DEMOS

FULL FACT

COMMUNITY DISORDER

HOW DO WE PREVENT AN INFORMATION EMERGENCY?

AN EPISTEMIC SECURITY NETWORK
POLICY BRIEFING

DEMOS AND FULL FACT - JULY 2025

Published by Demos July 2025

15 Whitehall, London, SW1A 2DD
T: 020 3878 3955
hello@demos.co.uk
www.demos.co.uk

INTRODUCTION

One year on from the Southport riots and less than 10 months before elections in Wales and Scotland, this briefing includes policy recommendations for information critical incidents in both election and non-election contexts. Please note that on 30 June Ofcom launched proposals for additional safety measures in the context of crises.[1]

RESEARCH SUMMARY

- The Southport riots were fueled by social media: the government's independent reviewer of terrorism legislation and Ofcom both described a clear connection between online activity and the violent disorder.[2,3]
- A false claim on LinkedIn suggesting the attacker was a migrant was seen 2 million times on social media.[4] Meanwhile an incorrect name for the Southport suspect, "Ali Al-Shakati", spread rapidly online,[5] alongside false claims he had recently come to the UK on a small boat,[6] or was Syrian.[7] Even after the police corrected the false information circulating about the attacker, the riots only intensified.[8]
- 'Community Notes', which uses a community-based approach to provide context or correction to false information online, was one of the systems of moderation that was in use on social media platforms such as X and YouTube during the riots. It is now also being trialled by Meta, alongside the company dropping its fact-checking programme in the US.[9]

Research by Demos and CFI[10] concerning the effectiveness of Community Notes as deployed on X concludes that they cannot be relied on as an effective measure against such instances of information crises in isolation. The research shows that, at the time of the riots:[11]

- **Community Notes were largely invisible to users during the riots, so could not prevent false and harmful information spreading:** The visibility of Community Notes is crucial to their effectiveness and only 4.6% (25) of posts in the dataset had Notes created by Community members during the Southport riots that were publicly visible during the same period, as these were the only Notes that achieved 'Helpful' status. 78.9% of posts had no visible Community Note, despite 424 having been created during the riots because they remained in the "Needs More Ratings" (NMR) status.

1 Ofcom consultation: https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/online-safety-additional-safety-measures pages 260-277.
2 Ofcom (2024) "Letter from Dame Melanie Dawes to the Secretary of State" https://bit.ly/4dv40mJ;His Majesty's Inspectorate of Constabulary and Fire & Rescue Services (2025). "Police ill-equipped to tackle impact of online content during serious disorder." http://bit.ly/3YEbw8w;
3 BBC (2024) "How a deleted LinkedIn post was weaponised and seen by millions" https://www.bbc.co.uk/news/articles/c99v90813j5o
4 Ibid
5 Full Fact (2024) "Incorrect name for Southport stabbings suspect circulates online" https://fullfact.org/online/incorrect-name-southport-stabbings-suspect/
6 BBC News (2024) "Did social media fan the flames of riot in Southport?"https://www.bbc.co.uk/news/articles/cd1e8d7llg9o
7 Sky News (2024) "Nigel Farage accused of being 'Tommy Robinson in a suit' over Southport stabbings comments"https://news.sky.com/story/nigel-farage-accused-of-being-tommy-robinson-in-a-suit-over-southport-stabbings-comments-13188129
8 Merseyside Police (2024, July 29) "Statement from Chief Constable Serena Kennedy following major incident in Southport". https://www.merseyside.police.uk/news/merseyside/news/2024/july/statement-from-chief-constable-serena-kennedy-following-major-incident-in-southport/
9 Meta (2025) "More Speech and Fewer Mistakes" https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/
10 https://www.lcfi.ac.uk/
11 Demos (2025) Researching the riots: an evaluation of the efficacy of Community Notes during the Southport riots. (Link to follow)

