Written Evidence submitted by the Free Speech Union (SMH0059)

SCIENCE, INNOVATION AND TECHNOLOGY COMMITTEE INQUIRY INTO SOCIAL MEDIA, MISINFORMATION AND HARMFUL ALGORITHMS

SUMMARY

The Free Speech Union (FSU) is a non-partisan, mass-membership, public-interest body that stands up for the speech rights of its members and campaigns for free speech more widely.

We have serious concerns about the framing of this inquiry. While we have structured our response to align with the consultation's questions, we believe these questions are grounded in assumptions and premises that are neither neutral nor sufficiently supported by evidence.

They appear to presuppose that concepts such as 'misinformation' and 'algorithms' are causing a degree of harm far exceeding what is empirically demonstrated, and imply an imperative for intensified regulatory or legislative action.

We reject these premises. In our response, we provide evidence to demonstrate that many of the fears underlying this inquiry are exaggerated or unfounded.

Additionally, we caution against relying solely on regulation or legislation as blunt tools to address what is, at its core, a nuanced and complex social phenomenon – and one which may not even constitute a significant problem.

We draw the Committee's attention to a growing body of research suggesting that such concerns amount to a 'moral panic', led by 'moral entrepreneurs' who display an unduly pessimistic view of their fellow citizens.

Weak evidence, flawed assumptions and questionable motivations have historically led to poor legislation, and we fear that this inquiry risks following a similar path.

To that end, we propose a set of recommendations that move away from a censorious approach. Instead, we advocate for strategies which empower users and are grounded in evidence, particularly where these approaches have been shown to mitigate the spread of misinformation and reduce social discord. Specifically, we recommend that the Committee prioritize the following empirically supported solutions:

- Avoid further over-correction
- Counterspeech
- Digital literacy
- Innovation
- Transparency

We urge the Committee to scrutinise the weak evidential basis for measures that curtail individual liberties and to focus on constructive, proportionate and proven interventions.

1. **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 most major social media platforms are fundamentally driven by advertising revenue. This creates two key dynamics: first, the imperative to collect extensive user data to deliver highly targeted advertising; and second, the financial incentive to funnel users toward content that maximises engagement and, ultimately, clicks on high-value advertisements.

However, user engagement alone is not sufficient for these platforms to thrive – advertising revenue is the cornerstone of their success.

In debates surrounding government and policy, it is too often simply assumed that the so-called 'attention economy' hinges exclusively on user-generated content, where a 'publish or perish' ethos inadvertently amplifies sensationalist or polarising clickbait.

Much less discussed is the extent to which this business model is reliant on content moderation practices such as censorship, shadow banning and content takedowns. This imperative arises from two distinct but interconnected pressures: the economic and the regulatory.

In recent years, major social media companies and search engines have implemented increasingly sophisticated, AI-based moderation systems.

This is, in part, a response to regulatory frameworks such as the Digital Services Act in the European Union, which mandates swift removal of 'inappropriate' content. Beyond compliance, there is also a strong economic rationale: moderation helps create a 'brand-safe' environment by removing lawful but potentially 'harmful' or 'offensive' content, shielding advertisers from associations with controversial material that could damage their reputation.

But in setting themselves up as arbiters of what is harmless (i.e., true, accurate or credible) and what is harmful (false, inaccurate or misleading), social media companies are taking a static 'snapshot' of important social debates, controversies and arguments that remain in process, unsettled and as yet unresolved.

There is an argument, then, for saying that in prematurely deciding on what to censor, social media platforms are hindering public debate and slowing progress towards the truth – or at least, towards social/medical/scientific consensus – by restricting information that may one day be seen as valid, while at the same time allowing certain types of information that may one day be considered 'harmful' to spread across platforms unopposed.

In other words, the line between "harmful content" and, say, "credible theory with explanatory power" is often defined only by the passage of time. This was eloquently expressed by the former Supreme Court judge Lord Sumption in an article in the *Spectator* about the shortcomings of the Online Safety Act:

> All statements of fact or opinion are provisional. They reflect the current state of knowledge and experience. But knowledge and experience are not closed or immutable categories. They are inherently liable to change. Once upon a time, the scientific consensus was that the sun moved around the Earth and that blood did not circulate around the body. These propositions were refuted only because orthodoxy was challenged by people once thought to be dangerous heretics.

In this context, there is a legitimate concern that platforms, by over-zealously moderating, inadvertently stifle the very debates necessary for societal progress and truth discovery.

The 'Lab-Leak Theory'

The social trajectory of one of the most prominent pandemic-era "conspiracy theories" – the lab-leak theory – demonstrates the complexities of distinguishing misinformation from plausible hypotheses. The theory holds that the SARS-CoV-2 virus was manufactured in – and then accidentally leaked from – a laboratory in Wuhan, China.

In October 2020, a Full Fact report, Conspiracy Beliefs, observed: "The Covid-19 pandemic also brought its own suite of conspiracy theories. The claim that 'SARS-CoV-2 was made in a lab' was believed by 30% of respondents in the UK, and almost as many (29%) in the US."

Similarly, the Institute for Strategic Dialogue (ISD) described the lab-leak hypothesis in April 2020 as a "popular conspiracy theory" that had been "repeatedly highlighted" by "conservative media".

In response, social media platforms aggressively censored discussions around the lab-leak theory. These included Twitter (now X) which also suspended the accounts of medical professionals who discussed it – as did YouTube. Meanwhile, Meta reported removing more than 20 million Covid-19-related posts from Facebook and Instagram by mid-2021.

This censorship was all the more outrageous, given that senior scientific gatekeepers themselves believed the lab-leak theory was probably true, but sought to suppress it lest it damage the relationship between Western scientists and their Chinese counterparts.

In February 2020, Sir Jeremy Farrar, the then Director of the Wellcome Trust, wrote to Dr Anthony Fauci, a public-health adviser to every US president since Ronald Reagan and Dr Francis Collins, who led the Human Genome Project. Sir Jeremy warned the two that Covid may have evolved in human tissue within a low-security laboratory, describing Wuhan research as akin to the "Wild West". In response, Dr Collins cautioned that public debate on the matter could harm "international harmony". In other words, the dismissal of this hypothesis as 'misinformation' was itself misinformation.

By August 2021, with the zoonotic theory of the pandemic's origins palpably failing to account for the facts, the head of the WHO's origins investigation team admitted to having been pressured into dismissing the lab-leak hypothesis to avoid political friction with China. Two months later, WHO Director-General Dr Tedros Adhanom Ghebreyesus wrote in *Science*: "Lab hypotheses must be carefully examined," emphasising that "a lab accident cannot be ruled out".

In 2022, a US Senate Committee released a comprehensive 304-page report after 18 months of investigation, concluding: "The preponderance of circumstantial evidence supports an unintentional research-related incident."

This illustrates how describing a plausible hypothesis as 'misinformation' for political reasons discredits senior scientists, public health authorities and large social media platforms, sowing distrust among the public. The danger, then, of encouraging social media platforms and their regulators to censor more content on the grounds that it's 'misinformation' is that you will end up making the public more sceptical of 'official' sources of information, not less.

Covid and Masks

During the Covid lockdowns, social media companies aligned their content policies with government and health authority directives. The UK Government, for instance, commended YouTube for censoring content that "contradicts the World Health Organization" (WHO). Facebook banned content that questioned mask efficacy.

Twitter flagged or removed tweets under its "Covid-19 misinformation" rules, which prohibited claims that masks were ineffective or harmful. YouTube similarly updated its policies, removing videos that contradicted official mask guidance and suspending repeat offenders.

Yet over time, the WHO itself reversed several key positions. Initially, it echoed Chinese claims that sustained human-to-human transmission was not happening, dismissing concerns from doctors in Wuhan as misinformation. On masks, the WHO and the British Government both advised against mask use for uninfected individuals until June 2020, citing a "false sense of security".

These reversals highlight a critical issue: open disagreement with prevailing authorities does not inherently constitute harmful misinformation – nor could it when those authorities start out saying one thing and then say the opposite. By censoring information based on a fixed 'snapshot' view of what is truth, social media platforms risk unintentionally shielding flawed policies from scrutiny. In doing so, they may delay the development of improved policies and more effective responses.

