ISD | Powering solutions to extremism, hate and disinformation

ISD Written Evidence to the Science, Innovation and Technology Committee Inquiry on Social Media, Misinformation and Harmful Algorithms

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org

Contents

Summary

This document contains written evidence compiled by the Institute for Strategic Dialogue (ISD) providing insight into the various harms associated with social media algorithms, the role these played in the 2024 UK riots, and regulatory approaches for effectively countering these harms.

This evidence examines the extent to which the business models of social media companies, search engines, and similar entities encourage the spread of harmful content and contribute to broader social harms. It explores how these companies use algorithms to rank content, the relationship between their business models and the dissemination of misinformation, disinformation and harmful content, and the role of generative artificial intelligence (AI) and large language models (LLMs) in creating and amplifying such content.

The evidence also considers the influence of social media algorithms on the riots that occurred in the UK during the summer of 2024, alongside an evaluation of the effectiveness of the UK's regulatory and legislative framework in addressing these challenges. This includes an assessment of the UK's Online Safety Act's potential impact, the need for further measures to tackle harmful content, and the roles of regulatory bodies such as Ofcom in preventing the spread of false and harmful content online. Additionally, the evidence addresses accountability for the spread of misinformation, disinformation, and harmful content arising from the use of algorithms and AI by social media and search engine companies.

Introduction

The 2024 summer riots starkly exemplified what has long been understood: the role of digital platform structures in facilitating the spread of harmful misinformation, the opportunities they provide for the normalisation and spread of hate speech, and the role of social media in facilitating offline violent mobilisation. These online harms serve not only to harm the safety of marginalised groups, but also create a chilling effect, threatening the integrity of rights, freedoms and democracy writ large. Beyond considerations around the mere platforming of such content, recommender algorithms share harmful content to wider audiences, curating online ecosystems which normalise hateful narratives and spreading misinformation to users who may not otherwise have seen it. Meanwhile, the manipulation of AI systems and content risks harming information integrity and trust in institutions. While the efficacy of the Online Safety Act towards these threats cannot be measured until after implementation begins in 2025, key gaps could be addressed in the short term, including crisis response and data access for researchers.

Case study: UK summer riots

Following the murder of three girls at a children's dance party in Southport on 29 July 2024, misinformation alleging the attacker was an asylum seeker and a Muslim spread quickly on social media. The violence witnessed in the subsequent riots exemplifies the real-world consequences of viral, unchecked misinformation on social media. ISD research in the immediate aftermath of the violence showed how false claims were uncritically amplified by pay-for-clicks users, including blue tick users with very large followings. For example, a post by blue-tick account 'Europe Invasion' which spread the false name of the attacker had been viewed more than 1.4 million times by 31 July. By 3pm the day after the attack, the false name had received over 30,000 mentions on X alone from over 18,000 unique accounts.

This false name spread both organically and recommended to users by platform algorithms. On X, the false name "Ali al-Shakati" was recommended to users in the "What's happening" sidebar. Similarly, when searching for "Southport" almost 9 hours after the policy confirmed the false nature of the rumoured name, the "Others searched for" sidebar recommended the false name of the attacker, "Ali-Al-Shakati Arrested in Southport" as well as "Tommy Robinson", the prominent extremist and amplifier of misinformation. Conducting the same exercise on 4 December 2024, analysts searched for "Southport" in the TikTok search bar. The first result recommended in the "Others searched for" bar was "southport cover up". When analysts clicked to see related videos, the first recommended video alleged a "Southport cover up" and the third video's thumbnail showed a headline: "Breaking News: Rumours link Keir Starmer to Southport killers dad as his lawyer in 2003" (the video goes on to clarify the lack of evidence to support this allegation).

Riot-related misinformation directly precipitated online . Across 45,000 messages posted to 55 British far-right Telegram channels in the 10 days after the Southport attack, anti-migrant hate rose 246% and anti-Muslim hate rose 276%. The use of anti-Muslim slurs on X more than doubled in this period, with over 40,000 posts containing one or more of these terms. The volume of hateful posts, which likely violate Terms of Service, calls into question both the

adequacy of moderation policies and their implementation during crisis contexts, whether by human moderators or automated mechanisms.