- **Community Notes were too slow to prevent false and harmful information going viral:** Community Notes must be visible quickly to have a chance of mitigating the misleading content of the post before it reaches a high volume of people. However, the daily average time it took between when a post was first created and when a Note was published to the public was 469 minutes (7.8 hours) rising to 1,193 minutes (19.8 hours) on 30th July - the day the riots began.[12,13] Posts associated with the Community Notes dataset received their highest engagement within the first 36-hours of being posted i.e. between 29 July and 30 July. To date, posts created over the period of the riots without a visible Community Note (despite one having been created, but not yet having found consensus), and that are both inaccurate *and* threatening to communities have been viewed 67.6 million times.
- **Community Notes did not prevent harmful, false rumours about the attacker amassing millions of views:** Posts that were false and relied on harmful stereotypes continued spreading without a Community Note, including posts that 'confirmed' the attacker was a Muslim (one post had 1.5 million views) or an illegal immigrant who had arrived in the UK on a boat (one post had 1.3 million views) - both false claims that have been debunked.[14]
- **Hate speech remained on X despite the use of both Community Notes *and* professional moderation teams:** Posts that incite racial hatred, and religious hatred, are illegal and against X's Terms of Service. Yet, posts that called for the permanent removal of Islam from the UK both lacked Community Notes *and* were not removed from the social media platform by professional teams, with one example receiving 1 million views. This demonstrates the pervasive and broader weaknesses of the professional moderation system on X, regardless as to whether the effectiveness of the singular moderation tool of Community Notes is increased for false information.

Research by CCDH and Maldita has also shown the following:

- 74% of accurate Community Notes on US election misinformation in 2024 were not shown to users.[15]
- In the 2024 European elections, only 15% of posts debunked by fact checkers had a Community Note[16]
- Community notes citing fact checkers become visible 90 minutes earlier than standard notes. The likelihood of a Note becoming visible rises from 8.3% to 12 citing fact checkers.[17]

12 By resolution, we mean for a Note to have received enough ratings to be considered either 'Helpful' or 'Unhelpful'. If it is rated 'Helpful' then the Note becomes visible with the post in Step 3. If the Note has been rated 'Unhelpful' then it remains invisible and no Note is shown. It is not clear from published information what happens with such Notes i.e. if you can continue voting on them and change their status or not.
13 Hope Not Hate (2024, 31 July) "The Far Right and the Southport Riot: What We Know So Far". https://hopenothate.org.uk/2024/07/31/the-far-right-and-the-southport-riot-what-we-know-so-far/
14 Note it is not possible to report the number of views within the period of the riots, only views at the point that the dataset was downloaded on 4 September 2024.
15 CCDH (2024) "Rated not helpful: How X's Community Notes system falls short on misleading election claims." https://counterhate.com/wp-content/uploads/2024/10/CCDH.CommunityNotes.FINAL-30.10.pdf
16 Maldita.es, "Faster, trusted, and more useful: The Impact of Fact-Checkers in X's Community Notes", February 2024, https://files.maldita.es/maldita/uploads/2025/02/maldita_informe_community_notes_2024.pdf
17 Ibid

POLICY OPTIONS

The following policy options have been developed in partnership between Demos and Full Fact and are designed to address each of the points of failure in the information supply chain during emergency scenarios, and are directed at government and the platforms themselves.

Full Fact and Demos are firm advocates of free speech and also of informed choice. In an age where polarised opinions, manipulated content and algorithms can push half truths or conspiracy theories from small online forums to millions of people in minutes, there's no freedom without reliable information.

