# Comments on the role social media played in the Southport Riots – the institute for strategic dialogue

## Summary

This document contains evidence from the Institute for Strategic Dialogue (ISD) on the role social media played in the Southport riots. This submission focusses on the following areas requested by the Oversight Board:

* The role social media played in the 2024 UK riots, including the spread of misinformation, in organizing riots and informing the public of relevant developments.
* Any documented links between anti-immigrant and anti-Muslim online speech and violence and discrimination in the UK.
* The role that imagery (i.e., pictures, graphics, memes, videos, including AI-generated) plays in online hate speech.
* Challenges for automation in assessing incitement or hate speech in imagery, especially AI-generated.
* Risks of overenforcement of Community Standards in such contexts and the potential negative impacts on the free expression rights of speakers and listeners.
* Whether Meta should consider content moderation measures short of removal for this type of content.

**About ISD**

[ISD](http://www.isdglobal.org/) is an independent, non-profit organization dedicated to safeguarding human rights and reversing the rising tide of polarization, extremism, and disinformation worldwide. Our work includes in-depth research and analysis identifying and tracking online manipulation, mis- and disinformation, hate, and extremism in real time. We also formulate, advocate and deliver evidence-based policy approaches and programming.

## The role social media played in the 2024 UK riots

The extremist riots which followed the 2024 Southport stabbings have been held up by many experts as a case study of the close relationship between online harms and offline violence. Analysis conducted by ISD illustrates how social media was central to the rapid spread of [disinformation](https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/) relating to the Southport stabbing which proved to be a catalyst for the riots; the [coordination](https://www.isdglobal.org/digital_dispatches/quantifying-extremism-a-data-driven-analysis-of-riot-related-far-right-telegram-networks/) of riots; expressions of support for rioters, including by well-known [extremist influencers](https://www.isdglobal.org/digital_dispatches/quantifying-extremism-a-data-driven-analysis-of-riot-related-far-right-telegram-networks/); the incitement of [violence](https://www.isdglobal.org/digital_dispatches/total-system-collapse-far-right-telegram-network-incites-accelerationist-violence-after-southport-stabbings/) and [hatred](https://www.isdglobal.org/digital_dispatches/evidencing-a-rise-in-anti-muslim-and-anti-migrant-online-hate-following-the-southport-attack/) targeting minority communities; and the amplification of [footage](https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/) of the riots which in turn served to promote riotous activity and lionise the rioters themselves. Our analysis showed that beyond this content being widely accessible on platforms, harmful content was amplified by [recommendation](https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/) algorithms. This activity took place in the context of a British [extreme right](https://www.isdglobal.org/digital_dispatches/the-foundations-of-violence-the-growth-of-far-right-hate-in-the-uk/) movement which has grown in confidence in recent years.

Following a stabbing attack in Southport on July 29th which left three girls dead, social media content which contained false information suggesting the attacker was a recently arrived asylum seeker and a Muslim rapidly spread online. These claims were amplified on X by a number of widely followed [‘blue tick’](https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/) users such as an account called ‘Europe Invasion’, which received 1.4 million views by July 31st. As the riots spread extremist accounts and channels engaged in frenetic activity, with far-right posting activity on Telegram rising by [327%](https://www.isdglobal.org/digital_dispatches/quantifying-extremism-a-data-driven-analysis-of-riot-related-far-right-telegram-networks/) in the 10 day after the attack.

A [constellation](https://www.isdglobal.org/digital_dispatches/quantifying-extremism-a-data-driven-analysis-of-riot-related-far-right-telegram-networks/) of Telegram channels played key roles ranging from organising activity to transnational amplification to target selection. A channel called ‘SOUTHPORT WAKE UP’ became a central organising group for riots and grew from only 44 members the afternoon after the attack to more than 15,000 members by the time it was removed a week later. Analysis of these channels provided a strong indication that they were used to draw attention to specific locations with the aim of directing hate, violence and riots, with mentions of the locations of riots being posted in these channels before they took place.

Platform recommendation infrastructure also played a role in [amplifying content](https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/). For example, a false name attributed to the attacker was recommended to X users in the “What’s happening” sidebar, whilst the “Others searched for” feature suggested the false attacker’s name as well as names of British extremists promoting disinformation. Similarly, the “Others searched for” feature on TikTok contained conspiracy theory content relating to the attack.

