DEMOS

RESEARCHING THE RIOTS

AN EVALUATION OF THE EFFICACY OF COMMUNITY NOTES DURING THE 2024 SOUTHPORT RIOTS

HANNAH PERRY
DR GIULIO CORSI
NAEMA MALIK

In partnership with:

CFI
LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

JULY 2025

Published by Demos July 2025

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

CONTENTS

ACKNOWLEDGEMENTS

We would like to thank: Jamie Hancock and Aidan Garner for their intensive work labelling and analysing the Community Notes dataset; James Ball, Demos Fellow, for his initial recommendation that we pursue this research; Elizabeth Seger for support in the set-up of this project and foundational thinking surrounding epistemic security; to Polly Curtis for her feedback at different stages of drafting this report; and to Chloe Burke, for design and publication support.

We would like to thank the following experts and partners for feeding back on our analytical approach:

- Dr Damian Tambini, Department of Media & Communications at LSE
- Dr Chris Allen, School of Criminology, Sociology and Social Policy at University of Leicester
- Milo Comerford, Henry Tuck and Isabelle Wright, Institute of Strategic Dialogue
- Azzurra Moores and Andy Dudfield, Full Fact
- Anki Deo, Hope Not Hate
- Fizza Qureshi, Migrant Rights Network
- DSIT
- Ofcom

Any mistakes are the authors' own.

Hannah Perry, Dr Giulio Corsi and Naema Malik

July 2025

ABOUT THIS RESEARCH REPORT

This research paper is the result of a partnership between Demos and Giulio Corsi, a Research Fellow at the Leverhulme Centre for the Future of Intelligence (CFI) at the University of Cambridge. Demos led the project, and the qualitative analysis of the posts and Community Notes. CFI led the quantitative methodology design and analysis.

Demos is the UK's leading cross-party think tank producing research and policies that have been adopted by successive governments for over 30 years. We exist to put people at the heart of policy making and to build a more collaborative democracy. Demos Digital, Demos's digital policy research hub, specialises in digital policy making to create a future in which technology is built for the good of people and democracy.

The Leverhulme Centre for the Future of Intelligence at the University of Cambridge is a highly interdisciplinary research centre exploring the nature, ethics and impact of artificial intelligence (AI). Funded by the Leverhulme Trust, CFI is based at the University of Cambridge, with spokes at Imperial College London and University of California, Berkeley, as well as close links with industry and policymakers. CFI brings together academics from a variety of disciplines as diverse as machine learning, philosophy, history, literary studies, engineering, media studies and design in order to explore the possibilities of AI in both the short and long term.

This research paper is a contribution to Demos's Epistemic Security programme. This programme aims to secure healthy and robust information supply chains within the UK and build resilience to adverse influence on our democratic processes. In the context of democratic backsliding and rising extreme populism, we are making the case this should be a central mission of this government.

EXECUTIVE SUMMARY

INTRODUCTION

The Southport riots shocked the country in the speed, intensity and ferocity with which they erupted in 2024. It is now accepted that these riots were fueled by social media: Ofcom has described a "clear connection" between online activity and the violent disorder.[1,2] Even after the police corrected the false information circulating about the attacker, the riots only intensified.[3] '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 one such social media platform, X, during the riots. It is now also being trialled by Meta and YouTube.

Over this same period, US-headquartered social media platforms have not just begun adopting community-based moderation methods, but have also cut back on their professional moderation teams and existing independent fact-checking interventions.

This report is the first to evaluate how effective Community Notes was at mitigating false information in the course of the Southport riots and, as such, is a major contribution to the wider understanding of the efficacy of the system in fast-moving events. While recognising the potential of such a system in everyday circumstances, this research demonstrates that this model of moderation is fundamentally unfit in a crisis context. Its challenges, exacerbated in polarised situations, will repeat in future crisis situations and demonstrate that such a model for moderation cannot be relied upon to mitigate false information and the violent disorder it can fuel from escalating. Such results underline the dependency on additional, complementary moderation mechanisms, such as independent fact-checking organisations and professional in-house teams, in such circumstances for effective content moderation.

This evidence demonstrates the need not just to strengthen the Community Notes moderation system itself, but to also mitigate the future risks posed to the UK's epistemic security by US social media companies' increasing reliance upon it whilst simultaneously undermining alternative moderation systems.[4]

METHODOLOGY

Demos and CFI have analysed a publicly available dataset of Community Notes together with the associated posts that were created in relation to the Southport riots during the period of 29 July (the day of the attack) to 11 August 2024. These posts were labelled based on their

1 Ofcom (2024) "Letter from Dame Melanie Dawes to the Secretary of State" https://bit.ly/4dv40mJ
2 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 Merseyside Police (2024, July 29) "Statement from Chief Constable Serena Kennedy following major incident in Southport". https://bit.ly/43eNbrr
4 Seger, Perry & Hancock (2025) "Epistemic Security 2029." https://demos.co.uk/wp-content/uploads/2025/02/Epistemic-Security-2029_accessible.pdf

accuracy and the potential threat they posed to communities and individuals. By analysing each post that the Community Note responded to and accessing the metadata about that post, we can assess the speed, scale and impact of their use and draw conclusions about the efficacy of the system.

FINDINGS

Our findings demonstrate that Community Notes, as deployed in July and August 2024, failed to mitigate the harmful, inaccurate information that fuelled the crisis.[5] The evidence shows that:

- **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. This suggests that challenges achieving sufficient consensus among community members prevented Notes becoming publicly available which meant they had no chance of stemming the tide of false information. Such a barrier is unsurprising in polarised, fast-moving and violent situations.
- **Where Community Notes did appear, they relied on traditional, independent fact checking:** 90% of the Community Notes that were rated 'Helpful' that remain visible for analysis (10) included links to mainstream news publications, such as the BBC, Sky News and CNN, that provided verified information that counteracted the claims made in the corresponding posts. Such an approach mirrors that taken by independent fact-checking organisations - the very approach that some social media platforms deploying Community Notes have critiqued and to differing degrees cut back on.
- **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.[6,7] 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) that are both inaccurate *and* threatening to communities have been viewed 67.5 million times.

5 We note that in October 2024, X announced an update to their Community Notes model indicating that they had found a solution to speeding up their publication. It is not clear what proportion of Notes that are created are able to achieve the 'sped up' version described and in what conditions. https://x.com/CommunityNotes/status/1851337944822325253

6 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.

7 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/

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.[8]

- **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 IMPLICATIONS

This paper presents new and rigorous evidence of the weaknesses of the Community Notes approach in a crisis context, and its failure to contain the false information that fuelled the Southport riots. It raises fresh concerns about the current approach to moderation by US headquartered social media platforms, which are increasingly relying on Community Notes while reducing support for independent fact checking organisations and professional in-house moderation teams. This poses serious future risks to the UK's epistemic security, creates harmful risks for the information environment and for citizens who rely on social media platforms as a source of news.[9]

Following the publication of this research, Demos will produce a policy briefing building on this evidence and setting out steps the government and other authorities should take to develop stronger responses to such information crises.

8 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.

9 Seger, Perry & Hancock (2025) "Epistemic Security 2029." https://demos.co.uk/wp-content/uploads/2025/02/Epistemic-Security-2029_accessible.pdf

INTRODUCTION

The violent riots that erupted during July and August 2024 following the devastating attack at a children's dance party in Southport, Merseyside, triggered shock and fear across the UK. These riots amplified the levels of racism, Islamophobia and levels of anti-migrant hatred lying latent within British society.[10] The rapid escalation of violence underlined the urgent task of tackling such hateful attitudes corroding the fabric of a safe, strong and healthy democracy.[11]

Ofcom has subsequently identified a "clear connection" between online activity and this violent disorder.[12] In the hours that followed the attacks, anti-migrant and anti-Muslim narratives began to spread online across multiple channels such as TikTok, Telegram, Meta and X. The false information surrounding the identity of the attacker - relying on harmful stereotypes that associated his violent behaviour with his falsely alleged Muslim faith or immigration status - fuelled rioters to protest the presence of Muslims and migrants in the UK, attacking mosques and attempting to set asylum seeker accommodation on fire. Such an offline uprising surged despite repeated police reports that corrected this information, confirming the attacker was a UK citizen born in Cardiff and that the name circulating online was incorrect.[13]

While the Online Safety Act was not in force during the Southport riots in July and August 2024 as it is today, the Secretary of State for the Department of Science, Innovation and Technology has stressed the need to learn lessons from this incident. In its Strategic Priorities for Online Safety, "the government is… clear that it expects platforms to take proactive steps to reduce the risks their services are used to carry out the most harmful illegal activity. This includes:...illegal disinformation and hate which incites violence towards specific individuals or groups, leading to societal fragmentation and disorder".[14]

However, social media platforms have defended their content moderation response during this period. Some have emphasised the size of the challenge they faced, telling Ofcom that misinformation seeking to whip-up hatred appeared "almost immediately" on their platforms and that they were "dealing with high volumes, reaching the tens of thousands of posts in some cases".[15] The challenges of scale and speed were repeated in a recent Science, Innovation & Technology Commons Committee Inquiry by a range of social media companies, including a statement by X that cross-functional crisis teams were set-up to work around the clock to tackle the problem.[16] The role of Community Notes was also emphasised by X as a mechanism that

10 Gohill (2024) "Record amount of anti-Muslim abuse reported in UK since 7 October attacks". https://www.theguardian.com/news/2024/oct/04/record-amount-of-anti-muslim-abuse-reported-in-uk-since-7-october-attacks
11 Runnymede (2024) "Six months since the riots, charities urge the government to take action on the growing threat of the far right." https://www.runnymedetrust.org/news/six-months-since-the-riots-charities-urge-the-government-to-take-action-on-the-growing-threat-of-the-far-right
12 Ofcom (2024) "Letter from Dame Melanie Dawes to the Secretary of State" 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
13 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/
14 DSIT (2024) "Draft Statement of Strategic Priorities for online safety". https://www.gov.uk/government/publications/draft-statement-of-strategic-priorities-for-online-safety/draft-statement-of-strategic-priorities-for-online-safety
15 Ibid
16 X (2025) "Written evidence submitted by X - X Submission to Commons Science, Innovation and Technology Committee Inquiry: 'social media, misinformation and the role of algorithms.' https://committees.parliament.uk/writtenevidence/133665/pdf/

could tackle 'misleading posts' with the statement that, "Notes in relation to these incidents have been viewed millions of times".[17]

At the time of the riots, the Community Notes system was in use on X in the UK. Whilst Community Notes was not in use by another social media platform, Meta, at the time, it is now being piloted for its US users and may be extended to the UK in the future.[18] Over the same period, both X and Meta have also made cuts to their independent fact-checking programmes explaining that they had concerns about the 'political bias' of fact-checkers, and citing research that US citizens had more trust in peer-led Notes, than in professional fact-checks.[19] Such choices made by executives in the US favouring Community Notes as an alternative to professional fact-checks have repercussions for UK citizens whose information diets are heavily affected by such decisions on an ongoing basis, as well as the vulnerabilities it may create for future crisis contexts.

This paper is the first to present analysis of the Community Notes dataset from the period of the Southport riots. It provides new insight into the efficacy of this moderation system in such a fast-moving crisis and highlights the opportunities and risks of its application for the UK.

This evidence:

- Provides insight for social media platforms, such as X, Meta and YouTube, who have adopted the Community Notes moderation mechanism to differing degrees
- Highlights the limits of the Online Safety Act in relation to the types of harmful content shared during the riots
- Provides lessons to inform the Additional Safety Measures and specifically the Crisis Response protocol consultation proposed by Ofcom on June 30
- Has broader implications for how the UK responds to rapid information threats to our democracy during moments of crisis for DSIT and the UK Government's Defending Democracy Task Force.[20]

As we set out in our recent paper *Epistemic Security 2029*, in the current geopolitical moment the UK faces an unprecedented coalescence of threats to the security of our information supply chain.[21] These threats stem from global democratic backsliding and authoritarian shifts, crumbling news ecosystems, and mass digitisation of communication. Epistemic security is about securing healthy and robust information supply chains and building resilience to adverse influences. This paper reflects the next stage in our Epistemic Security series examining and tackling UK ecosystem-level vulnerabilities in a crisis context.

