

No Ref in the Ring: Algorithmic Editorship and the Future of Social Media Regulation beyond the Online Safety Act 2023

Torin Holland

Abstract

Social media's arrival to the news-sharing sphere introduced capabilities for anybody with an internet connection to publish to millions, and the viral misinformation surrounding gold-medallist boxer Imane Khelif shows that mass communicative power can be abused. This article argues that media law must evolve to oversee platform system-design and challenge the algorithms which incentivise sensationalist or misleading content through virality. The editorial role assumed by algorithms is invisibly impactful, outmanoeuvring regulation under the guise of neutrality and creating an accountability gap which self-regulation has little incentive to fill. In evaluating the regulatory stance of the Online Safety Act (OSA) 2023, it becomes clear this legislation is unsuitable to reduce the viral spread of misinformation due to its hesitation to confront legal but harmful content. Despite OSA's shortcomings, its implementation marked a legal shift: system architecture is now subject to regulatory scrutiny. For a rapidly developing industry, system-design regulation offers the most future-proof and transparent avenue to ensure responsible dissemination of user-generated content. As such, three system-design proposals are explored, inspired by principles of existing media regulation, to assess their potential in translating onto social media platforms.

1 Introduction

The rise of social media has transformed the information age, ever since Facebook and YouTube emerged in the mid-2000s offering a platform for user-generated content. Over the last twenty years, social media has become a major site for news consumption across all age groups.¹ This shift reflects a loss of traditional media's monopoly as verified news appears alongside content from influencers and citizen journalists.² A turning point came in 2009 when Facebook and X (formerly Twitter) users reported Michael Jackson's death ahead of major networks, signalling a shift towards participatory journalism culture.³ Since then, journalism has undergone 'mediatisation', evolving into a 'false information Petri dish' where symbiotic relationships between online content and traditional media generate 'churnalism', news driven by 'click-bait' and sensationalist headlines.⁴ False news stories on X are now 70% more likely to be shared than true ones, leading the traditional media to criticise the internet as 'eroding the quality of the public's information environment'.⁵ As anyone can now publish globally with ease, the emergence of a 'Fifth Estate' raises questions about whether regulation should target the act of journalism, rather than the identity of the journalist. Using the misinformation surrounding Imane Khelif as a case study, this article examines how decentralised media power and engagement-driven algorithms challenge traditional frameworks of media accountability. It further evaluates the Online Safety Act 2023 and argues that media law must evolve beyond content-focused approaches towards oversight of the editorial structures shaping content dissemination on social media platforms. Finally, this article explores system-design interventions aimed at visibly disrupting the structural amplification of harmful online content.

¹ Lynn Schofield Clark and Regina Marchi, *Young People and the Future of News* (Cambridge University Press 2017), 29.

² Irini Katsirea, *Press Freedom and Regulation in a Digital Era* (Oxford University Press 2024), 41; Jennifer Alejandro, 'Journalism in the Age of Social Media' (2019) <<https://reutersinstitute.politics.ox.ac.uk/our-research/journalism-age-social-media>> accessed 8 June 2026, 41.

³ Jennifer Alejandro, 'Journalism in the Age of Social Media' (2019) <<https://reutersinstitute.politics.ox.ac.uk/our-research/journalism-age-social-media>> accessed 8 June 2026, 12.

⁴ Peter Coe, 'Tackling Online False Information in the United Kingdom: The Online Safety Act 2023 and Its Disconnection from Free Speech Law and Theory' (2023) 15(2) JML 217.

⁵ Uta Kohl, 'Toxic Recommender Algorithms: Immunities, Liabilities and the Regulated Self-Regulation of the Digital Services Act and the Online Safety Act' (2024) 16 JML 305; William H Dutton, 'The Fifth Estate Emerging through the Network of Networks' (2009) 27(1) P CSI 1, 9.

1.1 Case Study

This article examines the events surrounding Imane Khelif's gender eligibility controversy, to show the impacts of unchecked viral claims. Khelif, an Algerian boxer, competed in the 2024 Paris Olympics where her 46 second bout against Italy's Angela Carini drew intense online scrutiny. After stopping the match, Carini commented that her opponent was 'too strong', reigniting public debate over Khelif's disqualification from the 2023 Boxing World Championships over gender eligibility grounds.⁶ These previous ineligibility claims have been widely termed 'unsubstantiated', with the specifics of the eligibility tests remaining unclear.⁷ Khelif was cleared however, to compete at the Olympics after passing the International Olympic Committee's eligibility requirements and ultimately won a gold medal.⁸ Despite official backing, Khelif became the target of widespread cyberbullying, fuelled by media figures like Elon Musk and J.K. Rowling. With Musk being the first to post, Khelif expressed confusion as to why he 'led the attack', describing his behaviour as 'cruel'.⁹ Rowling amplified the controversy by sharing a picture of Khelif and Carini on X, claiming that Khelif was 'a male [...] enjoying the distress of a woman he's just punched in the head'.¹⁰ These posts triggered a surge of online transphobia and contributed to a prevailing narrative of misgendering Khelif, who reflected on the experience by saying:

My name and image have been used, unauthorised, to further personal and political agendas through the spreading and dissemination of baseless lies and misinformation.

–Imane Khelif¹¹

⁶ Shakiel Mahjouri, '2024 Olympics Boxing: Angela Carini Concedes Bout with Imane Khelif after 46 Seconds, Reigniting Gender Debate' *CBS Sports* (1 August 2024) <<https://www.cbssports.com/olympics/news/2024-olympics-boxing-angela-carini-concedes-bout-with-imane-khelif-after-46-seconds-reigniting-gender-debate/>> accessed 8 June 2026.

⁷ ITV News, 'Algerian Boxer at Centre of Gender Row Imane Khelif Wins Olympic Gold Medal' *ITV News* (9 August 2024) <<https://www.itv.com/news/2024-08-09/algerian-boxer-at-centre-of-gender-row-imane-khelif-wins-olympic-gold>> accessed 8 June 2026.

⁸ BBC News, 'JK Rowling and Elon Musk Reportedly Named in Imane Khelif Lawsuit' *BBC News* (14 August 2024) <<https://www.bbc.co.uk/news/articles/cm2njim4e2po>> accessed 8 June 2026.

⁹ Mouloud Achour, Interview with Imane Khelif, 'Imane Khelif, an Exclusive Interview for Clique' (2024) <<https://www.youtube.com/watch?v=-UaSFphxRug>> accessed 8 June 2026.

¹⁰ BBC News, 'JK Rowling and Elon Musk Reportedly Named in Imane Khelif Lawsuit' *BBC News* (14 August 2024) <<https://www.bbc.co.uk/news/articles/cm2njim4e2po>> accessed 8 June 2026.

¹¹ BBC Sport, 'Imane Khelif: Boxer Says She Is "Not Going Anywhere" Following Legal Action Announcement' *BBC Sport* (12 February 2025) <<https://www.bbc.co.uk/sport/boxing/articles/cwypzj4x0e6o>> accessed 8 June 2026.

