/1
[93] https://fullfact.org/policy/incidentframework/report/

Demos Digital recommends that Ofcom set out a much more expansive – but non-exhaustive – list to guide decision-making. This list should also avoid using the qualifiers that are attached to the current list of crises (namely "may" and "depending on the circumstances"). We also recommend that Ofcom should consider including situations that pose a risk to public safety which do not involve mass violence or hate. These other types of crisis could include environmental disasters and public health emergencies such as pandemics. The list of examples would not need to be exhaustive, but should be long and detailed enough to provide a better indication of the scope of the definition.

Our recommendation is based on the need to provide services, users, civil society, and policymakers beyond Ofcom with clarity about what the definition will apply to. This is crucial for providing regulatory certainty and ensuring that there is consistency in the application of the definition.

As an example of best practice, the Civil Contingencies Act 2004 sets out an extensive list of examples of emergencies that it applies to.[94] These include the "disruption of a supply of money, food, water, energy or fuel", "disruption of a system of communication", "disruption of facilities for transport," or the "disruption of services relating to health." Similarly, the EU DSA suggests that relevant crises may include "armed conflicts or acts of terrorism, including emerging conflicts or acts of terrorism, natural disasters such as earthquakes and hurricanes, as well as from pandemics and other serious cross-border threats to public health"[95].

Recommendation (3): define 'public safety' clearly and narrowly

Though 'public safety' is one of – if not the most – important terms in the current definition, the definition does not specify what is meant by this term. We strongly recommend that Ofcom should provide a clear disambiguation of what is meant by public safety.

Doing so would help ensure the definition is applied consistently across contexts. It would also help to avoid a 'know it when you see it' approach to identifying crises. Such ambiguity is not appropriate given the high-stakes context of a crisis situation, where there are serious risks associated with both misidentification and actions that would disproportionately affect users' rights to freedom of expression. A narrower definition would help mitigate the risk that the definition is applied overly broadly in a way which enables crisis protocols to be triggered unduly, which risks undermining users' free speech rights.

[94] HM Government (2025). Civil Contingencies Act 2004 Section 1. https://www.legislation.gov.uk/ukpga/2004/36/section/1
[95] https://www.legislation.gov.uk/ukpga/2004/36/section/1

One option for doing so could be to refer to the Civil Contingencies Act 2004's definition of 'emergency'[96]: either

(d) "an event or situation which threatens serious damage to human welfare in a place in the United Kingdom"

(e) Or "an event or situation which threatens serious damage to the environment of a place in the United Kingdom"

(f) Or "war, or terrorism, which threatens serious damage to the security of the United Kingdom.

Referring to the CCA would allow for consistency between guidelines for digital services' crisis responses and crisis responses in government.

Recommendation (4): Provide more detail regarding how an actor is expected to determine if illegal content has caused "a serious threat to public safety"

The current definition suggests that a crisis may be said to occur when "a serious threat to public safety" has arisen "as a result of a significant increase in relevant illegal content".

Unfortunately, such a direct causal relationship between specific instances of online content and offline threats to safety is generally hard to establish after the fact, let alone during a crisis. For example, research that seeks to identify links between harmful online content – such as hate speech – and offline violence tends to be limited to finding correlations and associations, rather than direct causation[97]. As a result, there is a risk that the current wording of the definition will unintentionally set a very high bar for what can be considered a crisis. It may be that crisis protocols simply are not triggered due to this high requirement.

To mitigate this risk, Ofcom should clearly set out the standards of proof that it expects services to meet when they identify a crisis according to the definition. This does not need to involve specifying what specific indicators or data a service should use. Rather, it would mean detailing the levels of risk and confidence that a service would have to meet in its assessment of its internal data.

Moreover, as we outlined in our response to Question 49, services should not be the actors with primary responsibility for determining if a crisis is underway. We propose that Ofcom should outline a mechanism whereby a democratically accountable body – such as Ofcom or a Secretary of State – is able to identify that a crisis is ongoing and able to direct services to trigger their protocols. The body responsible for this should be one that is subject to Parliamentary oversight. The notification to trigger the crisis protocol could be tailored to specific services that are affected by the crisis, rather than being directed at all relevant

[96] https://www.legislation.gov.uk/ukpga/2004/36/section/1
[97] For example, https://www.nature.com/articles/s41599-024-03899-1

regulated services. Furthermore, this mechanism would allow for a joined-up approach which addresses how illegal content often spreads between services during a crisis. The details of the proposed mechanism should be made public and the public should be informed whenever the mechanism is triggered.

