MIA
MEDIA INTERNATIONAL AUSTRALIA

Original Article

Media International Australia
1–20
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1329878X241236984
journals.sagepub.com/home/mia
Sage

Labelling, shadow bans and community resistance: did meta's strategy to suppress rather than remove COVID misinformation and conspiracy theory on Facebook slow the spread?

Amelia Johns
University of Technology Sydney, Australia

Francesco Bailo
University of Sydney, Australia

Emily Booth
University of Technology Sydney, Australia

Marian-Andrei Rizoiu
University of Technology Sydney, Australia

Abstract

In this paper, we ask how effective Meta's content moderation strategy was on its flagship platform, Facebook, during the COVID-19 pandemic. We analyse the performance of 18 Australian right-wing/anti-vaccination pages, posts and commenting sections collected between January 2019 and July 2021, and use engagement metrics and time series analysis to analyse the data, mapping key policy announcements against page performance. We combine this with content analysis of comments parsed from two public pages that overperformed in the time period. The results show that Meta's content moderation systems were partially effective, with previously high-performing pages showing steady decline. Nonetheless, some pages not only slipped through the net but overperformed, proving this strategy to be piecemeal and inconsistent. The analysis identifies trends that content labelling and 'shadow banning' accounts was resisted by these communities, who employed tactics to stay engaged on Facebook, while migrating some conversations to less moderated platforms.

Corresponding author:
Amelia Johns, Faculty of Arts and Social Sciences, University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007 Australia.
Email: Amelia.Johns@uts.edu.au

Keywords
content moderation, shadowbanning, misinformation, COVID-19, content advisory labels, conspiracy theory

Introduction

In March 2020, shortly after COVID-19 was declared a global pandemic, Meta's Vice President of Global Affairs and Communications, Nick Clegg, announced that the company was 'taking aggressive steps to stop misinformation from spreading' (Clegg, 2020). For Meta's flagship platform, Facebook, this entailed a mix of algorithmic recommendation and content moderation. The techniques employed have been described by Meta's own marketing team with reference to three priorities: 'remove, reduce, inform' (Meta Transparency Centre, n.d.). Following this framework, Clegg claimed that Meta would be identifying and removing false and misleading content that could contribute to 'imminent physical harm', such as false claims about cures and vaccines. For other types of problematic content, not in clear breach of Facebook's community standards (e.g. 'conspiracy theories about the origin of the virus'), the company would use: algorithmic recommendation to promote better quality information and counternarratives to popular misinformation; add content labels and warnings to information found by fact-checkers and other experts to contain misinformation (Gillespie, 2022: 1); and algorithmically demote or reduce the distribution of content that came up to but didn't breach community standards in users' news feeds, search and recommendations.

These techniques contain different underlying assumptions about human behaviour; with some taking a view that users' can be persuaded to trust better quality information if they are warned away from some content and recommended alternatives. Other assumptions are less generous, such as the claim that users are inevitably drawn to 'sensationalist and provocative content' (Zuckerberg, 2018) – a view that has justified and normalised less transparent approaches to content moderation to avoid attention being drawn to what gets moderated and why.

The transparency (or lack thereof) of these techniques, and the extent to which platforms are able to be held to public account, has become a controversial aspect of Meta's platform governance, with scholars, activists and human rights organisations repeatedly calling for the platform to be 'accountable to their own policies' (Gillespie, 2022: 8). Meta has reluctantly responded by submitting regular transparency reports and detailed metrics of enforcement actions. Nonetheless, critics have been quick to highlight what is missing from these reports. Firstly, there is limited transparency regarding the *effectiveness* or otherwise of these measures. This follows Suzor et al. (2019: 1527), who claim that merely providing aggregate statistics on content moderation enforcement does nothing to address whether it is working or not, and can even help platforms to evade responsibility by 'treating disclosure of information as a goal in itself'. Secondly, there is almost no reporting on decisions to demote content, leading to claims that one of the murkier backend content moderation techniques is not subject to transparency and accountability at all (Gillespie, 2022).

This selective transparency has been explained by internal modelling provided by Meta in 2018, with the release of its 'Blueprint for Content Governance and Enforcement' (Zuckerberg, 2018). As the blueprint explains, for content not in clear breach of Meta's community standards, decisions to remove are ineffective as they *increase* rather than *decrease* attention to the content, and incentivise content creators to find workarounds. Critics have been cynical of this justification (Gillespie, 2022; Matamoros-Fernandez, 2017), claiming that it is more about protecting Meta's commercial interests, as removing content and accounts that enjoy high engagement limits Meta's ability to gain commercially from their activity. Moreover, scholarship consistently shows that the rationale for moderating content without notifying users (referred to in lay language and industry parlance as 'shadow banning') has not produced the outcomes imagined by Meta. Content creators and community members are often made aware of 'shadow banning' through sudden, inexplicable reductions in engagement or users notifying account holders that their content is no longer visible in their feed (Are, 2021; Duffy and Meisner, 2023; Myers West, 2018; Savolainen, 2022; Suzor et al., 2019). This has contributed to rather than reduced attention and fuelled conspiracy theories that the platform is trying to silence conservative or non-mainstream thought (Myers West, 2018: 4374; Suzor et al., 2019).

