Written evidence submitted by Full Fact (SMH0047)

What are the links between social media algorithms, generative AI and the spread of harmful content online?

Summary

1. Full Fact is grateful for the opportunity to provide evidence to the committee's inquiry on the links between social media algorithms, generative AI and the spread of harmful content online, particularly following the riots this summer in the United Kingdom.

2. As the UK's leading independent fact checking organisation, Full Fact was actively involved in fact checking the riots this summer and ensuring misinformation was flagged and where possible, corrected. In this submission we have outlined some of the lessons learned, and what more we believe can be done to tackle harmful misinformation, but we would welcome the opportunity to discuss this further with the committee.

3. We believe that understanding the root problem of misinformation spreads online is a complex, multifaceted problem, but so is the solution to tackle it. While improving the regulatory system is essential, there is an urgent need for literacy to be factored into the solution, equipping the public with the skills to determine what is true and what is false.

About Full Fact

4. Full Fact fights bad information. We're a team of independent fact checkers, technologists, researchers, and policy specialists who find, expose and counter the harm it does.

5. Bad information ruins lives. It promotes hate, damages people's health, and hurts democracy. So, we tackle it in four ways. We check claims made by politicians, public institutions, in the media and online. We ask people to correct the record where possible to reduce the spread of specific claims. We campaign for system changes to help make bad information rarer and less harmful, and we advocate for high standards in public debate.

6. Full Fact is a registered charity. We're funded by individual donations, charitable trusts, and by other funders. We receive funding from both Meta and Google. Details of our funding can be found on our website.[1]

[1] Full Fact, Funding, 2022, https://fullfact.org/about/funding/

What role did social media algorithms play in the riots that took place in the UK in summer 2024?

7. Following the riots that took place in the summer of 2024, Full Fact shared an overview on the role on which misinformation played in contributing to the riots.[2] While we haven't investigated the origins of the claims, and other organisations including the BBC, Channel 4, Sky News and the Guardian have published their own analyses, Full Fact has undertaken a review of the scope of claims present during this period.[3]

8. The riots, triggered at least in part by false claims circulating on social media in the wake of the Southport stabbings, were a clear reminder of what can happen when online misinformation spills into the real world and the harm it can cause.

9. The X account of *Channel 3 Now News* (now deleted) was among the first to share the false name of the Southport suspect, though it may have been shared elsewhere too. Though the account had fewer than 3,300 followers at the time of the riots, it claimed to be a legitimate news outlet, which may have helped the claims spread, adding legitimacy to what was being said.[4]

10. While the suspect's name was falsely shared online, this was done alongside incorrect claims that the suspect had recently come to the UK on a small boat, or was Syrian. As we wrote in our fact check at the time, these claims were quickly rebutted by Merseyside Police.[5] Nevertheless, unrest broke out in Southport, with Merseyside Police reportedly saying people behind the violence had been fired up by social media posts. This view is shared by many, including by disinformation expert Marc Owen Jones[6] and the Guardian.[7]

11. Alongside what is being said on social media, a further worrying trend remains how this information is shared offline in messaging services, which are harder to track or regulate. The platform Telegram has been widely seen as contributing to the riots in the summer. Speaking to Full Fact following the riots, Professor Andrew Chadwick, Professor of Political Communication at Loughborough University and an expert in the spread of online misinformation, told Full Fact that false information soon ends up in personal messaging applications such as WhatsApp, "where news items circulate in smaller groups, often local,

---

[2] Full Fact, What role did misinformation play in riots after the Southport stabbings?, August 2024, https://fullfact.org/news/misinformation-southport-stabbings/

[3] Full Fact, What role did misinformation play in riots after the Southport stabbings?, August 2024, https://fullfact.org/news/misinformation-southport-stabbings

[4] Channel 3 Now has since removed the post with the false name and apologised.

[5] Full Fact, Incorrect name for Southport stabbings suspect circulates online, July 2024, https://fullfact.org/online/incorrect-name-southport-stabbings-suspect/

[6] Marc Owen Jones, X (Twitter thread), July 2024, https://x.com/marcowenjones/status/1818343641930514674

[7] The Guardian, Local. Left behind. Prey to populist politics? What the data tells us about the 2024 UK rioters, September 2024, https://www.theguardian.com/uk-news/2024/sep/25/local-left-behind-prey-to-populist-politics-data-2024-uk-rioters

perhaps family and friend networks", said Professor Chadwick. "There is always the risk they will have direct on-the-ground impacts in communities, and that's exactly what we saw," he said.[8]