Service providers would remain able to identify that a crisis is underway, using their internal data, and would communicate this to the relevant responsible body. If a service were to trigger its internal crisis protocol on its own, the service would then be accountable to Ofcom to justify its decision.

Recommendation (5): Provide more detail on the threshold criteria to be used to identify when situations become crises

The current definition and surrounding guidance does not provide detail on the precise thresholds at which an incident becomes a full-on crisis. While the lack of detail may be intended to provide flexibility, we suggest that the risks of ambiguity here outweigh the benefits. There is a risk that significant variations emerge between the thresholds that different services use to determine whether a situation has developed into a crisis. This could lead to inconsistencies between services, which could result in confusion and distrust amongst users in turn. Moreover, without stronger direction from Ofcom, there is a risk that services set a high baseline for what is to be considered 'normal' on their systems - thereby allowing them to justify choosing not to trigger their protocols.

To mitigate these risks, we recommend that Ofcom should provide more detailed guidance on:

(a) How services should determine what a 'normal' baseline is in regards to illegal content.

(b) What degree of deviation from this baseline would be considered a crisis.

(c) What kinds of threats to public safety this elevated level of illegal content should pose (see Recommendations 2-3).

Recommendation (6): Set out a a graduated definition of crisis severity

The current definition does not reflect the fact that crisis situations may be of differing levels of severity, or that some crises may require more intensive responses than others. Instead, the definition presents a binary view of crises: either a situation is a crisis or it is not.

Demos Digital recommends that Ofcom should reconsider this approach by introducing a graduated definition of crises based on levels of severity. One option for defining such severity levels would be to tie them to the risk of harm or degree of threat to public safety. These levels could then be used to set higher minimum standards for responses to higher

severity crises. For example, the definition could outline what would be considered to be crises with medium, high, and very high risks of harm to users and public safety.

Taking this approach would allow for greater flexibility in how services apply their crisis protocols, depending on needs of the situation, while also potentially raising minimum standards. Some crises are not as severe as others and will not require the same level of response. Rather than giving services total discretion about the intensity of response that is warranted, it would be best if Ofcom can provide a framework to guide and regulate such decisions.

It is common practice to use graduated definitions of crisis severity as a way of making crisis responses more effective. For example, the UK Health Security Agency uses four severity levels for its incident response plan: routine, standard, enhanced, and severe.[98] The National Cyber Security Centre (NCSC) uses a six-tier grading system to categorise the severity of cybersecurity incidents.[99] Likewise, it is our understanding that the government's Defending Democracy Taskforce (DDTF) and National Security Online Information Team (NSOIT) use a three-tier grading system for its information incident response. Finally, Full Fact's *Framework for Information Incidents* uses five tiers of severity[100]. All of these examples could provide inspiration for how Ofcom could approach its severity levels.

Recommendation (7): Ofcom should publish detailed information on how it will use its definition to identify a crisis for the purpose of assessing whether platforms should have initiated their protocols after an incident.

The current proposals indicate that Ofcom will not determine when and how services should trigger their protocols. However, it is our understanding that Ofcom intends to assess after the fact whether services have failed to trigger their protocols during crises that would have warranted them.

Demos Digital recommends that Ofcom should clarify when and how it will assess whether services should have triggered their protocols. In doing so, it should specify:

(a) When it will conduct these assessments.

(b) The procedures it will follow to evaluate whether a situation constituted a crisis that warranted action.

(c) The data and indicators it will use as part of these evaluations.

---

98 https://www.gov.uk/government/publications/emergency-preparedness-resilience-and-response-concept-of-operations/incident-response-plan

99 https://www.ncsc.gov.uk/information/categorising-uk-cyber-incidents

100 https://fullfact.org/policy/incidentframework/report/

(d) The actions it will take if it finds that a service has failed to trigger its protocol.

This recommendation is intended to address an ambiguity in the current proposals. Without further clarification, there is a risk that services and the public may be left unsure about Ofcom's procedures for ensuring that the proposed Code amendments are implemented.