To date, scholarly inquiry has addressed these debates by considering: how *platforms* have justified decisions to label or reduce rather than remove content; how these reduction techniques have been perceived and responded to by *users*, while a third aspect has questioned the *technical robustness of terminology* such as shadow banning, which some critics argue is often so vaguely applied that it risks contributing further to the misinformation problem (Duffy and Meisner, 2023; Keulennar et al., 2021; Nicholas, 2022; Savolainen, 2022). Not as many studies have sought to look at Meta's labelling and suppression strategies in terms of their effectiveness, (i.e. its impact on engagement), nor has there been detailed empirical investigation of right-wing and conspiracy communities' practices of collective *resistance* to these content moderation techniques, which, we hypothesise may challenge the modelling and assumptions which underpin them. To address this gap, in this paper, we propose an integrated analytical framework, drawing on mixed methods (qualitative, quantitative and computational), to answer the following questions:

> RQ1 - Have Meta's content moderation policies – focused on reducing the circulation of content rather than removing it – been effective in slowing the spread of vaccine misinformation and COVID related conspiracy theories on its flagship platform, Facebook?
>
> RQ2 – Do the communities impacted by these techniques behave in the way Meta's modelling and assumptions predict?
>
> RQ3 - Do they employ counter-strategies, and with what success?

We believe this research advances current scholarship and policy-making by bringing more nuance to research and debates concerning Meta's selective transparency in its approach to platform governance.

Literature and policy review

Meta's content moderation systems: what they do in the shadows

Tensions between Meta's role as a private, commercial enterprise, and as a mediator of public conversations are nowhere more apparent than in its content moderation policies. The need to moderate is an aspect of governance that platforms have performed reluctantly, partially informed by the ideologies guiding the early web, which were premised on fantasies of an 'open' platform which would 'host and extend all… participation, expression and social connection' (Gillespie, 2018: 5). But growing recognition that openness has allowed hate and misinformation to flourish and even become amplified by the commercial platform operators has dampened this enthusiasm, increasing calls for more regulation to limit this free for all. Nonetheless, discomfort with the necessity of imposing limits on 'free speech' is evident in platforms' repeated claims that they are 'not media companies' (Duffy and Meisner, 2023; Gillespie, 2018). Instead they have portrayed themselves as 'neutral' intermediaries who host and recommend user-generated content, while downplaying their role as ‘custodians’ who moderate and curate the content users see (Gillespie, 2018; Myers-West, 2018).

There are also economic reasons for this reluctance to moderate, or at least to be transparent about it. Platform business models, while still ad-driven enterprises, also monetise user expression, connection and engagement by building a ‘social graph’ of user participation, preferences, search data etc. which can be sold to advertisers and data brokers (Gillespie, 2018: 17). This aspect of the business relies upon platforms expanding their user-base and keeping users engaged. Removing content and banning or suspending users who are highly engaged with the platform, as some critics have highlighted, is bad for business and produces a conflict of interest that clouds platforms’ judgement on moderation principles.

These tensions are evident in the less than transparent content moderation policies of companies like Meta (Duffy and Meisner, 2023). But while attention to Meta’s content moderation policies has increased, much of this focus has been directed toward decisions to *remove* content and suspend accounts (Gillespie, 2022). This, it is argued, serves the interests of Meta, as decisions to remove content usually refers to clear, unambiguous breaches of terms of use or community standards. Much less focus is placed on moderation techniques that reduce the *visibility* and *reach* of content that doesn’t present a clear breach, but where this content remains hosted by the platform and made available to users who are determined to find it. Inattention to these decisions is desirable as they are often much more morally ambiguous, subjective and difficult to justify. As such, Meta has been willing to be transparent about their enforcement decisions relating to clear content violations, and less so with these other forms. In this paper, we want to shed light on two of these strategies ‘beyond removal’: the application of content advisory labels, and strategies to demote the distribution of content in feed, search and recommendation. We do so with the intention of addressing the effectiveness of these approaches, and querying whether the assumptions that underpin them actually hold.

Content advisory and labelling

The practice of adding content advisory labels to problematic content and misinformation during the COVID-19 pandemic has been noted in the multiple transparency reports Meta have produced. While this content continues to be hosted by the platform, Meta attaches labels advising users that fact-checkers have deemed it to contain misinformation, or that it offers support to ‘dangerous organisations and individuals’ (such as QAnon). Named ‘interstitial warning’ labels (Diaz and Hecht-Felella, 2021; Goldman, 2021; Guo et al., 2023) Facebook applies these labels in the form of a ‘cover to block and blur misinformation before people can see it’ (Guo et al., 2023: 2). Alternatively, users might encounter content appended with a contextual warning label, displayed as a banner. Banners often warn users but also recommend content, redirecting users to authoritative content about vaccines, for example (Goldman, 2021). Platforms justify the decision to warn users rather than remove content by claiming that some content may have public interest value, despite breaching community standards. There is also a belief in the educational value of allowing users to be shown misinformation, so they can be persuaded not to open and share it.

But findings have been mixed on the effectiveness of labelling. While there is strong evidence that supports the view that content warning labels generally reduce the likelihood of users believing or resharing fact-checked content, even in relation to politically contentious information (Martel & Rand, 2023), their limited effectiveness has been discussed in several studies. For example, there are claims that prior exposure to labelled content may increase the likelihood of users believing an information’s accuracy, despite labels later being appended (Pennycook et al., 2018). The same is true for repeated exposure to content formerly viewed with a label, but later viewed again without a label (Martel & Rand, 2023). Some of these diminished effects have been related to the high volume of content containing misinformation, presenting a problem for fact-checkers and platforms in effectively evaluating and labelling content at scale. But research has also shown that content warning labels may be inflammatory for some partisan users, reinforcing trustworthiness in the labelled content and increasing distrust in the moderating platform (Pennycook et al., 2018). Research has also assessed the validity of labelling content on video-sharing platforms by surveying ordinary users (Guo et al., 2023). Findings show that user perceptions of labelled content is varied, with labels that present more friction in user experience (i.e. labels that cover and block) being perceived to be more effective than contextual content labels that users become habituated toward seeing and often ignore.