PART 1 - POLICY OPTIONS FOR UK GOVERNMENT AND OFCOM

1. Cross-government readiness-enabling reforms

1a Clarify the UK government's definition of a crisis	The Defending Democracy Taskforce through its Joint Election Security Preparedness Unit and/or the National Security Online Information Team is already likely to have a set definition of what constitutes a critical information incident or crisis.[18] This cell already monitors what has been described as 'information incidents' and, as a result, the government has indicated there is no need for additional crisis protocols.[19] However, this process and what incidents trigger it currently lacks transparency. Definitions are needed to ensure proportionality and the use of evidence to justify the response measures. Full Fact has prepared a framework for defining and responding to such incidents, with indicative levels of evidence, following a process of consultation and feedback.[20] The framework defines an information incident as a cluster or proliferation of inaccurate or misleading claims or narratives - which can be sudden or have a slow onset - which relate to or affect perceptions of or behaviour towards a certain event or topic happening online or offline.[21] Certain events are likely to trigger information incidents and have a substantial and material impact on the people, organisations and systems that consume, process, share or act on information, towards good, neutral or bad outcomes. Full Fact identified eight illustrative categories of events or situations that require responses, including conflict, hybrid warfare and pandemics. The framework proposes five levels for the severity of an information incident: from business as normal (level 1) to a rare and severely high-impact incident (level 5). The severity at each level is determined with reference to a range of criteria, such as appearance on social media, search trends, influential sharing and coordinated behaviour. Having determined the severity and identified the key challenges, the framework asks users to determine their aims and appropriate responses to meet those aims,

18 Publictechnology.net (2024) "Government extends use of digital simulation for information incident crisis training."https://www.publictechnology.net/2024/07/01/education-and-skills/government-extends-use-of-digital-simulation-for-information-incident-crisis-training/
19 Joint Election Preparedness Unit (2025) Question for Ministry of Housing, Communities and Local Government https://questions-statements.parliament.uk/written-questions/detail/2025-01-08/HL3891
20 Full Fact, "How the Framework was created" https://fullfact.org/policy/incidentframework/about/
21 Ibid

	and to set evaluation criteria and an initial period within which to execute their responses. It calls for users to reconvene after that period expires, report on the effectiveness of their responses and consider any necessary adjustments.[22]
1b Introduce a cross-party, four nations, and civil society element to the Defending Democracy Taskforce's work.	The cross-government Defending Democracy Taskforce, set-up to ensure a whole Government approach to protecting the democratic integrity of the UK, with a particular focus on foreign interference, will be even more effective if it has four nation and English Mayoral representation to ensure it has the muscle of the wider state engaged. This will be particularly important ahead of elections in Scotland and Wales in 2026.
1c Establish multi-disciplinary advisory body to the DDTF	Civil society is also an invaluable partner in epistemic defense.[23] As a window into the citizenry, civil society organisations are often the first to detect where epistemic threats arise. The government should support and engage with Demos's multidisciplinary Epistemic Security Network as an advisory body to the Taskforce.
1d Strengthen and make transparent the UK government's crisis response protocols and procedures for responding to information threats, including foreign interference	Unlike the EU's Digital Services Act (Article 48, 'Crisis protocols'), the Online Safety Act (OSA) does not currently include a crisis or emergency response protocol, including for cases where false information threatens to mobilise offline violence. Ofcom has recently published proposals for 'Additional Safety Measures' for platforms to adopt their own crisis response protocol. However, the regulator cannot trigger the cross-government departmental response needed in an information crisis context. The Defending Democracy Taskforce should urgently consider developing transparent crisis protocols for incidents where the rapid spread of false information, such as that around the Southport riots, is causing offline disorder. The UK should introduce such crisis-specific mechanisms, including both crisis-specific risk assessments and crisis responses, in a manner which is clear, easy to understand, proportionate and open to public scrutiny. Indicative examples of such mechanisms include the Government of Canada's Critical Election Incident Protocol, the Global Internet Forum Counter Terrorism Content Incident Protocol and Full Fact's Information Incident Framework.[24] Any mechanisms should be transparent and will require carefully considered safeguards and high thresholds restricting their use. For example, they could be restricted to only being available whilst the UK is in a State of Emergency as defined by the Emergency Powers Acts, or following the issuance of a critical national security alert by the National Cyber Security Centre about an imminent or ongoing cyberattack. Moreover, the decisions made using these mechanisms should be challengeable in court and should be open to Parliamentary scrutiny.