17 Ibid
18 Meta (2025) "Testing begins for Community Notes on Facebook, Instagram and Threads". https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/
19 Meta (2025) "More speech and fewer mistakes". Meta. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/; Wojcik et al (2022) "Birdwatch: Crowd Wisdom and Bridging Algorithms can inform understanding and reduce the spread of misinformation." 10.48550/arXiv.2210.15723.
20 Ofcom (2025) "Statement: Protecting People from illegal harms online - Overview" https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/overview.pdf?v=387529
21 Seger, Perry & Hancock (2025) "Epistemic Security 2029." https://demos.co.uk/wp-content/uploads/2025/02/Epistemic-Security-2029_accessible.pdf

SECTION 1
BACKGROUND

COMMUNITY NOTES: A MODERATION MECHANISM EXPLAINED

Over the past 15 years, social media companies have relied upon professional, independent fact checking partnerships to tackle inaccurate content on their platforms. Extensive evidence has demonstrated that if social media platform users are provided additional context and given warning about a post's inaccuracy, they are less likely to believe it, spread it and provide a positive reaction to it.[22] However, professional fact-checkers have typically faced challenges of scale and speed given the sheer volume of content that needs fact-checking combined with the time needed to verify claims effectively.[23] More recently, in the USA, where many social media platforms host their headquarters, fact-checkers have faced increasingly intense accusations of bias particularly by right-wing voters and, in the last year, by social media platforms themselves.[24,25] In this context, a different model of moderation has been introduced called 'Community Notes'. As of April 2025, the system is used by X and YouTube and has recently been adopted by Meta.[26]

What is the Community Notes moderation system?

The Community Notes moderation system uses 'the crowd' (i.e. other social media platform users) to provide context to publicly posted content on their respective platforms.[27] Notes are intended to provide additional information to other users about the post to help subsequent readers orient and evaluate what they read. When first designed by Twitter in 2020, the moderation system was called 'Birdwatch'. It was rebranded to 'Community Notes' when Twitter came under new ownership in late 2022 and when Twitter was also rebranded to X.[28]

22 Porter, E. et al. (2024). "Factual corrections: Concerns and current evidence." Curr. Opin. Psychol., Vol 55, 101715. Elsevier.
23 Martel, C. & Rand, D. G. (2023) "Misinformation warning labels are widely effective: A review of warning effects and their moderating features." Current Opinion in Psychology 54, 101710
24 Walker and Gottfried (2019) "Republicans far more likely than Democrats to say fact-checkers tend to favor one side". Pew Research https://www.pewresearch.org/short-reads/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/
25 Meta (2025) "More Speech Fewer Mistakes" https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/
26 Meta (2024) "Testing begins for Community Notes on Facebook, Instagram and Threads" Meta Newsroom. Available at: https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/
27 Tom Stafford (2025) "Do Community Notes Work?" LSE Blogs. https://blogs.lse.ac.uk/impactofsocialsciences/2025/01/14/do-community-notes-work/
28 Stefan Wojcik et al 'Birdwatch: Crowd wisdom and Bridging Algorithms can Inform Understanding and Reduce the spread of Misinformation.' Available at: https://arxiv.org/pdf/2210.15723

Why does Community Notes matter to UK public policy now?

In addition to the rebrand, X expanded Community Notes beyond the USA into other markets, including the UK.[29] This rebranding and emphasis on its community-based moderation model occurred at the same time as a significant and rapid reduction of X's professional Trust and Safety Team responsible for determining what content should be removed from its platform. As a result, Community Notes became one of the core mechanisms through which X was moderating content on its platform.[30]

Last year, in 2024, an increasing number of US-based social media platforms began adopting the Community Notes moderation model. YouTube began piloting an identical system to X in June 2024, and now, since March 2025, Meta has begun using the same.[31] Both social media platforms allow users to opt in, contribute Notes to posts, and rely on an open-source 'bridging algorithm' to identify where users who have diverse perspectives have found agreement before adding the Note to the post.[32,33]

A number of concerns have been raised regarding the Community Notes method of moderation, particularly after posts on X that lacked Community Notes were referenced as contributors to the Southport riots in the UK. After these events, X announced they were exploring improvements to the system, including evaluating how to speed up the time it takes for a Note to be displayed on a post.[34,35]

Who gets a say in Community Notes?

Unlike professional fact-checking programmes, a Community Note can be created by any user who has opted-in to contribute and who meet certain criteria set by the respective social media platform. The following table indicates the criteria used by the different social media companies, demonstrating that they share only one common requirement: for all potential contributors to have been on their respective platforms for 6 or more months.

TABLE 1
ELIGIBILITY CRITERIA TO BECOME A COMMUNITY NOTES CONTRIBUTOR ON X, YOUTUBE AND META

	No recent violations of platform's policies	Joined the platform 6+ months ago	Has a verified phone number	Is not a supervised account	Account does not have multiple owners e.g. brand	Based only in the USA	Over 18 years of age
X[36]	✓	✓	✓	✗	✗	✗	✗
YouTube[37]	✓	✓	✗	✓	✓	✗	✗
Meta[38]	✗	✓	✓	✗	✗	✓	✓

29 Ibid
30 X (2023) "Code of Practice on Disinformation - Transparency Report - Report of Twitter for the period H2 2022". https://cdn.arstechnica.net/wp-content/uploads/2023/02/Twitter-January-2023.pdf; Iark, Lindsay "Elon Musk made 1 in 3 Trust and Safety staff ex-X employees, it emerges" The Register. :https://www.theregister.com/2024/01/11/elon_musk_twitter_safety_cull/
31 Meta (2024) "Testing begins for Community Notes on Facebook, Instagram and Threads" Meta Newsroom. Available at: https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/
32 Vanian, Jonathan (2025) "Meta's Community Notes will use tech from Elon Musk's X" NBC News. https://www.nbcnews.com/tech/tech-news/metas-community-notes-fb-instagram-will-use-tech-x-rcna196210
33 Ovadya, A. (2022) "Bridging-based ranking"https://www.belfercenter.org/publication/bridging-based-ranking
34 Community Notes "Introducing Lightening Notes" X. https://x.com/CommunityNotes/status/1851337944822325253?lang=en-GB
35 Hutchinson, Andrew (2024) "X Improves the Speed of Community Notes Being Displayed" Social Media Today. https://www.socialmediatoday.com/news/x-formerly-twitter-adds-faster-community-notes/731416/
36 Community Notes Guides "Signing up" X. https://communitynotes.x.com/guide/en/contributing/signing-up
37 YouTube Help. "Write notes on videos" Google. https://support.google.com/youtube/answer/14925346?hl=en-GB#zippy=%2Cwho-has-access-to-write-notes
38 Transparency Centre. "Community Notes: A New Way to Add Context to Posts" Meta. https://transparency.meta.com/en-gb/features/community-notes/

There are a number of notable differences between the social media companies' eligibility criteria. For example, on X and YouTube, contributors can be under 18. Furthermore, on X, contributors can be based anywhere in the world. Whilst this same flexibility is permitted on YouTube, the trial of Notes is limited to the US and therefore is likely to, like Meta's trial, be limited to just US contributors. This means that X is the sole social media company where Community Notes is both operational in the UK and where contributors can originate both from the UK and/or anywhere else in the world.

What do we know about the Community Notes membership?

Little is known about who actually makes up the Community Notes' community on X or Meta.[39,40] Contributors are not asked to confirm any identifying characteristics, such as gender, location or political view. It is uncertain whether the social media companies collect data on who contributors to Community Notes are, beyond the criteria outlined above and the data that is used to inform how the account is categorised by the bridging algorithm. This ambiguity makes it challenging to assess to what extent those contributing Community Notes reflect the social media platform's broader community or indeed the country in which that platform is being used.

What do Community members write Notes about?

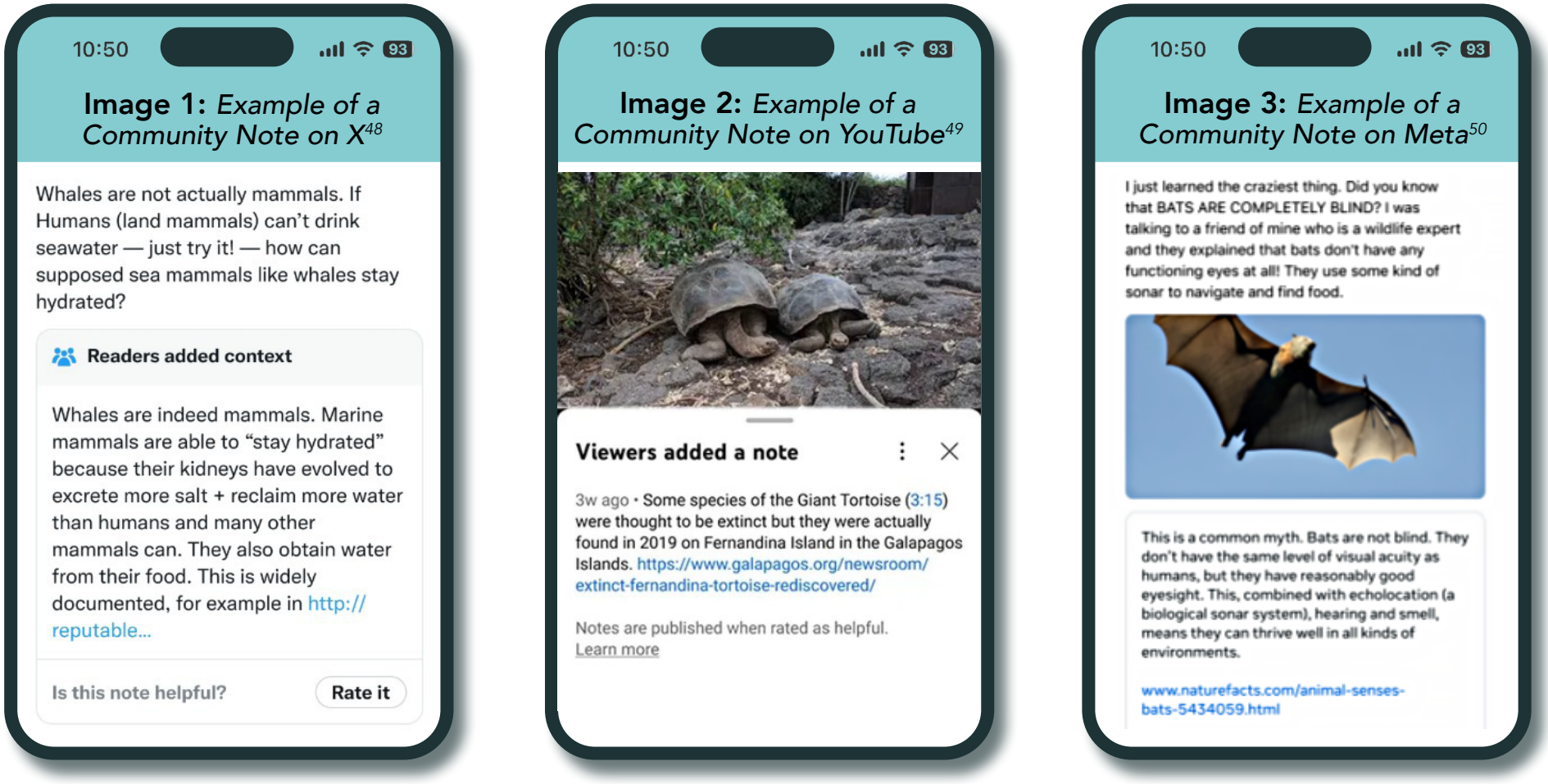
The fundamental aim of the Community Notes model is to allow contributors to provide context to misleading information. The Notes system on X, for example, describes their aim as 'to create a better informed world by empowering people on X to collaboratively add context to potentially misleading posts.'[41] Contributors on X can also categorise the post they are addressing using attributions such as 'manipulated media', 'factual error', 'unverified claim as fact' and 'missing important context'. Similarly regarding Meta, on its social media platforms Facebook, Instagram and Threads, it suggests Community Notes can 'add more context to posts that are confusing or potentially misleading'.[42] Furthermore, on YouTube, its aim is to 'provide relevant, timely, and easy-to-understand context' on videos.[43] It states that this could include clarifying when a song is a parody or when old footage is being presented as a current event.[44] The explicitly stated objectives above, though broad in their purpose, do not suggest social media companies anticipate using Community Notes as the moderation mechanism for violent or threatening content.