1.2 Research Rationale and Theoretical Framework

Social media journalism enables misinformation to reach mass audiences with alarming speed and ease. New speakers, guided solely by their own standards, can pose as legitimate sources and distort public discourse. This threatens well-informed democratic participation, and leaves victims of targeted online harms, like Khelif, with little recourse to reputational damage and cyberbullying. While rumours and gossip are age-old features of human nature, the speed of information sharing in ‘internet time’ is especially concerning.¹² ‘Media figures’ are defined here as individuals who routinely publish to large audiences regardless of their credentials, while ‘informational harms’ refer to the spread of misleading information to shape narratives. Despite the ‘cyberspace fallacy’, the idea that the borderless internet is beyond regulation, online activity falls within the jurisdictions of actors or communication equipment.¹³

The Khelif case exemplified social media’s global reach, but this article will turn focus to the UK’s regulatory responses to information harms. The law lags behind technological and social change, so considering its socio-legal application is crucial in identifying shortcomings. Fourth-wave feminist legal theory informs this work, especially regarding the Khelif case where misogynistic and transphobic content is platformed. This theory enables deeper interrogation of narrative manipulation and structural bias in moderation policies that reinforce existing inequalities.¹⁴ An intersectional lens reflects the true nature and lived experience of reputational attacks on minority individuals like Khelif, whose cyberbullying was fuelled by

¹² William H Dutton, ‘The Fifth Estate Emerging through the Network of Networks’ (2009) 27(1) P CSI 1, 8.

¹³ Ibid, 18; David R Johnson and David Post, ‘Law and Borders: The Rise of Law in Cyberspace’ (1996) 48 SLR 1367.

¹⁴ Nancy Levit and Robert RM Verchick, *Feminist Legal Theory: A Primer* (2nd edn., NYU Press 2016), 44.

transphobia. Women of colour are particularly vulnerable to transphobic attacks due to the global model on ‘femininity’ being rooted in white experiences.¹⁵

2 Regulation Beyond Traditional Journalism

2.1 Media Laws and Social Media

The scope of media regulation has skyrocketed alongside the diversification of information sharing technologies. Rowbottom notes the contrast between broadcast and print regulation, the latter holding greater independence from government through traditional self-regulation systems.¹⁶ As broadcast media is consumed more passively, Katsirea believes this distinction was required so that the content ‘invading our living rooms’ was held to higher standards of impartiality and fairness.¹⁷ A contending explanation might be that broadcast and print media emerged at different points in time, though this rationale fails to address any social concerns that we rely on such regulations to remedy. While fewer people consume print media, the deliberate choice of selecting and reading a newspaper likely prompts more relaxed regulation since the reader is in more control of their consumption choices. The underlying need for all media regulation is to hold mass communicative power to account, and with the expanding opportunities to publish digital content, more people than ever are subject to media laws.¹⁸ Social media has presented numerous new problems for media law, though the most pressing in this context is the giant platform it provides for unregulated speakers. Furthermore, digital media convergence facilitated by social media makes the separate regulatory systems an outdated concept since both forms of content are being consumed in the same way: through our phone screens.

Social media accounts owned by broadcast and print companies retain their standard publishing rules but are often not the loudest voices on these sites. Both established celebrities and complete newcomers have the ability to gain millions of ‘followers’, and the reach of some viral posts can amass hundreds of millions of views. It is argued that social media provides

¹⁵ Kimberlé Crenshaw, ‘Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics’ (1989) 1989 UCLF 156.

¹⁶ Jacob Rowbottom, *Media Law* (Hart Publishing 2018), 257.

¹⁷ Irini Katsirea, *Press Freedom and Regulation in a Digital Era* (Oxford University Press 2024), 25.

¹⁸ Jacob Rowbottom, *Media Law* (Hart Publishing 2018), 321.

significant publishing power for media figures, so the question arises whether the social media platforms' self-regulated content restrictions are sufficiently safeguarding against unchecked mass communication.

The catalysing platform of Khelif's media coverage, social media site X, has held a reputation for being a 'real-time' source for news as users often tweet alongside events as they are unfolding.¹⁹ Although this reputation has been questioned following Musk's takeover of the platform, X remains a stimulus for elevating news stories to national prominence. This reporting capability became apparent in 2009 when social media users announced Michael Jackson's death before major networks, and has since turned the journalistic process on its head as journalists scroll through social media to gather stories.²⁰ At the time of writing, Elon Musk and J.K. Rowling (mentioned in Khelif's cyberbullying complaint) have over 230 million followers between them on X, but journalistic involvement in X's feed means that the reach of their posts surpassed their follower counts. While the new challenges waved in by social media are present in all its various sites, X is particularly notorious for its intersection with journalism and its influence seeps into other social media platforms and media outlets.

2.2 The Case for Regulating 'Big Soapboxes'

In 1997, U.S. Supreme Court Justice Stevens observed that the internet enables any person to become 'a town crier with a voice that resonates farther than it could from any soapbox'.²¹ While some view this as an expression of digital free speech, it does not negate the fact that our governing media laws were developed on principles becoming less relevant in the social media context. The free speech and media freedom foundations were established for powerful mass media institutions, at a time when it was unforeseeable that just anybody could gain mass publishing power.²² Although media figures cannot publish directly in newspapers or mainstream broadcasts, they have gained a level playing field on social media as uploading content for a global audience resembles an act of publication comparable to that of professional institutions. The ability to publish misinformation or poorly informed views to such a large

¹⁹ Irini Katsirea, *Press Freedom and Regulation in a Digital Era* (Oxford University Press 2024), 40.

²⁰ Jennifer Alejandro, 'Journalism in the Age of Social Media' (2019) <<https://reutersinstitute.politics.ox.ac.uk/our-research/journalism-age-social-media>> accessed 8 June 2026, 12.

²¹ *Reno v ACLU*, 521 US 844 (1997) 870 (Stevens J).

²² Jacob Rowbottom, *Media Law* (Hart Publishing 2018), 308.

audience is a harm that media law has traditionally sought to prevent. Since these harms can be posed by anybody with an internet connection, journalistic obligations should now extend beyond the legacy journalism industry.

In *Reno v ACLU*, Justice Stevens reasoned that the internet does not face the access limitations of broadcast media.²³ Therefore, a major challenge is that potential targets of regulation are no longer clearly defined by their institutional affiliations. Widening the scope of media regulation beyond the formal press would require careful consideration as to who should be subject to regulation. Though pragmatic considerations are discussed in section 4, one solution could be basing it on the size of users' platforms as the best predictor of publishing reach. A predictable counterargument is that user platforms themselves vary greatly in source types, yet this only strengthens the case for reach-based requirements. This is because social networking sites gather a diverse collection of voices into one feed, including regulated and recognised news broadcasters as well as entertainment accounts of varying formalities, advertisements, accounts run by other organisations, and those owned by our family and friends. The House of Lords identified the problem of integrating greatly varying content into one feed, claiming that it created confusion for users with the nature of content becoming 'impossible to differentiate'.²⁴

This effect is supported by numerous psychological biases, particularly the disorientating nature of social media feeds which activate heuristic processing (mental shortcuts) in its readers. Kahneman reported on heuristics before the conception of social media, highlighting the bias-inducing impact of information which can be retrieved over again with ease.²⁵ Developing this, Ali and Zain-ul-abdin found that social media heuristics are especially impactful because users are exposed to an endless and overwhelming amount of information sharing, especially when large platform holders activate the 'authority figure heuristic' in their followers.²⁶ The social media environment creates impracticalities in terms of time and mental resources to carefully judge the information audiences are presented with.²⁷ This is likely a

²³ *Reno v ACLU*, 521 US 844 (1997) 870 (Stevens J).