22 Full Fact, "Using the Framework" https://fullfact.org/policy/incidentframework/how-to/
23 Written evidence submitted by the Home Office, April 2024 https://committees.parliament.uk/writtenevidence/129622/pdf/;
24 Government of Canada, Critical Election Incident Protocol https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/public-protocol.html GIFCT (2025) "Content Incident Protocol" https://gifct.org/content-incident-protocol/; Full Fact (2025). 'Framework for Information Incidents'. https://fullfact.org/policy/incidentframework/

1e Instigate fresh systems mapping and crisis scenario Red-Teaming research	Established expert groups from civil society, academia, regulators, and government departments should map the threat landscape in order to develop targeted tools to most effectively address vulnerabilities and counter threats. We recommend the use of extended hypothetical scenario-mapping and red-teaming exercises (deliberately exploring a scenario from an adversary's perspective) as conducted in the 2020 epistemic security report to help think more holistically and into the future.[25] Government should invest in building multidisciplinary epistemic security research groups and expert networks. Epistemic security experts are embedded within separate and diverse professions and often have limited capacity to respond to (or to help to pre-emptively mitigate) epistemic threats. To demonstrate the benefit of this Demos is currently undertaking work in this area.

2. New and renewed regulation

2a Platforms should assess and mitigate systemic risks, including to civic discourse, electoral processes and public security	Following a similar course to the EU's Digital Services Act (DSA), government should consider using the Elections Bill to incorporate into the Online Safety Act a requirement for Category 1 service providers (large online platforms) to assess the systemic risks - from the design, functioning or use of their services - of any negative effects on civic discourse, electoral processes and public security and to put in place reasonable, proportionate and effective risk mitigation measures tailored to the risks. Such a provision exists in Article 34 and 35 of the European Union Digital Services Act.[26] The Elections Bill should expand Schedule 8 of the Online Safety Act, to enable Ofcom to require regulated services to provide information related to the additional priority offences and systemic risks - underpinned by a Code of Practice. This could include: • Mis/disinformation that is harmful to democracy; • Viewing figures; • Processes to manage the risks; • Design and operation of algorithms which affect the display, promotion, restriction or recommendation of content relating to electoral processes; and • Political material that is AI-generated or manipulated.

25 Seger et al. (2020). 'Tackling threats to informed decision- making in democratic societies: Promoting epistemic security in a technologically-advanced world'. The Alan Turing Institute

26 https://www.eu-digital-services-act.com/Digital_Services_Act_Article_34.html; https://www.eu-digital-services-act.com/Digital_Services_Act_Article_35.html

2b Existing Terms of Service prohibitions against violent content should be consistent in all contexts - including for posts concerning the perpetrator of an attack or in its aftermath	In its Violent Content policy, X states that, in certain cases, content is allowed to remain on its platform, but be made less visible through restricting its reach if "the context is outrage or reactive against perpetrators of major harm". Such a provision suggests that *because* of the crisis context like that of the Southport attack, certain speech is allowable rather than treated as constituting higher risk.[27] Such an approach in Terms of Service presents an opportunity for posts that rely on false information and relate to the identity or motivations of the attacker, and are harmful to wider communities. The risk of such posts gaining high visibility is exacerbated given the opportunity for recommender and engagement-based algorithms to amplify them before they are identified by professional moderation teams. As a result, such permissive terms in the context of a crisis should be removed. Ofcom should require consistency with regards to illegal harms content as a minimum standard in social media platforms' Terms of Service, especially during crises.

3. Defend elections, protect candidates

3a Transparent and accountable systems for dealing with electoral information incidents, including foreign interference	The UK remains vulnerable to large-scale attempts at election interference by foreign powers, as has been seen in countries like the US and Romania. To challenge suspected foreign interference activities in its elections, the UK needs a timely, transparent, and democratically accountable mechanism for security services to notify the public about such incidents independently of the government of the day. In the event that such a notification becomes necessary, there would need to be procedures in place for the notification to be acted on with due propriety. Canada has a system in place called the Critical Election Incident Public Protocol from which the UK could take inspiration.[28] In a worst-case scenario – such as a critical cybersecurity attack – the result may require a delay in an election. Such a mechanism would require careful checks and balances to uphold democratic oversight and win public trust. For example, more transparency is needed from the security services regarding their work to safeguard elections; and about the roles, objectives, resourcing and activities of other bodies working on electoral security, including the Election Cell and the National Security Online Information Team. One solution could be to require the services to regularly publish electoral threat assessments for public viewing alongside summaries of steps being taken to prevent an incident. These would need to be published before, during, and after elections. Moreover, there must be pathways open for civil society to trigger foreign interference investigations. Under the National Security Act 2023 the public can write to the police alleging foreign interference for investigation by appropriate authorities, but the decision to prosecute remains with the Attorney General, a government appointee. If there is no decision to