Additionally, the riots corresponded with spikes in online [hate speech](https://www.isdglobal.org/digital_dispatches/evidencing-a-rise-in-anti-muslim-and-anti-migrant-online-hate-following-the-southport-attack/). Analysis conducted by ISD found that across 55 British extremist Telegram channels, anti-migrant hate rose by 246% in the ten days after the attack, whilst anti-Muslim hate rose by 276%. On X the use of anti-Muslim slurs more than doubled during this period after the attacks.

## The links between anti-immigrant and anti-Muslim online speech and violence and discrimination in the UK

Online activity presents an essential component of activities of British extremist movements targeting immigrants and Muslims. As demonstrated above, online communications, including hate speech, were central the anti-immigrant and anti-Muslim violence which the UK saw in July and August 2024. Extremist online communities where hate speech is normalised and accounts which frequently advance hateful material were central in advancing content which legitimised attacks against Muslims and immigrants, including providing lists of potential targets. These extremists targeted Muslims, migrants, critics of the far-right, and researchers tracking extremism with [explicit](https://www.thetimes.com/uk/crime/article/uk-riots-far-right-use-arson-manual-to-plot-attacks-on-lawyers-pk2bgtqhg) calls to violence. However, the role of discriminatory, hateful or violent online speech in driving anti-minority mobilization goes [beyond](https://www.isdglobal.org/digital_dispatches/the-foundations-of-violence-the-growth-of-far-right-hate-in-the-uk/) the 2024 UK riots, and ISD has tracked such activity extensively in the UK and globally for more than a decade.

Online platforms allow extremist movements to spread anti-Muslim and anti-immigrant ideologies, recruit members, and organize offline actions. Extremists occupy various digital spaces, from niche platforms like 4chan and Telegram to mainstream platforms such as X and Facebook. Recent ISD [research](https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-harms/2023/tangled-web.pdf?v=330163) conducted for Ofcom shows that British extremists can reach audiences in the hundreds of thousands across these fora. This study included analysis of 59 Facebook pages and groups, and 68 Instagram accounts which met our definition of extremism. On Facebook the top ten largest pages and groups had an average following of 258,986, and on Instagram 55,040.

Extremist actors exploit these platforms to promote violence and hatred against individuals and communities deemed to be their opponents, including minority communities. In a study for [Ofcom](https://www.isdglobal.org/wp-content/uploads/2023/11/Hate-of-the-Nation.pdf) ISD identified 394,753 hateful messages on Twitter, 26,085 hateful messages on 4chan, and 1,540 hateful messages on Facebook groups and pages over August 2022.

Surges in online hate speech often occur around major crises, correlating with spikes in offline hate crime. For example, following Hamas’s attack on Israel on October 7th 2023, [antisemitic](https://www.isdglobal.org/digital_dispatches/rise-in-antisemitism-on-both-mainstream-and-fringe-social-media-platforms-following-hamas-terrorist-attack/) hate speech increased 50-fold in YouTube comments, whilst anti-Muslim hate speech increased [43-fold](https://www.isdglobal.org/digital_dispatches/43-fold-increase-in-anti-muslim-youtube-comments-following-hamas-october-7-attack/) in YouTube comments and a [422%](https://www.isdglobal.org/digital_dispatches/use-of-words-phrases-and-hashtags-associated-with-anti-muslim-mobilisation-surges-amid-israel-gaza-conflict/) increase on X.

## The role that imagery plays in online hate speech

The use of image-based content is a key part of contemporary extremist culture and strategies for broadcasting their ideology and denigrating the communities they are opposed to. The use of this material is frequently shared casually but can be deployed in sophisticated [coordinated](https://www.isdglobal.org/isd-publications/gaming-and-extremism-the-extreme-right-on-discord/) harassment campaigns, as well as efforts to [disrupt](https://www.isdglobal.org/isd-publications/the-fringe-insurgency-connectivity-convergence-and-mainstreaming-of-the-extreme-right/) the democratic process.