12. The Institute for Strategic Dialogue has also issued similar concerns about Telegram's role in fueling the riots, claiming that "while Telegram lacks algorithmically-boosted exposure, it effectively operates as a safe space for extremists to coordinate activity and instigate violence."[9]

13. A recent report from the European Fact Checking Standards Network (EFCN), of which Full Fact is a member, shows that the fact checking community across Europe share the concern about Telegram's role in spreading misinformation. The report found that 84.9% of the fact-checking organisations who were surveyed by the EFCN were concerned about harmful disinformation on Telegram, with "75.8% agreeing that it plays a significant role in disseminating disinformation."[10]

How effective is the UK's regulatory and legislative framework on tackling these issues?

a) How effective will the Online Safety Act be in combatting harmful social media content?

14. The Online Safety Act (OSA) should have been a pivotal moment in the way the UK tackles the harms caused by misinformation. However, the final Act falls short of the former Government's original aim of making the UK "the safest place to be online."

15. There are currently no further plans to tackle the harms from online misinformation in the OSA and this continues to leave the public vulnerable and exposed to online harms. The only references to misinformation in the Act are about setting up the committee to advise Ofcom, and changes to Ofcom's media literacy policy.

16. The Act does not address health misinformation, which the Covid-19 pandemic demonstrated could be potentially harmful. It also does not set out any new provisions to tackle election disinformation (unless it is a foreign interference offence), nor misinformation that happens

---

[8] Full Fact, What role did misinformation play in riots after the Southport stabbings?, August 2024, https://fullfact.org/news/misinformation-southport-stabbings

[9] Institute for Strategic Dialogue, 'Total system collapse': Far-right Telegram network incites hate & violence after Southport stabbings, https://www.isdglobal.org/digital_dispatches/total-system-collapse-far-right-telegram-network-incites-accelerationist-violence-after-southport-stabbings/

[10] EFCN, Fact-checking and related Risk-Mitigation Measures for Disinformation in the Very Large Online Platforms: A systematic review of the implementation of big tech commitments to the EU Code of Practice on Disinformation, December 2024, https://efcsn.com/report-fact-checking-vlops-2024/

during ‘information incidents’ when information spreads quickly online, such as during terror attacks or during the August 2024 riots following the Southport murders. The OSA also does not extend to most harms from generative AI misinformation.

17. The riots across England illustrate the problem with focusing only on *illegal* content - in this case the Online Safety Act’s false communications offence. Very little of the misinformation circulating online in August 2024 related to the identity of the attacker would be categorised as a false communication, because it is very difficult to prove both intent to cause “physical or psychological harm” *and* definite prior knowledge that the information sent was false.

18. Specially in regards to misinformation, the Act introduces the requirement to create an Advisory Committee on Misinformation and Disinformation. At the time of drafting, the committee is still recruiting for members and will not be up and running until April 2025.[11]

19. However, even when implemented, the terms of reference set out for the creation of the committee don’t go far enough to combat misinformation and harmful social media content. We have recommended that the Committee should commence a review of the existing legislative and regulatory framework and make recommendations about any changes that might be needed to address harmful disinformation and misinformation effectively.

b) What more should be done to combat potentially harmful social media and AI content?

20. The previous government’s 2019 Online Harms White Paper, which had treated online disinformation and misinformation as a type of harm, had proposed that “companies will need to take proportionate and proactive measures to help users understand the nature and reliability of the information they are receiving, to minimise the spread of misleading and harmful disinformation and to increase the accessibility of trustworthy and varied news content.”[12]

21. However, the eventual Act is less ambitious than that. Other than the false communications and foreign interference offences, which are both likely to be very difficult to prosecute, wider misinformation provisions are limited to the Ofcom advisory committee on disinformation and misinformation, and specific media literacy powers.

---

[11] Ofcom, Advisory Committee on Disinformation and Misinformation, November 2024, https://www.ofcom.org.uk/about-ofcom/structure-and-leadership/advisory-committee-on-disinformation-and-misinformation/

[12] HM Government, Online Harms White Paper, April 2019, https://assets.publishing.service.gov.uk/media/605e60c6e90e07750810b439/Online_Harms_White_Paper_V2.pdf

22. Once the Online Safety Act has been implemented as drafted, an urgent review should take place to assess whether the Act can effectively combat the level of harm present on social media. As Full Fact has outlined, both in this submission and to the government directly, the harmful misinformation we've seen during Southport is only the tip of the iceberg.[13]

23. Regarding AI generated content, one specific area of interest to Full Fact relates to disclosure information. Many of the AI technologies currently being developed are focused on indirect disclosure tools. Indirect disclosure is a technical signal for defining whether a piece of media was created by AI. Direct disclosure is the other side of the coin: how these signals are then displayed to users alongside the content.