The 'Infodemic'

Misinformation in the UK has been shown to constitute only around 0.1 per cent of the population's news diet (Acerbi et al, 2022). Yet narratives about misinformation during the pandemic – ranging from obviously false claims (e.g., vaccines contain 5G chips to allow Bill Gates to mind-control the population) to the unfortunate consequences of incomplete and temporal scientific knowledge – often painted a far more exaggerated picture.

There also existed considerable misinformation about misinformation itself, chiefly in perpetuating the narrative of an 'infodemic' of vast quantities of misleading information pertaining to the outbreak. For instance, a persistent piece of misinformation attributes to Donald Trump the suggestion that people should drink bleach to kill Covid-19. This narrative was amplified by mainstream media and political opponents such as Joe Biden. (Austin American-Statesman, 2020). The US federal agency Centers for Disease Control and

Prevention contributed to this perception when it reported uncritically on a survey claiming that 4 per cent of Americans had tried "drinking or gargling diluted bleach solutions, soap water, and other cleaning and disinfectant solutions" (Gharpure et al, 2020). Media outlets widely picked up on the claim, giving it further traction (Reuters, 2020).

However, a closer analysis revealed significant flaws. At least 80-90 per cent of those surveyed were identified as "problematic respondents" – individuals who also made impossible claims, such as "recently having had a fatal heart attack" or "eating concrete for its iron content". Once these unreliable responses were discounted, there was no credible evidence of household cleaner ingestion occurring at meaningful levels.

The real-world effects of vaccine misinformation provide another revealing example. Exposure to false information about vaccines has been shown to have minimal, if any, impact on people's vaccination decisions (Litman et al, 2021). Some studies even suggest a small positive correlation between exposure to misinformation and vaccine uptake, though this effect remains weak (Saint Laurent et al, 2022).

Despite the significant attention devoted to the issue by governments and the media, empirical evidence consistently undermines the 'infodemic' narrative. The prominence of this idea may owe more to the pandemic context itself: the existence of a highly transmissible virus seemingly lent weight to the notion that people's minds can be infected by ideas in the same way bodies are infected with viruses (Altay et al, 2023).

This reductionist view of human communication offers an easy method of analysis and a convenient and sensationalist narrative – as well as allowing states to appear decisive in the face of crisis, providing a justification for policy interventions and censorship. But it is inconsistent with the empirical evidence about how people work (Simon & Camargo, 2023). It, too, constitutes a form of misinformation.

Puberty Blockers and Gender Dysphoria

In December 2024, the UK Government announced an indefinite ban on puberty blockers following an expert review that warned the drugs posed an "unacceptable safety risk" to children. The decision was grounded in evidence. Yet it came too late for many whose lives had already been irrevocably altered – victims of an untested medical approach where dissenting voices were silenced or punished for far too long.

For years, social media platforms censored criticism of the 'affirmative' approach to treating gender dysphoria in children and young people. Content questioning this model – whether from detransitioners, concerned parents, teachers or gender-critical campaigners – was frequently suppressed. Researchers and medical professionals also reported being censored or having their work removed under activist pressure. Posts were labelled "transphobic", accounts were suspended and platforms positioned themselves as arbiters of acceptable discourse.

The publication of the Cass Report in April 2024 marked a turning point in the debate. Led by eminent paediatrician Dr Hilary Cass, the report was the most thorough and systematic assessment of the "transgender care" pathway provided by the NHS-funded Gender Identity Development Service at the Tavistock and Portman Trust.

Dr Cass and her team delivered the damning conclusion that there was no robust evidence base to support the affirmative treatment model, including the prescribing of puberty blockers and cross-sex hormones to children.

The exposure of this medical scandal came at significant personal and professional cost to those who dared challenge the prevailing orthodoxy. Their commitment to medical ethics, evidence-led treatment and child safeguarding often met with hostility and suppression – much of it enabled by social media platforms eager to appease corporate or activist interests. And such social media censorship arguably held up scientific and medical progress by stifling important debates about the most appropriate treatments for gender dysphoria in young people.

Interestingly, the debate has now come full circle. In an article published in *The Times* on 13th December 2024, Dr Hilary Cass herself addressed this point, describing Labour MPs who oppose the ban as suffering from a "misunderstanding" of medical evidence and accusing them of spreading "misinformation". One of the main arguments in favour of the 'affirmative' approach is that refusing to affirm a gender-confused adolescent's self-diagnosis led to an increased suicide risk – a claim that itself turned out to be misinformation.

The irony is stark. The solution is not, of course, to censor those MPs – just as dissenting voices in the past should not have been silenced. The underlying issue lies in the flawed imposition of intermediary liability, which places tech platforms in the untenable position of arbiters of truth.

Ultimately, this case exemplifies the dangers of outsourcing complex and contested debates to such platforms, as well as expecting regulators to get these judgements right. Social media censorship did not protect vulnerable young people; it delayed the very conversations that might have protected them.

The dangers of intermediary liability

The current regulatory model under which platforms operate inherently favours complainants at the expense of online speakers. Faced with the risk of fines and regulatory intervention, platforms often default to removing content as the simplest and most economical course of action. In the absence of any recourse for a user who has been unfairly censored, this action eliminates legal risks for the platform, without the hassle of hiring a lawyer or even a minimum-wage employee to evaluate the acceptability of a post (Keller, 2018).

Granted, the law underpinning a complaint might allow for ambiguities as to how a given post is interpreted – e.g., edgy humour vs defamation; heated disagreement vs hate speech. However, platforms under regulatory duties that are focused on safety with little or no countervailing free speech obligations, cannot afford such flexibility and have no incentive to consider the nuances. Once they become aware of any risk, the rational course is to err on the side of removal.

This approach leads to widespread frustration and resentment among users. Content may disappear seemingly without cause, fostering perceptions of "censorship based entirely on unspecified ideological objection to the message or on the perceived identity and political viewpoint of the speaker", as former California Governor Pete Wilson remarked in a lawsuit regarding YouTube's taking down of videos by Right-leaning outlet PragerU. These included

content from Pulitzer Prize winners and a lecture by the renowned attorney Alan Dershowitz on the founding of Israel (Illinois Review, 2017).

Even platforms themselves acknowledge the inherent flaws in their moderation systems. A 2021 report into Facebook's systems revealed that the platform receives approximately three million reports a day. In roughly one in 10 cases – or 300,000 times daily – the wrong decision is made, whether by automated systems or human moderators (Koetsier, 2024). This means that hundreds of thousands of users potentially have their fundamental right to freedom of expression compromised every day.

To compound matters, users are often left in the dark as to why their content was removed. In many instances, platforms provide no explanation at all. This lack of transparency reinforces feelings of unfair treatment and deepens frustration, particularly when users discover that appealing the decision is nearly impossible due to limited or absent human review processes.

The potential for even deeper resentment when it is perceived that this injustice has been perpetuated at the behest of governmental or other authorities – who, as shown above, are far from immune from disseminating information which later turns out to be false – cannot be understated. This, in turn, undermines faith in institutions which by design ought to carry weight when it comes to the public's epistemic discernment. By alienating people, it also pushes them towards potentially more problematic avenues of expression, as will be discussed in the section about Question 6.

In any event, the exclusion of significant portions of the population from mainstream political discourse, whether through outright banning or because they are too fearful to speak their minds because of the possible consequences, represents a grave risk to political stability in Western democracies – driving them into the hands of marginal parties and other movements (Przeworski, 2019).

It is a risky and surely counter-productive strategy to force on tech platforms policies that foster such sentiments.

Ultimately, placing social media companies in the position of arbiters of truth – under threat of punitive regulatory action – creates a business model that suppresses dialogue, stifles dissent, and fosters resentment. Rather than protecting public safety or fostering meaningful debate, such measures risk deepening societal divides and undermining democratic norms.

Conclusion

These examples illustrate the dangers of labelling dissenting views as 'misinformation' and suppressing content under the guise of public safety. Social media companies, under pressure from governments, regulators and advertisers, risk acting as gatekeepers of truth based on incomplete or provisional knowledge – or simply censor content they believe to be true so as to curry favour with regulators.

