# BEYOND THE SAFE HARBOUR

## Navigating the 'Fake News' Conundrum

JYOTI PANDAY

*'From the thicket of relations between fiction and truth we have seen a third term emerge: the false, the non-authentic—the pretence that advertises itself as true.'*

— Carlo Ginzburg (2012)

Digital intermediaries, such as Facebook, YouTube, Instagram and Twitter, weave together countries and people in ways undreamed of even a few decades ago. Online platforms, while radically increasing our opportunities for cross-cultural communication, also amplify the potential consequences of our communication which are not always to beneficent effect. Following the 2016 US presidential elections, the manipulation and abuse of digital platforms by state and non-state actors in influencing critical social and political events has attracted global attention.

From the Brexit referendum (Wright, et al., 2019) to the campaign of ethnic cleansing against Myanmar's Rohingya Muslim minority (Mozur, 2018), a specific concern has been the rise of 'fake news' on popular social networking sites and messaging apps. While hoaxes, conspiracy theories and fabricated information have been documented throughout the history of communication, the centrality of platforms in our digitally connected worlds and the scale of 'fake news' presents an unprecedented challenge.

The label 'fake news' is popularly used to refer to a broad range of perceived and deliberately manipulated messages and information. Researchers working on the issue emphasise that the term inadequately captures the different ways in which

communication gets distorted, and prefer to use the term 'information disorder' (Wardle and Derakhshan, 2017).

The destructive impact of information disorder ranges from: (i) enabling foreign interference and propaganda; (ii) undermining trust in institutions that facilitate the functioning of democracy; (iii) the disruption of democratic deliberation; and (iv) contributing to violent ethnic and religious nationalism. While much of the discussion on the impact of false and misleading information has been focused on its political dimensions, information disorder muddies public discourse and trust on a range of issues.

A recent investigation by *The Guardian* uncovered that search results for information about vaccines on Facebook and YouTube were dominated by recommendations that steered viewers away from fact-based medical information toward deliberate misinformation meant to undermine trust in vaccines (Wong, 2018). In addition to hosting groups and channels promoting anti-vaccination propaganda, Facebook also accepted thousands of advertising dollars from organisations to target people interested in 'vaccine controversies' (Pilkington and Glenza, 2019).

In India, the Ministry of Electronics and Information Technology (MEITy) sent notices, ordering Google and Facebook to remove false and 'malicious' content on food safety, noting that 'such false propaganda...erodes trust in the global food system and potentially has far-reaching public health, social and trade implications' (Doval, 2019).

The complexities of information disorder raise the question who, if anyone, is responsible for the impact and consequence of the various modes of contamination in the information ecosystem, especially when those impacts and consequences are ethically problematic. In December 2018, MEITy announced its review of the regulatory treatment of digital intermediaries to hold them accountable under law for 'fake news'. The draft amendments present a striking contrast, both to the way in which the government has approached the regulation of intermediaries, and to the manner in which online platforms view their role regarding contentious content.

In this article, I argue that the focus on technological solutions outlined under the draft regulations are short-sighted interventions that fail to address the broader issue of platform responsibility.

Information disorder is a complex policy problem, and designing interventions to address it necessitates establishing accountability for the proliferation of false and manipulated content. However, this does not mean ceding control to unaccountable private companies when they seem ill-equipped to deal with the protection of speech.

I lay out my argument in three parts. First, I examine efforts to define and delineate types of 'fake news'. Next, I discuss the role of online platforms in facilitating the spread of contentious content online. In part three, I discuss proposals for increased liability for platforms, focusing on the viability and desirability of the changes being contemplated to the legal framework that governs intermediaries in India.

## NEITHER FAKE NOR NEWS

At the outset, one of the most challenging aspects of addressing the phenomenon dubbed 'fake news' is defining what constitutes fake news. The expression has been appropriated by politicians to dismiss disagreeable news coverage and is often used as a synonym for inaccurate journalism, propaganda, conspiracy theories, hoaxes, lies, fabricated pictures, and even Internet memes.