At the same time, platform justifications for labelling instead of removing content have been criticised by platform scholars, who claim that reducing rather than removing a content's visibility and reach – or introducing friction to user experience to make accessing the content more difficult – does not stop users who are highly motivated to consume and promote misinformation. This has led to claims that appending warning labels to content is a way for platforms to show they are moderating while not being seen to take a harder censorship line to politically fraught content, for example, by removing it. According to scholars like Diaz and Hecht-Felella, this 'amounts to protection for the powerful' (Diaz and Hecht-Felella, 2021: 12). Diaz and Hecht-Felella also highlight the unevenness in decisions to remove content by 'dangerous organisations and individuals'. For example, they note that QAnon, while listed as a 'dangerous organisation' by Meta, falls under what the company calls a 'tier 3' categorisation (p. 6). Rather than removing content and accounts found to support the group, this means some content is merely labelled. While this may act as a deterrent or warning for some users who innocently encounter this content, it is an action that also reveals a double-standard in how moderation is applied. This double-standard has also been highlighted by whistle-blowers from within the company (Diaz and Hecht-Felella, 2021; Karppi and Nieborg, 2021), who argue that the unevenness of applying content moderation policies challenges platform claims to impartiality.

Borderline content and 'shadow banning'

'Shadow banning' is a 'folkloric' term (de Keulenaar et al., 2021; Savolainen, 2022) that has been adopted by users to name another controversial platform moderation technique. Though often blurry in definition, it is frequently taken to refer to any action to algorithmically demote or hide the visibility of posts or account information to other users in search, ranking and recommendation, without the content creator or other users being aware (Are, 2021; Gillespie, 2022; Keulennar et al., 2021; Myers West, 2018; Nicholas, 2022).

Some of the controversy around the term arises from claims that it is misused to refer to a range of opaque platform mechanisms. This is particularly the case given that these techniques don't leave a 'trace' as do other actions like labelling or removal, and aren't included in platform transparency reports, making them difficult to detect or be explainable (Duffy and Meisner, 2023; Nicholas, 2022; Savolainen, 2022). Moreover, it is a technique that falls between content moderation and recommendation, with algorithmic reduction or 'suppression' being used to address problematic user-generated content, like misinformation, while it is also used to improve user experience by removing the visibility of spam, clickbait, bots, etc. (Gillespie, 2022: 9). These tasks have very different justifications and are actioned by different teams, so while users may feel that they have been 'shadow banned', it is often difficult to tell.

Nonetheless, platforms have been more forthcoming since 2018 in outlining where and why they choose to demote problematic content. For example, Meta clearly outlined their policies to demote rather than remove content in their 'borderline content' policy announcements in 2018 (Zuckerberg, 2018). In the announcement, the company outlined their use of techniques to reduce distribution and visibility of content which came up to but didn't clearly breach community standards. According to Meta's internal modelling, if borderline content, and the accounts distributing it, received less engagement, while still being hosted by the platform, this would improve the quality of information circulating through the entire platform without inflaming claims of bias or censorship (Zuckerberg, 2018: 29–31). The modelling also suggested that, if content was noticeably removed, it would 'incentivise' user efforts to find new ways to distribute the content. The underlying assumption here is that being *more* transparent would assist rather than minimise the circulation of problematic content and misinformation. Secretive or clandestine moderation of content then is justified as a means to an end.

This has fuelled a new focus in scholarship on platform transparency. For example, 'borderline content' policies have led to a spike in critical platform studies scholarship and user-focused research on shadow banning. The latter has investigated perceived shadow banning of content from diverse user communities, including: white nationalists, anti-vaccination communities, sex workers, Black Lives Matters activists, influencers and content creators, and user communities focused on mental health and self-harm (Are, 2021; Duffy and Meisner 2023; Gillespie, 2022; Gerrard, 2018; Suzor et al., 2019; Tiidenberg and van der Nagel, 2020). Platforms, if they acknowledge it at all, suggest downranking or demoting content cleans up public discourse and removes content that may offend or harm. But the vagueness around these explanations and what is and isn't defined as online harm has only accelerated claims that platforms are, at best, being inconsistent with the policy (Myers West, 2018; Suzor et al., 2019), and, at worst, are preferencing some narratives and content over others in a way that discriminates, often against marginalised communities (Duffy and Meisner, 2023; Gillespie, 2022; Nicholas, 2022)

Facebook's conservative and right-wing communities have joined this growing chorus, with claims that they too have been targeted and silenced by these suppression techniques (Myers West, 2018; Suzor et al., 2019). A common theme is a feeling that there is a political bias in play, with 'conservatives, Donald Trump supporters, "alt-right" figures, Gamergates, Bernie Sanders supporters, anti-vaccination campaigners […] and more' feeling that they have been subject to unfair algorithmic demotion techniques (Suzor et al., 2019: 1533).

In this paper, we hypothesise that user responses to perceived shadowbanning inverts Meta's modelling, with attention becoming hyper-focused on the lack of transparency with which content moderation techniques have been made, rather than such attention only ballooning around traces of removal. Following the literature, we also anticipate that users' receiving insufficient notification that they have been subjected to moderation would fuel rather than reduce the misinformation and conspiracy theories it is designed to limit.