27 X (2025) "Violent Content" https://help.x.com/en/rules-and-policies/violent-content
28 Government of Canada, Critical Election Incident Protocol https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/public-protocol.html

	prosecute then any investigation remains confidential and unpublished. Any process would require oversight by elected representatives in Parliament (when Parliament is sitting). This scrutiny could come via Select Committee hearings, Parliamentary debates, or similar procedures.
3b Extend risk assessment obligations of platforms to include deepfakes about electoral candidates and undue influence	Regulated services' already have risk assessment obligations in the Online Safety Act which covers illegal content such as terrorism content and hate. Based on these obligations, platforms must identify the risk of such content appearing on their platforms, assess the risk of harm, identify and implement measures to reduce the risk of harm, and report on their risk assessments on an ongoing basis, not just in the context of an election.[29] The Elections Bill should add the following priority offences to Schedule 7 of the Online Safety Act so that platforms can also assess, monitor and mitigate risks against them: • the offence of making or publishing a false statement of fact about an electoral candidate before or during an election for the purpose of affecting their return (section 106 of the Representation of the People Act), amended to expressly include deepfakes as recommended by the Electoral Commission,[30] and • the offence of undue influence, which includes forcing someone to vote in a particular way or not vote at all, or otherwise interfering with their free exercise of the franchise (section 114A of the RPA), as previously recommended by the Joint Committee on the Online Safety Bill;[31] and undue influence in relation to Scottish Parliament elections (Rule 77 of the Scottish Parliament (Elections etc.) Order 2015), and Senedd Cymru elections (Rule 81 of the National Assembly for Wales (Representation of the People) Order 2007).
3c Update legislation prohibiting threats, intimidation, and violence against electoral candidates to reflect online harms	The UK has seen a rise in violent and threatening online behaviour aimed at election candidates. The Speaker's Conference, launched to examine and recommend routes to tackling this issue, provided a number of robust recommendations for offline threats in June 2025. Online abuse will fall within the scope of its next inquiry.[32] A new system is required for the civil regulation of elections and online media to assess and mitigate the risks to candidates and other election participants. At this point, no comprehensive review of online risk to candidates has ever been published. A regulatory system could bring the Electoral Commission and the National Police Chiefs Council into the Online Safety Act regime to work with OFCOM and platforms to assess the risks of harm to victims and then put in place systems to mitigate those risks enforced by those regulators. The UK must introduce guidance and legislation to address online harms against candidates. To disincentivise such behaviours, these measures should go beyond the penalties in the Elections Act and could include fines or prosecution.

29 https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/quick-guide-to-online-safety-risk-assessments
30 Written evidence submitted by the Electoral Commission (2025) https://committees.parliament.uk/writtenevidence/141330/html/
31 Reported prepared by the Joint Committee on the Draft Online Safety Bill (2021), https://committees.parliament.uk/publications/8206/documents/84092/default/
32 Speakers Conference, 2024. https://committees.parliament.uk/publications/48116/documents/251907/default/

3d Increase investigative and enforcement powers for the Electoral Commission	The Elections Bill should give the Electoral Commission the powers: • To obtain information from any person outside of a formal investigation, including from the online platforms they do not regulate. This would enable it to better monitor and enforce the rules about how campaigners spend money to influence voters – which might include analysing bots, adverts paid for by overseas actors, and content that is being sponsored and boosted. • To share information with the police or other regulators. This would lead to faster and more straightforward collaboration with partner agencies, including Ofcom and the ICO. • To impose fines of £500,000 per offence or 4% of a campaign's total spend, whichever is higher. The Commission should be given sufficient resources to research and deploy effective public information campaigns about deepfakes and mis/disinformation during election periods. It should also be resourced to help raise media literacy and counter mis/disinformation – particularly but not exclusively for new younger voters – alongside Ofcom, civil society, grassroots organisations, educational institutions and others.