Claire Wardle has approached the categorisation of information disorder by distinguishing between messages that are true from those that are false, and messages that are created, produced and distributed by agents with an intent to harm from those that are not (Wardle and Derakhshan, 2017). The framework has been immensely useful in distinguishing between various forms of information disorder, and identifying factors contributing to the production and consumption of false information.

Under this categorisation, misinformation refers to false content being shared without an intent to harm. Although some misinformation can be relatively innocuous, it can still lead to harm: for example, when unverified information shared during public emergencies or calamities creates panic or disorder. People share misinformation for a variety of reasons, including as validation of their identity and belief systems, or as a form of civic duty. As a BBC study on fake news and citizens highlights, in some instances simply the act of seeking to validate the veracity of misinformation contributes to different types of false messages being shared widely within networks (Chakrabarti, et al., 2018).

Disinformation refers to scenarios where false information is knowingly created and disseminated to sow mistrust and confusion, or harm a person, social group or country. One striking example of disinformation is the fake video floated by a member of the Legislative Assembly from the then Opposition, and now ruling party in India, the Bharatiya Janata Party (BJP), which snowballed into the Muzaffarnagar riots of 2013 that claimed 60 lives and displaced more than 40,000 people (Ahuja, 2013). Alarmingly, Prime Minister Narendra Modi's official mobile application has emerged as a major hub for the dissemination of disinformation where supporters, including several BJP party members, often post objectionable content, including fake quotes from leaders of rival parties (Bansal, 2019).

The focus on truth assumes the existence of a single verifiable version or source of truth. In reality, the line between truth and untruth is often difficult to draw. This becomes immediately evident when we consider the growing problem of deep fakes or the phenomenon of hyper-realistic video or audio recordings, created with artificial intelligence (AI). This approach also assumes that people share falsehoods, because they do not have access to trustworthy information. However, a recent study from the Massachusetts Institute of Technology (MIT) has uncovered the novelty of false or manipulated information, and people's emotional reactions to it contribute to false content being shared more than accurate reportage online (Vosoughi, et al., 2018).

Similarly, the metric of intent creates the perception of objectivity in the process of evaluating information. In practice, identifying intent behind misinformation and disinformation, and taking action against such content entails taking messy and politically fraught decisions. For example, in 2016, Google amended its policy to restrict advertisements that 'misrepresent, misstate, or conceal information'. This move has resulted in reducing the number of channels that qualify for earning advertising revenue, a trend which impacts both long-term and new users of the platform. In the absence of clear guidelines or oversight, content creators are being demonetised without an explanation.

As the moderating practices of social networking platforms reveal, identifying truth or intent is not an easy task, since a lot of content lies in the middle ground between the intentionally deceitful

or misleading, and accurate, factual information.[1] Platforms' strategies, based on interpreting truth or intent, break down in environments where, for a significant part, users determine what content they upload, amplify and are exposed to. As the MIT study on rumour cascading on social networking platforms highlights, false information spreads because people share it. In this regard, the motivations of actors or organisations that engage with the content of a post, or amplify it, becomes as important as identifying the intent of the source of content (Helberger, et al., 2018).

The significance of social sharing becomes particularly evident when we contemplate the string of attacks across our country that have been fuelled by the circulation of fake messages on social networking sites. More than 30 people across 10 states became victims of vigilante justice, triggered by rumours and fake warnings of kidnappers or organ harvesters that were circulated on WhatsApp. Responding to the violence, the government has sought to allocate the responsibility of curbing the circulation of false information to online platforms.

By and large, the government's demands have been limited to seeking 'technological solutions', such as traceability and the use of AI and machine-learning systems, to track content. This approach towards regulating content on the Internet is as much due to the complexities of online content distribution, as it is to the fact that digital intermediaries qualify for 'safe harbour' and cannot be held liable for fake news or other types of contentious content on their platforms.