Scientific and social consensus evolves over time, requiring open debate, scrutiny and the challenging of orthodoxy. Suppressing these can only obstruct progress, undermine trust in institutions and reinforce flawed policies. A more measured approach, focused on empowering users through transparency and critical engagement, is essential to balancing the mitigation of harm with the fundamental right to free expression.

2. **How do social media companies and search engines use algorithms to rank content, how does this reflect their business models, and how does it play into the spread of misinformation, disinformation and harmful content?**

Search engines and social media platforms rely on ranking algorithms to highlight the most relevant and engaging content for users. This practice aligns with their business models: by providing a valuable user experience, platforms maximize advertising revenue.

It is important, however, to distinguish between misinformation and disinformation. The former is shared by individuals who believe it to be true – sometimes correctly, as later developments reveal. In contrast, disinformation refers to false or misleading content deliberately created and shared with the intent to manipulate, mislead or cause harm. This distinction is critical because people spreading disinformation are far likelier to exploit ranking algorithms than those inadvertently sharing misinformation.

Arguably, then, it is more precise to frame the issue not as algorithms "playing into the spread of misinformation and disinformation", but as "bad-faith actors manipulating algorithms to amplify disinformation". While subtle, this reformulation underscores that the algorithms themselves are not inherently problematic; rather, specific aspects of their functioning can be weaponised by malicious actors (e.g., Waissbluth et al, 2022).

For instance, disinformation actors deploy various techniques to manipulate ranking systems and amplify harmful content. These include:

- Keyword stuffing and link bombing, where networks of mutually reinforcing links are created to increase discoverability in search engine results (Tucker et al., 2018; Bradshaw, 2019).
- Exploiting search engine personalisation to target specific demographics with false information (Makhortykh et al., 2020).
- Manipulating autocomplete functions to suggest preferred search terms when users enter related queries (Wang et al, 2018).
- Utilising advanced SEO (Search Engine Optimization) methods to ensure disinformation ranks high in search results.
- Seeding malicious content so that it is indexed by search engines, thereby increasing the 'toxicity' of the information landscape (Invernizzi et al, 2012).

The point here is that if the manipulation of search engine algorithms is a key tactic used by those spreading disinformation – often in ways that are difficult for users to detect – the most sensible response is not regulatory oversight that risks tightening existing filtering systems in ways that will further curtail online freedom of expression. Rather, it lies in a renewed focus on enhancing the ability of social media users to spot false content via digital/media literacy initiatives.

Thus, a more effective response lies in empowering users to critically engage with online content. Researchers have long emphasised the importance of enhancing digital and media literacy to help people identify disinformation (Gamage et al., 2022; Shu et al., 2020; Hameleers et al., 2021). Improving the ability to spot false or manipulated content is a practical and achievable solution to mitigate the impact of disinformation without undermining free expression. As Joyner et al., 2023 argue, encouraging users to question,

verify, and evaluate the accuracy of information – and fostering their ability to do so – remains key.

It is also worth noting that algorithms struggle to evaluate the accuracy of content on complex or emerging topics where no clear expert consensus has been established. In such cases, the role of digitally literate users becomes even more crucial. By equipping them with the tools to critically assess information, platforms and policymakers can strike a balance between limiting the spread of disinformation and safeguarding the principles of open discourse.

3. What role do generative artificial intelligence (AI) and large language models (LLMs) play in the creation and spread of misinformation, disinformation and harmful content?

Generative AI is increasingly utilised by disinformation networks. Russian campaigns, for example, have deployed AI to recycle and rewrite genuine news items, posting them on superficially plausible sites populated with real stories. Within this camouflage of credible material, disinformation is harder for users to identify and dismiss (Warren and Linvill, 2024).

While nation-state actors currently lead in these techniques, it is far from inconceivable that other threat groups – transnational terrorists, domestic extremists or political organizations – could adopt similar strategies. However, the relative sophistication required to use AI means that for now such methods remain more prevalent among state-backed campaigns.

AI is also being used to manipulate audiovisual content (Onome, 2024). For instance, AI-generated voiceovers allow propaganda videos to be disseminated in target languages – as seen in both the Ukraine war and recent U.S. presidential campaigns (Marquardt et al, 2024).

In communication theory, the perceived credibility of a message often depends on the messenger's authenticity and authority. This underpins the use of deepfake videos, particularly those depicting politicians, as tools for spreading disinformation (Moreno, 2024). While current deepfake technology remains relatively identifiable with effort, it would be unwise to assume this will remain the case. Advances in AI suggest that future audiovisual representations may become indistinguishable from reality without technical assistance.

Despite these developments, the actual impact of AI-generated disinformation remains highly questionable. A major study into "fake news" exposure during the 2016 US election demonstrated that false content was overwhelmingly concentrated within small, pre-existing communities – for example, politically engaged Republican voters aged 65 and over (Grinberg et al, 2019). The stereotype of the 'Facebook MAGA boomer' sharing such material, while not wholly unfounded, highlights the limited reach of disinformation beyond these circles (Schroeder & Jungherr, 2021). Indeed, there is little evidence to suggest that such content significantly influenced voting behaviour (Eady et al, 2023).

Similarly, the UK Intelligence and Security Committee found evidence of Russian interference in both the 2014 Scottish independence referendum and the 2016 Brexit referendum. However, their findings indicated that these attempts had minimal, if any, effect on public opinion (Intelligence Select Committee, 2021).

Given the substantial effort invested in producing such propaganda, its limited immediate impact suggests that other motivations for its dissemination may exist – an issue addressed later in this report.

At a more basic level, generative AI is used to create images accompanying posts labelled as "harmful" – whether illegal or not – by individuals acting independently. The Facebook Oversight Board is investigating cases involving such images, particularly in connection with the summer riots in the UK (Oversight Board, 2024).

Here, the images often serve as low-cost, readily available illustrations, allowing posters to exploit algorithmic boosts given to multimedia content.

However, these tools are not inherently malign. AI-generated content also serves perfectly legitimate purposes, even when controversial. For instance, Amnesty International faced backlash for using AI images in its reports on the 2021 Colombian unrest (Taylor, 2023). Although Amnesty clarified that its intent was not to mislead, critics argued that fake imagery risked undermining the credibility of its findings and depriving photojournalists – who risked personal safety to document the protests – of deserved recognition and revenue.

Less controversially, generative AI allows small businesses and casual users to produce affordable graphics for advertising or creative expression – applications that might displace professional graphic artists but democratise access to design capabilities (Growcoot, 2023).

All of this underscores the point that, given the comparative newness of the technology, its wide range of sometimes surprising uses, and the ethical debates it provokes, any regulatory action seeking to curtail its use is likely to be premature and heavy-handed, allowing insufficient space for the nuances and social implications to be properly explored.

Finally, while generative AI and large language models (LLMs) facilitate the creation of false or misleading content, the most significant challenge for those seeking to spread disinformation remains the distribution of the material (MacCarthy, 2021). Ensuring disinformation reaches a broad audience requires significant resources and infrastructure, and AI tools currently do little to provide them

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

Despite Home Secretary Yvette Cooper's claim that social media platforms "put rocket boosters" under posts encouraging unrest, the precise role of social media algorithms in the UK riots of summer 2024 remains uncertain. While it is tempting to attribute a direct causal link between online activity and offline violence, the evidence suggests a far more nuanced relationship.

It is true that algorithms amplify highly emotive or sensational material, which in this case might have included posts about unrest. During the first week of August 2024, analysis of UK TikTok data by researchers at the University of Liverpool showed that "#riots" and "#riot" were the two most popular hashtags, with 10,000 and 5,000 posts respectively, coinciding with the Southport riots in England and Northern Ireland. Yet, while such platforms may facilitate the rapid spread of content, the researchers caution against assuming that algorithms alone can explain the complex interplay of factors behind civil unrest –

among them, traditional media coverage, political rhetoric and local conditions such as high levels of deprivation.

Studies from other contexts, meanwhile, provide mixed evidence about the influence of social media on unrest.