As Duffy & Meisner explain (2023: 287), responses to perceived shadow banning are potentially counter-productive to Meta's aims, with users often engaging in 'efforts to circumvent algorithmic intervention'. This may involve dropping content that is likely to be flagged, labelled or suppressed in comments rather than posts, owing to beliefs that this content is not as highly moderated (Moran et al., 2021). Users have also used 'social steganography' (for original usage see Duffy and Meisner 2023; Gerrard, 2018; Marwick and boyd, 2014) – using deliberate typos and encoded language– to evade algorithmic detection (see also Grondahl et al. 2018; Moran et al., 2021). Beyond this, account owners have directed their followers to less moderated platforms to share content likely to be detected on Facebook. By adding links in posts and comments to direct users to this alt platform content, there is a belief that account owners can avoid algorithmic moderation while continuing to maintain a presence on Facebook, which helps with legitimacy and credibility of their brand (Baker, 2022; Rogers, 2020; Zeng and Schafer, 2021). In the findings section, we explore these practices of subversion in more depth.

Methodology

The data included in the paper was sourced from a project which identified and mapped *user profiles* and *linked content* connecting 'far-right', 'right-wing populist' and 'anti-vaccination' communities across selected platforms (including Facebook) (Kong et al., 2022). The project included observation of active social media accounts and pages that shared misinformation, hate speech and conspiracy theory related to four topics. This paper focuses on one of these topics: anti-vaccination opinions and conspiracy theory. In this paper, we have narrowed our focus to Australian far-right and anti-vaccination communities on Facebook, both of whom were found to actively share misinformation and conspiracies related to vaccines during the height of the COVID-19 pandemic.

Sampling[1]

To collect the data corpus relevant to this paper, the research team used CrowdTangle to collect data at scale on a list of 21 Australian Facebook public accounts (pages and groups) found to be active in the observation stage (Step 1). We then used a mix of platform data sourced via CrowdTangle and ethnographic field notes to classify the accounts. Pages were labelled 'right-wing' if they demonstrated a prevalence of content and views which were anti-immigration, promoted climate change denial, promoted narrow definitions of sexuality and gender roles etc. Anti-vaccination accounts were classified based on prevalence of anti-vaccine content. Using CrowdTangle, we downloaded 38,704 public posts from the accounts between January 2019 and July 1 2021. Three were excluded as less than 1% of their posts contained terms related to vaccines or COVID-19 in the time period.[2]. This reduced the corpus to 18 accounts and 34,202 postings. A performance analysis was then conducted on the remaining accounts to address RQ1.

A next step (step 2) identified two accounts that were found to consistently over perform. We collected a set of 107 comment threads from these accounts by random sampling among all threads returned by CrowdTangle in the time period, using the same search terms as step 1 and then using web-crawling to collect 2842 comments from these threads, which is approximately 10% of the cumulative number of comments declared by Facebook. Posts and comments were then subject to exploratory topic modelling and quantitative content analysis, addressing RQ2 (see data analysis, Figure 1). We use the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003), a classic statistical topic model that describes how a set of observations (typically words in documents) can be explained by unobserved groups (topics) that explain why some parts of the data are similar.

A final step (step 3) involved purposively sampling three sets of comment threads from the 107 identified in step 2. All three were selected based on a codebook developed from the LDA topic modelling and quantitative analysis (see data analysis, Figure 1). The codebook also included terms related to content moderation and counter-moderation tactics employed by users, as found in the literature review. This included the terms 'shadow ban', 'censorship' and 'labelling'. After sampling the threads, we then web-crawled these accounts, obtaining 908 comments (or about 81% of the cumulative number of declared comments for the three threads, according to CrowdTangle statistics). Of the 908 comments, we randomly sampled 50 comments from each thread and performed a qualitative, thematic analysis on these comments ($N = 150$).

Data analysis

To analyse the performance of the Facebook pages sampled in step 1, we used two different metrics. The first metric is CrowdTangle's 'overperforming score',[3] which compares the performance of

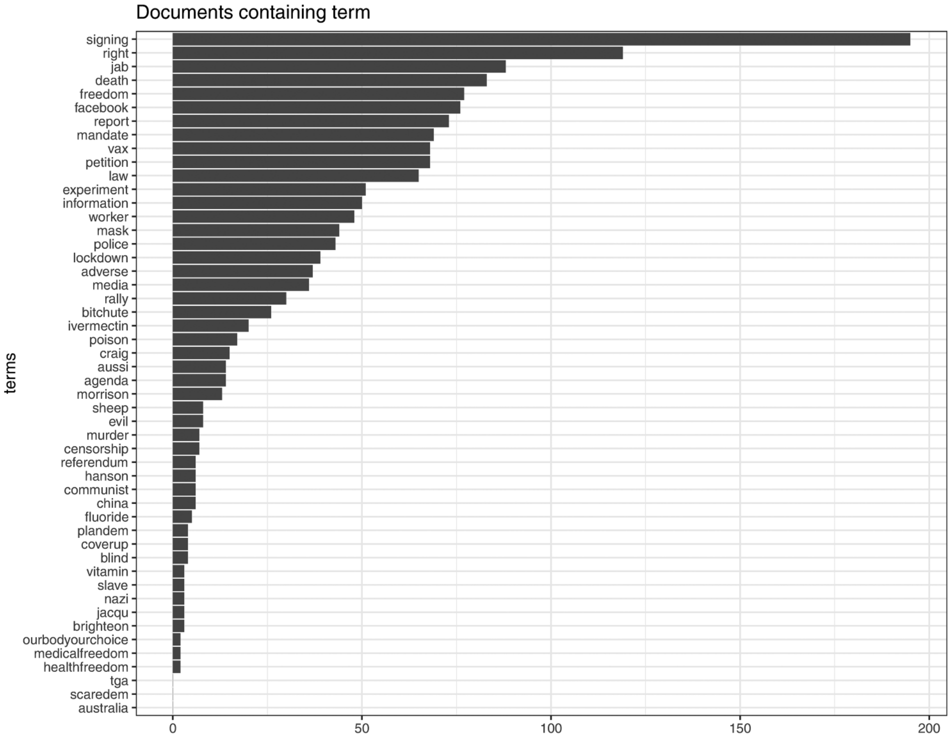


Figure 1. Overperforming pages' most frequent terms.