Interestingly, despite the intention to counter-balance concerns of perceived bias among professional fact-checkers, the provision of further context by community members appears to remain largely dependent on traditional fact-checking. Research by the Spanish fact-checking organisation Maldita found fact-checking organisations were the third most cited source globally when someone proposed a Community Note, behind X and Wikipedia.[45] The inclusion of references to fact-checking organisations' sources was also found to increase the effectiveness of the Note. Maldita found that whilst only 8.3% of proposed Community Notes are made visible on X, this rose to 12% when the Note cited a verification organisation, and to 15.2% when citing European fact-checkers.[46] Additionally, such Notes that did include a verified fact-check were found to be agreed upon by Community Notes members and therefore became visible 24 minutes quicker than general Notes that did not include a source from a fact-checking

39 The identity of contributors on both platforms is kept anonymous under an alias associated with a record of their previous contributions
40 Transparency Centre. "Community Notes: A New Way to Add Context to Posts" Meta. Available at: https://transparency.meta.com/en-gb/features/community-notes/; Community Notes Guide "Additional Review" X. Available at: https://communitynotes.x.com/guide/en/contributing/additional-review
41 X Help Centre "About Community Notes on X" X. Available at: https://help.x.com/en/using-x/community-notes
42 Meta "Introducing Community Notes" Meta. Available at: https://about.meta.com/technologies/community-notes/
43 The YouTube Team (2024) "Testing new ways to offer viewers more context and information on videos" YouTube. Available at: https://blog.YouTube/news-and-events/new-ways-to-offer-viewers-more-context/
44 Ibid
45 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/
46 Ibid.

organisation.[47] This suggests that whilst Community Notes draws on the time and enthusiasm of volunteers, its most successful operations depend on the work of professional fact-checking operations.

FIGURE 1
IMAGE-BASED EXAMPLES OF COMMUNITY NOTES



47 Ibid.
48 "Introduction" Community Notes Guide. Available at: https://communitynotes.x.com/guide/en/about/introduction
49 The YouTube Team (2024) "Testing new ways to offer viewers more context and information on videos" YouTube. Available at: https://blog.YouTube/news-and-events/new-ways-to-offer-viewers-more-context/
50 Meta (2025) "Community Notes: A New Way to Add Context to Posts" Transparency Centre. Available at: https://transparency.meta.com/en-gb/features/community-notes/

What decides if a Community Note about a post is ultimately published on a social media platform and becomes visible to public viewers of the post?

For all social media platforms, once an initial Note is posted, it is not yet public. Other opted-in Community Notes members must view the Note and rate it as 'helpful' or 'unhelpful'. Once the Note has been voted on, a 'bridging algorithm' decides which Notes are posted publicly. This is to ensure Notes reach 'cross-ideological agreement' with members of diverse perspectives reaching a consensus on the 'helpfulness' of the Note. Once a Note reaches a certain threshold of 'helpful' ratings, it is posted onto the social media platform attached to the original post for all users to view. If a Note is deemed 'unhelpful' that it remains unpublished. This 'bridging algorithm' is used by both YouTube and Meta.[51] This step-by-step process is summarised below in Table 2.

TABLE 2
STEP-BY-STEP OF WHAT HAPPENS TO THE POST AND COMMUNITY NOTE ON X WHEN A COMMUNITY NOTE IS FIRST CREATED ABOUT A POST

	Step 1	Step 2 - 'Needs More Ratings' (NMR)	Step 3 - 'Resolution'	
			Step 3A - 'Helpful' Resolution	Step 3B - 'Unhelpful' Resolution
Visible to any user	A post is posted and visible to all users	A post remains visible to all users.	A post that has received a 'Helpful' Note remains visible	A post that has received an 'Unhelpful Note' also remains visible
	No Note visible	No Note visible	Helpful Notes become visible	No Note is visible
Only visible to users registered to review Community Notes[52]	No Community Note created or associated with it	A Note is created and posted to a post. At this stage, this Note is visible only to registered users and is classified as 'Needs More Ratings' (NMR) Registered users can start voting on a CN to say if it is 'Helpful' or 'Not Helpful'	If a Note is given enough 'Helpful' ratings, then it is published alongside the post.	If a Note is rated 'Unhelpful' , then it is not published. It is not yet clear if users can continue rating the Note and change its status or if it's simply removed from view.

51 Meta "Introducing Community Notes" Meta. Available at: https://about.meta.com/technologies/community-notes/; The YouTube Team (2024) "Testing new ways to offer viewers more context and information on videos" YouTube. Available at: https://blog.YouTube/news-and-events/new-ways-to-offer-viewers-more-context/

52 Hutchinson, Andrew (2024) "X Improves the Speed of Community Notes Being Displayed" Social Media Today. Available at: https://www.socialmediatoday.com/news/x-formerly-twitter-adds-faster-community-notes/731416/

What is driving the increasing adoption of Community Notes as a moderation tool by US-headquartered social media platforms?

Politicised attitudes to professional fact-checking models in the US

Community Notes, when initially named Birdwatch, was first released as open source code to support social media platforms to mitigate what Twitter had identified at the time as an increasing amount of scepticism towards the neutrality and objectivity of existing professional fact-checking models. Research referenced by Birdwatch had identified a lack of trust among the American public in companies and the government's ability to moderate content.[53] In addition, US-based research found that existing fact-checking interventions were perceived differently depending on the user's political viewpoint which, in turn, had a knock-on impact on how useful or reliable they found the fact-check provided to be. For example, one study found that 70% of Republicans thought fact checkers favoured one side over the other.[54] In the face of these concerns, Twitter argued that the Birdwatch model was the most appropriate community-based solution to restore trust in fact-checking approaches to moderation, and to tackle the increasing political polarisation among US users in its perception.[55]

In January 2025, Meta's abandoning of independent fact-checking programmes and adoption of the Community Notes model was also explained by concerns surrounding bias and attributed to shifts in the government administration in the US.[56] In a video posted on Instagram, Zuckerberg explained that "recent elections… feel like a cultural tipping point" and that "factcheckers have just been too politically biased."[57]

Much like in the US, trust in mainstream media in the UK has also declined to its lowest levels in recent years.[58] However, there is limited evidence to suggest such sentiments are replicated towards fact-checking organisations or that there is a perception among the UK public that professional fact-checking interventions on social media are biased. The main fact-checking organisations in the UK, BBC Verify, Channel 4 FactCheck and Full Fact, all subscribe to the International Fact-Checking Networks' Code of Principles.[59] Furthermore, such organisations played a critical role during the Southport riots in effectively fact-checking the inaccurate and harmful claims circulating on social media when such platforms were struggling to mitigate the virality of the inaccurate claims circulating on their platforms.

Cost-saving factors associated with Community Notes

The Community Notes model, which relies on volunteers instead of paid staff, also enables social media companies to save on the moderation costs associated with employing in-house professional teams. While X maintained Community Notes (renamed from Birdwatch) following its acquisition in late 2022, it chose to cut 80% of its trust and safety engineers and 52% of its global content moderation team. This indicates there may be additional financial considerations

53 Stocking, Galen et al (2022) "The Role of Alternative Social Media in the News and Information Environment" Pew Research Centre. https://www.pewresearch.org/journalism/2022/10/06/the-role-of-alternative-social-media-in-the-news-and-information-environment/

54 Walker and Gottfried (2019) "Republicans far more likely than Democrats to say fact-checkers tend to favor one side". Pew Research https://www.pewresearch.org/short-reads/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/

55 Stefan Wojcik et al 'Birdwatch: Crowd wisdom and Bridging Algorithms can Inform Understanding and Reduce the spread of Misinformation.' Available at: https://arxiv.org/pdf/2210.15723

56 Meta (2025) "More Speech, Fewer Mistakes". https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/; BBC (2025) "Facebook and Instagram get rid of fact checkers". https://www.bbc.co.uk/news/articles/cly74mpy8klo Politico (2025) "Zuck goes full Musk, dumps Facebook fact-checking program". https://www.politico.eu/article/mark-zuckerberg-full-elon-musk-dump-facebook-fact-checker/

57 Zuckerberg (2025) Instagram post. https://www.instagram.com/reel/DEhf2uTJUs0/?igsh=dHVxbmdrbW9xMzg

58 Tobitt, Charlotte (2024) "Trust in UK media: UK drops to last place in Edelman survey of 28 nations" Press Gazette. https://pressgazette.co.uk/media-audience-and-business-data/trust-in-media-uk-edelman-barometer-2024/; Fletcher, Richard et al "Reuters Institute Digital News Report 2024" Reuters Institute. Available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-06/DNR%202024%20Final%20low-res-compressed.pdf

59 Edgington (2025) "Perceptions, power and polarisation: the political impact of UK fact-checking". Reuters Institute. https://reutersinstitute.politics.ox.ac.uk/perceptions-power-and-polarisation-political-impact-uk-fact-checking; Ofcom (2024) "Understanding misinformation: an exploration of UK adults' behaviour and attitudes." Ofcom.; Hawkins, A. (2016). International Factchecking Network Code of Principles. Full Fact https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/media-literacy-research/making-sense-of-media/dis-and-mis-information-research/mis-and-disinformation-report.pdf?v=386069

influencing the commitment to Community Notes as a method.[60]

What does this mean for the UK?

Whilst the research and key drivers for adopting a Community Notes model on US-headquartered social media platforms is likely to have been influenced by American socio-political and business model factors, the ramifications of these decisions impact the information diet of the UK public. The remainder of this report will focus on exploring the relative benefits and risks of alternative community-based moderation methods.

RESEARCH METHODOLOGY

Community Notes data collection

This study examines the dynamics of Community Notes by leveraging a comprehensive dataset obtained directly from the Community Notes website's Download Data page. The dataset encompasses every publicly available Community Note, including full text contributions, rating metrics, and complete status histories.

For this study, data were restricted to Notes generated in response to the Southport attack and riots, collected over a defined period from 29 July to 11 August 2024. This targeted approach enabled a focused analysis of Community Notes responses during the critical events that unfolded, including the attack on the morning of 29 July and the riots that followed. The data set was downloaded on 4 September 2024.

Search and filtering strategy

To isolate Notes discussing the Southport attack and subsequent riots, we developed an advanced search strategy based on composite keyword queries and regular expressions. This method was specifically designed to capture the various linguistic representations and descriptions associated with the events. By rigorously testing and refining the regular expressions, we account for terminology variations while maintaining precision in data collection, enabling us to comprehensively capture relevant Notes without collecting unrelated content. The complete list of search terms and patterns is provided in the Appendix ensuring transparency and reproducibility.

Post data integration and data preprocessing

After filtering notes specific to the Southport Riots, we used unique post identifiers (postIDs) associated with each Community Note to perform a secondary data collection phase, accessing and collecting the original post that received the notes in our dataset. For each post, we collected all relevant data and metadata, including creation timestamps, full engagement metrics including impression counts, the body of the post, its language, its status (live or removed), and information on the language used in the post. This integration established a complete timeline for each content piece and its corresponding Community Note, enabling a more complete analysis and understanding of the Community Notes dataset. Please see the Appendix for further detail on data pre-processing.

Analytical approach

The analytical framework employed in this study is grounded in quantitative methods designed to evaluate the performance and overall effectiveness of Community Notes during the Southport incidents. To understand the dynamics of consensus-building within the system, we first determined the proportional distribution of notes across distinct status categories, namely "Needs More Ratings," "Helpful," and "Not Helpful." This analysis provides a statistical

60 Austrailian esafety commissioner (2023) "Basic Online Safety Expectations" https://www.esafety.gov.au/sites/default/files/2024-01/Key-Findings-Basic-Online-Safety-Expectations-Summary-of-response-to-non-periodic-notice-issued-to-X-Corp.Twitter-in-June-2023.pdf

foundation for assessing how community evaluations reach a final status, and how common it is for notes to be rated as Helpful by the community.