PART 2 - PROPOSALS FOR PLATFORMS

4. Strengthen platform content moderation against hate

4a Existing Terms of Service prohibitions against violent content should be consistent in all contexts - including for posts concerning the perpetrator of an attack or in its aftermath	In its Violent Content policy, X states that, in certain cases, content is allowed to remain on its platform, but be made less visible through restricting its reach if "the context is outrage or reactive against perpetrators of major harm". Such a provision suggests that because of the crisis context like that of the Southport attack, certain speech is allowable rather than treated as constituting higher risk.[33] Such an approach in Terms of Service presents an opportunity for posts Demos that rely on false information and relate to the identity or motivations of the attacker, and are harmful to wider communities. The risk of such posts gaining high visibility is exacerbated given the opportunity for recommender and engagement-based algorithms to amplify them before they are identified by professional moderation teams. As a result, such permissive terms in the context of a crisis should be removed. We therefore recommend that social media platforms ensure that their provisions for illegal harms content in their Terms of Service are consistent in all contexts, including in the aftermath of an attack, and do not include such exemptions.

33 X (2025) "Violent Content" https://help.x.com/en/rules-and-policies/violent-content

4b Strengthen removal systems for racial and religious hatred, including when based on false information	Our results demonstrate that the racial and religious hatred that was allowed to remain on social media platforms relied on false information about the attacker - presuming a race and/or religion that was (later) falsified in news reporting. Given the stipulation that illegal hate must meet a threshold that demonstrates an intention of the poster to inflict harm or stir up religious or racial hatred, it is possible that such posts remained on social media platforms because of a gap in the time before the identity of the attacker was confirmed and thus such false assumptions were *proven* incorrect.[34] Whilst not a justification for this assessment, such a scenario highlights the complexity of assessing individual pieces of content for their illegality, particularly in a crisis context when the scale of the content proliferating on platforms escalates rapidly and when facts are not yet established. Given that the Online Safety Act now requires platforms to assess and mitigate the risk of users encountering illegal content such as hate speech, including racial and religious hatred: • Social media platforms should clarify their approach to racial and religious hatred that includes or relies on false or as yet unverified information about a violent offender in its Terms of Service. • Social media platforms should ensure that racial and religious hatred is removed regardless as to whether it relies upon false or as yet unproven information about a violent offender.
4c Prohibit hate and incitement to violence against migrants and asylum seekers on social media platforms	Ofcom's current Illegal Harms Codes and Risks Assessment Guidance does not include hatred towards migrants and asylum seekers as 'priority illegal content' because it is not listed in Schedule 7 of the Online Safety Act that is linked to the Public Order Act 1986. In the absence of clear legal frameworks, social media platforms should therefore voluntarily ensure that hate, including dehumanising language, and incitement of violence towards migrant communities and asylum seekers is disallowed on the platform. This voluntary arrangement should be accompanied by clear policies clarifying the difference between hate speech and free speech surrounding migration and asylum policy.

34 Ofcom (2024) Illegal Content Judgement Guidance. https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/illegal-content-judgements-guidance-icjg.pdf?v=387556