each post, in terms of interactions, to the average performance of the previous 100 posts published by the account. To avoid the metric being influenced by high- and low-performing outliers, CrowdTangle removes the top and bottom 25% of the posts based on performance (therefore using only 50% of the posts). We built the second measure as the average number of shares-per-post published by a single account over a 14-day moving window and compared this to a benchmark period (January 2019–December 2020). To avoid outliers, we removed the top and bottom 2.5% of the posts (that is, we use 95% of the posts). Finally, we mapped the performance analysis against key announcements and updates regarding Facebook's content moderation policies during COVID-19, as found in the policy review (see 'policy review'). As CrowdTangle does not archive removed accounts, we can be confident that the pages or accounts were not removed and, from this, can reasonably deduce that 'reductive' content moderation techniques (either labelling or demotion) were employed in the posts sampled for the analysis. Although it is not possible – without setting up an experiment – to disentangle and measure the effects of each of these approaches separately, we think it is reasonable to interpret significant increases or drops in performance (engagement) as likely produced by platform policies and 'reductive' content moderation techniques (considered in aggregate), which this study sheds light on.

The next stage of analysis was the quantitative and qualitative content analysis on overperforming accounts. On the corpus of comments and posts scraped in step 2, we conducted LDA topic modelling analysis (after removing stop words and punctuation) for exploratory purposes setting the number to 26 topics so as to obtain more precise and self-contained topics. We then used the key terms identified in the topic modelling to search the corpus for frequently used terms (see

Figure 1). This helped us to address RQ2 and RQ3 by providing insight to the conversation topics that dominated engagement on these pages.

In the final stage, two independent coders thematically analysed 150 randomly sampled comments from over performing accounts, identifying where frequently used themes in the LDA topic modelling were used in the construction of key discourses. This in-depth reading also identified how users interacted with labelled content and content believed to be subject to 'shadow bans', and finally it identified what tactics users employed to cope with or overcome these forms of platform governance, addressing RQ3. However, recognising that the randomly sampled comments did not necessarily reflect all the avoidance tactics of interest, a senior member of the research team manually scanned all 908 comments for relevant key terms and a further 19 comments were included in the thematic analysis.

Policy review

To analyse page performance of accounts against key content moderation and recommendation policy announcements, a member of the research team adapted Meta's integrity timeline: 2016–2021 (Meta, 2022) excluding policy announcements and updates between 2016 and 2020, and further excluding policies that (a) didn't refer to Facebook (instead referring to Meta's other platforms), and (b) did not make mention of reduction strategies relevant to this paper, but instead focused exclusively on removal or recommendation. Two policies referring primarily to removal are included for context. The key policies are summarised in Table 1 and have also been visualised against page performance in Figure 2.

Findings

Figure 2 shows the aggregated daily performance of the 18 accounts (median and mean shown with solid and dashed lines; and the 80% distribution – i.e. between the 10th and 90th percentile – shown by the grey dashed area). The figure showcases two divergent trends. Firstly, the median level consistently declines after March 2020. This indicates that most accounts suffered a decrease in performance due to what we have deduced are content moderation interventions aimed at reducing distribution and engagement with problematic content. However, both the mean line and the 80% distribution show increasing levels after November 2020. This indicates the existence of outliers that significantly outperform the baseline levers of 2020. In summary, while the majority of the monitored accounts *under*-performed, both in 2020 and in 2021, a few accounts instead *over*performed, and strongly so.

To examine what is happening in terms of engagement on high-performing pages, Figures 3 and 4 illustrate posting activity and engagement metrics for two accounts – the Informed Medical Options Party, and a community-based anti-vaccination page. To track the performance of these two accounts, we compared the evolution over time of four metrics: the frequency of postings, the number of page followers, the number of posts' shares by month and the posts' performance - CrowdTangle's 'overperforming score' - also by month. For readability, monthly statistics for both shares and performance scores exclude posts from both the bottom and top 10% quantile.

From the two figures, we observe that after February 2021, both pages see a significant increase in the number of followers as well as an increase in the number of shares and performance of their posts. From this, we can deduce that Facebook's policy changes introduced across the course of 2020, which peaked in 2021 with a seeming change of policy direction to remove rather than reduce false claims about vaccines (February 2021), had not only failed to severely dent engagement with some pages, but that, instead, these pages enjoyed a significant increase in followers and engagement.

Table 1. Review of relevant content moderation policy announcements and updates (Jan 2020–July 2021).

