Routledge
Taylor & Francis Group

Information, Communication & Society

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rics20

# De-platforming disinformation: conspiracy theories and their control

H. Innes & M. Innes

**To cite this article:** H. Innes & M. Innes (2023) De-platforming disinformation: conspiracy theories and their control, Information, Communication & Society, 26:6, 1262-1280, DOI: 10.1080/1369118X.2021.1994631

**To link to this article:** https://doi.org/10.1080/1369118X.2021.1994631

Published online: 28 Oct 2021.

INFORMATION, COMMUNICATION & SOCIETY
2023, VOL. 26, NO. 6, 1262–1280
https://doi.org/10.1080/1369118X.2021.1994631

Routledge
Taylor & Francis Group

OPEN ACCESS

# De-platforming disinformation: conspiracy theories and their control

H. Innes and M. Innes

Crime and Security Research Institute, Cardiff University, Cardiff, UK

**ABSTRACT**

Informed by two case studies of de-platforming interventions performed by Facebook against two high profile conspiracy theorists who had been messaging about Covid-19, this article investigates how de-platforming functions as an instrument of social control, illuminating the intended and unintended effects it induces. To help interpret the patterns in the data, two novel conceptual innovations are introduced. The concept of '*minion accounts*' captures how following a de-platforming intervention, a series of secondary accounts are set up to continue the mission. Such accounts are part of a wider retinue of '*re-platforming' behaviours*. Overall, the empirical evidence reviewed suggests that whilst de-platforming can constrain transmission of conspiratorial disinformation, it does not eradicate it.

**ARTICLE HISTORY**
Received 21 May 2021
Accepted 8 October 2021

**KEYWORDS**
De-platforming; disinformation; conspiracy theory; Facebook; social control; Covid-19

According to Facebook's own data, in the last quarter of 2020, action was taken against 1.3 billion fake accounts on the platform (Facebook, 2021) and more than 100 networks removed for engaging in co-ordinated inauthentic behaviour designed to manipulate public opinion (Rosen, 2021). Facebook estimates that of approximately 2.8 billion monthly active users in the last quarter of 2020 (Tankovska, 2021), 5% were fake accounts. Behind these headline statistics, the justifications and rationale for Facebook's actions shifted significantly during 2020, as they reacted to a toxic mix of coronavirus conspiracies interacting with political disinformation gravitating around the US Presidential election. This blend induced an extension to the more frequent removal of authentic accounts for repeated violation of community guidelines.

Challenged by the global health pandemic and the potential for prevalent misinformation about the causes and consequences of coronavirus (Donovan, 2020; Molter & DiResta, 2020), interacting with a US Presidential election campaign where disinformation was seemingly 'normalized' and delivered on an industrial scale (Election Integrity Partnership, 2021), Facebook altered both its logics and practices. Multiple de-platforming measures were applied to accounts assessed as posing risks to public safety in 2020, as defined by Facebook's Dangerous Individuals and Organizations Policy, including most infamously President Trump (Facebook, 2020). In the year to September 2020,

**CONTACT** H. Innes innesh@cardiff.ac.uk

Facebook took down more than 1 million groups for repeat violations (Alison, 2020), implementing new countermeasures intended to prevent the administrators of those groups from creating new ones. Where previously such interventions were implemented only following the detection of 'co-ordinated inauthentic behaviour', in the latter half of 2020 they were also introduced to control 'harmful content'. Similar strategic shifts were also made by Twitter and YouTube, amongst others.

De-platforming is Facebook's ultimate sanction. Their 'Community Standards' state accounts harmful to the community will be removed, including those that compromise the security of other accounts and Facebook services (Facebook, 2021, p. 17). Prior to any such action, repeated warnings and restrictions will be given for violations that pose severe safety risks, and non-compliance will lead to an account being disabled. De-platforming an account is thus positioned as the endpoint in an escalatory enforcement dynamic (Facebook, 2021a, p. 17):

> Because account-level removal is a harsh severe action, whenever possible, we aim to give our community a chance to learn our rules and follow our Community Standards. Penalties, including account disables, are designed to be proportionate to the severity of the violation and the risk of harm posed to the community.

Responding specifically to the Covid-19 crisis, Facebook prefaced their Community Standards with the following update:

> we're working to remove content that has the potential to contribute to real-world harm, including through our policies prohibiting the coordination of harm, the sale of medical masks and related goods, hate speech, bullying and harassment, and misinformation that contributes to the risk of imminent violence or physical harm. (Facebook, 2021a)

As this statement clarifies, there are some harms Facebook judge sufficiently serious that they warrant intervention and active regulation. Disinformation is defined as only one of these, but an especially interesting and challenging one. Since the discovery of the St Petersburg based Internet Research Agency's influence campaign to interfere in the 2016 US Presidential election, followed by similar subsequent revelations, how to control disinforming and misinforming communication is a challenge that has captured significant political and public attention (Benkler et al., 2018).

Somewhat surprisingly however, this policy interest has not been matched by the focus of academic work. There has been significant growth in studies explaining and documenting the causes and consequences of mis/disinformation. Inflected by methods and conceptual precepts with their roots in established intellectual traditions for analysing rumours (Shibutani, 1966), propaganda (Jowett & O'Donnell, 2012) and conspiracies (Hofstader, 1964; Albarracín, 2020), the more contemporary take on these issues tends to be organized around three main positions. The 'political economy' perspective attends primarily to how a range of social, political, economic and cultural forces have coalesced into a highly polluted media ecosystem (Benkler et al., 2018; Lance Bennett & Livingston, 2020). Studies focusing upon the 'pragmatics' of constructing and communicating mis/disinformation have illuminated the tactics and techniques used in authoring and amplifying misleading information (Innes, 2020;; Krafft & Donovan, 2020; Woolley & Howard, 2018). Finally, there is a more 'epistemic' strand of work concerned with the implications of these trajectories for the social ordering of reality, and how we are entering a 'post-truth' or 'post-factual' moment (Kakutani, 2018; Pomerantsev, 2019).

Collectively, such studies have done much to map how and why disinformation arises, but rather neglected how its impacts and influence can be managed and mitigated. A recent systematic review of social media countermeasures highlighted the lack of a robust evidence base about what works to control disinformation, and the relative efficacy of different 'supply-side' interventions (Courchesne et al., in press).[1] Far more research exists on 'consumer-facing' interventions like fact-checking and 'de-bunking' (Walter et al., 2020), whereas the review identified no empirical studies of 'de-platforming'. There are a small number of independent, published, empirical studies of account take-downs in respect of terrorism (e.g., Conway et al., 2019), but no equivalents for (dis)information campaigns.