Although recommender algorithms are less relevant to Telegram's functionality (although recent analysis from Southern Poverty Law Center has spotlighted potential risks from the service pushing users towards extremist content), the platform played an outsized role in mobilising offline action. Protests, many of which morphed into violent riots, were organised across both mainstream and alternative platforms. Far-right posting activity in the 10 days after the Southport attack rose by 327%. The lack of focus in the Online Safety Act (OSA) on smaller and more high-risk platforms such as Telegram and other extremist ecosystems will not adequately address these mobilisations, online as well as offline. We recommend Ofcom make full use of the flexibility within the OSA to assess risks based on functionality as well as size, ensuring smaller but high-risk platforms like Telegram are subject to appropriate duties, including transparency reporting. This approach would address gaps in oversight, enabling scrutiny of safety measures on platforms that facilitate extremism and hate, thereby mitigating harms before they escalate.

Platform algorithms

Recommender algorithms and algorithmic amplification

Platform recommender algorithms have served to promote harmful information, narratives and actors to users who may not otherwise have been amplified. YouTube has been found to have a powerful recommender algorithm which suggests misogynistic and in some cases extremist content. A study by ISD created 10 blank accounts of young men in Australia, some of which engaged with right-wing content. By clicking through recommended YouTube Shorts, analysts found that all 10 accounts were eventually recommended misogynist, manosphere and incel content. Similarly, ISD identified harmful recommendations on TikTok, with accounts and videos promoting Nazism algorithmically amplified on the platform's 'For You' page after a dummy account engaged with 10 videos containing Nazi content. The OSA's obligation for services to conduct risk assessments on illegal content and child safety along with mandatory transparency reporting will be a crucial first step towards scrutinising recommender systems and platform processes to ensure user safety.

Engagement-based ranking systems often prioritize borderline content that does not cross platform Terms of Service but may still be harmful. For a platform business model, a focus on maximising engagement can therefore result in algorithmic ranking optimising for more toxic, sensationalist, and borderline content. For instance, on X, ISD analysts identified a significant amount of pro-Hitler content, observing the rapid adaptation of recommender algorithms to serve similar posts. In the "For You" feed of one account, 10 of the first 19 posts (52%) served featured Hitler, praised Nazis or were clearly antisemitic. The OSA sets out to address the risks of amplification of violative content, although the efficacy of its implementation is still unknown. Independent third-party auditors—free from platform influence and with access to platform data and policies—are essential to assess algorithmic ranking systems, identify biases, and close remaining gaps in oversight to ensure harmful content does not evade accountability.

Autofill functions also served to recommend users new ways of reaching violent extremist, and in many cases illegal, online content. When searching on TikTok for videos of named IS figures using a sock puppet account of a minor, analysts found that some searches were blocked. However, the autofill search function recommended to the users misspelled terms to enable them to reach this content anyway. To address this issue, social media platforms must enhance their enforcement of existing policies and provisions around children's safety and the removal of illegal terrorist content. Equally, the stringent implementation of the OSA Illegal Harms Codes is essential, as well as making available data access for researchers to understand how platform functions may contribute to online radicalisation. Access to researcher data is also necessary to gain further insight into the nature, scale, and scope of (illegal) harms, ensuring that interventions are proportionate and enabling the measurement of the effectiveness of mitigation measures.

ISD research on misleading and manipulated content in the context of the Middle East conflict found that premium X accounts promoting mis- and disinformation during crisis events are algorithmically amplified and receive high levels of engagement. Under the OSA, platforms must take proportionate measures to reduce the risk of harm caused by illegal content and ensure robust systems for content moderation and verification. Ofcom, as the regulator, should rigorously enforce these duties, including holding platforms accountable for amplifying verified accounts that disseminate false or harmful information. Stronger requirements for verification processes and algorithmic transparency are critical to mitigate these risks.