A study examining protests in 16 countries during the Arab Spring of the early 2010s found that tweets with protest-related hashtags were associated with an increase in protests the following day. Acemoglu et al. (2018) analysed the same topic through the number of Twitter posts with keywords linked to Tahrir Square, as the centre of the Cairo protests – and found that they predicted protest turnout.

However, other academics have warned against overstating social media's role in civil unrest. A study analysing millions of tweets related to the 2011 English riots found that it was not used to organise or incite the riots. Dr Paul Reilly of the University of Glasgow notes that social media didn't cause either the violence that marred a peaceful protest against the shooting of Mark Duggan by the Metropolitan police in Tottenham on 6th August 2011, or the disturbances seen in Manchester or Birmingham days later.

On the contrary, the 2011 riots were driven by a "complex set of macro and micro factors, including anger at Mark Duggan's death, the disenfranchisement and disillusionment of young people as well as the 'wanton criminality' that has been the subject of many politicians' ire".

Similarly, Dr Reilly has argued that social media's role in the Southport riots has been overstated. Writing in The Conversation in 2024, he observed that most online activity surrounding the riots occurred after the events rather than before them, consistent with his previous research in Northern Ireland.

A comprehensive review by Patton et al. (2014) pointed to "major limitations with the existing studies" on the relationship between social media and youth violence, emphasising that findings are often inconclusive or overstated.

In short, the causal link between algorithms and unrest remains a subject of debate.

Part of the difficulty in assessing this link lies in the tendency in policy and government circles to view social media users as passive recipients of content shaped by algorithmic recommendations.

This perspective aligns with a theoretical framework reminiscent of B.F. Skinner's radical behaviourism, where behaviour is shaped entirely by external stimuli – in this case, algorithms. Such a model implies that individuals exposed to harmful content are inevitably conditioned to adopt the ideas or behaviours it promotes.

However, this overlooks the critical agency of users, who frequently approach media content with scepticism and discernment.

There is ample evidence to support this contention. Fenster (2008) observed that conspiracy theories often function as captivating narratives rather than credible claims, appealing to audiences as a form of storytelling. Highfield and Leaver (2016) suggest that users engage

with sensational or fictionalised narratives as a form of escapism, often with a full awareness of their exaggerated or fabricated nature. Phillips and Milner (2017) argue that digital culture thrives on playful forms of engagement, where content is shared as a form of creative expression rather than factual communication.

In other words, the implicitly behaviourist assumptions that underpin discussions of the 'toxic' or 'harmful' effects of mis- and disinformation are being challenged by academic research. (It's also worth noting that in a review of the literature, a recent UK Parliament briefing recognised that "people do not necessarily share mis/disinformation because they believe it".)

In a UK-based focus-group study, the research team challenged the notion that audiences passively consume misleading information. According to their findings, study participants adopted a "pragmatic scepticism" toward news and disinformation. Demonstrating an active and discerning approach to content curated by social media algorithms, users critically engaged with news by fact-checking and triangulating information served up by algorithms.

Such practices echo findings by Tandoc et al. (2018), who note that audiences authenticate information through personal judgment, verification from social circles or consultation with institutional sources.

Ultimately, then, while algorithms may have amplified content linked to the UK riots of 2024, they were neither the sole nor the most significant cause.

Policies addressing the risks posed by disinformation should prioritise enhancing media literacy and fostering critical engagement among active, critical and reflexive users, rather than relying exclusively on tightening algorithmic content moderation protocols, which risks overreach and the curtailment of lawful speech.

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

This question presupposes that "these issues" are significant enough to need "tackling" with regulation and legislation, and we would suggest that the best empirical evidence, as described above, strongly indicates that this is not the case. While we will look at the primary piece of legislation designed to address such issues, namely the Online Safety Act 2023, in further detail below, we lay out here by way of summary some key observations drawn from the evidence laid out so far.

- Targeting specific technologies from a legislative and regulatory point of view is likely at this stage to be a never-ending game of catch-up.
- Social media companies, facing punitive regulatory frameworks, are likely to err on the side of caution, adopting a 'safety-first' – or 'if in doubt, cut it out' – stance that risks censoring lawful but controversial content.
- Algorithmic moderation systems, which rely on datasets of previously flagged material, exacerbate this problem by struggling to interpret context, cultural references, nuance or satire. This could lead to the suppression of legitimate speech, the misinterpretation of cultural references and the removal of benign content.
- The opaque nature of AI decision-making processes also limits accountability and increases the likelihood of bias or error in this 'black boxed' system. The chilling

effect on online discourse is already evident, as users become hesitant to engage with certain topics for fear of being flagged by automated systems, ultimately stifling public debate and reducing the diversity of perspectives.

6. How effective will the Online Safety Act be in combatting harmful social media content?

The Online Safety Act 2023 imposes extensive legal obligations on social media platforms and search engines, requiring them to identify, mitigate and manage risks associated with illegal and harmful content. This includes not only the removal of offending material but also the implementation of systems to prevent its dissemination in the first place. Ofcom, tasked with enforcing the Act, has significant powers to impose penalties for non-compliance, including fines of up to 10 per cent of a company's global turnover – a figure that could exceed £10 billion for Facebook – and, in extreme cases, prison sentences for senior managers. While these measures aim to address concerns around harmful content, they raise significant questions about regulatory overreach, particularly regarding their impact on lawful speech and open debate.

At present, the full effects of the Act remain unclear. Many of its provisions have yet to come into force, and secondary legislation pertaining to the regulation of online service providers is still pending. Ofcom's guidance and codes of practice, which are critical for interpreting the Act's provisions, will not be fully published until 2025 and 2026 (Ofcom, 2023).

Premature calls to expand or amend the Act therefore risk exacerbating uncertainty, undermining the UK's appeal as a destination for tech investment, and compromising effective implementation. A period of careful observation and review is needed to determine whether the Act's framework can deliver on its objectives without causing undue harm.

Nevertheless, even at this stage, certain shortcomings and contradictions in the Act's design are apparent, particularly regarding new criminal offences and the broader implications of content moderation requirements.

Part 10 of the Act, which came into force in January 2024, introduces offences such as "false communications" and "threatening communications". Under these provisions, an individual may be held liable not only for creating harmful content but also for sharing or transmitting it – including via hyperlinks or oral communication, such as voice notes.

The threatening communications offence specifically criminalises messages that convey threats of serious harm, such as death, serious injury or financial loss, if the sender intends or is reckless about causing fear. Unlike the pre-existing offence under Section 127 of the Communications Act 2003, which covers menacing communications, this new provision allows for significantly harsher penalties, including imprisonment for up to five years.

This suggests that the threatening communications offence is designed to cover behaviour that is more serious than that covered by the s.127 offence.

These new offences expand an already draconian legal framework for 'hate speech' in England and Wales, which includes various pieces of legislation, predominantly Section 4A of the Public Order Act 1986; "stirring up offences" (Sections 18 and 29B) under the Public Order Act 1984; Sections 127(1) and 127(2)c of the Communications Act 2003; section 2 of

the Protection from Harassment Act 1997; Sections 31 and 32 of the Crime and Disorder Act 1998; and Sections 1 and 12 of the Terrorism Act 2000.

Expanding criminal liability further risks creating ambiguity, stifling legitimate expression, and deterring robust public debate.

In terms of the anticipated regulatory framework and Ofcom's duties, there is reason to suppose that the Act will not achieve that it will mean to, and may in fact be counter-productive, largely due to the futility of attempting to regulate the entirety of human online interaction. We offer some key examples here, showing two key paradoxes: how encouraging algorithmic suppression of certain types of "Priority Content" may end up undermining efforts to mitigate harm; and how suppression of extreme political rhetoric may give rise to even graver security concerns.

Tech platforms' content moderation rules vary, and Ofcom's approach will likely allow for such variation, provided it is in accordance with a properly conducted risk assessment and other measures. Some display more tolerance for content that may be regarded as 'harmful' within one or more of the categories above, whether due to a deliberate decision to allow it or through a broader choice not to moderate more heavily. This can result in users developing various workarounds. This is pertinent to considering Priority Content categories, where some platforms already flag content falling within that category, either for outright removal or for hiding behind a "content warning" label. This has resulted in the development of commonly used euphemisms, sometimes referred to as "algo-speak", designed to circumvent filters (Shams Ili, 2023). One commonly seen example is "unalive" for "die by suicide", as in say, 'the celebrity unalived himself after he was caught having an affair'.