## A SHIP IN HARBOUR IS SAFE

Although much of the recent discourse on information disorder in India has revolved round the question of whether platforms can be held accountable for content shared through them, as per Section 79 of the Information Technology Act, 2000 (IT Act), intermediaries are shielded from civil or criminal liability for any third-party content made available by, or hosted on, them. In return for this legal immunity or safe harbour, intermediaries have to comply with certain obligations, such as adopting statutory due diligence, or enforcing 'notice and takedown' procedures.

The expansive safe harbour regime in India provides little recourse for regulators, even in instances when platforms have

facilitated state-sponsored information warfare (Singh, 2018). A few countries have sought to frame information disorder as a cybersecurity issue, and introduce penalties on the platforms' failure to prevent or facilitate information warfare. However, this response targets one form of information disorder—state-sponsored disinformation campaigns—leaving other forms of strategic manipulation of information unaddressed.

Section 79 of the IT Act was challenged in the Supreme Court of India (SC), in *Shreya Singhal v. Union of India*[2] (Shreya Singhal case). The SC upheld intermediary safe harbour, and strengthened it by requiring a judicial or executive review of content removal requests. As a result of the standards set by the SC in the Shreya Singhal case, intermediaries are deemed to have no knowledge of unlawful content, and are not required to take down content until the receipt of government or judicial order. Importantly, intermediaries in India are not required to proactively monitor their platforms to track or remove contentious or harmful content.

The absence of legal requirements has not prevented digital intermediaries from regulating content on their platforms, and most large private companies have set up elaborate schemes for moderating commercial content, such as reporting tools, human moderators and automated systems. However, beyond such self-regulatory measures, private Internet corporations appear to be very reluctant to moderate misleading content and hate speech (Caplan, et al., 2018).

Companies such as Facebook, Google and Twitter argue that the information on their platforms is user-generated, and attempting to control it would make them the 'arbiters of truth' and infringe on the free-speech rights of their users. While there are merits to this argument, the reality is that almost all large private companies have developed elaborate content moderation and governance systems for deciding what kinds of speech are permissible, promoted and banned on their platforms. Such internal mechanisms have resulted in wide variations in how moderation is applied to different groups and content categories across platforms. There are numerous examples where platforms have been slow in their response to removing harmful content, as well as instances of over-broad censorship of legal and truthful information.

From the legal perspective, digital intermediaries rely on the host–editor dichotomy with the contention that since most of

the information on their platforms is user-generated content, their role is limited to providing the infrastructure for hosting content, unlike publishers or creators who take editorial decisions. Although content creators and users of platforms exercise some control over the information they engage with online, the social and technical architecture of digital platforms wields considerable influence in shaping access to information and user engagement online.

## OF THE PEOPLE, BY THE PEOPLE, FOR THE PLATFORM?

Digital platforms have expanded the ways in which Indian audiences discover and access information. The trend towards the greater availability of information obfuscates subtler transformations in the production, consumption and distribution of information online. Until recently, only special-effects experts could make realistic looking and sounding fake videos. But, today, cheap, sophisticated and user-friendly editing technologies allow people with little to no technical expertise to create deep fakes.

Tutorials, with step-by-step instructions for people who want to create such content, are widely available across multiple online platforms. Community-based forums, such as Quora, that have grown to become more influential than the websites of established media houses, have made it easier for non-expert influencers and opinion makers to share unverified, false or manipulated content with large audiences (Dasgupta and Sathe, 2018).

While diverse information is becoming more easily available, the distribution of information has become concentrated on a few platforms. The increasing reliance on social networking platforms—Facebook, Google and Twitter—for accessing information has led to the mass consolidation of audiences on these sites. For example, in 2018, Indian audiences spent nearly 47 billion hours on the country's top five video-streaming apps. WhatsApp, which counts India as its biggest market, with more than 200 million users, has reported that users spend more than two billion minutes each day on audio and video calls.

The concentration of information sources and audiences on opaque and largely profit-driven private platforms is particularly relevant to analysing the problem of information disorder. Platforms derive value from continued user engagement, and because of the economic incentive of retaining audience attention, they are less

likely to design or incorporate features that expose individuals to information that challenges their identity or world view. Arguably, by customising and tailoring information, based on users' histories and preferences, social networking and broadcasting platforms have exacerbated the polarised pluralism underlying the 'fake news' problem in India.