In particular, the use of internet [memes](https://www.isdglobal.org/explainers/memes-the-extreme-right-wing/) has, since the early 2010’s, become a core component of contemporary extreme right content, growing in prominence alongside the [“alt-right”](https://www.isdglobal.org/wp-content/uploads/2017/10/The-Fringe-Insurgency-221017_2.pdf) movement. The alt-right’s use of memes helped combine hateful white supremacist ideology with online forum, gaming and hacker cultures, creating a familiar aesthetic which condenses complex ideologies into an approachable format which can be readily disseminated to broader audiences. This use of imagery combined with the use of humour and irony facilitates the spread of online hate speech targeting a range of communities. These hateful memes range from the use of racist [caricatures](https://knowyourmeme.com/memes/happy-merchant) to content which explicitly calls for [violence](https://www.isdglobal.org/isd-publications/a-safe-space-to-hate-white-supremacist-mobilisation-on-telegram/) against minority communities. Hateful memes may influence [radicalization](https://icsr.info/wp-content/uploads/2021/01/Memetic-Irony-And-The-Promotion-Of-Violence-Within-Chan-Cultures.pdf) trajectories into violence, whereby prolonged exposure to the trivialization of hateful, violent, or racist beliefs can lead individuals to normalize the content and become gradually more tolerant of violent extremist ideologies.

Beyond image-based content, short and long-form video material are central communications tools for extremists which can be used to target minority communities with hate and inspire violence against them. As noted above, TikTok was widely [used](https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/) around the UK 2024 riots, both to announce riots and inspire others to engage in violent activity. This itself plays out against a backdrop of frequent extremist [use](https://www.isdglobal.org/isd-publications/tiktok-and-white-supremacist-content/) of the platform, both to radicalize individuals and spread hateful messaging.

Finally, extremists have rapidly adopted AI to create hateful imagery and propaganda, recognising both its utility in reducing the labour involved in content creation, and the novel opportunities it brings with it. An example of this is the use of AI to [translate](https://www.isdglobal.org/digital_dispatches/content-glorifying-hitler-surges-online-amid-growing-historical-revisionism/) Hitler’s speeches into English and generate accompanying video content, which ISD research found reached millions of users online. Additionally, analysis conducted by ISD’s Coalition to Combat Online Antisemitism found that AI tools are [vulnerable](https://www.isdglobal.org/wp-content/uploads/2024/01/The-Fragility-of-Freedom.pdf) to manipulation by extremists, with analysts finding that text to image tools can be easily prompted to generate material which distorts the Holocaust.

## Challenges for automation in assessing incitement or hate speech in imagery, especially AI-generated

Some of the most [powerful](https://www.isdglobal.org/isd-publications/hate-of-the-nation-a-landscape-mapping-of-observable-plausibly-hateful-speech-on-social-media/) tools developed for identifying and analysing hate at [scale](https://www.isdglobal.org/isd-publications/antisemitism-on-twitter-before-and-after-elon-musks-acquisition/) – such as Large Language Models and Natural Language Processing – are primarily focused on the detection of hateful text. Whilst image detection tools are increasing in power, less time has gone into their development and as such their limited sophistication acts as a barrier to their effective deployment in moderation efforts.

This is compounded by the fact that the [detection](https://www.isdglobal.org/isd-publications/hate-of-the-nation-a-landscape-mapping-of-observable-plausibly-hateful-speech-on-social-media/) of hate itself is challenging – reflecting the nuance which is needed to understand hate comprehensively. Whilst some types of content are clearly identifiable as hate speech - such as the explicit use of slurs or imagery which depicts violence against minorities – other content is less clearly identifiable and often occupies a ‘grey area’ where it’s hateful nature might not be easily recognisable. Identifying the deployment of content such as [conspiracy](https://www.isdglobal.org/wp-content/uploads/2024/06/Narratives-of-hate_Post-7-Oct.pdf) theories which villainise communities, or in-jokes which incite hatred requires in-depth understanding of extremist cultures and contemporary trends in online hate. Accordingly, it can be challenging to train models which reflect this nuance.

Additionally, questions remain around what an appropriate response to this content is. Whilst material which clearly incites violence may be illegal in the UK, and should be removed accordingly, it is often ambiguous whether ‘grey area’ visual content associated with hate breaches the law or even platform terms of service. Accordingly, additional approaches to addressing this content, such as masking it or providing content warnings should be considered.