It is with this gap in our knowledge that the current article engages. Informed by two empirical case studies of de-platforming performed by Facebook against two high profile conspiracy theorists promoting public health disinformation, the analysis investigates how de-platforming functions as an instrument of social control, illuminating the intended and unintended effects it induces. Framed in this way, the article has three principal aims:

- Investigating how de-platforming is organized and implemented on Facebook.
- Assessing the intended and unintended impacts of de-platforming in practice.
- Exploring the extent to which the criminological literature on social control provides conceptual insights into the workings of de-platforming strategies.

The next section elaborates some of the themes rehearsed above, about how Facebook constructs de-platforming as a mode of social control to help 'police' behaviour on their platform. The comparative case study design is then laid out, including a description of how data were collected, analysed and reported. This sets up a brief overview of the prevalence and distribution of Covid-19 related conspiracies, before the two empirical case studies are introduced. It is worth clarifying here that the analytic focus is not upon the substantive content of the conspiratorial ideas themselves, but how they are configured as suitable targets for de-platforming. The concluding section draws together and interprets the implications in terms of the intended and unintended impacts of de-platforming, and what this means for understanding it as a new modality of social control. The concept of 're-platforming' is introduced to describe behaviours performed by those targeted by the controls to subvert and circumvent their regulatory effects. The empirical evidence demonstrates how re-platforming responses enable the targeted actors to persist in their activities 'on-platform', albeit in a moderated form, whilst simultaneously driving a diversified 'cross-platform' presence, rendering surveillance more challenging. Thus, 'de-platforming' may be less impactful and less effective in controlling disinformation than previously supposed.

## The social control of disinformation

'De-platforming' is a term-of-art social media platforms use to describe a publicly visible and increasingly deployed countermeasure designed to control disinformation. It can be defined as where an account assessed to have engaged in problematic behaviour, in terms of authoring or amplifying malign or harmful content, is removed by a social media platform operator. Typically, imposition of this sanction is justified on the basis of a breach in compliance of the platform's Terms of Service or Community Standards. In this context, it is viewed by some as an antidote to harmful material emanating from particular online communities, but by others as an unacceptable, unilateral imposition of power by unregulated 'big tech' (Moynihan, 2021).

Countering and controlling disinformation is only one of several 'harms' that can trigger a de-platforming response. Others include online sexual abuse, hate crimes and extremist radicalization, where the risks and threats are often relatively clearly demarcated. Brian Fishman (2019) who leads on countermeasures against terrorist and hate organizations for Facebook, argues social media platforms require a range of deterrents to prevent misuse because of the wide array of functions provided to users (both benign and malign). Compared with some of these other problematic behaviours however, controlling disinformation and particularly conspiracy theories is especially challenging because the nature of the harms is often less clear-cut or contested – see for example Facebook v. CasaPound (Scuibba & Pasuetto, 2020). There are justified concerns about impinging freedom of speech rights central to the ethical precepts of liberal democratic politics, as well as the internet itself (Douek, 2021).

The limited number of studies conducted on the effects of de-platforming targeting different problematic behaviours have catalogued a range of intended and unintended impacts. Conway et al.'s (2019) assessment of Twitter takedowns of content posted by the terrorist group Daesh (aka Islamic State) is that it had a discernible effect on their presence on the platform. However, the researchers' reliance upon machine learning to detect and categorize accounts belonging to the group may have missed accounts where the signalling of belonging is more subtly encoded. Other empirical studies have suggested de-platforming may push 'extreme internet celebrities' and their supporters onto other platforms. Thus, triggering an influx of new followers to alternative, often encrypted, sites where posts are less moderated, and problematic behaviour can intensify (Blackburn et al., 2021). Notably, when the hard right 'influencers' Laura Loomer and Paul Joseph Watson were de-platformed from Facebook in 2019, the platform was criticized for pre-announcing the ban hours beforehand, allowing them to signpost alternative platform presences (Martineau, 2019). However, others like Milo Yiannopoulos, banned from Twitter and Facebook, claim it had a significant adverse effect on their income and audience reach.[2]

The so called 'Streisand effect' is where censorship backfires, hardening the ideological convictions of followers. Rogers (2020), for example, traced an alternative network of platforms used as replacements for YouTube, Facebook and Twitter, with personal websites and subscription services rapidly revived post de-platforming. Overall, however, on Telegram these extreme voices found their audiences reduced, their language became milder, and fewer hyperlinks were shared to Facebook and Instagram (compared to Twitter, YouTube and personal websites). However, especially pertinent to the current analysis is how, over the course of 2020, previously marginal platforms such as Telegram, have experienced rapid growth in user numbers (as evidenced by download statistics), in part because of user migration resulting from the increasing policing by Facebook and Twitter.

Focusing upon 53 distinct foreign state (dis)information operations conducted between 2013 and 2018, Martin et al. (2019) discuss how, whilst the targeted state may impose some retaliatory measures (for example diplomatic expulsions), the primary practical response is frequently delivered by private social media corporations. Where states do intervene more directly to censor misleading or undesirable content, similar patterns of multiple circumventions and 'work arounds' have been reported (Roberts, 2020). Significantly, 'de-platforming' is one of several countermeasures deployed to try and control conspiratorial disinformation and other online harms. Others include content suppression via algorithm adaptations; content labelling through inclusion of warnings; de-monetization by limiting advertising revenue; citizen information literacy campaigns and third-party fact-checking (Mena, 2019; Pennycook et al., 2020). However, the heterogenous nature of conspiratorial disinformation renders measuring intervention effects difficult (Schiffrin, 2020).

Framed in this way, de-platforming can be understood as a novel mode of social control. Cohen (1985) defined social control as comprising organized and programmed responses to conduct and behaviour perceived as problematic or troublesome in some manner. Applied to the regulation and governance of criminal behaviour, a conceptual distinction is routinely drawn between 'formal' and 'informal' social control. The former derives from law and the auspices of the state, the latter interventions conducted outside of such resources (Black, 1976). Consistent with several of the points made previously about de-platforming, analyses of contemporary social control have highlighted an increasing integration of formal and informal controls. For example, Garland's (2001) notion of 'responsibilization' delineates a process where state authorities increasingly require private companies and other organizations to assume front-line responsibility for the delivery of controls of deviant behaviour.

Cohen (1985) described how logics and practices of social control are consistently subject to exogenous and endogenous pressures of 'net widening' and 'net deepening'. The former references an expansion of scope, in terms of the range of problems the social control apparatus is used to engage with. The latter is the tendency to increase the intensity of sanctions applied to those problems. These perspectives perfectly describe the role de-platforming has come to play in the policing of a range of issues on social media platforms, from deliberately engineered hostile state information operations, through to more 'organic' episodes of misinformation. Thus, de-platforming can be understood as one of a number of new modalities of social control engaged in the regulation and governance of digital life.