Question 51: Do you consider these measures to be effective for services that are not large services? Please provide any evidence on the role of services that are not large services during crises.

For the purpose of this response, we will consider "services that are not large services" to mean Category 2B services as well as 'low reach' services with under or around 1% of the UK population as active monthly users, as defined in Ofcom's public statements regarding its 'Small But Risky Services Taskforce'[101].

It is important for any guidance on crisis protocols to be effective for these smaller services. During crisis situations, illegal content is often first created and distributed on smaller, non-mainstream platforms before being circulated on large mainstream ones [102]. This dynamic was observed during the Southport riots[103]. For example, according to the Institute for Strategic Dialogue (ISD), the smaller encrypted messaging service Telegram "played an outsized role in mobilising offline action" during the riots[104]. Smaller platforms like Telegram, Gab, and Rumble frequently play host to figures such as Tommy Robinson and Andrew Tate, who have played an active role in inflaming crisis situations such as Southport. [105]

Based on this evidence, Demos Digital supports Ofcom's proposal to include small services in the requirement to implement crisis protocols. Below, we have set out two recommendations to help ensure that these measures are effective.

Recommendation (1): Apply the measures to small platforms with medium risk of harmful content

Demos Digital recommends that the requirement should not just be applied to small services that are assessed to be at high risk of relevant harms, as is currently proposed.

---

101 https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/foi/2025/april/online-safety-small-but-risky.pdf?v=396599

102 https://committees.parliament.uk/writtenevidence/133348/html/

103 https://www.bbc.co.uk/news/articles/cvgrvw29x4jo.amp, https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/

104 https://committees.parliament.uk/writtenevidence/133348/html/

105 https://committees.parliament.uk/writtenevidence/132891/html/

Instead, the proposed measures should apply to services that are assessed to be of medium risk, which would match the requirements proposed for large user-to-user services. We recommend this to reflect the role that smaller services have historically played in producing and amplifying harmful content during crises. The change would help ensure that the measure reflects the broad risk landscape that small services pose.

Recommendation (2): Provide support for small services to identify and respond to crises

There is a risk that smaller services could lack the resources needed to undertake continuous monitoring to identify crises or to "increas[e] human content moderation resources" sufficiently to meet a crisis (Paragraph 20.36).

To mitigate this risk, Demos Digital recommends that smaller services should be supported by a government-funded independent body which conducts monitoring and research. This body should be able to access the services' data and flag potential crisis situations to them. The body should be accountable to Parliament and should publish reports on its activities at regular intervals. This recommendation aligns with our response to Question 50, where we recommend that Ofcom should outline a mechanism whereby a democratically accountable body – such as Ofcom or a Secretary of State – is able to identify that a crisis is ongoing and notify services accordingly.

Question 52: Is there any evidence of best practice in responding to a crisis that we have not identified? Please explain your reasoning, and if possible, provide supporting evidence.

We have identified seven best practices for crisis response which are not currently addressed by the proposals. These are mostly repeated from our answers in earlier questions, but repeated here for ease of reference:

- Best practice (1): Iterative testing of emergency protocols
- Best practice (2): Integrate civil society engagement
- Best practice (3): Use of crisis response thresholds
- Best practice (4): Implement crisis communication policies
- Best practice (5): Implement transparency and accountability measures
- Best practice (6): Implement human rights impact assessments and mitigation measures
- Best practice (7): Implement measures to promote reliable public interest information

Best practice (1): Test crisis protocols iteratively

It is best practice for services to develop and test their crisis protocols on an iterative basis. To do so, services should initiate trials, tests, or drills to ensure their protocols are effective.

As an example of a best practice policy for digital services, such testing is mentioned explicitly in Article 48 of the European Union's Digital Services Act: "The Commission shall encourage and facilitate the providers of very large online platforms, of very large online search engines and, where appropriate, the providers of other online platforms or of other online search engines, to participate in the drawing up, testing and application of those crisis protocols."[106]

Iterative testing is a standard component of crisis response policies in other contexts that involve risks to public safety. For example, the UK government's Exercising Best Practice Guidance for emergency preparedness recommends that organisations conduct regular tests of their procedures and systems[107]. These tests include tabletop exercises (TTX) to explore potential weaknesses, stress tests which provide a 'safe-to-fail' environment, and live play exercises (LIVEX) in which teams seek to replicate real emergency situations as closely as possible. It is best practice to regularly run such drills, evaluate the results, and update the emergency response policies accordingly.