These workarounds inevitably end up in an arms race of censorship, in which as euphemisms are identified and brought within the ambit of the content moderation systems, new ones are developed, with the only limit being that of human ingenuity.

But this does not only serve to render content suppression ineffective. Because algo-speak results in an argot that is used primarily in speech communities focusing on those particular topics, it may lead to a paradoxical effect. Searching for a specific bowdlerization of a particular term that has not yet been picked up by the algorithms – e.g., "an0r3xi4" instead of "anorexia" – may lead a user to content that actively promotes unhealthy behaviours, but does not trigger any automated flags that might, for example, signpost them to valuable health information or organisations that could provide them with help.

A further example of the paradoxical effect of regulation arises when radicalisation and extremism. Despite high error rates in identifying material which legitimately falls into this category, the scope for false positives disproportionately captures speech which is recognised as being deserving of particular protection, e.g., political discourse (vehement criticism of government), religious discourse (sermons by clerics suspected of extremism) and news reporting (broadcasts featuring clips of terrorist propaganda videos) (Keller, 2018). This has clear implications for the collateral damage that can be caused by content moderation that will inevitably – due to scale – be incapable of fully appreciating the nuance and context necessary to evaluate such matters, and is in any event incapable of providing the kind of in-depth consideration and adversarial testing of allegations necessary to map out the boundaries of acceptable speech.

But this topic can lead to even further difficulty. Merely censoring content does nothing to address the complex and multivariate psychological and sociological factors that lead to radicalisation. Removing the content from mainstream platforms will lead those who would seek it out elsewhere, resulting in them finding other platforms – often those that form significant echo chambers. This leads to two main issues.

First, so-called 'leaderless' jihadi movements – which give rise to the types of 'lone wolf' terror attackers who are often the most difficult for police and security services to trace – are particularly susceptible to 'soft approaches' to counter-radicalisation such as public correction of extremist narratives, often by senior members of the would-be radical's community (Sageman, 2009).

Daniel Kimmage, a senior analyst at Radio Free Europe/Radio Liberty, a vital institution for spreading the essential message of Western liberal democracy, has pointed out that social media platforms offer a unique opportunity for the message of extremist actors to be challenged in the comments sections, and this creates vulnerabilities for such ideologies (Kimmage, 2008).

However that potential advantage is lost if the messaging returns to more isolated platforms. This trend, of course, does not apply solely to the extremes of political though. It was witnessed following Donald Trump's 2021 ban from Twitter, when a significant number of his followers moved to his own Truth Social platform, and again in 2024 when a large number of left-leaning individuals left X for Bluesky in protest against platform-owner Elon Musk's support of Trump.

Second, and more troubling than the echo chamber effects of the latter two mass migrations, the platforms of last resort for extreme actors tend to be closed messaging groups such as those found on Telegram, as happened after the summer 2024 riots (Murphy, 2024).

Although tensions around this issue in the intelligence community are rarely considered in public from a strategic point of view, an example was to be found following a Pentagon-driven shutdown of a known Al Qaeda forum. This created a mass exodus to other platforms, resulting in CIA objections to a catastrophic loss of intelligence (Romanosky, 2016).

Most problematically of all, such platforms may make use of end-to-end encryption which significantly complicates security services' access to information. Given the positive security benefits of such encryption systems for the overwhelming majority of innocent users – as well as the political undesirability, technical challenge, and civil-liberties implications of attempting to mandate state backdoors to such systems – there is a clear advantage to allowing such discourse to take place on open channels.

Therefore, while it is acknowledged that the precise effects of the Online Safety Act have yet to be fully appreciated, there is strong reason for scepticism that it will be capable of addressing the mischiefs it set out to address, and good arguments to believe that in may, in fact, be counter-productive

7. What more should be done to combat potentially harmful social media and AI content?

This question rests on problematic assumptions. It presupposes a need to do more without first establishing whether there is a sufficient basis for action. The purpose of this inquiry is to assess the nature of the problem. If intellectually honest, that process must leave open the possibility that no further legislative measures are required.

The extension of legislative scope to regulate potentially harmful content also reflects an inappropriate reliance on the precautionary principle – the idea that a lack of certainty or evidence should not prevent regulatory action to mitigate perceived risks (Sands and Peel, 2012). While most commonly applied in environmental law – and controversial even there – the principle has been extended to other areas, such as the pandemic response, where it underpinned policies of dubious scientific merit (Wu et al, 2024).

The actual quantity of misinformation online remains vanishingly small, and engagement with it is equally limited. For example, Buzzfeed reported that the top 20 fake news stories on Facebook leading up to the 2016 U.S. election were engaged with nearly 9 million times. In the context of Facebook's 1.5 billion monthly users at the time, this accounted for a mere 0.006 per cent of activity – even on the overly cautious assumption that users interacted only once per day. Furthermore, engagement is not necessarily indicative of belief: users may ridicule, debunk, or critically examine such content. As discussed earlier, the persuasive effects of misinformation remain unproven and, in all likelihood, minimal (Watts et al, 2017).

We described above, talking of the riots, how the evidence shows that engagement is also not a good measure of people's attitudes to false information: they could be ridiculing it, debunking it, etc. We have also described how the persuasive effects of this information are unproven and likely small.

We would therefore draw the Committee's attention to the significant body of research which suggests that fears about misinformation represent a moral panic (Schroeder & Jungherr, 2021). This is a well-defined phenomenon in which a society becomes convinced that some issue, individual or group (the 'folk devil') poses a particular threat to its interests or values. In the classic text on the subject, Stanley Cohen described it thus:

> A condition, episode, person or group of persons emerges to become defined as a threat to societal values and interests; its nature is presented in a stylized and stereotypical fashion by the mass media; the moral barricades are manned by editors, bishops, politicians and other right-thinking people; socially accredited experts pronounce their diagnoses and solutions; ways of coping are evolved or (more often) resorted to; the condition then disappears, submerges, or deteriorates and becomes more visible.

At the heart of moral panics are the so-called 'moral entrepreneurs' – those who promote the perception of a crisis. Research has identified certain common beliefs among this group, particularly in relation to misinformation. In the UK, a tendency to overestimate the dangers of misinformation correlates with:

- Negative attitudes toward new technologies
- The belief that societal problems have simple solutions and clear causes
- A perception that misinformation is difficult to spot
- A conceit that other people are more gullible than oneself

Such findings lend empirical support to the view that advocates for further internet censorship often represent a technologically illiterate and condescending elite, incapable of appreciating the complexity of the modern information environment.

The moral panic around misinformation is not merely intellectual but also cultural. It reflects a profound discomfort with societal change. Traditional journalism and established information channels are under threat, political power has shifted, and politicians now face unprecedented levels of scrutiny. These disruptions are unsettling for those accustomed to gatekeeping public discourse, creating a strong impetus for them to reassert control. However, these cultural anxieties are not a suitable basis for legislation, particularly measures that threaten public liberty.

Against this backdrop, it is worth scrutinizing recent proposals for "strengthening" the Online Safety Act, as suggested by the Prime Minister and others. The Home Secretary, Yvette Cooper, has committed to a "rapid review of extremism" following recent violent disorder. Among the proposed next steps, Peter Kyle, Secretary of State for Science, Innovation and Technology, pledged to "strengthen the requirements for social media companies to take responsibility for the poison proliferated on their platforms". Critics such as Sadiq Khan have gone further, claiming – without evidence – that the Act is "not fit for purpose".

While such rhetoric is politically expedient, proposals to strengthen the Act raise serious concerns. Are critics suggesting that the "legal but harmful" clause – removed by the previous Government – be reinstated? This provision would have empowered ministers to define lawful content as "harmful" and mandate its removal. Alternatively, proposals to expand the Section 179 "false communications" offence to include misinformation would compel platforms to moderate an even broader range of content.