5. Strengthen platform moderation to tackle mis/disinformation

5a Use Fast Track Notes to head off critical threats to the information environment	Platforms should adopt a crisis protocol for community notes systems to triage high risk content and send it through a high priority funnel. To underpin Fast Track Notes, platforms should: • Use AI to identify higher risk content and allow contributors to filter for this • Develop a callout function to alert specific sets of contributors (e.g. fact checkers, humanitarian, digital forensics, infodemic managers) to the need for a Fast Track Note • Use generative AI to suggest counter-arguments and widely-recognised high quality sources such as fact checks • Adjust consensus thresholds so Fast Track Notes can be published quicker
5b Develop Super Notes - amalgamating notes into one published Super Note from those that are otherwise unpublished	Too often, notes on X do not get published, particularly when they relate to controversial or politically charged topics. Instead, many notes are left in limbo, often relating to the same claim or topic. Platforms should introduce a system to identify and automatically generate a single note amalgamating all the substantially similar notes left on a post/similar posts. This could apply when more than four posts have been added without agreement after 48 hours. This may require the development of technology such as: • Conflict resolution algorithms to determine when the AI should present multiple perspectives versus when it should synthesise a consensus view (for example source credibility weighting, and a public explanation of how the AI weighted and synthesised different viewpoints) • A human fallback mechanism for when AI synthesis is ineffective • AI-generated notes would need to be labelled as such, with the option for users to view the original constituent notes • Technology to identify and fuse multiple notes relating to the same post or claim and turn them into one Super Note (rewriting them using a GenAI trained to get sufficient votes due to clarity or balance)
5c Strengthen the number and quality of Community Notes datasets made available by platforms	Provision of the Community Notes dataset by X made this research possible. However, it is not clear if Meta plans to release their own Community Notes data in the same way. It will be particularly useful to be able to compare the applications of the same Community Notes models on different platforms to identify how different platform environments and applications may produce differential results. Platforms deploying Community Notes models and making their data publicly available should also enable the following details within the dataset to strengthen the quality of possible research:

	• When posts are removed - confirm if a post was removed by the user or platform and on what date. If removed by the platform, confirm on what rationale it was removed. Enable researchers to continue to view and analyse engagement metrics for a post despite being removed. • When post IDs are shared, add the further detail of the country location of the account as well as the Community Note that is created. This could facilitate assessments of the level and nature of foreign influence within a dataset of misleading posts and those seeking to correct them. As pervasive narratives become increasingly influential across borders, the origination and approach to tackling such influence becomes more important to understand.
5d Provide greater transparency into the size and shape of 'the community' needed to enable fast deployment of Community Notes	The preference for a community-based model of moderation is currently undermined by a lack of transparency into who the community adding Notes to posts actually is, or to what extent the community reflects the user-base on a given platform and/or the geographical context in which posts are being viewed. Platforms deploying Community Notes models should provide greater transparency and ideally publish datasets that clarify the characteristics and size of the community that is rating content on any one day, particularly highlighting if and when the community is lacking certain segments that are needed to ensure greater diversity of opinion and to facilitate stronger consensus-making that truly bridges divides. This data could be compared with the speed of Community Notes resolution to identify when the community is of a sufficient volume or diversity to enable fast decision-making and when it is too low to function effectively.
5e Provide faster, transparent insight into what proportion of Community Notes use links from fact-checking organisations and news media organisations, including for posts and Notes that are removed	There is an ongoing propensity for Community Notes members to draw on news media organisations publications and/or fact-checking organisations in their Community Notes, particularly those that achieve Helpful status identified in this research and that published by Maldita.[35] This demonstrates a potential dependency for the success of Community Notes on the work of fact-checking organisations and news media organisations which undermines the argument that community-based models offer a binary alternative choice to professional news media and fact-checking organisations. We were able to identify the proportion of Notes that included links to fact-checked sources through manual, qualitative analysis of each Note. But we could only do this for Notes that remained on the platform - reducing our dataset considerably when it appears that posts that do achieve a 'Helpful' Note are often deleted. By simplifying the recording of when a Community Note includes a fact-checking organisation or news media link in the dataset and enabling this data to remain in the dataset even when a post and therefore Note is removed, platforms could greatly assist researchers in rapidly assessing this relationship with a more robust dataset.

35 Maldita (2025) "Faster and more useful: the impact of fact checkers in X's Community Notes" Maldita.es. Available at: https://maldita.es/investigaciones/20250213/community-notes-factcheckers-impact-report/

Licence to publish

Demos – Licence to Publish

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

1 Definitions

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

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

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

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

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

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

2 Fair Use Rights

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

3 Licence Grant

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

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

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

4 Restrictions

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

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

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

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

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

5 Representations, Warranties and Disclaimer

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

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

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

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

6 Limitation on Liability

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

7 Termination

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

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

8 Miscellaneous

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

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

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

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

DEMOS

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

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

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

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

DEMOS

PUBLISHED BY DEMOS JULY 2025

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