²⁴ House of Lords Select Committee on Communications, *Media Convergence and Its Public Policy Impact* (HL 2012–13, 154).

²⁵ Daniel Kahneman et al., *Judgement under Uncertainty: Heuristics and Biases* (Cambridge University Press 1982), 192.

²⁶ Khudejah Ali and Khawaja Zain-ul-abdin, 'Post-Truth Propaganda: Heuristic Processing of Political Fake News on Facebook during the 2016 U.S. Presidential Election' (2020) 49 JACR 1.

²⁷ Shelley Chaiken and Durairaj Maheswaran, 'Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgement' (1994) 66(3) JPSP 460.

cause of declining media literacy amongst social media users as its effects can increase the perception of message credibility and make ‘fake news’ appear a lot more real.

In Khelif’s case, the harm caused by the spread of misinformation affected her deeply and she experienced significant fear surrounding the claims.²⁸ Khelif stated that ‘Elon Musk was one of the first to attack [her]’ online and attributes the spread of the ‘online campaign’ to his comments.²⁹ Since online public discourse cannot be prevented, it begs the question whether the misinformation ‘campaign’ would have been as impactful had it not been for the involvement of media figures like Musk and Rowling. Some posts can go viral without high-profile endorsements, especially when they tap into existing public anxieties like the participation of transgender athletes in sports. However influential figures’ involvement often lends credibility to the claim due to the authority figure heuristic.³⁰ Further, celebrities making speculation around Khelif’s gender eligibility would have gained prominence on the X feed since it prioritises ‘elite interactions’, making it easier for mainstream media to scoop and insert into their stories, increasing the virality of the controversy.³¹ It is compelling that media figure involvement in this ‘campaign’ boosted its virality to international levels and therefore some responsibility should be attributed to the owners of large social platforms that decided to write and publish these claims into the public sphere. Individuals acting irresponsibly with their large publishing powers may not acknowledge the influence they hold or could be aware and choose to publish regardless: journalist-like regulation will help to address both situations. By imposing basic standards of transparency and accuracy to news-like claims, it firstly signals that there can be real and harmful outcomes to social media publications and secondly creates a duty to prevent the spread of misinformation.

2.3 Freedom of Expression Concerns

Regulating speech and expression must be balanced carefully when trying to curb abuses because censorship damages the search for truth.³² This is not just crucial in protecting this

²⁸ Mouloud Achour, Interview with Imane Khelif, ‘Imane Khelif, an Exclusive Interview for Clique’ (2024) <<https://www.youtube.com/watch?v=-UaSFphxRug>> accessed 21 March 2025.

²⁹ Ibid.

³⁰ Khudejah Ali and Khawaja Zain-ul-abdin, ‘Post-Truth Propaganda: Heuristic Processing of Political Fake News on Facebook during the 2016 U.S. Presidential Election’ (2020) 49 JACR 1.

³¹ Irini Katsirea, *Press Freedom and Regulation in a Digital Era* (Oxford University Press 2024), 40.

³² Onora O’Neill, ‘The Rights of Journalism and the Needs of Audiences’ in James Lewis and Paul Crick (eds), *Media Law & Ethics in the 21st Century* (International Bar Association 2014), 38.

media function, but in ensuring Article 10 of the European Convention on Human Rights (ECHR) guarantees freedom of expression rights to all users. This right ensures the ability of every citizen to impart and receive information online and may only be restricted in limited circumstances, of which protecting health and morals, and preventing disorder and crime are most applicable in this context.³³ A broad application of freedom of expression principles is necessary for protecting and contributing to the communication which a democratic society functions on. Citizen journalists and online activists play a vital role in diversifying the voices of public discourse, as there is great value to voices which do not reflect corporate interests and close gaps in reporting where the professional media struggle to reach.

In line with this, attention must be paid to the audience's right to receive information founded on honesty and competence when truth claims are made.³⁴ This links to the dominant reasoning of media regulation being to protect the public interest, which Rowbottom suggests is best served by 'leaving the media alone'.³⁵ However, this view may be criticised because failing to protect the needs of audiences leaves them vulnerable to information dissemination primarily geared towards the interests of the speakers. Ensuring audiences receive information in a responsible and transparent manner is a cornerstone of the public interest, and so resisting this obligation undermines the reasoning of why we need media laws in the first place. Katsirea supports the notion that democratic discourse is best enabled when both the imparters and receivers of information have their respective rights protected, critiquing the free press model as a manifestation of scepticism around State involvement.³⁶ The State's role in media regulation is not solely about restricting expression, but also safeguarding it when market forces fail to do so, ensuring that freedom of expression is meaningful for all.

An important consideration in respect of Article 10 ECHR is that it only protects against restrictions from the State, whereas social media platforms are private companies. These companies often have their own self-regulatory codes that users must follow in order to post content to their sites, a comparable system to print media. If online information was restricted, an important distinction is required as to whether the restriction is coming from the social media

³³ Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, entered into force 3 September 1953) 213 UNTS 221, art 10, largely incorporated into English law under the Human Rights Act 1998.

³⁴ Onora O'Neill, 'The Rights of Journalism and the Needs of Audiences' in James Lewis and Paul Crick (eds), *Media Law & Ethics in the 21st Century* (International Bar Association 2014).

³⁵ Jacob Rowbottom, *Media Law* (Hart Publishing 2018), 259.

³⁶ Irini Katsirea, *Press Freedom and Regulation in a Digital Era* (Oxford University Press 2024), 13.

platform or from the State, since their powers and applicability will vary. In light of this, it may be asserted that social media sites would maintain greater control over restricting harmful content if they are left independent from State regulation, although concerns could be raised over the accountability of this control. Leaving social media platforms with greater powers to navigate the removal of harmful content could lead to uncertainties; since the category of ‘legal but harmful’ content can be especially vague, and without clear boundaries to detail what users should not post, they could experience chilling effects from regulatory overreach. Conversely, laws requiring sites to remove such content could lead to over-censorship to avoid legal sanctions, highlighting the need for clearly defined removal policies for content.

Due consideration must be paid to the freedom of expression rights of speakers, not just their audiences, which Rowbottom explains to be a good reason to protect the comments of ‘low-level’ speakers.³⁷ The majority of individual speech falls under Mill’s concept of ‘self-regarding’ conduct, only affecting the speaker, meaning their unrestricted individuality should prevail above all other considerations.³⁸ However, Mill recognises the creation of harm when opinions are presented as ‘instigations’ to malicious acts, showing that online speech can cross the harm threshold when aimed at others in a way which incites hostility, ostracism, or reputational damage.³⁹ Defining ‘harm’ in the context of social media proves to be challenging because the durability and searchability of digital messages could mean they are more harmful than the same comments made offline. This makes ‘intent’ an unclear basis for determining social media harms. The cumulative impact of widespread misinformation can blur the line of when low-level speech becomes viral harm, which is why the need to focus regulation on the most powerful media voices and the social media platforms themselves becomes more prevalent. Preventing harms generated and amplified on social media presents a sound argument for intervention.⁴⁰ Without intervention, the free market determines content visibility, tailoring it for engagement rather than for truth and the public interest. If media law centres on protecting the public interest, then good reasoning can be made for the law to offer protection regardless of where the publication comes from.⁴¹ To assert this with the least

³⁷ Jacob Rowbottom, *Media Law* (Hart Publishing 2018), 323.