The business models and commercial interests pursued by online platforms also encourage the production of content that is 'click-worthy' or engaging, irrespective of accuracy or truth. Fake news generates clicks, shares and social engagements, metrics that platforms build on to quantify individuals' social and political behaviour or gauge their preferences for precision targeting. This dynamic has been exploited by individuals and organisations to benefit financially or politically from spreading false or misleading content.

Considering the influence of digital intermediaries from this perspective, the host–editor dichotomy does not adequately cover the responsibility or accountability of companies for contentious content spreading on their platforms. Over the years, there has been a growing realisation amongst legislators, policymakers and opinion leaders that existing intermediary laws and regulations might be inadequate to address the technological and social changes ushered in by digital platforms.

A slow reconsideration of 'safe harbour' or conditional immunity frameworks that shield digital intermediaries from liability for user-generated content is underway. European lawmakers have established the Code of Conduct[3] as a way to pressure social networking platforms to crack down on hate speech. In 2017, the German Parliament passed a law requiring social networking platforms to take action against the spread of hate speech, criminal or false material directed at minorities being shared on their platforms.[4] In India, MEITy has released the Intermediary Guidelines (Amendment) Rules 2018 (Draft Rules) for public comments.[5] The draft amendments, while not yet finalised, have reignited the debate on the rights and responsibilities of digital intermediaries for content on their platforms.

## THROUGH THE LOOKING GLASS

Under the existing liability regime in India, digital intermediaries either rely on external sources to report contentious content,

or voluntarily regulate and manage content according to their policies and 'community' standards. The draft rules seek to dramatically expand the obligations of intermediaries with regard to user-generated content on their platforms. As policymakers assign responsibility to platforms, it is vital to think about how the changes being proposed to the liability framework will impact citizens' rights and democracy.

Consider the use of the term 'unlawful' in the draft rules. The vague term does not provide sufficient clarity on the types of content that should be restricted. In the absence of standards and guidelines for unlawful content, the draft rules ignore the complexities of assessing the legality of content. For example, in the content removal cases involving hate speech, satire, parody or defamation, the determination of legality or veracity is highly context driven, and nuances, which are crucial for the review of such content, are hard to code into technology designs. Facebook CEO Mark Zuckerberg, too, has acknowledged the company's AI struggle with language dialects, the context, and whether or not a statement qualified as hate speech, and that while it may be able to root out hate speech in five to 10 years, 'today we are not there yet' (Alba, 2018).

Leaving the determination of unlawful content to intermediaries goes against the SC's observations in the Shreya Singhal case: that an intermediary ought not to be placed in the position to decide the legitimacy or legality of information. Given the difficulties of identifying and interpreting information disorder, a more appropriate regulatory intervention would be for MEITy to enable key stakeholders to deliberate and negotiate shared understandings of what constitutes a genuine threat to public safety, and what will enable public access to trustworthy information through an open and transparent process (Helberger, et al., 2018).

Another aspect of concern of the draft rules is that the legislative approach to content regulation demonstrates an increasing reliance on technology to take decisions about the legality of online content. For instance, the draft rules direct companies to deploy 'technology-based automated tools' for proactively identifying, removing and disabling public access to unlawful content. Platforms have deployed automatic content detection tools to de-rank or remove a wide range of illegal content, including child pornography

and copyright violations, or to take action against suspicious accounts posting such content (Gerken, 2019).

While automatic content detection systems have had some success in removing illegal content at scale, such tools and technologies are very expensive. For example, YouTube invested more than $60 million to develop Content ID, its proprietary system for copyright and content management. In India, few digital intermediaries can afford to evolve such technological responses for content moderation. By requiring all types of platforms to implement these filters in order to qualify for safe harbour, the draft rules put young start-ups, which cannot afford to invest in such technologies, at a tremendously competitive disadvantage. Over the long term, this strategy will result in diminished innovation and diversity in Indian media markets, resulting in less choice for citizens.