Date	Link	Summary
1. Mar 25, 2020	https://about.fb.com/news/2020/03/combating-covid-19-misinformation/	'Aggressive' steps being taken to stop misinformation from spreading. Discussed in Introduction
2. Apr 16, 2020	https://about.fb.com/news/2020/04/covid-19-misinfo-update/	Once a piece of COVID-related content is rated false by fact-checkers Meta announces, it will reduce its distribution and show warning labels with more context.
3. Aug 19, 2020	https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/	QAnon is listed as a dangerous organisation and efforts made to restrict them from organising on Facebook. Measures include removal of accounts if they make violent claims, downranking accounts and content in newsfeed and search, limit in recommendations, prohibit the use of ads and sale of products in marketplace and prohibit the use of fundraising tools.
Oct 06, 2020	Update. See above (Aug 19) for link	Removal of QAnon accounts that promote the organisation, even if they don't contain violent content
4. Oct 21, 2020	https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/	Update to Dangerous Organisations policy: Meta announces that if someone searches Facebook using a term related to QAnon, a label will cover the content and they will be redirected to the Global Network on Extremism and Technology (GNET) initiative to combat violent extremism.
5. Feb 8 2021	https://about.fb.com/news/2020/04/covid-19-misinfo-update/	Update on claims that would now be removed after consultation with WHO. These include that COVID-19 is man-made or manufactured; vaccines are not effective in preventing disease; it's safer to get the disease than a vaccine; vaccines are toxic, cause harm or autism.
Mar 22 2021	https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/	Provides an update on misinformation reduction policies, specifically labelling, with one sentence dedicated to demotion strategies but no enforcement data
May 26, 2021	https://about.fb.com/news/2021/05/taking-action-against-people-who-repeatedly-share-misinformation/	Addition of content labels warning users who encounter accounts or pages that continuously share misinformation. The announcement also contains an update on efforts to notify account holders who continuously share misinformation of the steps taken against them. This includes notifying users when their content is demoted in newsfeed.

Content analysis

To seek further insight into these trends, we used LDA topic modelling and quantitative content analysis to identify key terms and themes arising across the two overperforming accounts, between February and July 2021.

The results (see Figure 1) show that topics related to vaccine misinformation and conspiracy theory were prevalent across both accounts. While the most frequently used terms indicate a

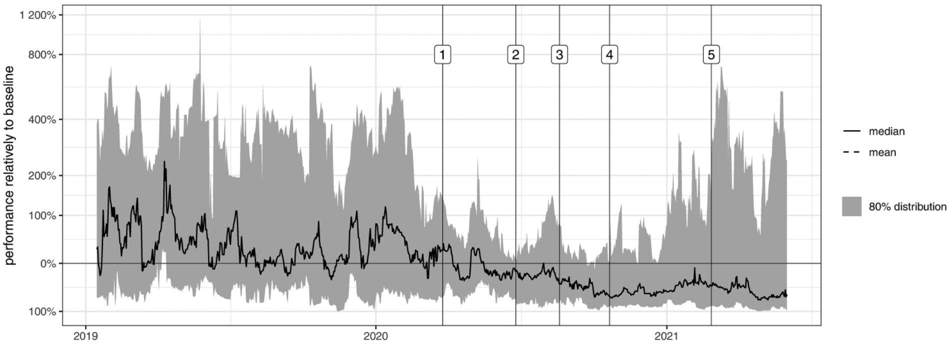


Figure 2. Daily performance of posts matched against key policy announcements: 1: Facebook vows more ‘aggressive action’ on COVID & vaccine misinformation; 2: labels added to content to show users the source of information; 3: expands lists of ‘dangerous organisations and individuals’ to include QAnon, and suggests suppression of content in newsfeed will result for any user or page that supports QAnon-related content; 4: Building on ‘dangerous organisations and individuals’ policy, users and pages sharing support for QAnon would be labelled violent extremists and users encountering content appended with this new label redirected to GNET counsellors 5: Facebook claims it will remove groups, pages and accounts that keep making false claims about vaccines.

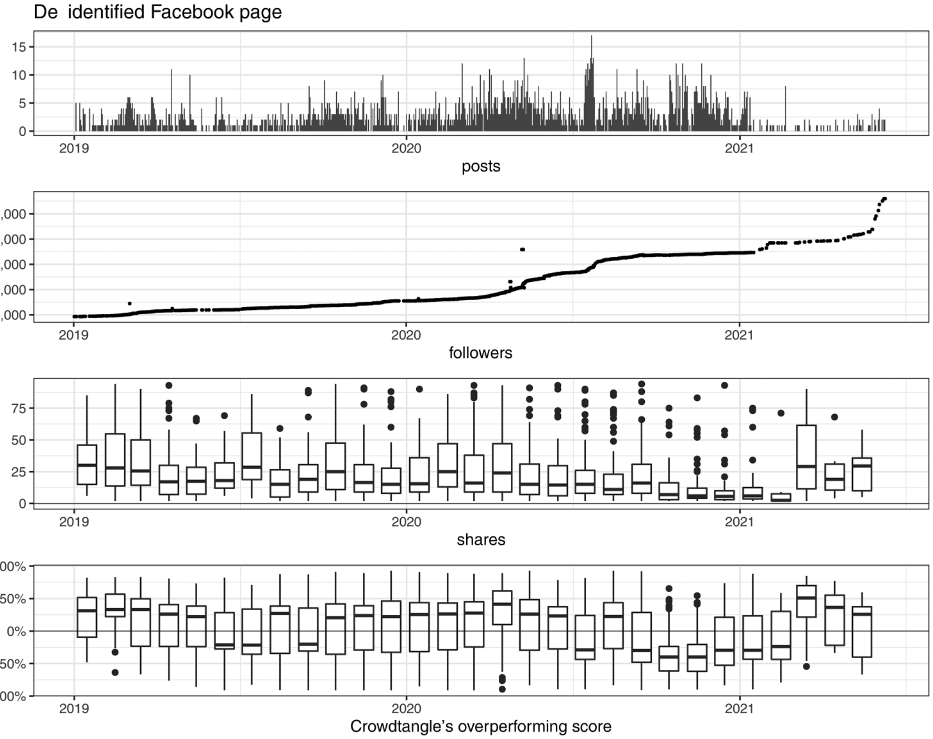


Figure 3. Posting activity and engagement metrics for a de-identified Facebook page.