We investigated temporal effectiveness by measuring two critical intervals. The first interval captures the delay between the original post's publication and the initial creation of a community note, while the second records the time required for notes to transition from the preliminary "Needs More Ratings" state to a definitive status. By applying rolling averages, we mitigated the influence of outliers and conducted a day-by-day temporal analysis to identify trends in system responsiveness throughout the incident period.

To assess the impact of Community Notes on engagement, we conducted a comparative analysis between posts accompanied by visible Community Notes and those that remained in the "Needs More Ratings" phase.

Lastly, our analysis extended to evaluating content persistence by examining the proportion of posts removed from the social media platform, with removal rates stratified by note status. This comprehensive approach enables a robust evaluation of how community-driven moderation processes perform under the pressures of a rapidly evolving, high-profile event.

Data coding and further quantitative analysis

All posts associated with the Community Notes dataset were further analysed and coded manually and inductively by Demos researchers. Posts were given descriptive codes that sought to achieve two things:

- To descriptively capture the posts' content in relation to (a) the potential threat or risk it might pose to individuals or communities; (b) its inaccuracy; and/or (c) risks it might pose to trust in societal institutions.
- To quantitatively and comparatively capture how explicit the language is and how confident the researcher was that such a post posed a threat or risk or that there existed proof that a claim was inaccurate.

While all data has been included in our analysis of the speed and visibility of Community Notes and their publication (discussed in Section 2), only data that met the highest tier of confidence has been included in sections that evaluate the threat posed to individuals and communities, inaccuracy and risks to trust in societal institutions (discussed in Section 3). All posts described in Section 3 have met the following criteria:

TABLE 3
THRESHOLD FOR POSTS TO BE INCLUDED IN OUR ANALYSIS IN SECTION 3.

Threat to individuals or communities	The threat was explicit and contained an incitement to violence and hate speech
Inaccurate	Researchers could find multiple, verified and reputable sources that proved a claim was inaccurate
Risks to trust in societal institutions	The claim explicitly stated that the government, news media or police were engaged in a deliberate cover-up

Please see the Appendix for more details on our coding process and the quantitative analytical framework used.

SECTION 2
RESULTS: THE EFFECTIVENESS OF COMMUNITY NOTES DURING THE SOUTHPORT RIOTS

In this section, we evaluate the different measures of effectiveness for Community Notes including the proportion that were visible during the riots, their speed of publication, their impact on the level of engagement with the posts themselves as well as other notable themes such as the level of removal of posts and the language of Notes themselves.

THE COMMUNITY NOTES DATASET

The Community Notes dataset comprises 673 Community Notes and 539 related unique posts addressing the Southport riots between Monday 29 July and Sunday 11 August 2024.[61,62] This demonstrates that more than one Note could be created about the same post irrespective as to whether one or more Note had already been created or published. Of the 539 posts present in the dataset, 166 posts have been removed since the events leaving a total of 373 live unique posts.[63] Of the posts that had been removed, we could analyse their ratings i.e. whether it had a Note that was Helpful or not, but not the post and Note itself or the levels of engagement with it.

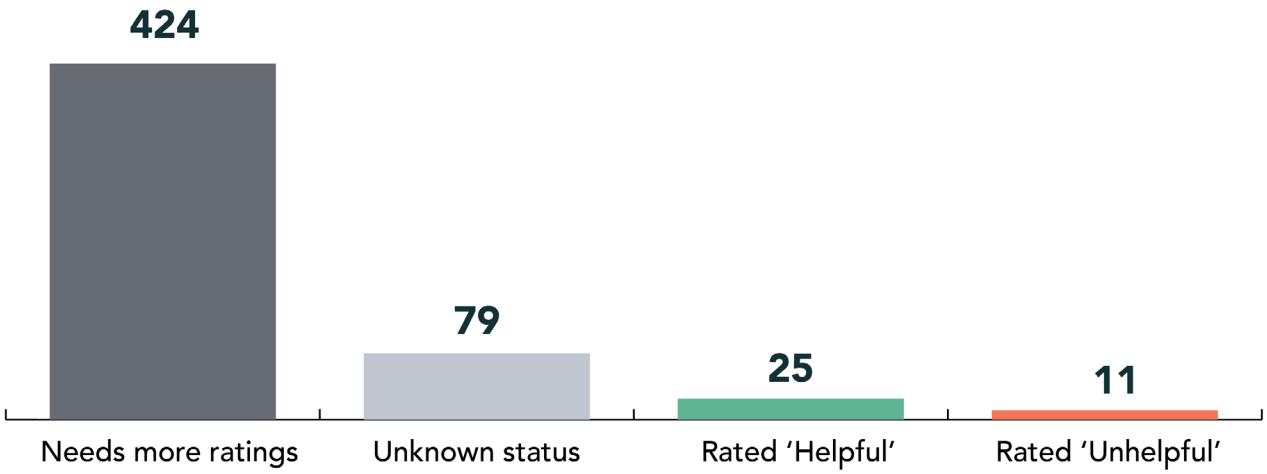
NOTES WERE LARGELY INVISIBLE DURING THE RIOTS, SO HAD NO CHANCE OF STEMMING THE TIDE OF FALSE RUMOURS

Just 4.6% (25 of all 539 posts) of posts had a visible Community Note, because only that small proportion achieved "Helpful" status and therefore could have been publicly visible during the period of the riots.[64,65] 78.7% of posts (424 of 539 posts) with Notes created during the riots

61 This means that some posts had more than one Community Note associated with it. However, the average number of Community Notes associated with each post is very close to 1.
62 Community Notes were filtered based on a string of keywords relating to the Southport riots.
63 There are also a number of posts in our dataset that are duplicates of other posts. Some of these duplicates have also been removed and so are described as 'removed', while others remained in our dataset.
64 In a small number of cases, some posts had more than one Note. Posts that had more than Note created about it and where one Note was rated 'Helpful' and another was rated 'NMR' have been counted as a post with a 'Helpful' Note. For example, if a post had one Community Note rated as 'Helpful' and 10 NMR Notes, only the Helpful Note has been counted. In posts counted as NMR, none of these posts had a Community Note that received a 'Helpful' rating.
65 By visible, we mean visible after it achieved 'Helpful' status, which could have been during the riots and in the period running up to when we downloaded the dataset on 4 September 2024. See section below on the length of time taken for a Note to achieve 'Helpful' status for further detail on when visibility might have been possible.

remained in the "Needs More Ratings" (NMR) status and therefore had no visible Note and a further 2% (11 of 539 posts) had Notes that were flagged as "Not Helpful" and so also remained invisible.

FIGURE 2
PROPORTION OF POSTS WITH COMMUNITY NOTES THAT WERE RATED 'HELPFUL' AND THEREFORE WERE VISIBLE TO THE PUBLIC (INCLUDING POSTS THAT WERE REMOVED)



The fact that just 25 Notes i.e. 4.6% of those created were published during the whole two-week period of the Southport riots demonstrates significant challenges with relying on Notes as a mechanism for moderation during fast-moving events.

Such a result is worse, i.e. a smaller proportion, than a comparable study of the efficacy of Community Notes conducted between 2022 and 2023 which used a much longer observation period and thus a much larger dataset and found that just 12% of Notes were displayed to users.[66]

COMMUNITY NOTES WERE TOO SLOW TO PREVENT FALSE AND HARMFUL INFORMATION GOING VIRAL

A central question of this project is to understand the speed at which Notes are successfully shown to the public under potentially misleading and/or threatening content.

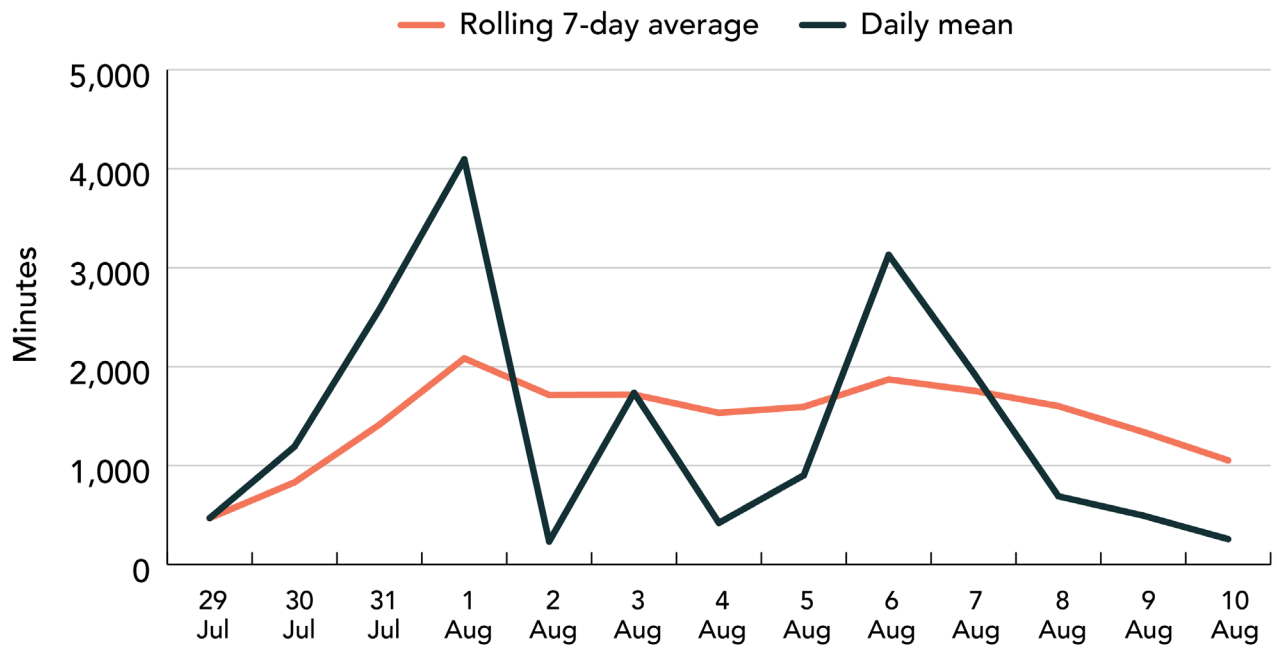
We found:

1. An initial Community member response lag: Community members took an average of 634 minutes (10.5 hours) to post the first Note to a post after a post had been created over the period of the riots.
2. Public visibility lag: Community members then required a further 1,325 minutes (22 hours) on average over the period of the riots to find sufficient consensus to transition a Note from Needs More Ratings to a definitive status (Helpful/Not Helpful) i.e. for the post/ Note become either visible or remain suspended and thus invisible.

66 Chuai, Y., Pilarski, M., Lenzini, G., & Pröllochs, N. (2024). "Community notes reduce the spread of misleading posts on X." https://doi.org/10.31219/osf.io/3a4fe

While the speed time for Notes to be created and consensus fluctuates considerably over the period of the riots, delays remained significant. The average daily resolution time for Community Notes on the day of the attack took 469 minutes (7.8 hours) rising to 1,193 minutes (19.8 hours) on 30th July - the day riots began.[67,68] For posts posted on 31 July, the day disorder broke out in Hartlepool, County Durham and Aldershot in Hampshire, Community Notes took 2,580 minutes (43 hours) to be created and made visible. These delays highlight systemic limitations in addressing violent or inaccurate content during time-sensitive crises, where viral content can solidify narratives long before corrective actions are initiated and implemented.