In addition to tracking and tracing these master patterns in the organization of social control, concepts that attend to the outcomes specific episodes of control are designed to deliver are especially helpful in mapping the potential consequences of de-platforming interventions. 'Deterrence' has a prominent position in discussions of social control, as it holds out the promise of prevention and reduction. Criminological formulations of deterrence, contrasted with those maintained in the discipline of International Relations, differentiate between specific and general deterrence (Kennedy, 2009). Specific deterrence aims at dissuading perpetrators from repeating their own crimes in the future, whereas general deterrence is more concerned with discouraging others from engaging in similar behaviours. Where deterrence incorporates a preventative orientation, the concept of 'displacement' is more concerned with harm reduction logics, attempting to shift the targeted anti-social behaviour into more manageable forms. For example, displacements targeting the methods malign actors use can require them to have to develop new ones; temporal displacements can shift when opportunities for harm arise, and spatial displacement the location where they take place. Advocates of displacement as a tactical option suggest that a temporary effect during an especially sensitive time is an important victory, even if it is only momentary.

'Disruption' is a newer addition to the social control toolkit, focusing upon creating impediments to the continuation of harmful behaviour. It has become an increasingly important tactical option to counter organized crime and terrorism (Innes et al., 2017), and involves increasing a bad actors' costs or level of effort, to reduce the frequency, intensity or scale of their activity. Disruption can be a cost-efficient and timely intervention when the scale of problematic behaviour outpaces control resources – a challenge obviously relevant to countering online disinformation. By taking steps to disrupt an activity, a social media company (or police) can address a problem without launching a full spectrum response.

## Materials and methods

Data for the two empirical case studies presented in this paper were accessed through CrowdTangle, a public insights tool owned and operated by Facebook. The tool tracks interactions on public content from Facebook pages, groups and verified profiles, and is consistent with our focus on the mainstream reach of conspiratorial material to audiences. It does not include activity on private accounts, advertisements or posts made visible only to specific groups of followers.

Search queries in CrowdTangle used the full names of the conspiracy thought leaders to identify public post mentions in languages using the Latin alphabet. Note that, because Facebook have removed the accounts of interest, none of the post mentions come from these sources. Rather, the data represent the 'digital footprint' the public profiles had, and continue to have, before and after the accounts were de-platformed (with the caveat the data will not include other content and accounts taken down by Facebook, or users, between posting and data retrieval in December 2020).

Given the absence of large datasets on de-platformed accounts, two case studies centred on charismatic, high profile conspiracy influencers were selected for in-depth analysis. Both constructed coronavirus narratives during 2020 and saw resulting growth in their audiences. The cases allow us to compare and contrast the effects of de-platforming for: (1) a well-established conspiratorial actor, David Icke; and (2) a relative newcomer, Kate Shemirani, whose profile grew in prominence during the pandemic.

Thematic analysis of mentions for Icke are based on 11,877 public posts in May 2020, for the four weeks following the de-platforming of his account, yielding some 2.2 million user interactions. This was supplemented by manual creation of a list of Icke affiliate pages and groups on Facebook as at 10/12/20, their creation dates, language and audience size ($N = 104$) which supported a CrowdTangle analysis of the activity of the largest and most active ones. For Shemirani, analysis was for 1636 post mentions throughout 2020, tracking her evolution and growing momentum on the platform before and after her account was de-platformed on 4 September 2020. For both profiles, post mentions were analyzed over time and by user engagement and audience size metrics. Data on the type of links shared in posts were also thematically coded to distinguish their source. Consistent with other research of this nature, the content of posts is summarized rather than reproduced as screenshots, and account names anonymized.

## Results: COVID conspiracies

Onset of the global pandemic early in 2020 fuelled numerous online conspiracy theories, disinformation and misleading speculation about the causes and consequences of the novel virus. A rapid review conducted by the authors identified that by May, at least 53 distinct coronavirus centred conspiracy theories could be identified. Some reheating and reworking established tropes and motifs from long-established conspiracies, where others were new and more bespoke in their fit to coronavirus themes. A brief selection of some of the themes are summarized in Table 1.

Many of these conspiracies, either explicitly or implicitly, subvert and contest public health directives and guidance, and found significant connections with new, larger audiences via social media (Donovan, 2020). Blending with worries about the social and economic impacts of repeated 'lockdowns' and social distancing regulations (which have themselves increased the amounts of time people are spending online), adherents to these conspiracies frequently engaged in off-line public protest events.

From May 2020 onwards, London along with many other European cities, was the site for anti-lockdown protests against public health rules on mask-wearing, social distancing and government control over everyday life. These protests were supported by the creation of new online pages, events and groups, where like-minded individuals formed communities and reinforced their worldview by sharing from a growing pool of disinformation. Pivotal to the presentation and propagation of this content, advocating conspiracies and real-world action, were a series of charismatic influencers or 'thought leaders'. These included long-established 'conspiracy celebrities' such as Icke. Significantly, the pandemic also established the reputations of several new conspiracy thought leaders, such as Kate Shemirani, who became influential not least because they presented their medical qualifications and experience as trustworthy credentials.

Coronavirus mis/disinformation intermingled and interacted with the equally polluted political campaigns associated with the 2020 US Presidential election. Consequently, on 19 August 2020, Facebook announced takedowns of any groups identifying with militarized social movements (defined as 'militias or groups that support and organize violent acts amid protests') and any associated with 'violence-inducing conspiracy networks' that included QAnon (Facebook, 2021). The following month, content was downgraded by Facebook via algorithmic adaptations for groups whose content had been restricted but not removed. Enforcement was further strengthened on 6th October

**Table 1.** Examples of health-related conspiracies on Facebook (Mar-May 2020).

| |
|---|
| Magna Carta guarantees the rights of individuals to protest under common law |
| Covid is caused by 5G |
| Covid tests have an 80% False Positive rate |
| Vaccines contain human DNA |
| Mask wearing is harmful |
| The Jewish Rothschild banking dynasty are responsible for coronavirus |
| Agenda 21 is a sinister, satanic plan by governments worldwide to use a fake pandemic to restart a new world order. |

2020 with the de-platforming of any group representing these movements, with or without threats of violence in their posts. Facebook justified this on the grounds that QAnon conspiracies, for example, were instigating real-world harm in other ways.

According to Facebook's figures, by January 2021, these interventions had identified 890 militarized movements that had been operating on the platform, removing approximately 3400 Pages, 19,500 groups, 120 events, 25,300 Facebook profiles and 7500 Instagram accounts (Facebook, 2020). Figure 1 shows the largest number of removals for militarized social movements were Facebook profiles, followed by groups. QAnon affiliated pages and groups were more likely than militarized social movements to create Facebook events associated with offline activity and were more evident on Instagram.

That accounts or groups subject to de-platforming might be re-inventing themselves, and/or re-appearing on the same or different platforms, is critical for the conceptual interests of this article. In social control terms, it focuses the analysis upon what the envisaged purpose and practical function of de-platforming is intended to be. Is it designed to disrupt an actor's 'on-platform' activities, or induce deterrence? Alternatively, are such countermeasures intended to effect displacement of troubling activity towards less-used platforms? Such considerations shift our attention to the relationship between intended and unintended outcomes, and these are investigated through the two case studies.