The call for expensive technology-based interventions stem from the belief that digital communications platforms are neutral mediums and will advance sophisticated, technical tools in accordance with policymakers' specifications, as long as they have an obligation under the law to do so. In reality, far from being neutral, predictive models and algorithms to identify and filter content are vulnerable to the biases of their creators and users. This is because while social media companies rely on data-driven signals to determine importance or relevance, false content continues to not only exist on these platforms, but also to trend (Caplan, et al., 2018). The producers, consumers and amplifiers of harmful content with economic or political motivations can also easily shift their tactics to avoid detection by automated content moderation systems.

Automatic content moderation technologies also make it difficult for users to understand and challenge content removal decisions. The lack of transparency or accountability in automated systems is particularly relevant, as there is no way to measure how much content platforms remove or deprioritise using such tools, or the impact of their use on limiting contentious content. For instance, Facebook claims to be able to remove 99 per cent of ISIS- and al-Qaida-affiliated content using AI-powered algorithms and human content moderators (Matthews and Pogadl, 2019). The company's claims have not been independently investigated, and with little to no public information about the workings of the moderation system,

there is no way of knowing whether AI or humans are the key to the strategy's success.

By giving too much weightage to the power of automatic content detection, the draft rules have overlooked the problems of over-broad censorship, the arbitrary removal of content, or the abuse of detection tools of platforms for extortion (Geigner, 2019). The growing evidence of the biases and discrimination in automated content moderation systems serve as an important reminder that proactive filtering technologies neither can, nor should, become an industry-wide standard.

In the absence of measures to strengthen transparency and accountability, the draft rules hand over the power to set and enforce the appropriate boundaries of public speech to profit-driven technology companies. Information disorder is a complex policy problem, and addressing it requires accounting for the role of both platforms and users in organising cross-cultural communication. The lack of accountability in centralised liability regimes and gaps in the self-regulatory frameworks suggest that new forms of governance, based on a more distributed approach to the allocation of responsibility, are required.

## TOWARDS DISTRIBUTED RESPONSIBILITY

The rapid evolution/expansion of information disorder, and the associated risks to society, require comprehensive and durable solutions to meaningfully address the problem. The complexities of information disorder necessitate allocating responsibility to human, technological and institutional actors that contribute to the creation, distribution and circulation of false and misleading information (Helberger, et al., 2018).

In our current media environment, the distribution of information has become concentrated on platforms such as Facebook, Google and Twitter. From deciding the information that is most accessible to audiences, including type or format of content, to controlling financial incentives for influencers and opinion makers on their platforms, the private ordering of online platforms has considerable influence in organising and shaping communications. Although digital intermediaries have evolved in their role as facilitators of digital communication, laws and regulations have not kept pace with their shifting role.

The conditional immunity regime under Section 79 has ensured that digital intermediaries can set standards and processes to govern speech allowed on their platforms without being required to have any editorial control or liability for that content. While technology companies have benefitted from an open, unconstrained regulatory environment, the safe harbour provision and the requirement of removing content upon the receipt of government or judicial order has also shielded technology companies from confronting targeted manipulation, such as information disorder enabled by the socio-technical design of their platforms and services. The lack of overview of the platforms' information moderation systems, the high costs of approaching courts, and the glacial speed at which the judicial system operates, render judicial and executive orders as an insufficient recourse for tackling harmful and illegal information online.

The scale of information disorder, and the dynamic context of abuse and misuse of digital platforms, suggest that the current liability regime may be untenable. A re-examination of the different types of intermediaries and their concomitant duties and obligations will bring a measure of accountability to online platforms and service providers. Designing interventions that seek to expand the legal liability of intermediaries for contentious content gets complicated, because even though platforms shape user engagement, they do not determine it (ibid.).

Various types of individual and institutional actors create, distribute and amplify false and misleading information for financial, political and social reasons. Policy interventions to tackle information disorder must put in place systems to identify different state and non-state actors, and hold them accountable for targeting and manipulating public opinion.