24. Of course, not all AI generated content will be misinformation or disinformation. However, it seems likely that those concerned with maintaining trust in information online will decide to directly disclose content that has been generated by AI. This may be a helpful step, but it will need to be backed up with investment in research that seeks to understand: how to present labels in a way that does not degrade trust in information that is not labelled but is nevertheless high quality or accurate; how to actively support users' understanding of the content they are consuming; and how to maintain users' privacy by not sharing identifiable information about individuals unnecessarily.

25. Another pertinent example that needs to be addressed is health misinformation. When it spreads it can introduce confusion, make it harder to distinguish truth from falsity, and distract from or undermine medical evidence. This was exemplified during the Covid-19 pandemic. Full Fact saw in real time the risks that can come when people do not understand how a virus is transmitted, or how to protect themselves from it.

26. Later in the pandemic we saw confusion and concern about the safety of the vaccine. For example, the initial lack of information about the safety of vaccines for pregnant women had lasting effects, with both women and vaccination centres receiving mixed messages, and pregnant women not being given second doses or thinking they needed to start their course again.

27. Ideally the Online Safety Act would be amended specifically to bring harmful health misinformation into scope, but until such a time as that happens, the government must encourage platforms to take immediate responsibility, by establishing clear and transparent policies for the effective treatment of harmful health misinformation, and applying those policies consistently.

---

[13] Full Fact, The Online Safety Act and Misinformation: What you need to know, https://fullfact.org/policy/online-safety-act/

28. There are understandable and justified concerns that tackling online misinformation will come at the expense of freedom of speech. But in our 2023 report on health misinformation, we argued that it is possible to balance the two.[14] There are content neutral methods available to regulators to reduce the harm from misinformation, which means that removing content should rarely be necessary. These include promoting good information, such as the Covid-19 information centres on Facebook; having initiatives which introduce friction, such as read-before-you-share prompts introduced by Twitter (now X); and highlighting independent fact checking. This principle of finding the right balance should be central to any new legislation or amendments, and should be front of mind for Ofcom.

c) What role do Ofcom, and the National Security Online Information Team play in preventing the spread of harmful and false content online?

29. Ofcom is limited in their ability to regulate the Online Safety Act based on the scope of the Act. However, the government could better support Ofcom by learning from the experiences of other countries and should consider whether it is investing wisely in a cohesive regime to tackle the full range of harmful content that lies outside the scope of current legislation.

30. Ofcom's Advisory Committee on Disinformation and Misinformation should play an important role in preventing the spread of misinformation. As a committee of experts in this issue, they will play a vital role in assessing whether the Online Safety Act is sufficiently able to tackle false content.

31. Full Fact has produced a number of recommendations for the Advisory Committee, but we believe one of its first tasks should be to undertake research itself, or call for and be able to commission research into misinformation and disinformation on regulated services, and the effect that it has on the public (which the Committee should then advise on). As well as producing an assessment of whether there should be a dedicated Ofcom code of practice on misinformation and disinformation.

32. Ofcom has a further role to play when it comes to information threats, and this should be made more clear in the text. ("Services should also remain live to emerging information threats, with the flexibility to quickly and robustly respond, and minimise the damaging effects on users, particularly vulnerable groups"). We believe that the regulator is the logical coordinator of a centralised framework for information incidents.

33. Full Fact has developed such a draft framework and we commend the approach to government.[15] Ofcom's role in responding to information incidents should include introducing

[14] Full Fact, 'Online health misinformation in the UK', April 2023, https://fullfact.org/media/uploads/online_health_misinformation_in_the_uk_full_fact.pdf.
[15] Full Fact, Policy Incidents Framework, https://fullfact.org/policy/incidentframework/

a system whereby emerging incidents can be publicly reported, and different actors such as fact checkers, news organisations, community representation groups and service providers can request that Ofcom bring together a response group to discuss severity and response.

Which bodies should be held accountable for the spread of misinformation, disinformation and harmful content as a result of social media and search engines' use of algorithms and AI?

34. Social media companies should take the ultimate responsibility for the content that exists on their platforms. While the government can try and enforce good practices and extend regulations, platforms retain the final responsibility.

35. In parallel to this, media literacy duties which have been extended to Ofcom should also be seen as a vital way to equip the public with the tools to determine what is true and what is false.

18 December 2024