³⁸ John Stuart Mill, *On Liberty, Utilitarianism and Other Essays* (Oxford University Press 1991) 56.

³⁹ *Ibid*, 55.

⁴⁰ David Lyons, ‘Liberty and Harm to Others’ in Gerald Dworkin (ed), *Mill’s On Liberty* (Rowman and Littlefield Publishers, Inc 1997), 116.

⁴¹ Jacob Rowbottom, *Media Law* (Hart Publishing 2018), 322.

obstruction on user expression, the audience needs to be able to assess reasons for claims they read, a responsibility that should be placed on social media platforms to provide.⁴²

3 Assessing the Current Response to Informational Harms

3.1 User-driven Solutions

Some social media users challenge misinformation, through adding comments or reporting posts to moderators. However, such responses lack the speed and impact necessary to curb the rapid spread of misinformation. To highlight corrections from other users in a designated feed position, X has incorporated ‘Community Notes’ into feeds whereby users submit context they believe is needed on the original post, with their responses voted on by more users to determine its usefulness. Meta is following in their footsteps, claiming their initiative aims to restore ‘more speech and fewer mistakes’.⁴³ Meta argues that moderation efforts went ‘too far’, censoring harmless posts as collateral to removing harmful content.⁴⁴ Despite being designed and implemented by social media platforms, this policy relies on users to do the heavy lifting, and risks recalibrating their own feeds since algorithmic associations with users’ engagement is geared to continue displaying similar content.⁴⁵ Generating public debate and reasoning around misleading posts can be a positive step to redress harmful claims, but may also peddle the virality of these posts to appear on more users’ feeds. Pasquale describes the media power to include, exclude, and rank as the power to determine the longevity of public impressions, and by X and Meta handing some of this power to its users, it can be argued that these platforms are delegating a huge responsibility to a group with even less accountability.⁴⁶

Meta’s initiative also relaxes their independently set restrictions on speech relating to immigration and gender identity. These examples were given in Meta’s statement as they are recognised to prompt controversial discourse online. Justified as harmonising Meta’s standards

⁴² Onora O’Neill, ‘The Rights of Journalism and the Needs of Audiences’ in James Lewis and Paul Crick (eds), *Media Law & Ethics in the 21st Century* (International Bar Association 2014), 42.

⁴³ Joel Kaplan, ‘More Speech, Fewer Mistakes’ (*Meta* 2025) <<https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>> accessed 8 June 2026.

⁴⁴ Ibid.

⁴⁵ Sang Ah Kim, ‘Social Media Algorithms: Why You See What You See’ (2017) 2 GLTR 149.

⁴⁶ Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015), 61.

to those ‘on TV or the floor of Congress’, this approach appears naive to the snowball effect of online discourse, especially when sharing ideas that users may be discouraged from sharing in real life.⁴⁷ Transgender women’s identities are regularly belittled by the British media, normalising such attitudes and contributing to readers feeling more ‘affected’ by transgender people online than in real life.⁴⁸ Controversial topics gather more digital footfall by platforming the most discourse because the social media business model focuses on engagement-metrics (whether positive or negative), creating a feedback loop boosting polarising narratives.

Kim underscores the importance of high-engagement content to social media platforms, driving user engagement and therefore exposing them to more paid advertisements.⁴⁹ The risk of widespread misinformation is boosted by the amplification of negative interactions. Since scroll-worthy content is often controversial, ‘harmful’ ideas and views gain unprecedented virality, such as the hatred stirred up online ahead of the Southport Riots in 2024.⁵⁰ Such effects can also be seen by platforming dangerous claims regarding Khelif’s gender identity, accusing her of committing a crime by her country’s standards and exposing her to the severe effects of transmisogyny. Stepping away from content-based regulation, Meta are upholding their mission of reducing freedom of expression infringements, however their algorithm still serves this intensely profit-generating purpose. Therefore controversial topics are likely to be exploited for views at the cost of communities the discourse targets.

The most concerning aspect of Meta’s policy changes is the use of ‘less obtrusive labelling’ on posts which have been flagged as requiring fact-checking, removing full-screen warning overlays and making additional information more conspicuous, only to be found by ‘those who want to see it’.⁵¹ This could be viewed as a step backwards in the fight against misinformation as it exploits users’ heuristic processing by making context harder to find in an already endless stream of information. By relying on user-driven solutions, X and Meta are leaning on system-based regulation (changing the design of their feeds), but in a way that will generate further discourse and therefore profit. Adding to the disorientation of the social media feed, scaling

⁴⁷ Joel Kaplan, ‘More Speech, Fewer Mistakes’ (*Meta* 2025) <<https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>> accessed 8 June 2026.

⁴⁸ Shon Faye, *The Transgender Issue: An Argument for Justice* (Penguin Books 2021), 90.

⁴⁹ Sang Ah Kim, ‘Social Media Algorithms: Why You See What You See’ (2017) 2 GLTR 149.

⁵⁰ Andy Gregory, ‘How Lies about Southport Knife Attack Suspect Led to Riots and Clashes with Police’ (*The Independent* 31 July 2024) <<https://www.independent.co.uk/news/uk/crime/southport-attack-riots-far-right-social-media-b2588628.html>> accessed 8 June 2026.

⁵¹ Joel Kaplan, ‘More Speech, Fewer Mistakes’ (*Meta* 2025) <<https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>> accessed 8 June 2026.

back its current safeguards only fuels the ‘battle for mindshare’ and honours the economic functions that facilitate irresponsible information sharing.⁵²

3.2 Profit-driven Algorithms

Approximately 90% of social media users are passive and not engaging with content through likes, comments, or shares.⁵³ To curate personable content streams, sophisticated metrics were required to measure users’ scroll-depth and total reading time to monetise time spent looking at content and advertisements.⁵⁴ Schofield and Marchi argue this has driven journalism organisations to offer an unprecedented amount of content, employing ‘aggressive social media strategies’ in competition for views.⁵⁵ This has also prompted the rise of ‘click-bait’, where posts are intentionally provocative or misleading to attract digital footfall to their websites. As an extreme form of sensationalism, click-bait heightens the sense of competition for views, having a relegating effect on credible news stories as they appear less captivating in the feed. The extent to which ‘shareworthiness’ itself triggers news stories to be sensationalised is up for debate, yet it could be argued that unregulated information sharers are taking advantage of this effect because they can create a news-like product which is tailored to succeed in the social media landscape.⁵⁶ This competitive edge over their formal counterparts allows unverified content to soar to unrivalled prominence, calling into question the legitimacy of social media as a trustworthy information-hosting platform. By prioritising controversy for engagement, platforms create a disproportionate focus on marginal issues and disrupt the flow of contextually grounded information.