Yet both proposals hinge on the technical operationalisation of inherently subjective terms such as "misinformation". In practice, this would require platforms to algorithmically moderate content based on vague or politically charged definitions of what constitutes false or harmful information.

This approach raises several problems.

First, "misinformation" frequently serves as a euphemism for dissent – views that challenge the policies of the day or question mainstream orthodoxy. Suppressing such speech does not eliminate its underlying causes – but instead drives it to the margins, fostering resentment and alienation. Far from reducing harm, this risks escalating online frustrations into real-world consequences.

Second, the problem of temporality – stressed during previous sections of this submission – further complicates efforts to regulate misinformation. Consider, for instance, the case of public health officials during the Covid pandemic. At the time, their statements were considered authoritative and their critics dismissed as spreaders of misinformation. Yet with hindsight, some of official messaging is now classified as misinformation. Algorithmic systems were – and despite technological advances, remain – ill-equipped to handle such evolving truths, risking censorship of emerging dissent that may later prove valid. As a result, platforms attempting to differentiate misinformation from information risk amplifying the former while stifling the latter.

Third, there remains scant evidence of significant harm resulting from misinformation. Calls for expanding regulatory powers are therefore disproportionate, particularly given the collateral impact on free expression.

Another concerning proposal involves granting Ofcom "emergency response" powers to tackle misinformation during crises, such as national security threats or public safety incidents.

In August, the Centre for Countering Digital Hate (CCDH) hosted a closed-door meeting under the Chatham House rule to discuss the role of social media in civil unrest. The meeting included officials from DSIT, the Home Office, Ofcom and other organisations. CCDH's subsequent policy recommendations included amending the Online Safety Act to enable the Secretary of State for DSIT to grant Ofcom additional "emergency response" powers to fight "misinformation" that poses a "threat" to "national security" and "the health or safety of the public".

CCDH's proposal would involve amending the section 175 "special circumstances" directive created by the Online Safety Act to enable Peter Kyle, to issue a "directive" to Ofcom to ramp up its censorship powers if the Government feels there is a threat to national security or to the health and safety of the public (both, notably, constituting exemptions under Article 10(2) of the ECHR which empower states in certain circumstances to curtail the liberties of their citizens).

The stipulation that such powers would apply only in an "emergency" is hardly reassuring. Historically, terms like "crisis" and "emergency" have proven elastic, readily expanding to suit political agendas.

The epistemological challenge is equally clear: the distinction between "misinformation" and a "plausible hypothesis" – such as the lab-leak theory of Covid-19's origins – is often determined only with time. Empowering governments to define harmful content in real time carries significant risks for democratic debate.

The implications of these powers are particularly troubling in contentious areas such as climate change.

For example, CCDH has categorised "climate denial" as arguments used to undermine climate action. Given Mr Kyle's past advocacy for declaring a "climate emergency," it is conceivable that he could direct Ofcom to remove dissenting views on climate policy under the pretext of public safety.

While some might argue that such views are harmful and should be removed, the reality is that there is legitimate scientific and political debate on how to address climate change.

Different solutions to tackling climate change are informed by different values and recommending one approach over another inevitably involves making a political choice. There is no such thing as an apolitical, "scientific" solution and, therefore, it is a dishonest sleight of hand to categorise dissent from one particular solution as "misinformation".

Efforts to suppress misinformation – whether through expanded legal provisions, emergency powers or algorithmic systems – risk conflating complexity with harm, dissent with danger

and debate with disorder. Such measures are not only disproportionate but also counterproductive, fostering resentment, alienation and unintended consequences that undermine the very goals they seek to achieve.

8. What role do Ofcom and the National Security Online Information Team play in preventing the spread of harmful and false content online

We have significant concerns about the ongoing role of the National Security Online Information Team (NSOIT), previously known as the Counter Disinformation Unit (CDU), particularly where its remit extends beyond countering disinformation, as "false and/or manipulated information".

Equally troubling is the fact that the unit continues to operate without parliamentary oversight or sufficient democratic accountability. These concerns have intensified following Peter Kyle's decision to task NSOIT with monitoring online activity in response to the Southport riots.

This decision comes just months after the House of Commons Culture, Media, and Sport Committee raised serious concerns about "the lack of transparency and accountability of [NSOIT] and the appropriateness of its reach".

The Committee recommended that the Government commission an independent review of the unit's "activities and strategy", with a report due within 12 months. Such scrutiny is long overdue, given the potential risks posed by a unit operating largely outside formal democratic processes.

To be clear, the FSU has no objection to NSOIT monitoring "harmful and false content online", where that phrase is intended to mean content that incites violence or deliberately disseminates false or manipulated information. The critical question is whether the unit will go beyond this remit to monitor and flag lawful online posts for removal.

NSOIT's origins as a unit established within government to combat disinformation might appear reasonable at first glance, given that disinformation is generally understood as "false information spread in order to deceive people". Insofar as it applies to content promulgated by serious threat actors and nation-state adversaries, monitoring such material is undoubtedly necessary.

However, in practice, the use of the word "disinformation" seems to have been peculiarly malleable.

During the pandemic, the unit's remit expanded to include the "inadvertent sharing of false information", effectively conflating disinformation with misinformation. This shift widened the scope of its activities considerably, raising questions about whether there were any meaningful limits to who or what the CDU would scrutinise.

Conservative MP and former Home Secretary David Davis, for example, was cited in CDU files as being "critical of the Government" after he questioned the mathematical modelling behind Imperial College's pandemic projections.

Similarly, Dr. Alexandre de Figueiredo, the Statistics Lead at the Vaccine Confidence Project, attracted the unit's attention after publishing research suggesting that vaccine passports could negatively impact vaccine confidence.

The CDU also targeted journalists and commentators.

TalkRadio's Julia Hartley-Brewer was flagged for sharing a clip from her show in which a listener detailed the devastating consequences of lockdown measures, such as the closure of her fiancé's business, the cancellation of her father's cancer treatment and her grandmother's fear of leaving home.

Similarly, Silkie Carlo, director of Big Brother Watch, was logged for comments on TalkTV opposing vaccine passports, which she described as "a vision of checkpoint Britain". The rationale for flagging these posts often appeared tenuous. A spokesperson for Logically, the AI-based monitoring firm contracted by the Department for Digital, Culture, Media, and Sport (DCMS) during this period, explained that even legitimate posts might be included in reports if there was potential for such narratives to be "weaponised" by critics of Government policy.

This raises profound concerns about the boundaries of NSOIT's activities.

While the previous Conservative government claimed that NSOIT lacks the power to compel social media platforms to remove flagged content, the unit nevertheless benefits from DCMS's "trusted flagger" status. This designation gives its reports significant weight, making it highly likely that platforms will comply with its recommendations, even when the flagged content is lawful.

This informal yet influential mechanism effectively circumvents established legal protections for freedom of expression.

There are other, equally troubling, legal implications in NSOIT's activities.

Article 10 of the European Convention on Human Rights (ECHR), for instance, guarantees the right to freedom of expression and specifies that any interference by the state must be clearly defined in law. This is a particularly weak protection of fundamental speech rights, containing as it does a variety of exemptions which allow states to censor their citizens. Of particular relevance here, given NSOIT's name, is national security, which seems to be allowed a particularly expansive interpretation in this context.

Nonetheless, even though NSOIT's supporters may point to these exemptions as offering justification for its activities, there are restrictions. It is well established that any interference with fundamental rights must be proportionate. As was made clear by the Supreme Court in *Bank Mellat v HM Treasury* [2013] UKSC 38, the following criteria must be met:

> (1) whether the objective of the measure is sufficiently important to justify the limitation of a protected right;
>
> (2) whether the measure is rationally connected to the objective;

(3) whether a less intrusive measure could have been used without unacceptably compromising the achievement of the objective;

(4) whether the measure's contribution to the objective outweighs the effects on the rights of those to whom it applies.

We would suggest that none of these tests are met. The effect of misinformation is known to be minuscule. Therefore the objective (constraining it) cannot possibly be important enough to limit fundamental rights. The extent to which the activities of NSOIT are "rationally connected" to that objective are equally questionable, given the overreach of the sort described above. It is undoubtedly the case that less intrusive measures could be taken. The severity of the state monitoring and logging private citizens' perfectly lawful speech is high, and therefore we suggest that any potential balancing act could not possibly justify the excesses we have seen.