FIGURE 3
DAILY AVERAGE TIME (MINUTES) AND ROLLING 7-DAY AVERAGE BETWEEN A NOTE BEING CREATED AND ACHIEVING FIRST NON-NMR STATUS



POSTS RAPIDLY GAINED OVER 100 MILLION VIEWS IN THE FIRST 36-HOURS AFTER THE ATTACK

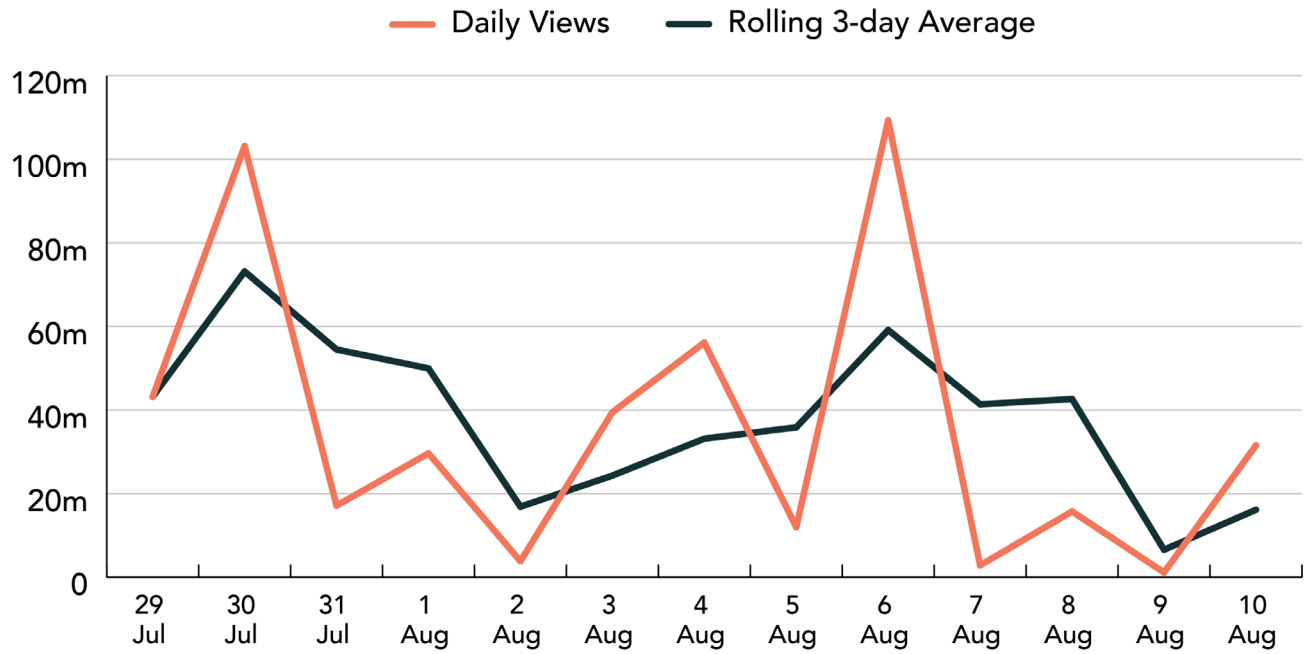
The volume of engagement with posts in the minutes and hours after they are posted demonstrates why the speed of mitigating actions is so important. Posts posted in the first 36-hours received the highest engagement i.e. between 29 July and 30 July: 146,302,406 views - 31% of the total views accumulated by associated posts posted over the whole 14 day period. After this date and until Sunday 11 August 2024, there was a steady decline in engagement across the dataset as a whole, consistent with typical crisis event patterns.[69] The following chart that combines the daily views with a rolling 3-day average illustrates the volatility of the sheer and sudden spikes in views, particularly in the first 36 hours and on the 6th August, interspersed with sudden declines. This demonstrates the importance of moderation methods that can spring into action fast in order to have a chance of being effective in the minutes and hours that follow a post being posted.

67 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. Resolution is calculated as the difference between 'noteCreatedAtTime' and timestamp 'TimeOfFirstNonNMRStatus' which are two data entries we have in the status dataset.

68 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/

69 Yury Kryvasheyeu et al. (2016) "Rapid assessment of disaster damage using social media activity." Sci. Adv.2,e1500779(2016). DOI:10.1126/sciadv.1500779

FIGURE 4
DAILY VIEWS AND ROLLING 3-DAY-AVERAGE DAILY VIEWS ACROSS THE WHOLE DATASET BETWEEN MONDAY 29 JULY TO SUNDAY 11 AUGUST 2024

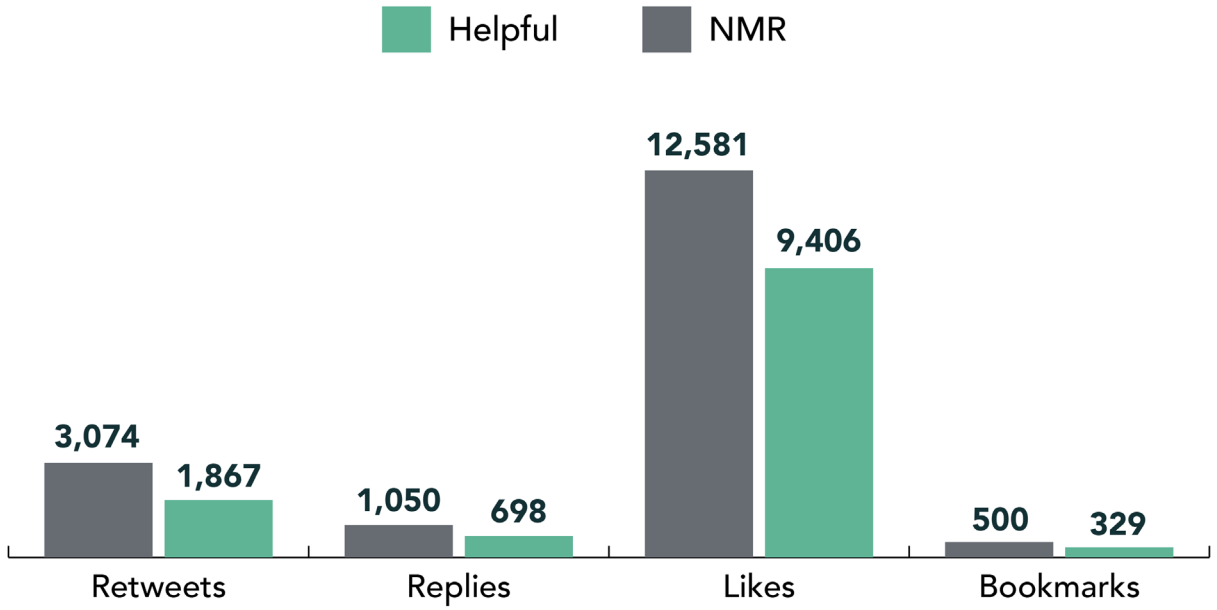


Overall engagement with the posts associated with the Community Notes dataset was very high. The analysed posts obtained a total **465 million views** (averaging over 1 million views per live post), 3.9 million likes and 0.96million reposts, reflecting the viral nature of the event.[70,71] However, 99% of the total views of posts across the whole dataset (445 million) reflect views of posts without visible Community Notes.[72] Posts without a visible Community Note that have been classified as both inaccurate *and* threatening to communities amassed 67.6 million views.

A critical component of this analysis examined engagement metrics between posts with visible Community Notes (**25 notes**) and those left without a visible Note to the public (**556 notes that had NMR status**).[73] The findings reflect significant differences in engagement at the time of measurement.[74] Posts with visible (i.e. 'Helpful') Community Notes showed consistently lower engagement across all key metrics relative to NMR posts. For example, NMR posts (i.e. those without a visible Community Note) received on average 1,487,620 views compared to posts with visible 'Helpful' Community Notes which received an average of 645,454 views (57% lower on average). The following chart also highlights the differences in other engagement metrics including reposts, replies and likes, (39% lower, 33% lower, 25% lower and 34% lower on average respectively).[75]

70 'Views' and 'impressions' are used interchangeably throughout our reporting. Most platforms call views 'impressions'. It's not known precisely how views or impressions are calculated on X, but it is normally assumed that 'an impression' refers to the number of times a post appeared in a user's timeline. View calculations are based on when the data was collected on September 2024.
71 Non-rounded numbers include: total views, 464,971,212 views; 959,058 reposts, 333,226 replies, 3,978,286 likes, and 156,098 bookmarks.
72 The 465m views is based on 373 posts and therefore does not include the impressions of posts that were removed 30.80% of posts (166 of 539) prior to our access to the dataset
73 This comparison used static, cross-sectional data. It is also important to note that the sample of posts with Helpful notes is very small (25 Notes), and this should be kept in mind when interpreting these findings.
74 4 September 2024
75 Please note these metrics do not include those corresponding to the posts that had subsequently been removed prior to our download of the dataset.

FIGURE 5
DIFFERENCES IN AVERAGE ENGAGEMENT METRICS PER POST BETWEEN THOSE THAT HAD A VISIBLE COMMUNITY NOTE (I.E. HAD RECEIVED A 'HELPFUL' RATING) COMPARED TO POSTS WITHOUT A COMMUNITY NOTE (I.E. THAT HAD 'NEEDS MORE RATINGS' (NMR) STATUS)



It is also useful to understand the maximum and minimum ranges of engagement for each post that can be disguised by average figures. Posts without visible Notes (NMR) achieved a far higher maximum number of Likes, with ranges extending **2-3x beyond posts with visible Notes.** This pattern holds across all metrics. The effect is particularly visible for replies—a metric often correlated with polarised debate—where posts with a visible note received a maximum of ~2,500 replies, while posts without a visible Note received up to 41,000 replies.[76] These findings align with prior research on the efficacy of Community Notes that used a different methodology using time series analysis and with a much larger dataset than has been analysed here.[77] However, as noted earlier, delays in Notes being published undermines this potential, particularly during critical early phases of virality.

Beyond raw metrics, the data might suggest a positive impact of visible Community Notes when they are eventually made public in terms of reducing engagement with posts that could be misleading or including factual errors. However, this pattern may also reflect correlation rather than causation. Given that such a significant proportion of Notes that didn't manage to achieve a Helpful or Unhelpful rating, posts that have achieved a 'Helpful' Community Note status are potentially more likely to be easily identified as inaccurate and therefore attract less engagement because users, like community members, can themselves discern its lack of reliability.

WHEN POSTS HAVE BEEN REMOVED, IT IS UNCLEAR WHO MADE THE DECISION

Across the entire dataset, **30.80% of posts (166 of 539) were removed**. Of the 166 posts associated that have been removed prior to our research being conducted, it is unclear *who* removed these posts. They are simply labelled 'removed' rather than 'removed by original poster' or 'removed by platform team'.

76 For studies highlighting the correlation between the volume of replies and the polarisation of the topic, see: Garimella et al (2017) "The Effect of Collective Attention on Controversial Debates on Social Media". arXiv:1705.05908;
77 Chuai, Y., Pilarski, M., Lenzini, G., & Pröllochs, N. (2024). "Community notes reduce the spread of misleading posts on X." https://doi.org/10.31219/osf.io/3a4fe

Removal rates varied significantly based on Note status: posts with **Helpful notes** saw a 44% removal rate (11 of 25), while those with **NMR notes** had a removal rate of 29.20% (124 of 424). Notably, none of the posts flagged with "Not Helpful" notes (0 of 11) were removed. This disparity suggests that users may be more inclined to delete posts after receiving a visible Community Note, potentially reflecting self-correction or discomfort with public moderation. However, this is speculative, as removal reasons (e.g., user deletion, social media platform enforcement, account suspensions) are not labelled in the dataset and so remain unknown to researchers.

NOTES APPEAR IN DIFFERENT LANGUAGES, PARTICULARLY SPANISH AND FRENCH

English dominated the language of posts though it was possible to identify Spanish (36 posts) and French language too (28 posts) demonstrating the international engagement with the crisis. It was notable that Community Notes would also be written in their corresponding language. For example, one Helpful Community Note included a reference to a corresponding Spanish-language newspaper that included the correct information and debunked the false claim in the original Spanish language post. The presence of foreign language Notes demonstrates the utility of enabling participation of an international community in the Community Notes user base, especially when debunking claims made in foreign languages about UK-based events that are garnering international attention and generating further false rumours overseas.

SUMMARY

In this section, we have shown that Community Notes failed at a time when strong moderation was most needed to prevent the spread of false information leading to large-scale offline violent disorder. Due to the context of a highly polarised, violent event, it was not possible for community members to agree on the Notes that needed to be published quickly to counter the spread of harmful content during a crisis. Just 4.6% of Community Notes created during the riots were visible to the public. On the day the riots began (30th July), the average resolution for a Community Note was 1,193 minutes (19.8 hours). This indicates that the Community Notes system was significantly hampered by challenges which are intrinsic to both the way that the Notes model is designed (to find consensus) and core to the patterns of a crisis situation that involves violent disagreement. These challenges will repeat in future crisis situations and demonstrate that such a model for moderation is simply unfit for preventing violent disorder from escalating.