## *Case study 1: de-platforming David Icke*

David Icke is a prolific author and global speaker, with a significant online presence built by espousing 'new age' conspiracy theories and prophecies since the 1990s. Icke identified public figures – among them the British Royal family and the Clintons – as belonging to a controlling, hybrid elite, linked to paedophilia and satanism, and leading humanity to a global fascist state or 'New World Order'. Arguably, this discourse shaped

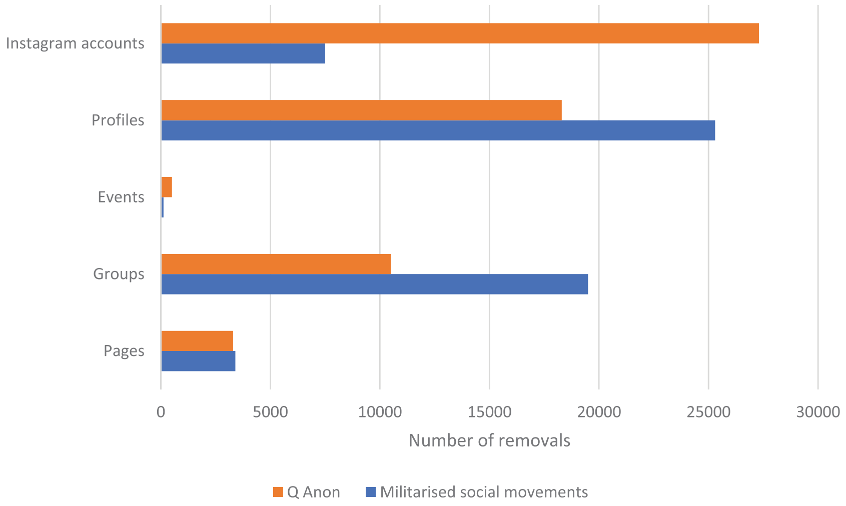

**Figure 1.** Volume and type of content de-platformed under Facebook's Dangerous Individuals and Organizations Policy, 19/08/20 to 19/01/21.

the contemporary conspiracy theories associated with the QAnon movement (Lawrence & Davis, 2020)).

The Icke 'brand' of conspiracism has developed a network of monetized websites, with regular video content streamed on multiple platforms and links shared across many more. In early 2020, Icke rapidly assimilated the pandemic, and global responses to it, into the belief system he and his followers already shared. Claims of a New World Order were already established but were supercharged by public fear and uncertainty about a new virus. His messaging during 2020 incorporated and integrated multiple conspiracy narratives about vaccinations, microchips, mind control and 5G. His digital persona is of particular significance then for a study of the causes and consequences of de-platforming.

Following repeated violations of Facebook policies on harmful disinformation (BBC, 2020), Icke's official page with 800,000 followers was removed from Facebook at the end of April 2020 for publishing 'health misinformation that could cause physical harm'. Around this time, Icke was among those calling for offline protests about COVID restrictions in the UK.

Icke reacted to Facebook's de-platforming decision on Twitter, where he maintained a verified account with 324 K followers until November 2020, when it too was removed for violation of rules on COVID misinformation (Spring, 2020). A tweet at this time signposted followers to alternative digital presences: his main website; and an intermediary Facebook page that continues to stream his content. The latter evidences how removing Icke's verified page, whilst curtailing his direct connection to his followers, did not prevent him reaching large audiences on Facebook's platform via accounts performing this function on his behalf.

Seven months after Icke's de-platforming, there were 64 active Facebook pages and 40 active Facebook groups using his name. Thirty-three of these were using Icke's image as their avatar. Confirming the international resonance of the Icke 'brand', one-third of pages were in non-English languages. In the seven days following his account removal on 30 April, his public mentions on Facebook increased by 84%, from 2833 to 5220. In the short-term this suggests one effect of de-platforming was to boost his profile on the platform, both in terms of a reaction from supporters and media articles shared about it – the aforementioned 'Streisand effect'.

Examining creation dates for active Icke-affiliated profiles on Facebook (see Table 2) shows some are long-lasting, with 31 profiles created more than five years ago, albeit not necessarily active at the current time. However, there are also signals of new page and group creation on Facebook around the time Icke's verified page was removed: of 18 public Facebook pages created in 2020, ten were created in May; the same month saw four new private groups in Icke's name. This suggests this de-platforming episode was

**Table 2.** Icke-affiliated Facebook profile types by dates.

| Group creation dates | N public pages | N public groups | N private groups |
|---|---|---|---|
| 2020 | 18 | 2 | 9 |
| Last five years (2015-2019) | 42 | 7 | 9 |
| >5 years (pre-2015) | 24 | 5 | 8 |
| Total | 64 | 14 | 26 |

associated with a compensatory 'blowback' reaction, whereby multiple new profiles representing the banned persona were created.

Private groups have the largest audiences on Facebook out of all Icke-profiles, with an average of 6358 members and an estimated audience of 165,000, far greater than for public Icke groups (Table 3). That private group creation was far more common during 2020 implies a user-reaction to greater public and platform surveillance of their activity on Facebook.

These data suggest that, in terms of its social control outcomes, there is little empirical evidence that Facebook's de-platforming had a deterrence effect. It may have had a modest disruption impact, although this is over-shadowed by the user counter-reaction which proliferated the number of new groups and pages, both public and private. Critical in explaining these patterns is the emergence of what we label 'minion accounts'. These are clearly associated with the de-platformed 'leader' and continue to perform their ideological mission, albeit not under their personal direction and control.

This interpretation is supported by examining the largest Icke-affiliated public Facebook page (Figure 2). It shows a peak in the volume of posting (n=71) and user interactions (n=46,294) around the time Icke was de-platformed, with signs of a 'second wave' in the volume of posts coinciding with Icke's offline participation as a speaker at the lockdown protests later in the year. That said, for this particular page, there has been a notable decline in post interaction, suggesting the de-platforming had some impact over time.

Facebook groups affiliated with Icke have remained active and continued to grow following the intervention. The most active public group in 2021 posts, on average, 59 times per day to an audience of just over 7000; an increase of 47% in its membership since Icke was de-platformed. Figures 3 and 4 track audience numbers for two of the most currently active Icke-affiliated public Facebook groups over time. One is an established group of five years (labelled 'Group 1 est' in both figures), the other newly created in May 2020 (labelled 'Group 2 new' in both figures). The audience for both increased between June 2020 and February 2021; by 62% for the established group and 5329% for the group created once Icke was de-platformed.

