REPHRAIN

Protecting citizens online

REPHRAIN response to the Science, Innovation and Technology Committee's Call for Evidence: Social Media, Misinformation and Harmful Algorithms

Specifically, Prof. Stephan Lewandowsky (Chair in Cognitive Pyschological Science, University of Bristol) and Josie Curtis (Policy Engagement Associate, REPHRAIN, University of Bristol) contributed to the formulation of this response.

December 2024

UK Research and Innovation · University of Bristol · UCL · The University of Edinburgh · King's College London · University of Bath

REPHRAIN response to the Science, Innovation and Technology Committee's Call for Evidence on 'Social media, misinformation and harmful algorithms'

Thank you for the opportunity to provide our response to this call for evidence. We are writing on behalf of REPHRAIN, the National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online. REPHRAIN is the UK's world-leading interdisciplinary community focused on the protection of citizens online.

Led by the University of Bristol and partnered with University College London, King's College London, the University of Edinburgh, and the University of Bath, REPHRAIN unites experts across disciplines such as Computer Science, Law, Psychology, and Public Policy to explore how to keep people safe online while enabling full participation in digital technologies. Announced by UKRI in October 2020, REPHRAIN now has over 100 experts from 23 UK institutions, working across 50+ research projects to address our missions:

- Delivering privacy at scale while mitigating its misuse to inflict harms
- Minimising harms while maximising benefits from a sharing-driven digital economy
- Balancing individual agency vs. social good.

This is a submission from the REPHRAIN centre. Specifically, **Prof. Stephan Lewandowsky** (Chair in Cognitive Psychology, School of Psychological Science, University of Bristol) and **Josie Curtis** (Policy Engagement Associate, REPHRAIN, University of Bristol) contributed to the formulation of this response.

To what extent do the business models of social media companies, search engines and others encourage the spread of harmful content, and contribute to wider social harms?

The business models of social media platforms are fundamentally rooted in the attention economy, where user engagement directly drives advertising revenue. To maximise engagement, these platforms actively curate and amplify content through algorithmic systems designed to capture and retain attention. These algorithms prioritise content that elicits strong emotional responses – such as outrage, anger, or surprise – because such reactions keep users engaged for longer (McLoughlin *et al.*, 2024).

These dynamics often amplify harmful content, including misinformation, because such content is typically novel and emotionally provocative. As Lewandowsky and Kozyreva (2022) observe, the attention economy incentivises the spread of divisive and manipulative material, as the very algorithms that promote user engagement inadvertently undermine trust and public discourse.

For example, in cases of public unrest or riots, misinformation that exploits stereotypes or entrenched biases often goes viral, further escalating tensions. This was the case for the 2024 summer riots in the UK, a direct consequence of platforms optimising for content likely to trigger strong emotional reactions and engagement (TIME, 2024). These algorithmic practices not only spread harmful content but also entrench polarisation and erode democratic norms (Lewandowsky and Kozyreva, 2022).

Thus, the current business models of these platforms are at odds with the principles of a healthy democracy. Addressing these harms requires systemic changes, such as implementing transparency measures, mandating algorithmic accountability, and incentivising platforms to prioritise societal well-being over short-term engagement metrics. Regulatory frameworks must

encourage platforms to adopt ethical design principles and reduce the amplification of harmful content.

What roles do algorithms and generative AI play in the spread of misinformation, disinformation and harmful content?

Generative AI and algorithms play critical roles in the dissemination of misinformation and harmful content. When paired, these technologies create a potent feedback loop that accelerates the spread of false narratives. Algorithms governing online platforms prioritise engaging content, which amplifies AI-generated misinformation and creates significant risks to public trust and discourse.

Our research highlights that generative AI can be employed to manipulate individuals by creating personalised, persuasive messages tailored to their personality traits. Simchon, Edwards, and Lewandowsky (2024) demonstrated that AI-generated messages could more effectively influence people when messages were targeted at the recipients' psychological profiles. This has significant implications for how disinformation campaigns could exploit these capabilities to influence public opinion or behaviours at scale.

In addition, the European Commission's report *Technology and Democracy* (2020), led by Lewandowsky and Smillie, underscores how technological advancements amplify misinformation when combined with algorithmic systems designed to prioritise engagement over accuracy. These systems often promote sensationalist or polarising content, fuelling the rapid spread of misinformation.

The combined power of generative AI and algorithms underscores the urgent need for robust governance mechanisms and ethical guidelines to mitigate the risks associated with their misuse. Without intervention, these technologies will continue to disrupt public discourse and erode trust in information ecosystems.

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

The UK's current regulatory and legislative framework, including the Online Safety Act (2023), Ofcom, and the National Security Online Information Team, shows limitations in effectively addressing the challenges posed by harmful online content.

The riots provide a concerning example of this framework's ineffectiveness. Harmful and misleading content related to the riots was widely disseminated on social media platforms, often remaining online for extended periods before being flagged or removed (BBC Bitesize, 2024). This allowed the content to reach and mobilise significant audiences, exacerbating tensions and undermining public safety.

The Online Safety Act has inadequately addressed misinformation for various reasons. First, Ofcom's phased approach to rolling out the Act prioritises measures against illegal harms, such as child abuse content and terrorist material, leaving pervasive societal harms from misinformation largely unaddressed in the early phases. Platforms are not currently mandated to comprehensively mitigate misinformation unless it directly overlaps with illegal activities. This has allowed misinformation to persist without systemic accountability measures in place.