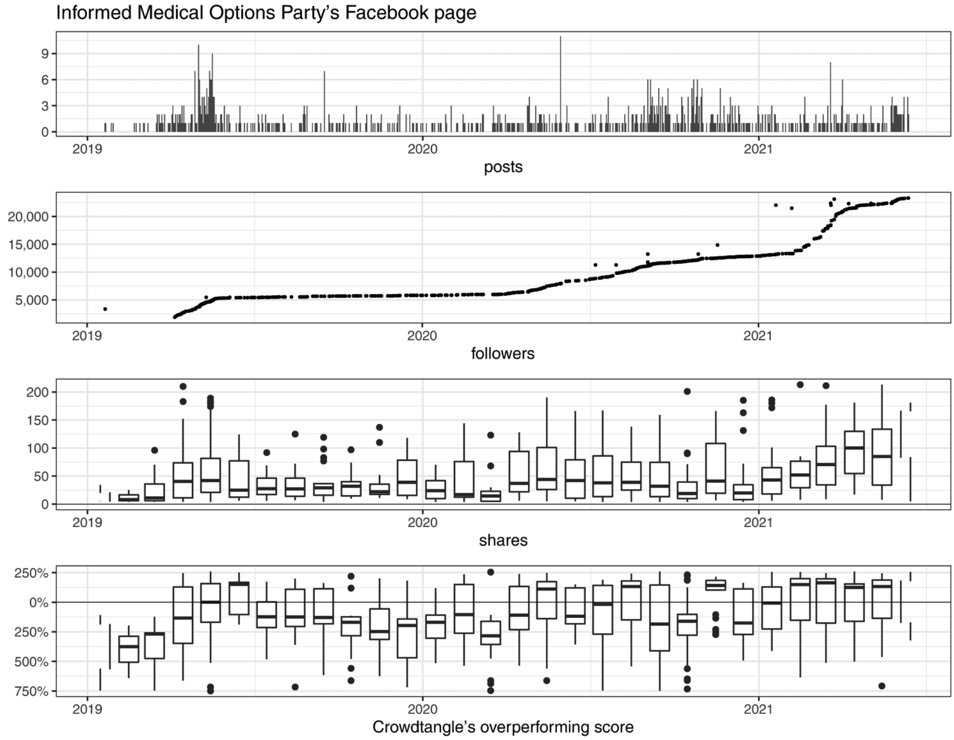


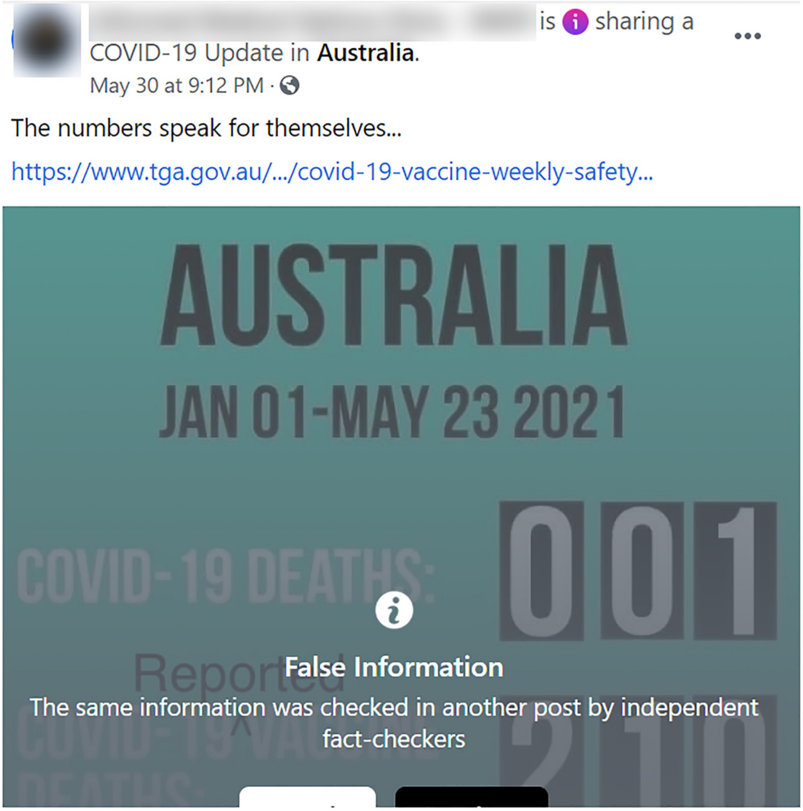
Figure 4. Posting activity and engagement metrics for Informed Medical Options Party.

focus on mainstream tactics to raise the community's voice and influence policy, (e.g. signing petitions), the equally frequent use of terms such as 'jab', 'death' and 'experiment' indicates continued engagement and sharing of vaccine misinformation and conspiracy theories, with these terms often being raised in discussion around government cover-ups of deaths related to the vaccines (as shown in the thematic analysis). Associated tactics of sharing misinformation to cause doubt about the vaccines' approval for use in Australia was also noted with frequent use of terms such as 'experimental'. Although terms like ivermectin, plandemic and coverup – all of which indicate the sharing of higher end conspiracy theory and vaccine misinformation – are toward the lower end of the scale in terms of frequency, the next stage of the thematic analysis sheds light on the use of social steganography (lexical variation and encoded terms) to dog whistle to community members a belief in these conspiracy theories while limiting algorithmic detection and moderation.