Of 18 Pages created using Icke's persona in 2020, ten were signposting users to Icke content on other platforms. This included an established Icke video streaming channel with more than 72000 subscribers, and a new streaming website created in August 2020 hosting banned content. Most directional activity was towards a main webpage, from where it is straightforward to repost a URL back on to Facebook. At the time of writing, 2965 public group posts – including 31 groups using the Icke name – had posted the webpage link, yielding 8494 interactions in total. The most popular recent content from the Icke website was a June 2020 article (thus after the de-platforming) authored

**Table 3.** Icke-affiliated Facebook profile types by audience size.

| Audiences | Range | Total audience | Average |
|---|---|---|---|
| Public Pages | 24–49,900 | 135,662 | 2120 |
| Public Groups | 1–6800 | 22,700 | 1621 |
| Private Groups | 31–29300 | 165,306 | 6358 |

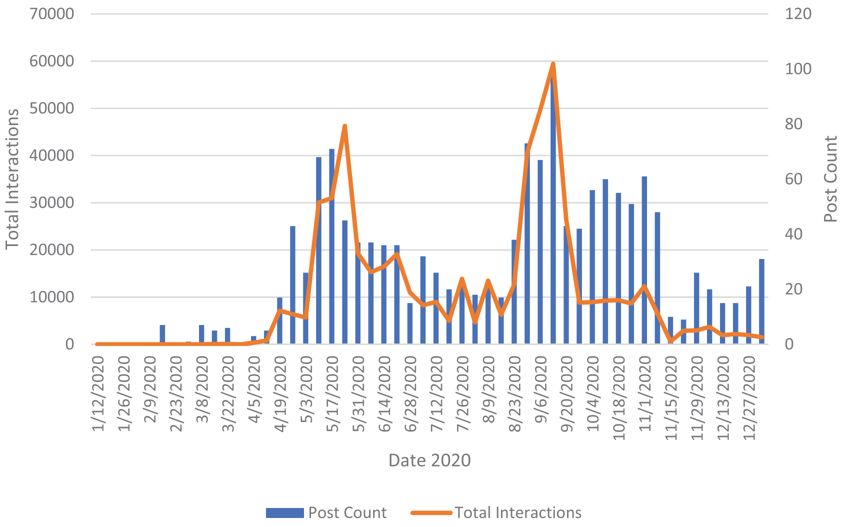

**Figure 2.** Posting and interactions over time for largest Icke-affiliated page.

by Icke about vaccine-induced infertility. This was shared two thousand times on Facebook and shows that sharing outside links remains a key method by which such content continues to circulate on the platform. In fact, on 7 May 2020, shortly after his removal from Facebook, Icke used his website to appeal directly to his audience to share his videos and stories with five people each day, who should then share with five others, and so on. This he identified operates 'outside the control of Silicon Valley psychopaths', permitting a reach of tens of millions (Icke, 2020). This was co-opted by Icke pages on Facebook where the same message became '#ShareIcke10Times challenge', generating 35 K interactions on Facebook pages and groups during May 2020. An example is reproduced below, which also signposted to video streaming platform BitChute and WhatsApp:

> Join the #ShareIcke10Times challenge. Once you have shared in 10 places which could be posted on your page in a group, a friend on WhatsApp or messenger. Share it via email

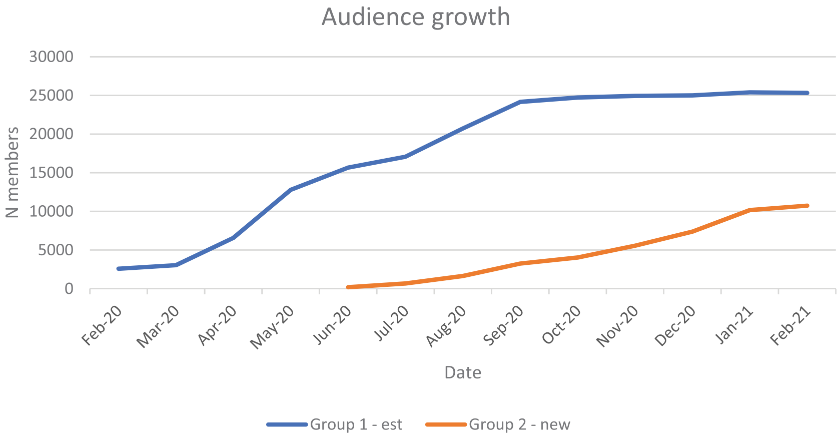

**Figure 3.** Facebook audience growth for most active Icke-affiliated groups.

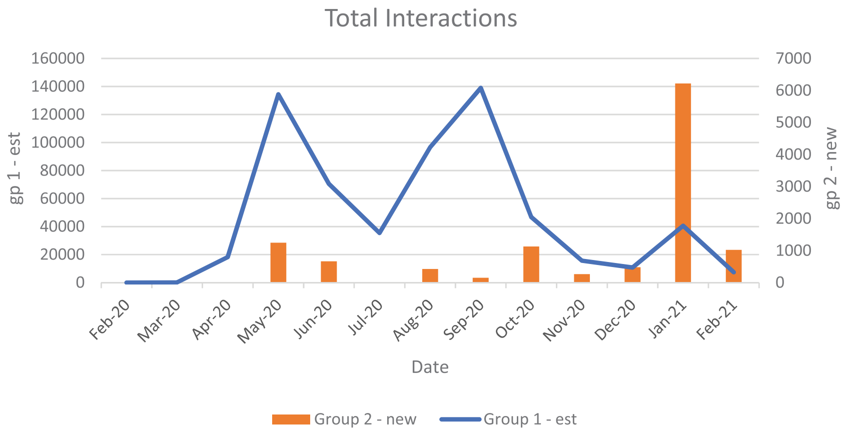

**Figure 4.** Facebook user interactions for most active Icke-affiliated groups.

> text however we can. When you shared it in 10 places comment done and that will also help the post go better. The more you like and comment the further it goes so go crazy 🤩🤩🤩 Let's get the info out loud and clear [link to BitChute]

Analyzing 221 posts from Icke-affiliated pages and groups in the month following his de-platforming shows the vast majority (94%) included links to other sources and platforms. Thus, one unintended outcome of the de-platforming intervention may have been to increase the digital resilience of the targeted conspiratorial thought community; displacing the activity onto other platforms, driving content diversification and the capacity to offset any attempts at disruption. The following posts made shortly after Icke was de-platformed, show supporters immediately using Facebook for displacement activity, advertising links to alternative platforms:

> Join me on our new video platform on LBRY … it is censorship free, and you get to make crypto and transfer it into your local currency. Screw Youtube! David Icke Alex Jones, many of the truthers are here. I will be moving all of my 200 videos onto this platform beginning today. Time to move on from FB CIA operative page.

> David Icke was banned on Fascistbook last week, and his last LIVESTREAM interview with XXXX was removed by #NaziTube after it reached 9,000,000+ views in a few days. This was LIVESTREAMED at 9am PST this morning [links to BitChute].

Equally important is how, for conspiratorial communities, de-platforming can be constructed as a 'badge of honour' because their insights worried mainstream 'big tech' sufficiently to act against them. This reframing of de-platforming as censorship and 'proof' is neatly encapsulated in the following post by an Icke supporter:

> … they have de-platformed anyone who has a large audience who goes against their mainstream narrative … . Luckily for us XXXX is a brave man with a strong following who have backed him financially, so he has created his own streaming platform, where he can interview guests uncensored … .