Reigning in information disorder will require going beyond deputising platforms to proactively remove information, based on their own standards and interpretation of unlawful content. This approach not only entrenches the dominance of a few platforms, but without oversight or transparency, technology companies will continue to take opaque content decisions without any potential consequences. The technological solutions outlined under the draft rules do not fix the broader problem: that the design of these socio-technical architectures is geared towards user engagement and rapid content consumption. As Zeynep Tufekci has emphasised, the ability

to control the attention of the people is much more powerful than outright censorship (2017).

Instead of leaving private platforms to continue scaling content moderation practices and technologies, policymakers need to intervene to establish responsibility for the business practices of digital intermediaries. Policymakers and regulators should guide technology companies to provide more information about their content moderation practices, including details on the organisation and functioning of internal or external content review bodies. Similarly, platforms should also be required to report on content removal appeals, and establish a right to be heard for users whose information has been removed, including by automated systems. Structural interventions, aimed at introducing transparency and due process, would go a long way in clarifying the role of various types of digital intermediaries with regard to the management of the wide variety of contentious content on their platforms.

In addition to these measures, there are several possible avenues for legislation to improve accountability in the enforcement of content moderation policies, and to encourage platforms to take on responsibility for the protection of consumer rights, public safety and security of communications. Interventions to determine enforceable standards and to guide technology companies to better serve the laws of our country should be developed through an open, multi-stakeholder process. In the absence of such collaborative and nuanced deliberation, there is a real danger that policymakers will wind up with over-broad or obsolete solutions. Regulators and policymakers must tread carefully in this complex policy area and pursue nuanced, incremental improvements over reactive, catch-all solutions. The release of draft rules provides an opportunity for a thorough discussion on the social responsibility of platforms. Whatever the future of the regulation of digital intermediaries, the consideration of responsibility of platforms will only be one component in a larger effort to combat information disorder.

◆

**NOTES**

1. Facebook. 2018 'Facing Facts' https://newsroom.fb.com/news/2018/05/inside-feed-facing-facts/.

2. *Shreya Singhal vs. Union of India*. 24 March 2015. Writ Petition (Criminal) No. 167 OF 2012. https://indiankanoon.org/doc/110813550/.
3. In 2016, the European Union launched the online 'Code of Conduct' to fight hate speech, racism and xenophobia across Europe. Facebook, Twitter, YouTube and Microsoft were involved in the creation of the Code and have signed up to it, although the terms are not legally binding. The Code establishes 'public commitments' for the companies, including the requirement to review the 'majority of valid notifications for removal of illegal hate speech' in less than 24 hours, and to make it easier for law enforcement to notify the firms directly. https://ec.europa.eu/info/sites/info/files/code_of_conduct_on_countering_illegal_hate_speech_online_en.pdf.
4. In 2017, the Bundestag, Germany's Parliament, passed a Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG) that allows authorities to fine social media companies which fail to remove hate speech posts, fake news and terrorist content that violate German law within 24 hours. In cases that are more ambiguous, Facebook and other sites have seven days to deal with the offending post. If they don't comply with the new legislation, the companies could face a fine of up to 50 million Euros ($57.1 million). https://www.dw.com/en/eu-hails-social-media-crackdown-on-hate-speech/a-47354465.
5. Ministry of Electronics and Information Technology (MEITy). 2018. 'The Information Technology [Intermediary Guidelines (Amendment) Rules] 2018.' http://meity.gov.in/content/comments-suggestions-invited-draft-"-information-technology-intermediary-guidelines.

## REFERENCES

Ahuja, R. 2013. 'Muzaffarnagar Riots: Fake Video Spreads Hate on Social Media', *The Hindustan Times*, 10 September. https://www.hindustantimes.com/india/muzaffarnagar-riots-fake-video-spreads-hate-on-social-media/story-WEOKBAcCOQcRb7X9Wb28qL.html.

Alba, D. 2018. 'Why Facebook will Never Fully Solve its Problems with AI', *Buzzfeed News*, 11 April. https://www.buzzfeednews.com/article/daveyalba/mark-zuckerberg-artificial-intelligence-facebook-content-pro.