Over 90% of the Community Notes that did achieve 'helpful' status (and were therefore made visible to users) referenced professionally fact-checked content from sources such as Sky, the BBC and CNN. Whilst this datapoint is based on a small sample of available posts and Notes, because such a small proportion achieved Helpful status and remain available for analysis (10), this indicates that where Community Notes did help correct false information on X during the Southport riots, the dependency on referencing professionally fact-checked sources appears to have been key. This demonstrates the importance of maintaining and financially sustaining access to fact-checking work to facilitate the Community Notes moderation model during crises.

SECTION 3
RESULTS: THREATS TO COMMUNITIES AND INDIVIDUALS

Robust links have been drawn between a range of social media channels and the offline violence that followed the Southport attacks.[78] This section provides more insight into the threats posed by posts remaining on X specifically, most of which, as highlighted in the previous section, lacked visible Community Notes because they had not achieved resolution.

We demonstrate first that Community Notes were not just created in relation to misleading posts, such as the false claims about the attacker's identity, the attack and its causes. Community Notes were also created about posts that built from those inaccurate and harmful assumptions and constituted racial and religious hatred as well as hatred towards migrants and asylum seekers.

We then go on to highlight the threats posed by certain posts to individual members of the public as well as politicians. We also highlight the complex relationship between harassment and abuse of individual political figures and critiques of democratic institutions, including the government.

FALSE POSTS WITHOUT NOTES THREATENED MARGINALISED COMMUNITIES

The targets of the posts that threatened communities overlapped heavily with the subjects of the violent riots that broke out across the country in the first 36-hours after the attack. In July and August 2024, rioters falsely described the attacker as a Muslim, asylum seeker, immigrant or 'foreigner' and physically targeted mosques as well as hotels where asylum seekers. 20.2% of the dataset (109 posts) was found to be explicitly threatening to these communities.

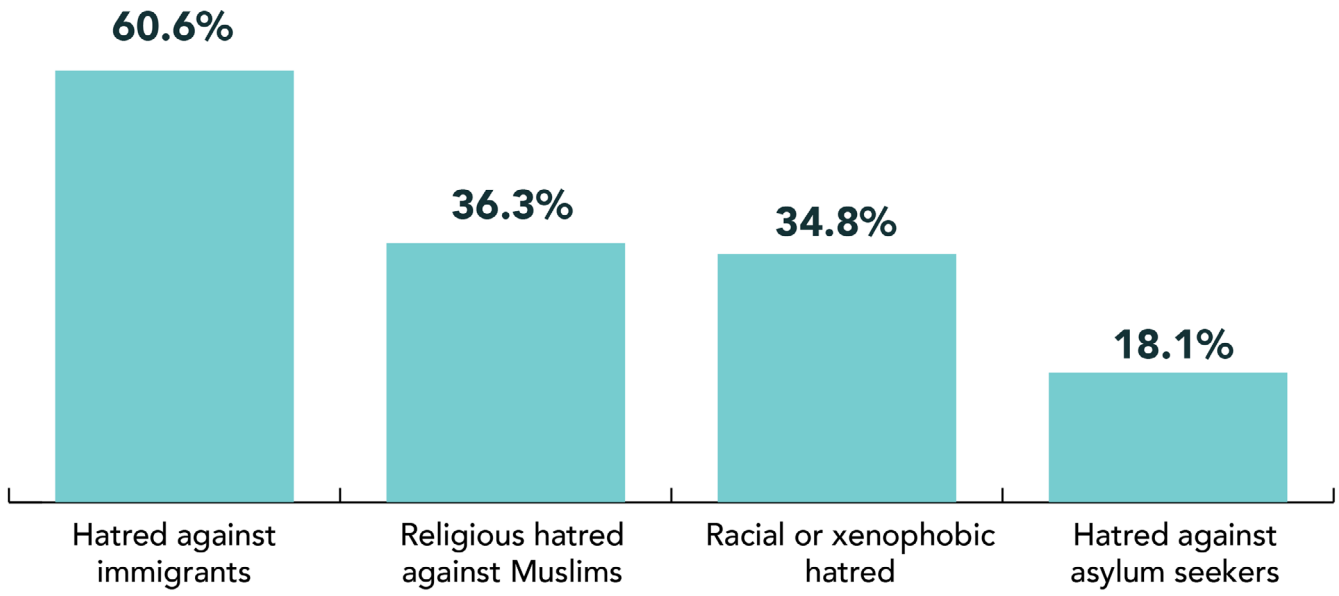
78 Spring (2024, 31 July) "Did social media fan the flames of riot in Southport?" BBC. https://www.bbc.com/news/articles/cd1e8d7llg9o. amp; ISD (2023) "From rumour to riots: How online misinformation fuelled violence in the aftermath of the Southport attack". ISD Digtial Dispatches. https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/ ; Hope Not Hate (2024) "The Far Right and the Southport Riot: What We Know So Far." Hope Not Hate. https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/

For example, we identified:

- Posts that confirm the attacker was a Muslim (one post received 1,503,953 views) or an illegal immigrant who had arrived in the UK on a boat (one post received 1,296,980 views) - both false claims that have been debunked.
- Posts that claim that attack was racially motivated (one post received 21,086,519 views)
- Posts that call for the permanent removal of Islam from the UK (one post received 1,079,500 views)

As illustrated in Figure 6, over half (58.7% or 64) of the posts included in this subset of the dataset, associated with a Community Note, reflected explicit hatred towards 'immigrants' specifically and a further fifth (18% or 20 of the posts) reflected explicit hatred against 'asylum seekers' specifically.[79] These have been distinguished from posts that refer to 'foreigners' or people from Rwanda which fall into the 'racial or xenophobic hatred' category.

FIGURE 6
PERCENTAGE OF THE THREATS TO COMMUNITIES BROKEN DOWN BY THE TARGETS AND NATURE OF THOSE THREATS



36.7% of the dataset or 40 posts include explicit threats to Muslims which can be classified as religious hatred. Such posts included, for example, claims that rely on the discriminatory stereotype that Muslims are intrinsically violent or terrorists, that Muslims are a threat to children, that the white community is under attack by Muslims, that use derogatory slurs or dehumanising language towards the Muslim community or that call for the eradication or expulsion of Islam from the UK.

We also found 34 posts that contain threats towards 'foreigners', people of African heritage or specifically from Rwanda which have been classified as 'racial or xenophobic hatred'. Examples of such posts include claims that people of African heritage are all criminals or violent. These latter themes are likely to be associated with the false assumption made by some rioters that

79 Data points included here have met the highest confidence rating from our research team (Tier 3), where the post presented an explicit threat and/or could be proven to be inaccurate using the evidence available.

the attacker was not born in the UK and was actually from Rwanda. These posts have been categorised as 'racial or xenophobic hatred'.

In some instances, the speech we categorised as racial, religious or xenophobic hatred and hatred against migrants included rhetoric used by prominent politicians, particularly during debates on boat migration in Parliament and the Safety of Rwanda Bill. For example, in the context of false assumptions regarding the attacker's identity, we found a number of derogatory memes depicting racist, stereotypical images of migrants and refugees arriving via boat and wider calls replicating the political slogan frequently affixed to the former Prime Minister, Rishi Sunak's podium, "stop the boats".[80]

Is this content allowed under social media platforms' Terms of Service or Community Guidelines?

Even in the context of social media platforms' Terms of Service, it is difficult to decipher if the types of content described above ought to have been removed according to the social media platforms' own rules.

Some social media platforms only consider communities with protected characteristics in the context of their Hate policies which would exclude migrants and asylum seekers. For example, X's policy states that you may not directly attack other people on the basis of race, ethnicity, national origin and religious affiliation'.[81] It does not include immigration status here. On this basis, X might permit hate towards migrants and asylum seekers, but not to Muslims. Whereas Meta, whilst stressing that users can critique immigration policy, does make provisions for 'refugees, migrants, immigrants and asylum seekers from the most severe attacks' which include 'allegations of serious immorality and criminality', including 'violent criminals (including but not limited to: terrorists, murderers)'.[82]

Second, some social media platforms make exceptions for certain hateful speech depending on the context. For example, on X, hateful references, incitement, slurs and tropes and dehumanisation are all considered violations of the policy. Yet the policy also suggests that X would not necessarily remove such content, and only 'limit its reach' if the moderator deems the post to use 'coded language' or if the 'target is unclear'. Such language is difficult to interpret in this context.[83] Furthermore, in its violent content policy, X would also take 'proportionate action' based on the 'severity and likelihood of harm' and that in certain cases content could remain online, 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 a crisis context, like being in the aftermath of a violent attack, certain speech is more likely to be allowed than it being constituted a higher risk.[84]

In contrast to X, Meta's policy in the UK appears to assess risk based on such a context. For example, it recognises that misinformation can be dangerous 'in a specific context' and so would remove misinformation if 'expert partners have determined are likely to directly contribute to a risk of imminent violence or physical harm to people'.[85] These examples demonstrate the contrast between what speech is acceptable on social media platforms during crises like the aftermath of an attack and the ambiguity in how such Terms of Service or Community Guidelines might be applied.

80 Doherty (2023) "'Stop the boats': Suank's anti-asylum slogan echoes Australia's harsh policy." https://www.theguardian.com/uk-news/2023/mar/08/stop-the-boats-sunaks-anti-asylum-slogan-echoes-australia-harsh-policy
81 X, (2023) "Hateful Conduct". https://help.x.com/en/rules-and-policies/hateful-conduct-policy
82 Meta, (2025) 'Hateful Conduct' https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/
83 X, (2023) "Hateful Conduct". https://help.x.com/en/rules-and-policies/hateful-conduct-policy
84 X (2025) "Violent Content." https://help.x.com/en/rules-and-policies/violent-content
85 Meta (2025) "Misinformation" https://transparency.meta.com/en-gb/policies/community-standards/misinformation/

Is this content allowed under the Online Safety Act and Ofcom's current guidance?[86]

While the Online Safety Act was not in force during the riots, based on Sections 72(3), Ofcom is now able to enforce against the illegal harms duties. However, the additional duties for Category 1 services, such as X, are not yet in force and are unlikely to be until 2026 at the earliest.[87] An official letter that followed the Southport riots indicated that Ofcom would be 'concerned if services do not have clear and consistent terms of service prohibiting illegal hateful content.'[88] However, based on the persistent ambiguity apparent within such social media platform policies surrounding what speech they do permit, it is already clear how challenging enforcement will be for Ofcom, particularly when the Online Safety Act also does not allow Ofcom to set any minimum standards for Terms of Service.

However, through its Illegal Harms Codes and Risks Assessment Guidance, Ofcom is now able to mandate social media platforms to assess and mitigate their risks in relation to 'priority illegal content'. Notably, this includes religious and racial hatred, but does not include hatred towards migrants and asylum seekers.[89] This is based on the offences listed in Schedule 7 of the Online Safety Act that are linked to the Public Order Act 1986. This means that Ofcom can mandate that social media platforms have systems in place to reduce the risk that users are exposed to posts inciting religious or racial hatred, but not to posts inciting hatred directed towards migrants and asylum seekers.

Furthermore, while the Codes require social media platforms to have systems or processes to take down illegal content and to prioritise that which is viral, they do not - as is the case in the draft Protection of Children codes - mandate that algorithms are changed to downrank such content on recommender feeds.[90] By design, such algorithms have a significant impact on the proportion of people who will view the content it recommends and therefore disabling such a system in the context of illegal content would significantly reduce its exposure and thus impact.

FALSE POSTS ALSO PRESENTED A THREAT TO POLITICIANS AND INDIVIDUAL ORDINARY PEOPLE

Within the dataset, 3.34% (17 posts) could be considered threatening towards individuals particularly in the context of such violence offline. Such posts could be broken down into that which posed a threat to members of the public compared to public figures such as national politicians. Of threats to individuals, 82.35% (14) reflected a threat to politicians whereas 17.64% (3) posed a threat to an every-day member of the public.

Such posts included images of politicians and stating that they are to blame for the attacks, that they have the blood of British children on their hands, and that they had fought for people like the alleged attacker to gain access to the country (based on the false assumption that the attacker was an immigrant). Such posts can be argued as political opinion and critique that is allowed in the UK, but can also be viewed as threatening when they multiply and are shared by those who are also behaving violently.