Whilst de-platforming may constrain the mass reach of conspiratorial ideas, it can simultaneously have the effect of reinforcing the bonds and sense of belonging between already 'devoted' believers.

## Case study 2: de-platforming the natural nurse

The second case study focuses on someone who, like Icke, is an influential conspiracy theorist, but differed in their trajectory into the public eye. Kate Shemirani's Facebook profile benefited directly from the unfolding COVID-19 pandemic, bolstered by her background as a medical nurse. Prior to the pandemic, she was not an established conspiracy thought leader.

Her profile had 54,000 followers when removed on 4 September 2020 by Facebook for repeated violation of policies against harmful misinformation. This followed an earlier 10-day suspension in April 2020. Shemirani – who presents online as a 'Natural Nurse in A Toxic World' stresses her medical qualifications and is linked to multiple anti-vaccine and antisemitic narratives and symbols that have been widely shared on social media (Harpin, 2020). For example, one Facebook post linking to a video streaming site showed a thumbnail image of a Nazi symbol in the centre of the British flag under the header 'COVID tyranny'. Shemirani has also expressed support for QAnon theories of satanic cults and global elites abusing children (Lawrence, 2020). In August and September 2020, she was a key figure galvanizing off-line protest, compering a large London protest not long after sharing a stage with Icke amongst others at another event. First suspended, and now removed from the Nursing and Midwifery Council register for spreading COVID-19 misinformation, Shemirani gains continued credibility by association: in the last quarter of 2020, her name was listed alongside 58 other medical professionals worldwide winning public influence by 'speaking out' against the official narrative.

Unlike Icke, Shemirani's Facebook profile did not have an established network of support pages and groups, nor a website. Analysis of public post mentions of her name shows prior to March 2020, there were only a handful of mentions on Facebook (see Figure 5). This changed from April onwards, when 81 Facebook posts shared YouTube links to her content (Figure 6), including '#5G will be Genocide' and 'UK Nurse Explains the Corona Virus PCR Test Fraud' (both sources now removed by YouTube). Amplifying this content were anti-5G Facebook groups (including a large regional UK network of groups), alongside other conspiracy and 'truther' groups.

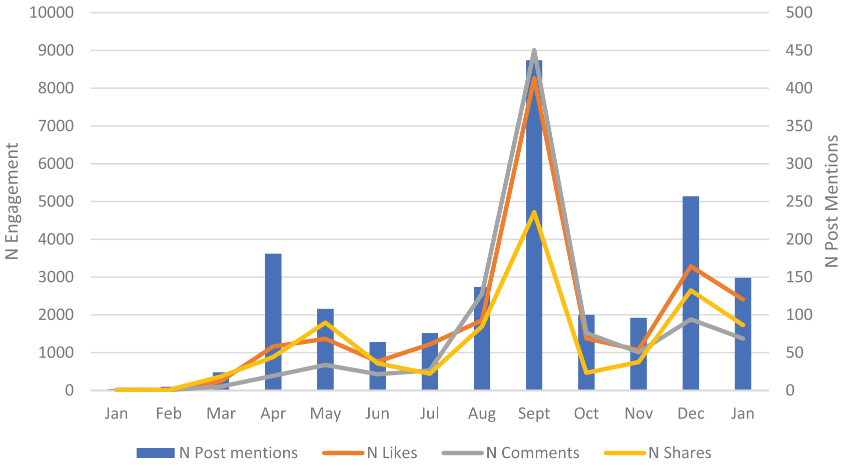

**Figure 5.** Facebook post mentions and engagement for Shemirani.

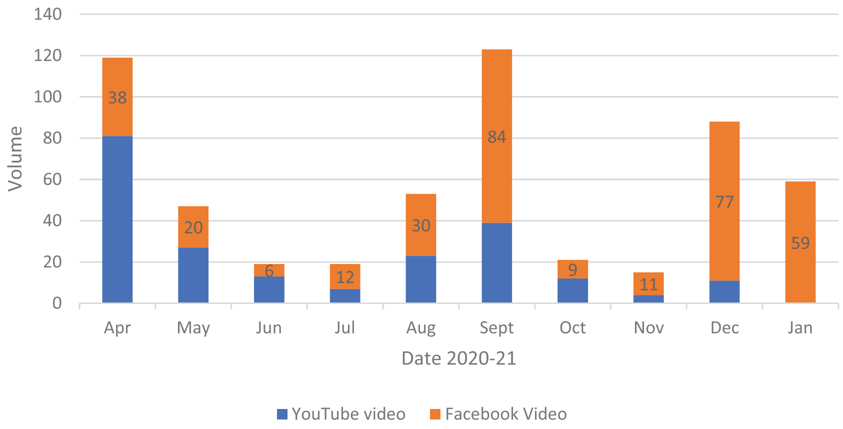

**Figure 6.** Source and volume of video links shared with post mentions of Shemirani.

Growth in Shemirani's Facebook presence coincided with her offline activity. Post mentions and engagement metrics of shares, comments and likes, all sharply peaked when she featured prominently in the London protests and was arrested by police on 5 September 2020. Facebook had only acted to de-platform her account the day before, suggesting their decision was influenced by her rapidly escalating online and offline visibility. It is difficult to disentangle the impact of her arrest and the removal of her account on her boosted profile during September 2020, but the most popular content immediately after, was a video of her arrest posted by an Icke-affiliated page on Facebook that received 28,000 views.

In the two months following the take-down, post mentions and user engagement on Facebook decreased markedly, with a sizeable reduction in the number of links being shared on the platform in her name, both from other Facebook pages and YouTube (Figure 6). This suggests that de-platforming was effectively disrupting her connection with audiences on the platform. However, the suppression effect appeared to be temporary, with signs of revival from the end of 2020, where the number of Facebook video shares increased from approximately ten in October and November 2020 to over 60 in the next two months.

A growing number of video links were sited on the Facebook platform indicative of an 'on-platform' displacement effect. Whilst this is likely to reflect the removal of content by YouTube, it also illustrates the continuing ease with which video links are shared by supportive 'minion accounts' linked to a banned account. These amplify and promote Shemirani's narratives independent of her presence on the platform. The main minion account can be readily identified as an associate of a US disinformation media platform (where Shemirani is their stated health and wellness expert), but three Icke-affiliated pages also perform this function.

Shemirani has further sustained her online profile via a series of tactical collaborations with others, through podcasts and interviews posted to an array of alternative streaming platforms. One, for example, hosts 28 of her videos posted by nine accounts since August 2020 and links from any can be posted onto Facebook. The lack of any discernible deterrence effect is evidenced by that fact that Shemirani was briefly able to return to Facebook as part of an anonymous new account in November 2020, with a simultaneous presence on Twitter. These accounts publicized harmful content and offline action, including a claim that the UK government were accountable for genocide from vaccine harm, shared across 48 public Facebook spaces. At the time of writing, this video content was available on BitChute, YouTube, and an Instagram account associated with her profile remained active.