Best practice (2): Integrate civil society engagement

Civil society organisations have valuable insights that should be used to shape crisis protocols to be effective. Community organisations, faith groups, fact checkers, research organisations, and other civil society organisations can sense-check protocols as they are developed and can help ensure their implementation is successful. Moreover, civil society engagement would help establish public trust and build up legitimacy of protocols. One option could be for services to create oversight boards for their protocols which feature civil society involvement.

We understand that large services do already engage with civil society on an informal basis. Likewise, civil society organisations already submit information and flag content to platforms, even if they do not have special status. As stated in the October 2024 letter from Ofcom's Chief Executive Dame Melanie Dawes to the Secretary of State for Science, Innovation and Technology:

> "In one instance, a service proactively reached out to a civil society organisation focused on anti-Muslim hatred requesting training to improve their moderation

[106] https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

[107] https://www.gov.uk/government/publications/exercising-best-practice-guidance/exercising-best-practice-guidance-html

> systems and took down potentially illegal content based on a referral from law enforcement agencies."[108]

Meanwhile, Article 48 of the European Union's Digital Services Act states that "the Commission may, where necessary and appropriate, also involve civil society organisations or other relevant organisations in drawing up [voluntary] crisis protocols" for services.[109]

Such community engagement is an example of best practice which other services should follow if they have the resources to do so. We recommend that these arrangements should be formalised through the crisis protocol guidance.

Best practice (3): Use of crisis response thresholds

It is common practice to use graduated definitions of crisis severity as a way of making crisis responses more effective. For example, the UK Health Security Agency uses four severity levels for its incident response plan: routine, standard, enhanced, and severe.[110] The National Cyber Security Centre (NCSC) uses a six-tier grading system to categorise the severity of cybersecurity incidents.[111] Likewise, it is our understanding that the government's Defending Democracy Taskforce (DDTF) and National Security Online Information Team (NSOIT) use a three-tier grading system for its information incident response. Finally, Full Fact's *Framework for Information Incidents* uses five tiers of severity[112].

Demos Digital recommends that Ofcom should follow this best practice by adopting a graduated understanding of crises based on levels of severity. One option for defining such severity levels would be to tie them to the risk of harm or degree of threat to public safety. These levels would then be used to set higher minimum standards for responses to higher severity crises. For example, the definition could outline what would be considered to be crises with medium, high, and very high risks of harm to users and public safety.

Taking this approach would allow for greater flexibility in how services apply their crisis protocols, depending on needs of the situation, while also potentially raising minimum standards. Some crises are not as severe as others and will not require the same level of response. Rather than giving services total discretion about the intensity of response that is warranted, it would be best if Ofcom can provide a framework to guide and regulate such decisions.

---

108 https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/public-correspondence/2024/letter-from-dame-melanie-dawes-to-the-secretary-of-state-22-october-2024.pdf?v=383693

109 https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

110 https://www.gov.uk/government/publications/emergency-preparedness-resilience-and-response-concept-of-operations/incident-response-plan

111 https://www.ncsc.gov.uk/information/categorising-uk-cyber-incidents

112 https://fullfact.org/policy/incidentframework/report/

Best practice (4): Implement crisis communication policies

Well-managed public communications are a best practice for any crisis response: by sharing key information, selecting language carefully, and helping to set expectations, crisis communications policies can mitigate threats to trust. Without well-managed crisis communication, there is a risk that a service's crisis response could create distrust. For example, during the Southport riots, services such as TikTok[113] and X[114] appear to have implemented crisis protocols without clear communication to their users about the measures.

Therefore, Demos Digital recommends that Ofcom should require services to implement crisis communication policies as part of their crisis protocols. These policies should cover the service's communications objectives, key audiences, and guiding principles. Where possible, they should identify key messages and communications channels. The policies should also identify the teams responsible and ensure these will receive sufficient support.

As an example of best practice, the UK Government Communication Service has published a Crisis Communications Planning Guide.[115] The Guide uses a STOP framework (Strategy, Tactics, Organisation, and People) to give government departments a template for their crisis communications policies.