The influence of algorithm design on shaping social media narratives cannot be overstated. An exemplar event was Facebook’s algorithm change mid-2016 where posts from family and friends were boosted and mainstream news stories deemphasised. This prompted two concerning outcomes: a dramatic 26-47% fall in traffic to top news publisher’s websites, and the ease of sensational fake stories to circulate far more than legitimate news during the 2016

⁵² Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015).

⁵³ Sang Ah Kim, ‘Social Media Algorithms: Why You See What You See’ (2017) 2 GLTR 149.

⁵⁴ Lynn Schofield Clark and Regina Marchi, *Young People and the Future of News* (Cambridge University Press 2017).

⁵⁵ Ibid.

⁵⁶ Irini Katsirea, *Press Freedom and Regulation in a Digital Era* (Oxford University Press 2024), 37.

U.S. Presidential campaign.⁵⁷ In the absence of verified content, misinformation can escalate in social media feeds and Facebook's redesign underscored the editorial function of social media platforms to shape prevailing narratives through their curation of content feeds. This editorial role assumed by platforms demonstrates a legally significant activity when considering their responsibilities in supporting public discourse.

Katsirea terms content moderation as the service which sets social media platforms apart from the open web, and a great deal of trust is placed in these platforms to provide this impartially.⁵⁸ Although extensive editorial efforts happen in newsrooms, this authority is legitimised through regulatory compliance and adherence to basic standards of transparency and neutrality. Both traditional and social media rely on profit, but social media's vast content pool fuels a more aggressive competition for attention. It is counterintuitive that the media body with unparalleled amounts of content and users is faced with the least statutory regulation. User-driven solutions could be relatively obsolete when algorithms are programmed to platform low-quality content, raising the question of whether purely profit-incentivised systems can be a responsible medium for information dissemination. Facilitating change to the profit system of a \$250 billion industry is far from straightforward, and as self-regulatory initiatives are proving insufficient, the introduction of statutory duties for these media intermediaries is proving ever more necessary.

3.3 Online Safety Act 2023

Determining the role social media platforms should play in content regulation has been long disputed in the courts. Initially viewed as technology companies, it became apparent that providing and profiting from the use of their service required an assumption of responsibility.⁵⁹ The approach to determine this responsibility in England and Wales follows *Bunt v Tilley*, where the court must assess the company's level of complicity in moderating content.⁶⁰ This establishes whether the site had 'knowing involvement' of the harmful content published, a high threshold given the volume of new content shared every day.⁶¹ Rowbottom suggests this

⁵⁷ Lynn Schofield Clark and Regina Marchi, *Young People and the Future of News* (Cambridge University Press 2017).

⁵⁸ Irini Katsirea, *Press Freedom and Regulation in a Digital Era* (Oxford University Press 2024), 46.

⁵⁹ Jacob Rowbottom, *Media Law* (Hart Publishing 2018), 329.

⁶⁰ *Bunt v Tilley* [2007] 1 WLR 1243

⁶¹ *Ibid*, [23].

could deter social media platforms from taking an active role in content moderation to avoid the perception of assuming legal responsibility.⁶² However, intermediaries are generally granted civil and criminal immunities for third party content, and the European Union's Digital Services Act (DSA) reaffirms that voluntary initiatives undertaken in 'good faith and in a diligent manner' will not compromise intermediary immunities.⁶³ The OSA 2023 remains silent on these immunities, which presents the first significant shortfall. Despite a broadly adopted understanding that platforms curate the flow of online content, the immunities regime appears frozen in the early perception of platforms as 'passive providers of a neutral infrastructure'.⁶⁴ The OSA and DSA seek to restrain these harmful algorithms, yet this aim appears to be greatly undermined by the intact immunities regime. Kohl writes that content shared by individuals may be unsavoury but only creates harm when directed in a 'clustered and escalating way to users, revealing a missed opportunity for OSA to shift accountability to those truly responsible'.⁶⁵

The main thrust of OSA's system regulation is section 9, which requires social media platforms to conduct risk assessments on the ease of illegal content dissemination on their sites.⁶⁶ This includes assessing the role that the platform's business model, governance, and proactive technology has in illegal content shareability. Although compelling platforms to undertake these risk assessments is a spearheading provision, the findings of these assessments could vary widely because OSA allows platforms to use subjective interpretations of risks.⁶⁷ Ofcom has previously highlighted this form of approach is flawed unless supported by transparency reports, independent audits, guidelines, and training.⁶⁸ With failure to implement these requirements, risk assessments may not provide an accurate reflection of risks posed by the platform's system design because they lack any incentive or statutory requirements to do so.

⁶² Jacob Rowbottom, *Media Law* (Hart Publishing 2018).

⁶³ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1, art 7.

⁶⁴ Uta Kohl, 'Toxic Recommender Algorithms: Immunities, Liabilities and the Regulated Self-Regulation of the Digital Services Act and the Online Safety Act' (2024) 16 JML 305, 303.

⁶⁵ *Ibid*, 307.

⁶⁶ Online Safety Act 2023, s.9.

⁶⁷ Victoria Nash and Lisa Felton, 'Treating the Symptoms or the Disease? Analysing the UK Online Safety Act's Approach to Digital Regulation' (2024) 16(4) P&I 818.

⁶⁸ Ofcom, *Ofcom's First Year of Video-Sharing Platform Regulation* (2022) <<https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/vsp/reports/2022/2022-vsp-report.pdf?v=328583>> accessed 8 June 2026.

Further, not extending section 9 risk assessments to ‘legal but harmful’ (LBH) content appears off-target because the majority of harmful content amplified by algorithms falls under the LBH bracket. Illegal content has a far smaller risk of being misleading or given prominent placement in users’ feeds. The Online Harms White Paper (OHWP) sought to address the amplification of LBH content, yet was scaled back over censorship concerns.⁶⁹ Although the category’s subjectivity could lead to unnecessary content removals, the government’s move to substitute LBH provisions with section 15’s ‘User empowerment duties’ does not provide adequate safeguards around LBH content.⁷⁰ Providing adults with content control features should be welcomed, but offering harm reduction as an option and not a site-wide policy appears overcautious. Social media platforms are required to provide ‘clear and accessible’ instructions for how users can utilise the s. 15(7) control features, but these could be buried in the terms of service and inconvenience users.⁷¹ Such tools need to be eye-catching because the curation of LBH content is intentionally disorientating and emotion provoking, so it must be able to target those with the weakest media literacy skills who fall vulnerable to it.

The most significant shortcoming is the Act retaining the legacy media framework which regulates our traditional institutions. Given social media’s distinct nature from legacy media, the shift to system-based regulation needed to be more comprehensive. To regulate system design is to regulate the editorial role that social media platforms assume. The main pushback on system design came from the House of Commons over concerns for delay and ‘legal confusion’, yet this represents a failure to produce a relatively future-proof piece of legislation, which could have been secured with intuitive system design policy.⁷² Content-focused regulation only addresses the surface issues of the distorted information pool, whereas system-focused regulation can tackle the root cause. Were the OSA to be wholly system-based, there would be stronger protection of freedom of expression principles since the value judgements associated with traditional moderation are not required to be given on pieces of content.⁷³ The reluctance to regulate the spread of LBH content is rooted in the legacy framework requirement to assess the piece of content concerned, but fully embracing system-based regulation

⁶⁹ Peter Coe, ‘Tackling Online False Information in the United Kingdom: The Online Safety Act 2023 and Its Disconnection from Free Speech Law and Theory’ (2023) 15(2) JML 217, 229.