These two case studies illuminate some of the complex impacts induced by de-platforming. Whilst there was some disruptive impact, it was not sustained. Likewise, some displacement did occur, but much of this was to minion accounts that 're-platformed' the content on Facebook itself. Over the mid-term, removing Shemirani's direct connection with audiences on Facebook has probably increased her resilience as a messenger with multiple alliances spread across multiple other platforms linking back to Facebook. She has secured a position within an 'alternative influence network' as identified by Lewis (2018), whereby her ideology supplements a broader reactionary conspiratorial base.

## Conclusion

'Who watches the watchers?' is a classic dilemma for the conduct of social control in liberal societies, reflecting a fundamental tendency for unsupervised power and authority to be corrupted by self-interest. This is especially relevant for the social control of mis/disinformation where social media companies have been 'responsibilized' to lead the imposition of countermeasures. The findings and insights set out in this article show de-platforming interventions by social media companies constitute a recurring and reinvented modality of social control, although there is a lack of independent evidence for their effectiveness.

Informed by empirical case studies of two de-platforming episodes targeted towards charismatic coronavirus conspiracists, this analysis has tracked and traced the intended and unintended effects that flow from such control strategies. The available evidence suggests that some intended impacts can be observed, in terms of disrupting and suppressing problematic behaviour and content. Equally however, multiple unintended consequences were also detected thanks to the 're-platforming' activities of the adherents of the conspiracies. Notably, this included spawning 'minion accounts' on the platform that perform the role of spreading the disinforming material being produced by the 'prime' influencer 'off platform.' A second unintended consequence is that such measures may ultimately increase the resilience of the target group by encouraging them to diversify their cross-platform presence. There is certainly little evidence that de-platforming has triggered deterrence, or large-scale disruption or displacement away from Facebook.

The reasons for this are two-fold. First, it is relatively easy to circumvent such controls by posting hyperlinks to signal back to where a presence, either existing or new, is retained. Second, de-platforming does nothing to address motivations. Many adherents of conspiracy communities, as with other extremist groups, display traits of what Attran (2016) dubs 'devoted actors'. These are individuals who subsume their self-identities into a collective identity oriented to an extreme political cause. The social-psychological investment of such individuals in these politicized collective identities makes it difficult to deflect them from their shared aims and aspirations using external control interventions. Indeed, some of the evidence reported herein suggests one unintended consequence of de-platforming is a kind of 'blowback effect', where it is perversely calibrated as a 'badge of honour', confirming commitment to a collective identity perceived as illegitimately targeted.

There appear to be at least four 're-platforming' behaviours regularly activated by targets of de-platforming control measures:

(1) *Moderate* – 'shelter in place' on platform, but produce less overtly problematic content, albeit this can contain coded messaging.
(2) *Multiply* – develop a network of alternative and minion accounts both on the platform, and across others, signalling their presence.
(3) *Migrate* – account shifts to one or more different platforms.
(4) *Mingle* – link to other groups and ideas that collectively constitute an alternative influence network.

Attending to these permutations and documenting their prevalence suggests de-platforming may be less impactive on conspiratorial thought communities than is perhaps expected. Certainly, metrics that companies use to assess such measures, such as volumes of hashtag posting, seem poorly calibrated to capture the diversity of unintended responses and reactions documented across the two case studies.

Precedents for thinking in such terms can be found in the wider literature on social control. Erving Goffman's (1961) account of 'total institutions' highlighted how, even in situations where there is a pervasive and 'wrap around' surveillance architecture, there are always gaps and crevices that can be exploited by those subject to the regime. As a result, resistance, pushback, and unintended consequences inevitably arise. In policy terms this is especially consequential for our understanding of the formulation and role of de-platforming in controlling disinformation. Individual companies are incentivized to get bad actors off their platforms, but that does not necessarily result in them desisting from malign activity overall. Moreover, this study of Facebook suggests that whilst such interventions may have some limited short-term effects, there is little reason to suppose that over the medium-term they control the flow of disinformation.

## Notes

1. This 'supply' and 'demand' side distinction can be traced back to Schiffrin (2017).
2. See: Beauchamp, Z (2018) Milo Yiannopoulos's collapse shows that no-platforming can work. *Vox*, 5 December. https://www.vox.com/policy-and-politics/2018/12/5/18125507/milo-yiannopoulos-debt-no-platform and also Rogers (2020).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

The authors and research were not funded by Facebook.

## Notes on contributors

*Helen Innes* (PhD) is a Research Fellow at the Crime and Security Research Institute (CSRI) at Cardiff University where her work contributes to the Open Source Communications Analytics Research (OSCAR) programme on disinformation [email: InnesH@cardiff.ac.uk].

*Martin Innes* is Director of the Crime and Security Research Institute and the Universities' Police Science Institute at Cardiff University and a Professor in the School of Social Sciences. He leads a major international research programme (OSCAR) to understand the causes and consequences of distorting and deceptive digital communications [email: InnesM@cardiff.ac.uk].

## References

Albarracín, D. (2020). Conspiracy beliefs. In R. Greifeneder, M. E. Jaffé, E. J. Newman, & N. Schwarz (Eds.), *The psychology of Fake News: Accepting, sharing, and correcting misinformation* (pp. 196–219). Routledge. https://doi.org/10.4324/9780429295379

Alison, T. (2020, September 17). Our Latest Steps to Keep Facebook Groups Safe. *Facebook.* https://about.fb.com/news/2020/09/keeping-facebook-groups-safe/

Attran, S. (2016). The devoted actor: Unconditional commitment and intractable conflict across cultures. *Current Anthropology*, *57*(S13), S192–S203. https://doi.org/10.1086/685495

BBC News. (2020, May 1). Coronavirus: David Icke kicked off Facebook. https://www.bbc.com/news/technology-52501453

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Black, D. (1976). *The behavior of Law*. Academic Press.

Blackburn, J., Gehl, R. W., & Etudo, U. (2021, January 15). Does 'deplatforming' work to curb hate speech and calls for violence? 3 experts in online communications weigh in. *The Conversation.* https://theconversation.com/does-deplatforming-work-to-curb-hate-speech-and-calls-for-violence-3-experts-in-online-communications-weigh-in-153177

Cohen, S. (1985). *Visions of social control*. Polity Press.

Conway, M., Moign Khawaja, M., Lakhani, S., Reffin, J., Robertson, A., & Weir, D. (2019). Disrupting Daesh: Measuring takedown of online terrorist material and its impacts. *Studies in Conflict & Terrorism*, *42*(1-2), 141–160. https://doi.org/10.1080/1057610X.2018.1513984

Courchesne, L., Lihardt, J., & Shapiro, J. N. (in press). *Review of social science research on the impact of countermeasures against influence operations*. Harvard Kennedy Misinformation Review.