Among politicians, the Prime Minister and Secretary of State for the Home Office received the

86 It was not possible to include a thorough analysis of the recent proposals made by Ofcom for Additional Safety Measures published on 30 June prior to publication of this report. However, we hope the evidence shared in this research paper will be useful for those preparing their submissions.
87 Ofcom (2025) 'Ofcom's approach to implementing the Online Safety Act.' https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/roadmap-to-regulation
88 Ofcom (2024) 'Letter from Dame Melanie Dawes to the Secretary of State, 22 October 2024.' 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 Ofcom (2024) "Protecting people from illegal harms online: Register of Risks". https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/register-of-risks.pdf?v=390983
90 https://assets.publishing.service.gov.uk/media/680a04f7532adcaaab3a2718/FINAL_-_Protection_of_Children_Code_of_Practice_for_user-to-user_services__2025_Parli_AC.pdf

greatest share, 47% and 29% respectively. However, the inclusion of posts that relate to the Co-Leader of the Green Party and the Member for Hackney North and Stoke Newington is striking and relates to the former's position in favour of immigration and the latter being a frequent target of racial and misogynistic abuse on social media platforms, irrespective as to whether the policy topic in question bears any relation to the MP's politics.[91]

Is this content allowed under social media platforms' Terms of Service?

Under X's Terms of Service, targeting of and threats towards individuals are prohibited under the Abuse and Harassment policy and may additionally violate X's policies against Hateful Conduct and Violent Content depending on the nature of the post.[92] For example, if a post calls others to target people with abuse or harassment online or behaviour that urges offline action such as physical harassment, then that would break X's Terms of Service.

However, the exception in X's Abuse and Harassment policy is for posts deemed to be a "critique of institutions, practices and ideas" because it is considered "a fundamental part of the freedom of expression".[93] As a result, a post that attacks an individual public figure could be considered acceptable by X because it could be treated as critical political commentary. This demonstrates the higher threshold of threat a post must meet in order for it to be removed if discussing a public figure.

Similarly under Meta's Community Guidelines for Bullying and Harassment, public figures who meet certain criteria definitively receive fewer protections than ordinary users or indeed those with a smaller public profile.[94] Meta similarly permits attacks and harassment of prominent public figures so long as it is not considered 'severe' or contains death threats that such figures are 'purposefully exposed' to. All other threats to national government officials are permitted. Similarly, Meta's requirement that threats be "credible" or that national public officials are purposefully exposed creates a higher bar for removal.

These Terms of Service suggest that X and Meta all leave significant discretion to their employed moderators regarding what posts meet the threshold of removal. Specifically, X and Meta's moderation teams must consider if a post should be considered too severe or is simply a "critique" rather than a threat.

Is this content allowed under the Online Safety Act and Ofcom's current guidance?[95]

The Online Safety Act does not make specific provision for politicians and simply refers to 'users' who should be protected. Politicians are highlighted by Ofcom as potential victims of three potential priority offences listed in the Online Safety Act, including being victims of harassment, stalking, threats and abuse, foreign interference offences and false communications, but do not indicate any special provision or threshold that they too must meet.[96] This absence of clarification as to in what ways it is acceptable for social media platforms to be facilitating these different thresholds for abuse and harassment creates significant unhelpful ambiguity.

91 Amnesty Global Insights (2017) Unsocial Media: Tracking Twitter Abuse among Female MPs. https://medium.com/@AmnestyInsights/unsocial-media-tracking-twitter-abuse-against-women-mps-fc28aeca498a

92 X Help Center (2024) Abuse and Harassment. https://help.x.com/en/rules-and-policies/abusive-behavior; X Help Center (2023) Hateful Conduct. https://help.x.com/en/rules-and-policies/hateful-conduct-policy; X Help Center (2025) Violent Content. https://help.x.com/en/rules-and-policies/violent-content

93 X Help Center (2024) Abuse and Harassment. https://help.x.com/en/rules-and-policies/abusive-behavior

94 Meta (2025) Bullying and Harassment https://transparency.meta.com/en-gb/policies/community-standards/bullying-harassment/

95 It was not possible to include a thorough analysis of the recent proposals made by Ofcom for Additional Safety Measures published on 30 June prior to publication of this report. However, we hope the evidence shared in this research paper will be useful for those preparing their submissions.

96 Ofcom (2024) "Protecting people from illegal harms online. Register of risks." https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/register-of-risks.pdf?v=390983

SUMMARY

In this section, we have highlighted results that demonstrate not just weaknesses in the Community Notes system, but also failings in other forms of moderation which should be protecting communities from violent content on social media platforms. We have shown that harmful content is appearing in the Community Notes dataset demonstrating the failure of professional in-house moderators to remove such content. While provision is made for protections from racial and religious hatred in the Online Safety Act, Illegal Harms duties and protections had not yet come into force at the time of the riots. It is clear that the moderation systems in place at the time for removing such content did not respond fast enough and therefore will require rapid strengthening by social media platforms in order to meet these new duties.

However, there also remain clear gaps in protections in Ofcom's Illegal Harms Codes guidance from the types of online hate that also caused significant harm during the riots, that which was directed towards migrants and asylum seekers. These gaps are reflected also in a lack of protections for these communities within some social media platforms' Terms of Service, such as X's.

Finally, it is clear that there remain significant ambiguities in social media platforms' Terms of Service which could also be used to facilitate certain dangerous content remaining online and which can send confusing or, at worse, permissive signals for such speech to users. For example, terms that make exceptions for violent or hateful content in the aftermath of a violent attack and/or when it is targeted at a politician. In the next section, we will discuss recommendations for next steps.

SECTION 4
RESEARCH IMPLICATIONS AND FUTURE RESEARCH DIRECTIONS

Our research sheds new light on the vulnerabilities of the Community Notes moderation system in an information crisis context. It demonstrates why community-based moderation systems cannot be relied upon as a replacement for or an alternative to historical moderation mechanisms such as independent fact-checking programmes or professional, in-house moderation teams as appears to be the current approach taken by US headquartered social media platforms. Instead, Community Notes must be used in tandem and as a complementary approach. This research demonstrates that pervasive weaknesses in moderation systems remain, enabling ongoing risks to marginalised communities in the UK in a crisis context- as was clearly the case during the Southport riots.

Given the reality that such an incident and resulting information crises will occur again in the UK and while systems of social media platform moderation remain materially unchanged, all parties must urgently review how to mitigate the risks of information disorder exacerbating the threats to marginalised communities.

Community Notes remains a useful *additional* tool within a suite of other moderation mechanisms, particularly alongside independent fact-checking and in-house professional moderation teams, but cannot be relied upon in a crisis. However, further research and evidence is needed to understand its full potential as a moderation mechanism before it can be relied upon as an effective moderation method.

Our policy recommendations will be provided in a standalone report. However, the following recommendations highlight opportunities for future research on Community Notes:

1) 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.
2) 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.

3) 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.[97] 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.
4) Fund research to evaluate user perceptions of community-based models of moderation in the UK context	Given that the existing research used by platforms to demonstrate the need for a community-based model of moderation has relied solely on attitudes data in the US context, it is important to evaluate if such hypotheses stand in the UK.[98] DSIT and/or Ofcom should commission research to evaluate and explore the drivers of trust in different models of moderation for misinformation, including a comparison between community-based models and those led by credited fact-checking organisations. Options that blend the two i.e. where community-based models relied on fact-checked sources should also be tested.
5) Evaluate the Community Notes model in different contexts and on different topics, including in relation to an issue or news story that is not high profile	This research has demonstrated how the context in which Community Notes is used has an impact on our understanding of its efficacy, particularly on speed of publication. This research has focused on evaluating Community Notes in the context of a high profile news event both on and offline when there was a high level of awareness among platform users and a high level of journalist and fact-checking activity focused on evaluating the facts of the issue. It would be prudent to evaluate the Community Notes model in additional similar contexts as well as entirely different contexts, when, for example, the topic is not as high profile and when Community Notes may take even longer to achieve a sufficient number of ratings from community members due to a lack of awareness of or interest in the topic.

97 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/

98 Walker and Gottfried (2019) "Republicans far more likely than Democrats to say fact-checkers tend to favor one side". Pew Research https://www.pewresearch.org/short-reads/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/

6) Evaluate the efficacy of Community Notes using alternative methodologies including measuring engagement with posts before and after appended Community Notes have been posted.	This research has drawn its findings on the efficacy of Community Notes based on the proportion and speed of Community Notes that were (or were not) made publicly visible in relation to misleading and/or harmful posts. We also compared engagement levels with posts with and without Notes. However, there are additional ways in which the impact of Community Notes can also be assessed - for example - by measuring the level of engagement with the same post both before and after a Community Note has been published. This approach is useful for assessing the impact of Community Notes on slowing the level of visibility of misleading posts, if and when a sufficient number are able to reach publication.

CONCLUSION

Overall, this paper has provided the first evaluation of the efficacy of Community Notes during the Southport riots - a case study of its effectiveness in a crisis situation. It has demonstrated that Community Notes failed to mitigate the false information that fuelled the violent disorder. As such, it offers strong evidence that the system is not fit for purpose in crisis situations.

Whilst Community Notes offers a fresh solution to moderating misleading posts in every-day contexts, its reliance on consensus makes it fundamentally unfit to respond fast enough in polarised situations. An insufficient number of Community Notes were made visible (4.6%) during the riots and those that were published took too long to have a chance at being impactful (1,193 minutes (19.8 hours average resolution time)). The system failed at a time when strong moderation was needed most: to prevent the spread of false information fuelling large-scale, offline, violent disorder. The fact that the Community Notes system failed *because* of the context of a highly polarised, violent context and *because* community members couldn't find agreement fast enough, demonstrates why Community Notes intrinsically cannot succeed in a crisis-setting and why additional, independent fact-checking and professional, in-house moderation teams are so crucial to keeping our society safe.

X has updated its Community Notes system since the research fieldwork period (in October 2024) stating that it is faster than was clearly the case during the Southport riots.[99] However, our results have demonstrated that speed of Notes creation and resolution is strongly dependent on the context in which it is operating and the topic in question. Our results demonstrate that the system does not work in polarised, crisis contexts and so reporting of 'what is possible' in best case scenarios is weak evidence for relying on such systems even in non-crisis contexts.

Furthermore, this paper also spotlights flaws in more traditional moderation systems, beyond Community Notes. Our results have demonstrated how the same Islamophobic, racist and xenophobic hate that fuelled the violent disorder have remained on the platform, amassing millions of views, despite constituting what we assess to be illegal, harmful content - that which should be removed by professional, in-house moderation teams. This exposes the broader weaknesses of the social media platforms' moderation systems whose ambiguous Terms of Service policies and recent cuts to their in-house teams in favour of a greater reliance on community-based moderation systems have allowed this content to proliferate and marginalised communities to suffer its consequences.

Now that the Online Safety Act is in force, platforms must ensure illegal harms content, particularly that which relates to religious and racial hatred, as well as hatred towards migrants and asylum seekers, is removed swiftly from users' feeds by professional moderation teams; and to ensure the risks intrinsic to crisis contexts are reflected in stronger and more consistent Terms of Service policies.

These results should sound the alarm for the Defending Democracy Taskforce, DSIT and Ofcom that our information supply chains need serious and urgent attention to not just prevent these spikes of violent disorder, but also the longer term corrosion of the fabric of our society and thus the health of our democracy.

99 X (2024) 'Introducing Lightening Notes' - https://x.com/CommunityNotes/status/1851337944822325253

APPENDIX

METHODOLOGY

Search and filtering

To isolate Notes discussing the Southport incidents, we developed an advanced search strategy based on composite keyword queries and regular expressions. This method was specifically designed to capture the various linguistic representations and descriptions associated with the events. By rigorously testing and refining the regular expressions, we account for terminology variations while maintaining precision in data collection, enabling us to comprehensively capture relevant Notes without collecting unrelated content. The complete list of search terms and patterns is provided below ensuring transparency and reproducibility.