⁷⁰ Irini Katsirea, *Press Freedom and Regulation in a Digital Era* (Oxford University Press 2024), 63.

⁷¹ Online Safety Act 2023, s.15(7).

⁷² HC Deb 12 September 2023, vol 737, col 807.

⁷³ Victoria Nash and Lisa Felton, ‘Treating the Symptoms or the Disease? Analysing the UK Online Safety Act’s Approach to Digital Regulation’ (2024) 16(4) P&I 818.

acknowledges the platforms' culpability in spreading online harms and seeks restraint on these mechanisms.

Many freedom of expression concerns over the OSA likely stem from its co-regulatory model, extending State-level limits onto social media moderation.⁷⁴ This is disjointed from the 'hands-off' approach system-based design takes on subjective content censorship, but the government's inability to harness this could suggest severe pragmatic concerns around implementation. Coe warns that by making services responsible for content curation, the OSA could enhance their ability to undertake an invisible editorial role. Nevertheless these platforms already engage in such practices, revealing a gap in editorship regulation which the OSA is beginning to address.⁷⁵ However as long as social media systems incentivise 'churnalism', there is no assurance that we will find quality information on their platforms.⁷⁶ We have not yet seen the Act's full manifestation, but it is already clear that the OSA is not the 'radical panacea' for online harms that it was purported to be.⁷⁷

4 Design-based Solutions to Disrupt Informational Harms

4.1 The Case for Legal Intervention

The economic incentives behind social media content curation suggest self-regulation cannot protect the public interest. O'Neill argues that self-regulation has been neither effective nor ethically sufficient.⁷⁸ States have grown more willing to hold major technology companies accountable, prompting efforts to expand their corporate responsibilities.⁷⁹ The OSA is an early

⁷⁴ Uta Kohl, 'Toxic Recommender Algorithms: Immunities, Liabilities and the Regulated Self-Regulation of the Digital Services Act and the Online Safety Act' (2024) 16 JML 305, 333.

⁷⁵ Peter Coe, 'Tackling Online False Information in the United Kingdom: The Online Safety Act 2023 and Its Disconnection from Free Speech Law and Theory' (2023) 15(2) JML 217, 239.

⁷⁶ *Ibid*, 217.

⁷⁷ *Ibid*, 240.

⁷⁸ Onora O'Neill, 'The Rights of Journalism and the Needs of Audiences' in James Lewis and Paul Crick (eds), *Media Law & Ethics in the 21st Century* (International Bar Association 2014), 44.

⁷⁹ Victoria Nash and Lisa Felton, 'Treating the Symptoms or the Disease? Analysing the UK Online Safety Act's Approach to Digital Regulation' (2024) 16(4) P&I 818.

attempt to regulate system design, but gaps remain in addressing the rapid spread of harmful content.

Since algorithm design causes real-world harm, like Imane Khelif's controversy, an analogy can be drawn: just as product manufacturers must prevent foreseeable harm to its users, algorithm designers should owe a similar duty. Although less tangible, these harms resulting from service use still present a case for provider liability. Perrin and Woods' paper behind the Online Harms White Paper (OHWP), proposed an overarching duty of care for online service delivery, which was included in the OSA but later dropped as a main principle.⁸⁰ Had this duty been widely adopted, the OSA could have created stronger preventative measures to force social media platforms to assess all forms of harm, not only illegal content. While the OSA seeks to 'rebalance corporate conduct in favour of the public interest', it fell short of implementing broader statutory duties.⁸¹

A statutory basis behind social media regulation could enable stronger sanctions, while reducing censorship concerns by targeting platform design instead of users' content.⁸² Rowbottom cautions that while State regulation can be a sign of legitimacy, concerns over political influence make transparency essential.⁸³ One argument could be that the explosion of online voices has diluted old media power, calling for less regulation. But this assumption overlooks the shift in power from speaker to platform editor. Recognising social media platforms are no longer neutral intermediaries is key to effective user protection.⁸⁴ These proposals are underpinned by the aim of disrupting the invisible editorial control of platforms by introducing transparency and labelling requirements, making feeds less disorientating. This could be implemented by amending s. 9 OSA to include all content in risk assessments, allowing Ofcom involvement in the oversight of feed transparency. This would improve market

⁸⁰ Lorna Woods and William Perrin, 'Online Harm Reduction – a Statutory Duty of Care and Regulator' (*Carnegie Foundation* 2019) <<https://carnegieuktrust.org.uk/publications/online-harm-reduction-a-statutory-duty-of-care-and-regulator/>> accessed 8 June 2026.

⁸¹ Victoria Nash and Lisa Felton, 'Treating the Symptoms or the Disease? Analysing the UK Online Safety Act's Approach to Digital Regulation' (2024) 16(4) P&I 818.

⁸² Onora O'Neill, 'The Rights of Journalism and the Needs of Audiences' in James Lewis and Paul Crick (eds), *Media Law & Ethics in the 21st Century* (International Bar Association 2014).

⁸³ Jacob Rowbottom, *Media Law* (Hart Publishing 2018), 270.

⁸⁴ Uta Kohl, 'Toxic Recommender Algorithms: Immunities, Liabilities and the Regulated Self-Regulation of the Digital Services Act and the Online Safety Act' (2024) 16 JML 305.

fairness between media figures and regulated news outlets, helping high-quality journalism regain visibility over viral misinformation.

4.2 Separate Feeds for Verified News Sources

The first proposal introduces a separate feed for content from reputable news organisations. This aims to combat epistemic disorientation experienced when viewing mixed content types, offering an anchoring source of truth. Broadcasting requirements in the Communications Act offer precedent, where platforms like Freeview and Sky must ensure prominence for public service broadcasters, such as the BBC.⁸⁵ However, extending this principle to social media raises two key challenges. First, determining which news outlets qualify as reputable sources brings difficulties in practice. Ideally independent regulators, like Ofcom, would assess eligibility on standards of fairness and impartiality. This would involve close cooperation between Ofcom and platforms, as too much discretion from either body risks politicising source selection. Further, this supports traditional outlets competing with the shareability of informal news because scaling down their regulation to level the field only results in a ‘race to the bottom’.⁸⁶ If the aim is to heighten standards, unregulated high-level speakers would have to meet regulatory criteria to break through to this reputable publishing space.

Secondly, most vitally, separating content risks intensifying poorly informed ‘echo chambers’. Whilst this proposal highlights verified information sources, removing them from the main feed could accelerate the spread of misinformation due to claims losing direct contention. As demonstrated in section 3, quietening mainstream news on social media allowed sensationalist falsehoods to circulate Facebook in 2016, showing that ‘fake news’ thrives in the absence of factually based counterpoints.⁸⁷ Separating, but still displaying, verified sources could have similar effects as this feed might only engage users actively seeking news stories, rather than the majority who encounter it passively. Further, Lewandowsky and others show that factual sources are often ineffective at challenging false perceptions, even when placed closely to the

⁸⁵ Communications Act 2003, s.310.