Content moderation algorithms

Content moderation – both automated and human-led - is a long-standing approach employed by the majority of services to address online harms. However, platforms have failed to effectively implement their Terms of Service and remove harmful, hateful, and in some cases, illegal content. One reason for this issue is the algorithmic ranking practices employed by these platforms, which prioritise content with high engagement rates, even if that content is borderline or illegal. While the OSA addresses algorithmic practices to some extent, research conducted by ISD emphasises the importance of comprehensive measures and increased transparency, such as through data access for researchers, to effectively assess the challenges posed by algorithmic ranking in content moderation.

The variety, extent and prevalence of online harms demonstrates that content moderation measures and associated user reporting tools are necessary but not sufficient in effectively mitigating risks. Therefore, as a primarily reactive measure, content moderation should not be over-relied upon over proactive safety-by-design efforts to prevent or discourage illegal content or activity from occurring in the first place.

Services' content moderation efforts have also too often been characterised by a lack of genuine transparency, both for individual users and at a macro level, to enable independent external assessments of their proportionality, consistency and effectiveness. In our view, many services' existing transparency reports often rely on self-selected metrics and measures of success that do not present an objective assessment. We would therefore recommend that Ofcom introduces baseline expectations and consistent measures to assess their impact in mitigating risks and reducing harms and allow for cross-industry comparisons.

Artificial Intelligence

Content generation

Bad actors have adopted generative AI technologies to create content designed to radicalise individuals, spread harmful views, and denigrate their opponents, harnessing the aestheticisation of digital ecosystems. AI-generated content can also be misused to influence political communication and media trust, including through non-consensual intimate AI-generated content (non-consensual intimate deepfakes (NCID)). While the OSA takes a welcome approach in prohibiting the sharing of NCID content, it does not explicitly mandate the labelling of generative AI content. The OSA does impose duties on service providers to mitigate risks associated with illegal and harmful content, including that produced by generative AI. This implies a need for transparency and accountability in how such content is managed, though not necessarily through explicit labelling. Given the rapid evolution of AI technologies and their integration into online platforms, as ISD's research shows, there is a need for clearer guidelines and potential regulatory measures concerning the labelling of AI-generated content in the UK. ISD research suggests that digital services should also be required to implement measures to distinguish AI-generated (political) content from authentic content, by labelling it appropriately. It is essential to strike a balance and avoid mislabelling content as AI-generated, to avoid exaggerating the technical capabilities and persuasive power of those disseminating disinformation.

Dissemination and targeting

Recent ISD research has shown how platforms fail to label and remove AI-generated and manipulated election content, in contravention of their own commitments and policies. In the lead up to the US elections, ISD found 154 instances of unlabelled or unremoved election-related AI-generated or otherwise manipulated content across Facebook, Instagram, YouTube, X, and TikTok, which accumulated over 51 million impressions. While this example comes from outside of the UK, it nevertheless underlines the ease at which harmful content spreads around carefully monitored events such as elections and the issues around inconsistent AI content labelling. While the OSA does not specifically mandate labelling AI-generated content, the broader obligations to protect users from harmful material should incentivise platforms to adopt practices like labelling to enhance transparency and user awareness. It is worth noting that other jurisdictions are considering or have implemented specific measures regarding AI-generated content. For instance, the European Union's AI Act imposes transparency obligations on users of AI systems that generate or manipulate content

resembling real persons, requiring disclosure that the content has been artificially generated or manipulated.

Information environment architecture

Manipulated imagery has also been harnessed by hostile actors in the ongoing Israel/Gaza conflict. Within seven hours of Iranian drones being launched towards Israel on 13 April 2024, 34 false, misleading, or AI generated images and videos claiming to show the ongoing conflict were posted and received over 37 million views on X over the next day. This content was either repurposed from previous conflicts or was AI generated, including images of "supersonic missiles from Iran" and President Biden wearing military fatigues, which served to foster panic. Most of this content was spread by 'verified' paid premium users, allowing it to receive algorithmic amplification by the platform. These findings underline the need for social media platforms to ensure that 'verified' status cannot be given without adequate checks to prevent the amplification of accounts that disseminate misinformation. Additionally, further focus should be on monitoring and moderating the content shared by these verified users, especially during critical events. Finally, the OSA does not explicitly mandate crisis protocols for platforms, unlike similar regulation such as the EU's Digital Services Act. More explicit crisis response measures could provide the focus and mandate for platforms to ensure enhanced monitoring and action of illegal and harmful content, and collaboration with authorities in these instances. However, safeguarding of human rights and procedural accountability mechanisms must also be in place.