Further, ECHR jurisprudence is clear that any interference must be "prescribed by law". NSOIT, like its predecessor, operates without statutory authorisation, judicial oversight or a formal law enforcement function. The practice of flagging entirely lawful content at the discretion of unaccountable civil servants creates an unacceptable risk to free expression. Such a system allows for the suppression of lawful speech outside the safeguards of parliamentary scrutiny, leaving individuals vulnerable to censorship with little recourse to challenge decisions.

The risks posed by NSOIT's lack of statutory footing were evident during the pandemic, when journalistic activity was monitored and logged under the guise of addressing disinformation.

Even if the unit refrained from more intrusive actions at that time, its unregulated nature prompts the question: "What is to stop it from going further in the future?"

These concerns are not merely theoretical. The gradual creep in the definition of "disinformation," coupled with the discretionary power granted to NSOIT, threatens to erode the distinction between harmful, illegal content and legitimate, lawful expression. By operating without statutory authorisation or democratic oversight, the unit poses a clear and present danger to freedom of expression in the UK.

NSOIT may well have a legitimate role in addressing illegal content and deliberate disinformation, but its activities must be rigorously defined, transparent, and subject to parliamentary oversight.

Without robust legal safeguards, there is nothing to prevent the expansion of NSOIT's remit to more actively interfere with lawful speech during future crises.

An independent review of NSOIT's strategy and activities is urgently needed to ensure that it remains within the bounds of its original remit and does not become an unchecked tool for suppressing dissenting voices.

9. 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?

As the above evidence shows, we do not believe that the correct solutions to such problems as exist will lie in further imposition of liability on platforms, nor on tighter regulation. Therefore we view the above question as entirely misplaced and based on false premises.

Rather, we propose a set of solutions here which are evidence-based and robust, focusing on the positive side of human potential, empowering people rather than restricting them, and respecting fundamental rights and liberties. We believe that these solutions will not only make the UK more secure and our institutions healthier, but would also contribute to the economy and help the UK retain its place as a world leader in cutting-edge technology.

10. Recommendations

Avoid further over-correction

Such was the authority that governments and health agencies granted themselves during the pandemic that they actively discouraged the consumption of information from 'non-approved' sources. Perhaps most notoriously, New Zealand's then Prime Minister, Jacinda Ardern, urged the public to regard her government as a "single source of truth", adding that "unless you hear it from us it is not the truth" (NZ Herald, 2020).While this was in the context of encouraging the public – quite rightly – to be sceptical of what they saw on social media, the above examples of where governments got it wrong on Covid demonstrate that their pronouncements should also be accorded a relevant level of scepticism by the public; indeed this is healthy, and a necessary part of the democratic process.

We have shown how, especially in a fast-moving crisis, knowledge is provisional, liable to be falsified and surpassed as new facts emerge. Hubris on the part of authorities, state media organisations, and others in pretending that they are not susceptible to the same epistemic vulnerabilities as anyone else may also lead the public to distrust them. We have further described how, in turn, this leads to potential rights violations as well as a broader polarisation and breakdown of trust in institutions.

Seeing as we know that nation-state sponsored disinformation has vanishingly little utility in actually changing opinions to conform with the content of the disinformation itself, but that adversaries such as Russia still devote considerable resources to disseminating such material, it is reasonable to look at alternate explanations of why they may be doing so.

One answer may be found in the doctrine of reflexive control theory. Dating back to the Soviet era, but still in heavy use today in guiding Russia's information warfare doctrines, this strategy involves manipulating the information environment in order to cause an enemy to make decisions unfavourable to himself (de Goeij, 2023). It is a simplistic assumption to regard the target of these disinformation campaigns as the consumers of the disinformation itself: as already noted, this forms a tiny proportion of the overall information environment, and the content is largely engaged with by those already persuaded (Kapoor & Narayanan, 2023).

This doctrinal understanding of the Russian approach to disinformation allows one to see that in many respects, the true targets of these campaigns are not the citizenry, but the decision-makers. In this respect, their operations have enjoyed an unprecedented level of success. Western governments have engaged in a programme of suppression and trampled upon the

liberties of their citizens on the basis of not very much substance at all, truly using a sledgehammer to crack a nut.

This serves to further heighten the distrust in Western institutions, and work towards destabilising our democracies. It is far more dangerous than someone who was already going to vote for Trump anyway potentially seeing some nonsense on Facebook about Kamala Harris being an alcoholic – and to view such superficial effects as being the real objective of the action is almost pathetically naïve.

We have handed this triumph to them on a plate. In turn, they play upon the very distrust that has been created by Western over-correction, and the justifiable perception of rights violations and draconian rule. For example, the Storm-1516 network, currently one of the Kremlin's leading 'propaganda factories', has established a front organization called the 'Foundation to Battle Injustice', purportedly a 'human rights organisation' seeking to expose illiberal actions by Western governments, interwoven with further manufactured stories that further undermine confidence in institutions (Warren and Linvill, 2024).

Whether or not the claimed scientific basis for reflexive control theory is sound, it has been astonishingly successful in its effects. Our own failings are being used against us. The childlike over-simplifications resulting from the cognitive biases of the 'moral entrepreneurs' play right into the hands of opponents who have been perfecting these techniques for decades. This needs to end.

Counterspeech

Perhaps one of the most prominent examples of the use of counterspeech in combatting misinformation is X's 'Community Notes' feature. This is a mechanism by which users are able to suggest corrections or context to be appended to misleading posts, which are then ranked by a subset of the user community selected (via knowledge obtained by user profiling) to be representative of a broad range of opinion. Those debunkings ranked highly are then prominently displayed alongside the post in question, providing an alternative to content removal.

This creates a strong incentive to ensure accuracy in what one posts. The effect is magnified for public figures, whose credibility will be damaged by regularly posting material that is subject to such a process, and of course the more prominent an account is (and therefore the further its reach), the more likely it is that a misleading claim will be picked up on and 'community noted'.

Although there is some research that Community Notes do not significantly reduce engagement with misleading posts, as described above this is a poor measure of the effective reach of misinformation. For example, in such cases it is reasonable to assume that people could be engaging with the post to criticise it rather than endorse it.

The phenomenon of the crowd seeking to counter misinformation arises spontaneously as well, outside of organised structures to facilitate it. Rather than being helpless victims of deceit and propaganda, social media users overwhelmingly tend to be proactive in seeking truth and correcting others who mislead (Schroeder & Jungherr, 2021).

What is more, it can lead to a positive feedback loop, where not just accuracy but civility can be self-reinforcing and raise the standard of conversation, and this is even beneficial to platforms by increasing engagement. In 'below the line' discussion threads on newspaper articles, it has been found that factual and polite responses to uncivil comments raise the tone of the discussion and increase participation rates (Ziegele & Jost, 2020).

Just as human beings in real life rely not on ubiquitous mechanisms of 'content moderation' to arbitrate daily interactions, many of the same norms and social pressures that tend to reward civility and punish dishonesty and hostility are applicable to online interactions, despite the potentially lower threshold for negative behaviours.

For example, evidence also suggests that counterspeech can be an effective means of tackling hate speech. In a study involving Slovakian Facebook posts exhibiting anti-Roma slurs and stereotypes, it was found that people posting pro-Roma comments in turn led to others with a pro-Roma attitude joining the conversation (Miškolci et al, 2020).

While such counterspeech may be unlikely to persuade the original poster, it can piggyback off their reach, shape the tone and content of the following conversation, and have an influential effect on the post's audience. There is further evidence to suggest that censure from a high-status member of one's own group can reduce a poster's usage of ethnic slurs and other forms of racist harassment, indicating that rather than moderation which merely removes the problematic content from sight, such measures may in fact have a positive influence on future user choices and behaviour (Munger, 2017).