⁸⁶ Jacob Rowbottom, *Media Law* (Hart Publishing 2018), 327.

⁸⁷ Lynn Schofield Clark and Regina Marchi, *Young People and the Future of News* (Cambridge University Press 2017).

initial misinformation.⁸⁸ This raises doubts about the success of the reputable feed to anchor quality sources, or whether the split-feeds would alter users' perceptions on misinformation.

The split-feed rationale is undermined when factual news content is inevitably shared on the main feed by journalists or content creators with no desire to move to the reputable feed. Fierce contention over source selection could create a two-tiered information landscape, unfairly excluding emerging reporters and resulting in a more disorientating social media experience to the one at present. Despite the aim of rebalancing attention economy around quality journalism, too many risks arise.

Had this been in effect during the summer of 2024, misinformed claims about Khelif's gender identity would have been restricted to the unregulated feed. Although journalists would undoubtedly report on the speculation, bringing the discussion into the regulated feed would have added vital context. This might have made clearer the falsity of speculative claims by offering reputable outlets a designated spot to respond. Yet the reputable feed would still fall outside of most users' primary experiences, limiting its impact to those already engaged with verified news. This would be poorly aimed, since this group is already less susceptible to misinterpreting or spreading misinformation. The proposal's ambition to restore trust and structure fails to amount to worthwhile redress because its success hinges on unpredictable factors like user behaviour and platform collaboration.

4.3 Source Tags for High-Reach Posts

The second solution involves tagging high-reach posts to denote the content's source type. This could directly address interpretive confusion and prime content expectations based on source type. Labels could include: opinion, entertainment, factual reporting, and advertisements. Media convergence on the social media feed has collapsed content genres, so source tags would help distinguish tone, purpose, and factual basis. Responsible information dissemination requires enabling audiences to assess reasons behind claims, and gauge the interests it may serve.⁸⁹ Source tags expand upon online advertising disclosures set by the Advertising

⁸⁸ Stephan Lewandowsky et al., 'Misinformation and Its Correction: Continued Influence and Successful Debiasing' (2012) 13(3) PSPI 106.

⁸⁹ Onora O'Neill, 'The Rights of Journalism and the Needs of Audiences' in James Lewis and Paul Crick (eds), *Media Law & Ethics in the 21st Century* (International Bar Association 2014), 42.

Standards Authority to uphold users' rights to clear and honest information. If the OSA seeks to promote user empowerment, then guaranteeing such clarity is a more direct method than content filtering options. Source tags would also promote accountability in exchange for visibility, an important media law principle. Requiring tags on posts surpassing a high impression threshold means that viral content must disclose its basis. With careful implementation, tags could serve broader transparency obligations expected in modern feed curation.

A key challenge is determining how 'high-reach' is defined. A fixed numerical base could offer certainty to content creators but may disproportionately impact smaller social media platforms. To ensure fairness, the threshold could vary with platform size and user base. This follows a core principle of the OSA: that regulation is proportionate to the scale and potential harm of the service.⁹⁰ Whilst fewer impressions could quickly trigger tagging on smaller sites, it ensures that source transparency is given when posts gain prevalence on their original platforms. Since mistaken beliefs often persist after correction, tagging once a threshold is reached provides no remedy for the initial audience.⁹¹ This supports lowering the threshold, but a balance must be struck to not unfairly burden users. Another consideration is determining who assigns the tags. If assigned by the content creator, greater regard for freedom of expression is ensured but sites must provide appeal or correction routes to prevent abuse. This could repurpose X and Meta's 'Community Notes', involving users in source-type correction rather than the onus of providing interpretive context. To prevent bias in tag assignment, both users and the platform should be involved in its review, though this burden is limited by the high-reach criteria. Since tagging definitions cannot rest with a small group of editors, coordinating all impacted parties is this proposal's greatest challenge.⁹² As a media literacy aid, tags must be prominent, consistent, and educational. Tagging high-impact posts could counter the decreasingly reliable way to 'define boundaries of separate media industries and interests'.⁹³

⁹⁰ Ofcom, *Risk Assessment Guidance and Risk Profiles* (2024) <<https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/risk-assessment-guidance-and-risk-profiles.pdf?v=390984>> accessed 8 June 2026.

⁹¹ Lewandowsky and others (n 88) 109.

⁹² Onora O'Neill, 'The Rights of Journalism and the Needs of Audiences' in James Lewis and Paul Crick (eds), *Media Law & Ethics in the 21st Century* (International Bar Association 2014), 48.

⁹³ House of Lords Select Committee on Communications, *Media Convergence and Its Public Policy Impact* (HL 2012–13, 154) <<https://publications.parliament.uk/pa/ld201213/ldselect/ldcomuni/154/154.pdf>> accessed 8 June 2026.

Critics may view the tags as implied value judgements, especially ‘opinion’ and ‘factual’. This distinction is central to most content, so will be unavoidably contentious. Clear definitions and transparent tagging criteria could help to mitigate this, but the system must be framed as a contextual, not evaluative, tool. A valid concern is that multiple highly similar posts could reach high-visibility without individually activating tags. Decentralised posting strategies could bypass tagging requirements and undermine the effectiveness of impressions-based thresholds. While this complicates implementation, cumulative reach monitoring is no novel burden as it already exists in disinformation campaign oversight.⁹⁴ However, regulatory burden would increase, so refinements are required to address content replication that could circumvent tagging.

If this policy had been active during the 2024 Olympics, viral speculative posts would have needed to disclose their unproven nature to remain visible. This would have notified users about the nature of content, reducing misinformation by acting as a ‘pre-exposure warning’.⁹⁵ Helping the audience contextualise motivations behind viral content could have lessened reputational harm from unverified claims, restoring the interpretive lens often lost in social media convergence. Source tags offer a promising mechanism for contextualising content, but the need for clear definitions, appeal routes, and monitoring of cumulative reach shows this system requires refinement. With a transparent and appealable structure, this literacy-based tool strengthens conditions under which users encounter information, signalling clear responsibility onto those with unchecked communicative power.

4.4 Algorithmic Transparency Tags

The final proposal introduces tags denoting the platform’s reason content is shown on feeds. Algorithmic Transparency Tag (ATT) categories could include: ‘interacting with similar content’, ‘recommended based on personal data’, ‘sponsored’, ‘follower-based’, or ‘trending in your area’. The media industry promotes transparency, but lacks the disclosure duties of

⁹⁴ Samantha Bradshaw, Hannah Bailey and Philip N Howard, ‘Industrialized Disinformation’ (*Oxford Internet Institute* 2020) <<https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2021/02/CyberTroop-Report20-Draft9.pdf>> accessed 8 June 2026.

⁹⁵ Stephan Lewandowsky et al., ‘Misinformation and Its Correction: Continued Influence and Successful Debiasing’ (2012) 13(3) PSPI 106, 116.

other public institutions, so this proposal directly disrupts algorithmic opacity.⁹⁶ ATTs empower critical engagement, reshaping content interpretation and challenging assumptions of fairness, through visible signals of behind-the-scenes decision making. This directly aligns with Article 27 of the Digital Services Act (DSA), disclosing recommender systems at the user interface level, rather than buried in terms of service.⁹⁷ Real-time explanations of profiling can further contextualise influence, particularly in political, social, or identity-based content where algorithmic amplification can structure prevailing narratives.⁹⁸ To provide meaningful context, ATTs require expandable short explanations of content placement and recommender systems in use. Transparency tools are difficult to oppose on principle and may pressure platforms to construct algorithms more responsibly, as ATTs signal shortcomings in providing a reliable information environment.