AI has been used in 2024 to translate Hitler speeches into English and mimic videos of speeches in German, facilitating the accessibility of his ideas to new audiences. ISD identified over 140 Instagram reels reaching wide audiences, with one such audio clip garnering 229,000 likes and 4.9 million views. Threatening the integrity of Holocaust memory and facilitating revisionism, generative AI systems have been successfully prompted to create images of children happily playing in concentration camps. ISD's research demonstrates an enforcement gap in content moderation policies. While many platforms have established content moderation policies, they lack consistent enforcement and are often quickly outdated given technological advances. Platforms should review their practices on a regular basis and monitor the effectiveness of their policy enforcement, including the training of content moderation staff on new trends and technological developments, such as generative AI. The OSA's requirement for digital services to conduct annual transparency reporting must be an essential tool to assess the measures and processes taken by social media platforms.

Transparency and data access

While qualitative investigations have uncovered the significant role of platform recommender systems in exacerbating online harms, such as in the aftermath of the Southport attack, the research sector at large has been unable to gather large-scale evidence due to lack of data access. Where meaningful data access is not mandated under the OSA, third party independent bodies are unable to build a systematic understanding of the online threat landscape or the role of platform systems such as algorithms. Data access has deteriorated in

recent years through the closure of Meta's CrowdTangle tool in August 2024 and the prohibitive costs of API access on X for services which were previously free of charge. However, access to platform data is imperative for independent researchers to understand the scale, scope and nature of online harms, as well as to measure the impact, efficacy and proportionality of interventions. It thus provides transparency mechanisms essential for digital services accountability. Without granting Ofcom the authority to mandate data access—unlike the EU's Digital Services Act (DSA)—UK social media users will remain less informed than their EU counterparts regarding the true nature of online harms, nor the effectiveness or proportionality of interventions. Establishing a solid evidence base is essential for effective compliance with the OSA and for monitoring its impact. The proposed Data (Use and Access) Bill presents a valuable opportunity to align with the minimum data access requirements outlined in Article 40 (12) of the DSA. We welcome the consultation of the Bill to reflect this.

Transparency gaps exist in the construction of recommender algorithms, the training of platform content moderation LLMs, and the training received by those who build them. Thus, policymakers should prioritise improving transparency in the development of recommender algorithms and the training of content moderation LLMs. An essential element to greater transparency could be independent audits of digital services at the minimum, as laid out in the EU under Article 37 of the DSA. Currently, such audits are not envisaged under the OSA. Additionally, it is crucial for regulators to invest in research aimed at detecting LLM-generated content and identifying LLM-driven social bot networks. The ongoing competition between detection methods and evasion tactics highlights the need for a comprehensive approach to effectively tackle these emerging threats and explore further strategies to address these challenges.

Besides independent audits for digital services, it is crucial for Ofcom to effectively utilise its information-gathering powers to ensure the successful implementation and enforcement of the OSA. As the OSA is still in its early stages and there are long-standing limitations on data access coupled with a lack of transparency, the existing evidence regarding online harms and the effects of service design is often fragmented and insufficient. While proportionality is an important factor, all platforms presenting significant risks to online safety should be required to fully comply with information requests. Ofcom should also assist these services in meeting their compliance obligations. Furthermore, in the absence of more robust provisions for researchers to access data it is essential for Ofcom to maximise its information-gathering capabilities, especially as many services either restrict or do not provide independent access to data. This approach will help address the data access challenges faced by researchers in the online ecosystem, as highlighted in our previous research.

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org