There is also evidence to suggest that 'official' fact-checking is less effective than spontaneous crowd-sourced methods. In a study of tweets pertaining to Covid misinformation, it was found that professional fact-checking tweets had limited engagement, whereas there would be an increase in tweets refuting misinformation in direct response to particular misinformation tweets gaining prominence – and that these tweets had better reach than the professional ones (Micallef et al, 2020). This social correction extends to Facebook and WhatsApp (Rossini et al, 2021).

Although lacking some of the appeal of the 'organic' forms of counterspeech contemplated above, a further option exists in the form of attaching notes from state or other authorities to content dealing with particular topics. This was widely employed for material dealing with Covid-19, where, for example, YouTube videos covering the topic would be accompanied by a prominent link to WHO information. While this is vulnerable to the distrust that may exist about such bodies, that trust can (and should) be rebuilt (Acerbi et al, 2022). The same method is also applicable to issues of radicalisation and extremism, as described above and further seen, for example, in early implementations of Google's Jigsaw programme which provided counter-messaging alongside search results (Meleagrou-Hitschens & Kaderbhai, 2017).

Digital literacy

Ofcom defines media literacy as "the ability to use, understand, and create media and communications in a variety of contexts" (e.g., Ofcom, 2023; Ofcom, 2024).

In theory, this definition positions Ofcom as a key player in equipping individuals to navigate and contribute to the digital world. In practice, its media literary initiatives concentrate

heavily on the consumption of online content (i.e., the ability to use and understand media and communications), while largely neglecting the equally important dimension of content creation (i.e., the ability to create media and communications).

For instance, its 2024 report, *Understanding Misinformation: An Exploration of UK Adults' Behaviour and Attitudes*, emphasises the importance of teaching users to identify misleading information and avoid unreliable sources (Contents list: "1. Attitudes towards information and news; 2. Encountering mis and disinformation; and 3. How people characterise and deal with mis and disinformation").

Before the 2024 UK General Election, Ofcom targeted new voters aged 18-24 with a campaign to help them detect disinformation during a crucial democratic period. Ofcom collaborated with social enterprise Shout Out UK and the Electoral Commission to launch the Dismiss campaign – which aimed to increase awareness of existing and emerging misinformation/disinformation tactics.

In its three-year media literary strategy, *A Positive Vision for Media Literacy*, published in October 2024, a similarly safetyist narrative emerges. Ofcom will "heighten public awareness and understanding of how people can protect themselves and others online", "encourage the development and use of technologies and systems so that users of regulated services can protect themselves and others online", "signpost users to resources, tools or information that can raise awareness about how to use regulated services to mitigate harms", and so on. The word "create" features only once, when the regulator cites its definition of "media literacy".

These efforts remain narrow in scope. They are firmly rooted in a consumption-oriented approach, focusing almost exclusively on helping users recognise and avoid harmful content rather than empowering them to actively shape the digital environment by creating media. Subtly, the emphasis shifts to knowing what is 'illegal' or 'distasteful' online content, rather than what is strictly 'lawful'.

By failing to address content creation, Ofcom overlooks a crucial avenue for tackling disinformation: enabling individuals to produce their own lawful and responsible content, grounded in an understanding of the limits and protections of free expression.

This oversight is particularly significant because content creation is not just a technical skill. It is an essential component of active digital citizenship. Without training in how to responsibly contribute to online discourse, users are left as passive consumers of information, vulnerable to the very forces of disinformation that Ofcom digital literacy initiatives seek to combat.

A comprehensive digital literacy program should include education on the UK's legal framework for freedom of expression, such as the Human Rights Act 1998 (covering Article 10 of the European Convention on Human Rights), the Equality Act 2010 (anti-discrimination provisions), the Defamation Act 2013 (protection from reputational harm caused by false statements), and the Communications Act 2003 (governing harmful and offensive online communications). Such training could help individuals understand not only their rights but also their legal responsibilities when creating online content. This could include topics like avoiding defamation, understanding privacy laws under GDPR, and recognising hate speech or harassment.

Furthermore, issues like contempt of court – a critical area of online legal responsibility, as witnessed by the online fallout from the Southport stabbings – are absent from Ofcom's initiatives. Under the Contempt of Court Act 1981, individuals can inadvertently interfere with judicial processes by posting or sharing prejudicial material about active legal cases. Training users to understand and avoid such pitfalls is an important aspect to any well-rounded package of digital literacy initiatives.

If Ofcom were to adopt a more balanced approach that included content creation, its media literacy strategy could better reflect the realities of digital participation. Empowering users to create lawful and impactful media would allow individuals to contribute constructively to online spaces, reducing the dominance of disinformation while fostering a more diverse and inclusive digital environment.

As it stands, Ofcom's initiatives do little to equip users with these critical skills. Its current strategy prioritises mitigating harm through consumption-oriented campaigns, while largely ignoring the empowering potential of responsible media production. This limited vision undermines its ability to fully address the challenges posed by online misinformation and disinformation.

A more ambitious approach – one that recognises the importance of content creation – could transform Ofcom's role from a reactive regulator to a proactive force in shaping a healthier, more balanced digital landscape.

Fostering innovation

The crowdsourcing of fact-checking in such programmes as Community Notes is an example of a key innovation that has huge potential and has dramatically altered the information space. It is a strong solution that is unlikely ever to have been dreamt up by a regulator, and in any event could not work within that milieu, relying as it does on the voluntary participation.

Similarly, there is a great deal of research being undertaken into how to address problems of AI-generated deepfakes at a technical level. For example, there are nascent technologies using blockchain and related methods to provide authentication for videos, which AI-generated fakes would be unable to bypass due to the cryptographic security provided. An example of a company operating in this space is the London-based OpenOrigins startup, which specifically focuses on building trust in what is real, rather than in censoring what is not, and whose product would therefore serve to reinforce digital literacy efforts.

The UK is in a good position to be able to foster the development of such technologies, which provide a long-term and secure means of dealing with these types of threats. However, premature regulation focusing purely on content removal and blocking may dissuade companies from developing such methods by reducing the incentives and in some cases even rendering such technologies outside the regulatory framework. This would be a missed opportunity.

Government effort should therefore focus instead on the creation of further incentives for companies to be able to work on the problems and innovate.

Transparency

Given the importance of trust in creating and maintaining a healthy information environment, we consider that efforts should be made to ensure that this is rebuilt.

This refers not just to trust in media organisations and the state, which has been severely damaged over the last decade, and most notably during the pandemic. Citizens should also be able to trust the mechanisms by which the information environment is regulated and constrained, and understand how, and when, it is manipulated by third-party actors.

There is a growing understanding of the importance of the principle of transparency in building this trust (MacCarthy, 2021). This includes allowing for the understanding not just of algorithms which manipulate feeds, but also moderation tools and processes, including (perhaps especially) where this is automated.

The EU has incorporated provisions relating to transparency into its Digital Services Act (European Commission, 2023). This would give accredited academic researchers access to data of very large online platforms (VLOPs) and search engines (VLOSEs). We consider that incorporating similar provisions into the UK's information infrastructure, possibly linked to Category 1 platforms under the current regulatory system, would be a positive step. We further believe that affording access to this data not just to academics but also to human rights organisations would improve scrutiny and increase accountability.

Just as we can learn from what is being done on the Continent, so we can learn from what is being done across the Atlantic. The new administration in the US has advanced a number of principles underpinning a 'digital bill of rights' for users of online platforms (Times of India, 2024). These include:

1. A prohibition on the government collaborating with private organisations to restrict *any* lawful speech;
2. Making platform immunities contingent on speech protections for users – if they exercise editorial control in removing lawful content, they must thereby be held liable;
3. The threshold for the removal of content should be set high, including requiring court orders where there is no evidence of immediate and significant harm;
4. When users have accounts or posts removed, shadowbanned, throttled, they should have the right to:
    a. be informed that it's happening
    b. a specific explanation of the reason why
    c. a timely appeal considered by a human.

We consider that establishing legally enforceable rights for UK users in line with the above would be a further positive step in improving transparency and building trust. The stakes could not be higher. The less accountable platforms are in terms of removing content, and the more they are seen to be collaborating with the state on repression and censorship, the greater the threat will be to our democratic institutions and liberal traditions.

18 December 2024