A central risk is that ATTs could worsen cognitive overload in already information-dense content feeds. Users may ignore or misinterpret tags, unless paired with education and intuitive designs that enable informed decision making. With clear, brief, and visible ATTs, users could gain greater control over their content feeds. Where interacting with content means sensationalised topics still gain virality, ATTs could help users reduce repeated exposure. I believe this supports user empowerment by providing tools to directly respond to the algorithmic forces that shape users' experiences. ATTs therefore function as both algorithmic disclosures and mechanisms for user intervention.

Effective monitoring is crucial because platforms will be reluctant to reveal exploitative editorial patterns. Opening platforms up to public scrutiny may undermine trust in their service, which could lead them to use vague tags like 'Suggested for you' to avoid backlash and generic labels fail to meaningfully inform, undermining the intent of ATTs. Robust oversight involving field experts is essential to monitor fair implementation and reduce the impact of platform self-interest. The ATT approach confronts algorithmic choices directly by exposing and challenging their role in incentivising informational harms.

⁹⁶ Onora O'Neill, 'The Rights of Journalism and the Needs of Audiences' in James Lewis and Paul Crick (eds), *Media Law & Ethics in the 21st Century* (International Bar Association 2014), 45.

⁹⁷ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1

⁹⁸ Khudejah Ali and Khawaja Zain-ul-abdin, 'Post-Truth Propaganda: Heuristic Processing of Political Fake News on Facebook during the 2016 U.S. Presidential Election' (2020) 49 JACR 1.

If implemented before the Paris Olympics, this policy would have fostered online transparency and exposed mechanisms that amplify prejudicial content. Real-time contextualisation may have slowed the viral momentum of baseless accusations, by prompting users to reassess credibility. Achieving this outcome depends on striking a balance between the conciseness of tag explanations and sufficient detail to maintain their usefulness. By showing why posts repeatedly appear on users' feeds, ATTs reveal feedback loops and in turn empower users to engage intentionally. ATTs may not prevent vulnerable groups, such as transgender individuals, from being drawn into sensationalist headlines. However, reducing user engagement with such content lowers their visibility and profitability, reducing the incentive to utilise inflammatory content for engagement-driven gain. With strong oversight and public accountability, ATTs offer a promising recalibration of online responsibility. Despite requiring more educational groundwork, ATTs could ensure long-term redress of the systematic imbalances fuelling widespread digital harms.

Each solution presented draws on existing media policy and could enhance online experiences. By targeting system design, they circumvent the usual enforcement issues when the perpetrator is overseas, a key hurdle in Khelif's cyberbullying complaint.⁹⁹ Despite growing awareness, misinformation remains too opaque for audiences to successfully differentiate fact from fiction.¹⁰⁰ This presents a case for regulatory intervention because although worldview shapes individual interpretations, visible interference to platform design could foster a 'healthy degree of scepticism', contributing to a better-informed population.¹⁰¹ Implementing harm redress directly onto feeds affirms the necessity of legal intervention, not for censorship, but to compel structural accountability in a relatively unchallenged industry.

5 Conclusion

5.1 Summary of Findings

Through social media's rise into the information sharing sphere, high-reach publishing power is no longer exclusive to the media industry. This power has been decentralised and is now

⁹⁹ Jacob Rowbottom, *Media Law* (Hart Publishing 2018), 318.

¹⁰⁰ Stephan Lewandowsky et al., 'Misinformation and Its Correction: Continued Influence and Successful Debiasing' (2012) 13(3) PSPI 106, 109.

¹⁰¹ Ibid, 124.

shared to anybody with internet access. A new body of unregulated speakers, with unprecedented publishing capabilities, have been exploiting these platforms over the last decade, but the law has not caught up. Citizen journalists and influencers can mimic news products without the corresponding regulatory scrutiny, as was shown in the Khelif case where misinformation spread like wildfire under the guise of honest reporting. This highlights a critical gap in media law frameworks whereby substantial control of the public narrative is assumed by speakers with no obligation to act transparently or fairly. Section 2 also explored the freedom of expression implications for addressing this new body of speakers, underscoring the need to balance their exercise of free speech with the audience's right to high-quality information which supports democratic processes.

Section 3 expanded upon the reports of social media creating a heuristically disorientating environment; by analysing the ways certain content is amplified for higher user engagement. Controversial and misleading content, by its very nature, attracts the most user engagement which in turn brings the most views to paid advertisements. The platforming of such topics is disproportionate to the marginal issues they often represent, and the amplification of reaction-provoking misinformation shows little regard from social media platforms for the public interest. Profit incentives are continually prioritised over informational integrity, allowing algorithmic systems to assume an editorial role without the editorial responsibilities imposed on the rest of the media. This structure enabled unverified speculation around Khelif to soar to mainstream reporting heights, causing reputational harm and instigating alarming amounts of transmisogyny. Such profit incentives make self-regulation an inadequate safeguard against viral harms, and whilst the OSA 2023 introduced a shift to system-based regulation, it remained silent on the more evasive legal but harmful content or to enforce any meaningful changes in platform conduct. The OSA's reluctance to fully embrace system-based regulation limits its longevity in a rapid-paced area of technological development.

Finally, section 4 explored three design-based solutions to be implemented under comprehensive system regulation. Viral misinformation cannot be remedied through content restrictions without raising censorship concerns or being limited to reactive measures, which is not effective on social media's rapidly moving platforms. Instead, structural transparency tools offer recalibration of the digital information environment with less intrusion on free expression. Algorithmic transparency tags, as a highly visible extension of existing social media policy, provide purposeful user empowerment tools and disrupt the invisible curation of

the social media feed. These tags would have slowed the spread of misinformation about Khelif, by contextualising its virality and supporting users to make informed judgements over content credibility. The law must reclaim oversight of social media platforms' editorial powers, by imposing transparency and accountability at the design level.

Ultimately, this article has argued that the decentralisation of media power must be addressed through the social media platforms responsible for the aggregation and amplification of low-quality information. The social media environment allows any source to purport credibility, a regulatory gap exposed by the rapid virality of misinformation about Imane Khelif. While the OSA marked a crucial step towards system-based regulation, it did not hold social media accountable for its systems which structure and distort public discourse.

5.2 Future Directions

Looking beyond this article, future research could examine how long-established concepts of editorial responsibility and public interest may evolve into obligations for algorithm-driven services, particularly in assessing their relevance within social media as the Fifth Estate. Comparative analysis of social media regulation would further support in holding platforms accountable across jurisdictions, reflecting their inherently cross-border nature promoting more responsible platform architecture. Additional research might evaluate the effectiveness of the proposed interventions in mitigating heuristic processing among social media users, or explore legal pluralistic approaches that extend beyond statutory frameworks. Finally, while these proposals are intended to function as media literacy tools, exploring media education initiatives would further empower users and help reduce the persuasive credibility of misinformation.