One possible approach to the automated detection of visual content which has been effective in facilitating the moderation of terrorist and child sexual abuse material is the use of image hashing databases. It is possible that a similar approach could be used to facilitate enforcement against some types of hate material. However, this is challenging as hateful imagery – particularly memetic content – rapidly evolves over time, something which is accelerated by the ease with which novel content can be made using AI.

Accordingly, models which automate hateful imagery will need to be consistently re-trained to stay abreast of trends in online extremist culture, requiring ongoing input from subject matter experts. One potential solution here would be reserving the deployment of automated image detection models for crisis moments, only deploying them around viral hateful material which has the potential to incite violence – such as the content described above which incited individuals to riot following the Southport stabbings.

## Risks of overenforcement of Community Standards in such contexts and the potential negative impacts on the free expression rights of speakers and listeners.

Reflecting the nuance associated with hate speech outlined above, it is possible that moderators which are unfamiliar with internet culture, and trends in contemporary extremism and hate speech may either under or over-enforce against hate speech content. This itself is recognized by extremists, who have notably made efforts to troll reporters and watchdogs into declaring absurd items – such as a pint of [milk](https://www.wired.com/2017/05/alt-rights-newest-ploy-trolling-false-symbols/) – as hate symbols in an effort to waste time and promote over-enforcement. This itself has the potential to negatively impact on free expression of internet users who inadvertently use these symbols. Equally, however, given the nuances of hate speech, including the use of coded language and imagery, there is a risk that unfamiliar moderators may allow content which actively promotes hatred against communities to remain online.

Ultimately, however, due to diminishing data access afforded to researchers it is becoming increasingly challenging to independently make conclusive statements around the efficacy of enforcement or the extent of over-enforcement. The provision of meaningful data access to researchers is an essential component if independent assessments are to be made on whether terms of service are applied consistently.

## Whether Meta should consider content moderation measures short of removal for this type of content

In moments of crisis – such as the 2024 riots – image-based content which helps incite violence, facilitate the coordination of extremist activity, or target minority communities with hatred can spread rapidly. In rapidly evolving situations it may not be immediately clear whether content crosses thresholds in terms of service, particularly if it is viewed out of context, or uses coded language. This itself highlights the importance of crisis response mechanisms – as [recommended](https://www.onlinesafetyact.net/analysis/disinformation-and-disorder-the-limits-of-the-online-safety-act/) by Ofcom – which would be highly relevant in situations like the Southport stabbings and 2024 riots. Existing protocols such as those provided by the Christchurch Call and the GIFCT currently only cover terrorist incidents rather than the kind of violence and rapidly accelerating hatred which were observed in August. However, as demonstrated by the UK Riots non-terrorist incidents at crisis moments are nevertheless capable of harming communities and inciting mass violence and disorder and are still deserving of rapid response. Accordingly, it is recommended that similar protocols are developed for febrile moments when extremists are particularly active, or when hateful material is rapidly and widely disseminated, although these protocols should incorporate considerations around the best ways to safeguard human rights and implement procedural accountability mechanisms.

Additionally, it is important to underline that content moderation should not be the only tool deployed to counter hateful and extremist content – particularly when there is ambiguity whether material breaches a community threshold.

As ISD [research](https://www.isdglobal.org/wp-content/uploads/2022/12/Understanding-How-Algorithmic-Ranking-Practices-Affect-Online-Discourses-and-Assessing-Proposed-Alternatives.pdf) underscores and the case study of the UK riots highlights, algorithmic ranking systems play a significant role in accelerating the visibility and reach of hateful and extremist material. Accordingly, platforms like Meta should explore proactive measures to limit the amplification of extremist and borderline content through algorithmic adjustments. Additionally, these systems remain largely opaque to researchers, regulators, and the public. Greater transparency into how these algorithms prioritize, recommend, or demote content is essential for transparency and trust building with users. Platforms should disclose the metrics and testing processes behind algorithmic decisions, particularly during crises, alongside regular reporting on the effectiveness of interventions such as "quality-focused" ranking strategies.