Bansal, S. 2019. 'Narendra Modi App has a Fake News Problem', *Medium*, 27 January. https://blog.usejournal.com/narendra-modi-app-has-a-fake-news-problem-d60b514bb8f1.

Caplan, R., L. Hanson and J. Donovan. 2018. 'Dead Reckoning: Navigating Content Moderation after Fake News', *Data and Society*, 21 February. https://datasociety.net/output/dead-reckoning/.

Chakrabarti, S., L. Stengel and S. Solanki. 2018. 'Duty, Identity, Credibility: Fake News and the Ordinary Citizen in India', BBC. http://downloads.bbc.co.uk/mediacentre/duty-identity-credibility.pdf.

Dasgupta, P and G. Sathe. 2018. 'After Ruining Twitter, Indians are Turning Quora into a Troll-Fest', *Huffington Post*, 28 December. https://www.huffingtonpost.in/entry/twitter-indians-quora-politics_in_5c24c958e4b08aaf7a8e0eb1.

Doval, Pankaj. 2019. 'Remove "Fake" Content on Food Quality, Govt. tells Facebook', *Economic Times*, 21 January. http://timesofindia.indiatimes.com/articleshow/67617097.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst.

Gerken, T. 2019. 'YouTube's Copyright Claim System Abused by Extorters', BBC, 14 February. https://www.bbc.com/news/technology-47227937.

Geigner, T. 2019. 'YouTube's Content ID System being Repurposed by Blackmailers', *Celebrity Access*, 14 February. https://celebrityaccess.com/2019/02/14/youtubes-contentid-system-being-repurposed-by-blackmailers/.

Ginzburg, Carlo. 2012. *Threads and Traces: True False Fictive*. US: University of California Press.

Helberger, N., J. Pierson and T. Poell. 2018. 'Governing Online Platforms: From Contested to Cooperative Responsibility', *The Information Society*, 34 (1): 1–14. DOI: 10.1080/01972243.2017.1391913. https://www.tandfonline.com/doi/pdf/10.1080/01972243.2017.1391913.

Matthews, K. and N. Pogadl. 2019. 'Big Tech is Overselling AI as the Solution to Online Extremism', *The Conversation*, 17 September. http://theconversation.com/big-tech-is-overselling-ai-as-the-solution-to-online-extremism-102077.

Mozur, P 2018. 'A Genocide Incited on Facebook, With Posts from Myanmar's Military', *The New York Times*, 15 October. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

Pilkington, E. and J. Glenza. 2019. 'Facebook Under Pressure to Halt Rise of Anti-vaccination Groups', *The Guardian*, 12 February. https://www.theguardian.com/technology/2019/feb/12/facebook-anti-vaxxer-vaccination-groups-pressure-misinformation.

Singh, P 2018. 'Planet-scale Influence Operation Strikes at the Heart of Polarised Indian Polity', 26 November. https://pukhraj.me/2018/11/26/planet-scale-influence-operation-strikes-at-the-heart-of-polarised-indian-polity/.

Tufekci, Z. 2017. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. US: Yale University Press.

Vosoughi, S., D. Roy and S. Aral. 2018. 'The Spread of True and False News Online', *Science*. 9 March. http://science.sciencemag.org/content/359/6380/1146.

Wardle, C. and H. Derakhshan. 2017. 'Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making.' https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Wong, J. 2018. 'Don't Give Facebook and YouTube Credit for Shrinking Alex Jones' Audience', *The Guardian*, 5 September. https://www.theguardian.com/commentisfree/2018/sep/04/alex-jones-infowars-social-media-ban.

Wright, M., R. Mendick, C. Hope and G. Rayner. 2019. 'Facebook Paid Hundreds of Thousands to Host Anti-Brexit "Fake News"', *The Telegraph*, 18 January. https://www.telegraph.co.uk/news/2019/01/18/facebook-accused-pumping-fake-news-running-ads-claiming-endangered/.

◆◆