"Southport", "Southport Riots", "Southport (civil unrest|disturbances)", "Southport (protests|demonstrations)", "Southport (clashes|confrontations)", "Southport (violence|violent incidents)", "Southport (public disorder|civil disobedience)", "Southport (social|community) (unrest|tensions)", "Southport (police|law enforcement) (clashes|confrontations)", "Southport (protesters|demonstrators) (clashes|confrontations)", "Southport (riot police|crowd control)", "Southport (arrests|detentions) during (riots|protests)", "Southport protest escalation", "Southport riot aftermath"

The Community Notes dataset

The Community Notes data provided by X is quite extensive, and includes the Note Text, as well as comprehensive metadata, which comprises timestamps of note creation, note status history, rating metrics. It also includes the misleading classification categories, which provide detailed information into which category of misleading content a post belongs to.

Post data integration and data preprocessing

After filtering Notes specific to the Southport riots, we used unique post identifiers (postIDs) associated with each Community Note to perform a secondary data collection phase, accessing and collecting the original post that received the notes in our dataset. For each post, we collected all relevant data and metadata, including creation timestamps, full engagement metrics including impression counts, the body of the post, its language, its status (live or removed), and information on the language used in the post. This integration established a complete timeline for each content piece and its corresponding Community Note, enabling a more complete analysis and understanding of the Community Notes dataset.

Prior to analysis, the raw dataset underwent a systematic preprocessing pipeline to enhance data quality and analytical validity. This process include standard pre-processing steps, such as deduplication and text cleaning, as well as restructuring the data into organised JSON files, merging the datasets so that for each Note we had a clear time-stamped understanding of its status history as well as the history of the post it relates to.

Data protection

The personal data of those who created posts contained in the dataset, including usernames, personal names, profile images and bios were deleted upon download of the dataset and were not stored or analysed.

However, the personal name of certain individuals mentioned within posts were retained if it fell into one of the following two types: if they were acting in a public capacity at the time of the incident, such as the name of the Prime Minister or Home Secretary, or if it referred to the name of the alleged attacker or was the assumed name of the attacker.

As the identity of the attacker was reported to be a key area of mis/disinformation during the riots i.e. the fact that a high volume of people on X were repeatedly misidentifying the attacker, including posts that refer to the name of the attacker or the assumed name of the attacker, it was determined that this was a crucial aspect of the proposed dataset and research.

Under UK GDPR Article 6, our approach met the 'consent' and 'public task' legal thresholds given that the data is made publicly available on X and that the processing of this data is for research purposes to inform government policy on tackling information threats.

The anonymised dataset will be stored in the UK on the Centre for Future Intelligence's local servers for a maximum of 1-year after which point it will be deleted.

Data coding

All posts associated with the Community Notes dataset were analysed and coded manually and inductively by Demos researchers. Posts were given descriptive codes that sought to achieve two things:

Descriptively capture its content in relation to the potential threat or risk it might pose.

Quantitatively and comparatively capture how explicit and confident the researcher could be that such a post posed a threat or risk or that there existed proof that a claim was inaccurate

1. Descriptive coding for types of threat, inaccuracy or risk

Researchers provided initial high-level descriptions of the posts in the dataset and then iteratively refined these during the coding process. Posts could be given more than one code where content overlapped. For example, a post could also be inaccurate and pose a threat to communities, who could be asylum seekers and/or Muslims. These codes were then evolved into sub-codes and subdivided by three categories introduced below. These categories were relabelled following consultation with experts to distinguish the 'threat to individuals, communities and buildings' and the 'risk to trust in democratic institutions' as these were found to reflect distinct types of potential harm that would consequently require distinct analysis from a policy and regulatory perspective.

2. Coding for confidence in how threatening, risky or inaccurate a post was

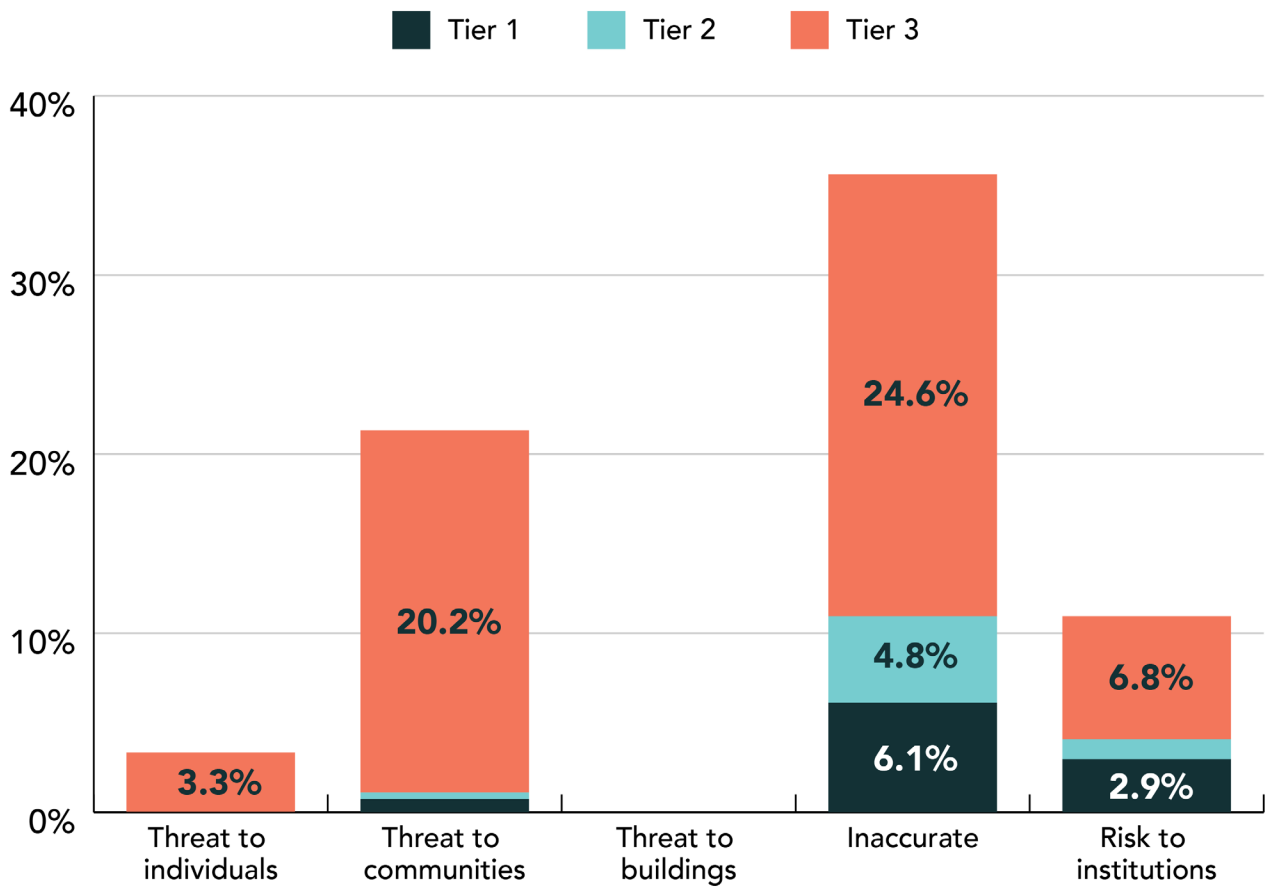
Researchers gave each post and its descriptive code a quantitative rating between 0 and 3. Ratings across all labels were peer-reviewed by another researcher and were tested for inter-

rater reliability. Tier 3 posts, those posts that have been included in this research report, had to meet the following criteria:

Threat to individuals or communities	The threat was explicit and contained an incitement to violence and hate speech
Inaccurate	Researchers could find multiple, verified and reputable sources that proved a claim was inaccurate
Risks to trust in societal institutions	The claim explicitly stated that the government, news media or police were engaged in a deliberate cover-up

The following chart indicates what proportion of the dataset could be included in one more of the categories broken down by Tier to demonstrate what proportion of the posts were afforded what confidence status.

FIGURE 7
OVERVIEW OF POSTS CLASSIFIED BY WHETHER THEY COULD BE CONSIDERED A THREAT TO INDIVIDUALS OR COMMUNITIES, INACCURATE AND A RISK TO INSTITUTIONS



Quantitative analytical framework - threat, inaccuracy and risk to institutions

The risks and threat posed by the posts associated with the Community Notes dataset were then quantitatively analysed based on the three different categories introduced through the coding process. First, to assess the potential threat such posts could pose to individuals, communities and to physical buildings, such as mosques and hotels where asylum seekers were staying. Second, to assess the level of inaccuracy of the posts and therefore the risk of being misleading to audiences. And third, the risk the posts could pose to trust in our societal institutions such as the UK government, the police and news media. These categories were chosen based on an inductive analysis of the posts and in consultation with a range of experts.

TABLE 4
QUANTITATIVE ANALYTICAL FRAMEWORK - THREAT, INACCURACY AND RISK TO INSTITUTIONS

The categories are summarised below:

1. THREATS TO INDIVIDUALS, COMMUNITIES AND BUILDINGS	1a Threats to individuals	Threats to members of the public
		Threats to public figures
	1b Threats to communities	Religious hatred
		Racial and xenophobic hatred
		Hatred against immigrants
		Hatred against asylum seekers
	1c Threats to buildings	
2. INACCURATE	2a Inaccurate descriptions of the attacker	
	2b Inaccurate descriptions about the attack	
	2c Inaccurate information about causes of the attack	
3. RISKS TO TRUST IN INSTITUTIONS	3a Risks to trust in government	
	3b Risks to trust in the news media	
	3c Risks to trust in the police	

Please note that a number of additional sub-codes were also included in the data coding process, but were not chosen to be prioritised for the quantitative analysis included in this paper.

Total views

The following table captures a breakdown of the number of views data referenced in this report for posts that meet the Tier 3 criteria only.

	TOTAL VIEWS OF POSTS WITHOUT A VISIBLE NOTE THAT MEET THIS CRITERIA	TOTAL VIEWS OF POSTS WITH A VISIBLE NOTE THAT MEET THIS CRITERIA
1a Threats to individuals	10,798,120	0
1b Threats to communities	92,737,667	1,090,322
2 Inaccurate	85,928,540	17,163,228
3 Risks to institutions	51,155,870	17,163,228

Posts that are both threatening to communities *and* inaccurate: 67,568,754

Posts that are threatening to communities or inaccurate (including those that are both threatening to communities *and* inaccurate without double counting): 111,097,453.

No associated status data

12% of Notes (81) in the dataset had no associated status data at all. At the time of writing, it is unclear to the research team what this signifies.

DIVERSITY, INCLUSION, EQUITY AND JUSTICE STATEMENT

As part of Demos's ongoing efforts to facilitate greater diversity, inclusion, equity and justice in all areas of our work, we assess and publish our approach to meeting our goals in each of our papers.

In this project, our team identified that a variety of communities could be impacted by undertaking and publishing a research study exploring the efficacy of Community Notes during the Southport riots and by specifically analysing posts that included racial and religious hatred, and hatred towards immigrants and asylum seekers. We also identified risks of bias and inconsistency in our approach to analysing and coding such posts.

We sought to mitigate the impact of bias as well as to manage the impact of analysing harmful content on members of our project team with a number of steps. First, we increased the number of analysts engaged in qualitatively analysing the material to three, and held regular meetings to evaluate and test the codes inductively developed to label the posts in question. Second, we employed a cap on the number of posts that could be reviewed per day to ensure regular breaks and reduce the chance of fatigue among the project team when developing and deploying labels. Third, all post labels were peer-reviewed by another member of the analysis team, and a sample was tested by the team leader for consistency. Finally, we invited expert assessment of the coding framework that was developed to label the posts from a diverse set of experts, including those with expertise in false information and independent fact-checking, migrant rights and far right extremism, as well as social media regulation and illegal harm judgments. Experts in Islamophobia were invited to contribute their expertise, but were unable to during the required research period.

We plan to ensure that the evidence generated and presented in this paper is made widely available, through Demos's owned channels, including our website and newsletter, as well as proactively circulating it to those who are engaged in media and digital policy, and/ or civil society and community organisations impacted by false information as well as illegal harms content online. This will include those who shared their expertise as part of this research study; are members of the Online Safety Act Network and/or those who support communities who were disproportionately affected by false information during the Southport riots; and those who continue to be affected by illegal harmful content posted online including hate speech, such as minority ethnic and/or religious communities and migrant and asylum seeker communities.

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