Best practice (5): Implement transparency and accountability measures

Transparency and accountability measures are necessary to ensure that the implementation of crisis protocols does not harm public trust, which can undermine the crisis response in turn. If a protocol is triggered without sufficient transparency, this could trigger concerns about shadowy forces, censorship, and double standards. Concerns already abound regarding a lack of transparency in digital services' content moderation policies. Such opacity has been found to "lead to mistrust, backlash and uncertainties surrounding platforms' policies and operations"[116]. During the Southport Riots, for example, allegations of inconsistent content moderation and censorship arose which threatened to undermine the credibility of the government's response[117]. As Demos found in our research on X's Community Notes system during the riots[118], a lack of data access can hamper accountability: We were only able to conduct our research because X provided access to

[113] https://committees.parliament.uk/writtenevidence/134454/html/
[114] https://committees.parliament.uk/writtenevidence/133665/html/
[115] https://www.communications.gov.uk/publications/crisis-comms-planning-guide/
[116] https://policyreview.info/articles/analysis/transparency-content-moderation
[117] https://www.bbc.co.uk/news/articles/cr548zdmz3jo
[118] https://demos.co.uk/research/researching-the-riots-an-evaluation-of-the-efficacy-of-community-notes-during-the-2024-southport-riots/

data on this system, yet no data is available on how X's other content moderation systems performed in the period.

For comparison, the EU DSA requires that services with crisis protocols should "report to the Commission by a certain date or at regular intervals specified in the decision, on the [crisis risk] assessments referred to in point (a), on the precise content, implementation and qualitative and quantitative impact of the specific measures taken [...] and on any other issue related to those assessments or those measures, as specified in the decision" (Article 36).[119] We think this offers an example of best practice that Ofcom should also replicate.

Best practice (6): Implement human rights impact assessments and mitigation measures

To mitigate the risk that services' respond to crises by implementing measures which disproportionately limit users' right to freedom of expression, privacy, and other rights, Demos Digital recommends that services should be required to conduct human rights impact assessments as part of their crisis protocols. Such assessments should identify how their crisis response policies would affect users' rights and set out appropriate mitigations.

As an example of best practice, the EU DSA's measures for platforms to implement crisis protocols include a requirement for services to "take due account [...] of the actual or potential implications for the rights and legitimate interests of all parties concerned, including the possible failure of the measures to respect [users'] fundamental rights" (Article 36). The DSA requires the EU Commission to set out how its voluntary crisis protocols will include "safeguards to address any negative effects on the exercise of the fundamental rights enshrined in the Charter, in particular the freedom of expression and information and the right to non-discrimination" (Article 48).[120]

Best practice (7): Implement measures to promote reliable public interest information

Services' crisis protocols should include provisions to ensure that important public interest information is displayed prominently. During crisis situations where there is an elevated amount of illegal, false and unreliable content, it is important as a way to ensure that the public is able to access critical information from public bodies such as the NHS, Fire Services, police, and local government.

Public services can struggle to communicate effectively with the public over social media in crises[121], where their messages may be drowned out by more attention-grabbing sources. In public health communications, for example, health misinformation may easily win out in

[119] https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng
[120] https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng
[121] https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/12/public-communication-scan-of-the-united-kingdom_6c3acae1/bc4a57b3-en.pdf

users' feeds[122]. During the Covid-19 crisis, this information environment presented significant challenges for the NHS[123]. In response, the Government Communications Service has stated that it worked closely with Google, Facebook and Twitter to "ensur[e] that public health campaigns were promoted through reliable sources" on their services during the pandemic.[124]

Question 53: Do you agree with our assessment of the impacts (including costs) associated with this proposal? Please provide any relevant evidence which supports your position.

N/A

[Response ends]

[122] https://www.tandfonline.com/doi/full/10.1080/13648470.2024.2386887#abstract
[123] https://committees.parliament.uk/writtenevidence/132776/html/, https://www.communications.gov.uk/wp-content/uploads/2020/10/COVID-19_Communications_Advisory_Panel_Report.pdf
[124] https://www.communications.gov.uk/wp-content/uploads/2020/10/COVID-19_Communications_Advisory_Panel_Report.pdf