Donovan, J. (2020). Concrete recommendations for cutting through misinformation during the COVID-19 pandemic. *American Journal of Public Health*, *110*(October), S286–S287. https://doi.org/10.2105/AJPH.2020.305922

Douek, E. (2021). The free speech blind spot: Foreign Election Interference on social media. In D. Hollis, & J. Ojlin (Eds.), *Defending democracies: Combatting foreign election interference in a digital age* (pp. 265–291). Oxford University Press.

Election Integrity Partnership. (2021). *The long fuse: Misinformation and the 2020 election. Stanford Digital Repository: Election Integrity Partnership. v1.2.0.* https://purl.stanford.edu/tr171zs0069

Facebook. (2020, August 19). *An update to how we address movements and organizations tied to violence.* https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/

Facebook. (2021). *Community standards enforcement report December* 2020. https://about.fb.com/wp-content/uploads/2021/02/CSER-Q4-2020-Data-Snapshot.pdf

Facebook. (2021a). *Community Standards. 17 Account Integrity and Authentic Identity*. https://www.facebook.com/communitystandards/misrepresentation

Fishman, B. (2019). Crossroads: Counter-terrorism and the internet. *Texas National Security Review*, *2*(2), 83–100. https://doi.org/10.26153/tsw/1942.

Garland, D. (2001). *The culture of control*. Oxford University Press.

Goffman, E. (1961). *Asylums: Essays on the condition of the social situation of mental patients and other inmates*. Anchor Books.

Harpin, L. (2020, September 10). Suspended nurse at the centre of anti-lockdown protests called NHS 'the new Auschwitz'. *The Jewish Chronicle*. https://www.thejc.com/news/uk/suspended-nurse-at-the-centre-of-anti-lockdown-protests-called-nhs-the-new-auschwitz-1.506453

Hofstader, R. (1964). *The paranoid style in American politics. Harper's Magazine*, 77–86.

Icke, D. (2020, May 07). I Need Your Help (No – Not Money!) With Bypassing The Censorship. Here's What You Can Do … Memes and Headline comments by David Icke. https://davidicke.com/2020/05/07/need-help-no-not-money-bypassing-censorship-heres-can/

Innes, M. (2020). Techniques of disinformation: Constructing and communicating 'soft facts' after terrorism. *British Journal of Sociology*, *71*(2), 284–299. https://doi.org/10.1111/1468-4446.12735

Innes, M., Roberts, C., & Lowe, T. (2017). A disruptive influence: Prevent-ing problems and countering violent extremism policy in practice. *Law and Society Review*, *51*(2), 252–281. https://doi.org/10.1111/lasr.12267

Jowett, G. S., & O'Donnell, V. (2012). *Propaganda and persuasion*. SAGE Publications.

Kakutani, M. (2018). *The death of truth*. Harper Collins.

Kennedy, D. M. (2009). *Deterrence and crime prevention: Reconsidering the prospect of sanction*. Routledge.

Krafft, P., & Donovan, J. (2020). Disinformation by design: The use of evidence collages and platform filtering in a media manipulation campaign. *Political Communication*, *37*(2), 194–214. https://doi.org/10.1080/10584609.2019.1686094

Lance Bennett, W., & Livingston, S. (2020). *The disinformation age*. Cambridge University Press. https://doi.org/10.1017/9781108914628

Lawrence, D. (2020, August 28). The UK's emerging conspiracy theory street movements. *Hope Not Hate*. https://www.hopenothate.org.uk/2020/08/28/the-uks-emerging-conspiracy-theory-street-movements/

Lawrence, D., & Davis, G. (2020). In *Q Anon in the UK the growth of a movement* (pp. 16–18). Hope not Hate Charitable Trust. https://www.hopenothate.org.uk/wp-content/uploads/2020/10/qanon-report-2020-10-FINAL.pdf.

Lewis, R. (2018). *Alternative influence: Broadcasting the reactionary right on YouTube*. Data & Society Research Institute.

Martin, D. A., Shapiro, J. N., & Nedashkovskaya, M. (2019). Recent trends in online Foreign influence efforts. *Journal of Information Warfare*, *18*(3), 15–48.

Martineau, P. (2019, May 2). Facebook bans Alex Jones, other extremists – But not as planned. *Wired*. https://www.wired.com/story/facebook-bans-alex-jones-extremists/

Mena, P. (2019). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & Internet*, *12*(2), 165–183. https://doi.org/10.1002/poi3.214

Molter, V., & DiResta, R. (2020). Pandemics and propaganda: How Chinese state media creates and propagates CCP coronavirus narratives. Harvard Misinformation Review. https://misinforeview.hks.harvard.edu/wp-content/uploads/2020/06/Ipedits_FORMATTED_PandemicsandPropaganda_HKSReview.pdf

Moynihan, A. (2021). Deplatforming trump puts big tech under fresh scrutiny. *Chatham House*. https://www.chathamhouse.org/2021/01/deplatforming-trump-puts-big-tech-under-fresh-scrutiny

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, *66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Pomerantsev, P. (2019). *This is not propaganda: Adventures in the war against reality*. Faber & Faber.

Roberts, M. (2020). Resilience to online censorship. *Annual Review of Political Science*, *23*(1), 401–419. https://doi.org/10.1146/annurev-polisci-050718-032837

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, *35*(3), 213–229. https://doi.org/10.1177/0267323120922066

Rosen, G. (2021, March 22). How We're tackling misinformation across our apps. *Facebook*. https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/

Schiffrin, A. (2017). How Europe fights fake news. Columbia Journalism Review (26/10/17). https://www.cjr.org/watchdog/europe-fights-fake-news-facebook-twitter-google.php

Schiffrin, A. (2020). *Mis- and Disinformation online: a taxonomy of solutions*. [Doctoral dissertation University of Navarro] Columbia School of International and Public Affairs. https://www.sipa.columbia.edu/

Scuibba, C., & Pasuetto, I. V. (2020, February 17). Misinformation, Media Manipulation and Anti-Semitism. *Columbia News*. https://news.columbia.edu/news/misinformation-media-manipulation-anti-semitism-event-marks-holocaust-remembrance-day

Shibutani, T. (1966). *Improvised news: A sociological study of rumor*. Bobbs-Merrill.

Spring, M. (2020, November 4). Twitter bans David Icke over Covid misinformation. *BBC News* https://www.bbc.com/news/technology-54804240

Tankovska, H. (2021, February 2). Number of monthly active Facebook users worldwide as of 4th quarter 2020. *Statistica*. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide

Walter, N., Cohen, J., Lance Holbert, R., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Woolley, S. C., & Howard, P. N. (2018). *Computational propaganda: Political parties, politicians and political Manipulation on social media*. Oxford University Press.