Another of the Act's shortcomings is that it does not fully incorporate provisions for tackling harmful misinformation that is not outright false but creates false impressions. This is evidenced in the 'False communications offence' (Section 179), which states that an individual

commits this offence if they send communication that they "know to be false", which is intended to "cause non-trivial psychological or physical harm to a likely audience". This approach has led to large gaps in how misinformation spreads and is managed online. For instance, misinformation that manipulates emotions or biases, without technically being false, is still widespread, especially when platforms rely on algorithmic amplification based on user engagement rather than content accuracy (Van der Linden, Ecker and Lewandowsky, 2024).

Furthermore, the Act places too much responsibility on individuals for spreading misinformation, with the 'False communications offence' focusing on holding individuals accountable. This ignores the role of platforms in curating, amplifying, and spreading such content. And while the Act includes an advisory committee on misinformation to tackle the problem, this approach is insufficient. The committee can only provide advice, not enforce actions, and lacks the power to mandate platform responsibility.

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?

To effectively tackle the spread of misinformation, disinformation, and harmful content online, accountability must be distributed across multiple levels, with platforms bearing primary responsibility due to their pivotal role in content dissemination.

Accountability of platforms

Online platforms should be held principally accountable because they do not merely serve as passive conduits for user-generated content; rather, they actively curate and amplify information through algorithmic systems designed to maximise user engagement. As mentioned, research (e.g. McLoughlin *et al.*, 2024) has demonstrated that these algorithms prioritise emotionally charged content, which contributes to the proliferation of harmful and misleading content.

Platforms profit directly from these processes, creating an economic incentive to maintain systems that exacerbate the issue. Yet, historical evidence highlights their ability to reduce misinformation when motivated to do so. For example, prior to the 2020 U.S. Presidential Election, Facebook adjusted its algorithms to minimise the spread of election-related misinformation. However, it is argued that this measure happened too late (in October 2020), and if they had acted earlier, they could have prevented an estimated 10.1 billion views for top-performing pages that repeatedly shared misinformation (Avaaz, 2021). This demonstrates that platforms have the technical capability to act responsibly, but they often choose not to for commercial reasons.

Accountability of individuals

While platforms play a central role, individuals must also be held accountable for sharing harmful content, particularly in cases where such actions violate existing laws. However, platform accountability is paramount because they facilitate the amplification of harmful content on a scale that individual actors could not achieve independently.

Recommendations

1. **Expand the scope of misinformation in the Online Safety Act:** Currently, the Online Safety Act includes provisions on harmful content, but it falls short in addressing

misinformation comprehensively, especially when it comes to how algorithms amplify false content, or content intended to create false impressions. This would include setting clear standards for detecting and removing misinformation proactively, not just relying on reactive measures such as media literacy campaigns and advisory committees.

2. **Stronger regulatory oversight:** Ofcom should hold platforms accountable for the content amplified through their algorithms. Platforms should be required to provide transparency reports on algorithmic decisions, content moderation practices, and the prevalence of flagged and removed content, whilst they should also be mandated to design algorithms that actively minimise the spread of harmful content or misinformation. Independent audits of algorithms should be conducted regularly to assess their societal impact.
3. **Legal and financial penalties:** Ofcom must enforce penalties for platforms that fail to act on harmful content or misinformation, or whose algorithms are found to amplify such content.

References

Avaaz (2021) *Facebook: From Election to Insurrection*, Available at: https://secure.avaaz.org/campaign/en/facebook_election_insurrection/ (Accessed: 13/12/24).

BBC Bitesize (2024) *Timeline of how online misinformation fuelled UK riots*. Available at: https://www.bbc.co.uk/bitesize/articles/zshjs82 (Accessed: 18/12/2024).

European Commission: Joint Research Centre, Lewandowsky, S., Smillie, L., Garcia, D., Hertwig, R. *et al*., (2020) *Technology and Democracy*, Available at: https://data.europa.eu/doi/10.2760/709177.

Lewandowsky, S. and Kozyreva, A. (2022) *Algorithms, lies, and social media*. Available at: https://www.niemanlab.org/2022/04/algorithms-lies-and-social-media/ (Accessed: 13/12/2024).

McLoughlin, K. *et al.* (2024) 'Misinformation exploits outrage to spread online', *Science,* 386, pp. 991–996. Available at: 10.1126/science.adl2829.

Simchon, A., Edwards, M., and Lewandowsky, S. (2024) 'The persuasive effects of political microtargeting in the age of generative artificial intelligence', *PNAS Nexus*,3:2, Available at: 10.1093/pnasnexus/pgae035.

TIME (2024) *Online Misinformation Stoked Anti-Migrant Riots in Britain*. Available at: https://time.com/7007925/misinformation-violence-riots-britain/ (Accessed: 18/12/2024).

Van der Linden, S., Ecker, U. and Lewandowsky, S. (2024) *Misinformation is a threat to society – let's not pretend otherwise*. Available at: https://blogs.lse.ac.uk/impactofsocialsciences/2024/10/08/misinformation-is-a-threat-to-society-lets-not-pretend-otherwise/ (Accessed: 13/12/24).