Thematic analysis

The thematic analysis allowed more in-depth analysis of practices of content moderation evasion related to specific techniques of content moderation and key policy announcements. The first two of these threads were published on the page of IMOP on 30 May 2021 and 3 July 2021 (Figures 5 and 6), while the third post was published on the community-based anti-vaccination page on 26 June 2021 (Figure 7).

'Please Read the Comments': labelling, social steganography and conspiracy seeding. The first commenting thread contained responses to an infographic (see Figure 5) to which Meta had added an



[redacted] is ⓘ sharing a COVID-19 Update in **Australia**.
May 30 at 9:12 PM ·

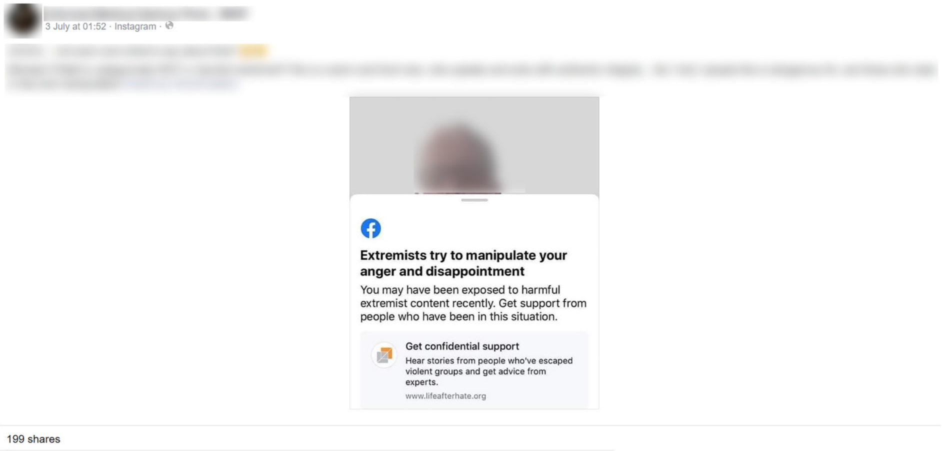
The numbers speak for themselves...

https://www.tga.gov.au/.../covid-19-vaccine-weekly-safety...

False Information

The same information was checked in another post by independent fact-checkers

Figure 5. Post of an infographic detailing 'vaccine deaths', which is a misrepresentation of data from the Therapeutic Goods Administration.



[redacted]
3 July at 01:52 · Instagram ·

[redacted]

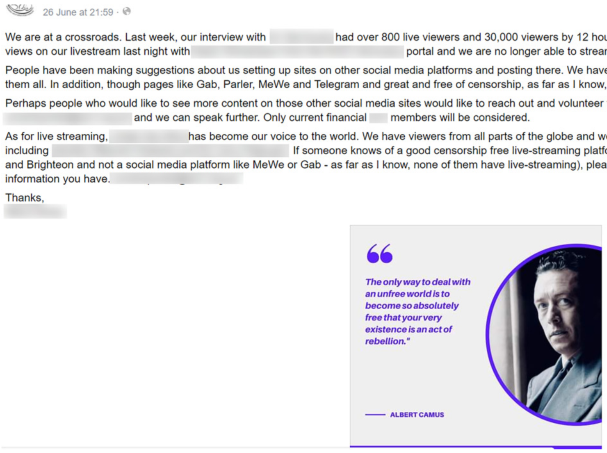
Extremists try to manipulate your anger and disappointment

You may have been exposed to harmful extremist content recently. Get support from people who have been in this situation.

Get confidential support
Hear stories from people who've escaped violent groups and get advice from experts.
www.lifeafterhate.org

199 shares

Figure 6. Screenshot of a Facebook video where content labelled 'extremist' is reposted to Facebook newsfeed.



26 June at 21:59 ·

We are at a crossroads. Last week, our interview with [redacted] had over 800 live viewers and 30,000 viewers by 12 hours after it played live. Now, due to Facebook's censorship, we had 27 views on our livestream last night with [redacted] portal and we are no longer able to stream to this page.

People have been making suggestions about us setting up sites on other social media platforms and posting there. We have multiple social media pages but not enough hands on deck to man them all. In addition, though pages like Gab, Parler, MeWe and Telegram and great and free of censorship, as far as I know, they do not have live-streaming as an option.

Perhaps people who would like to see more content on those other social media sites would like to reach out and volunteer their time to run these pages for us? Contact us at [redacted] and we can speak further. Only current financial [redacted] members will be considered.

As for live streaming, [redacted] has become our voice to the world. We have viewers from all parts of the globe and we have some incredible interviews coming up in the very near future including [redacted] If someone knows of a good censorship free live-streaming platform (not a place where recorded videos can be uploaded like Bitchute and Brighteon and not a social media platform like MeWe or Gab - as far as I know, none of them have live-streaming), please contact me as a matter of urgency and let me know what information you have. [redacted]

Thanks,
[redacted]

The only way to deal with an unfree world is to become so absolutely free that your very existence is an act of rebellion."
ALBERT CAMUS

Figure 7. Post from vaccine critical community admin seeking advice on livestream options to avoid perceived shadow banning.

interstitial label, blurring the content, and indicating it contained false information. The contents of the infographic claimed that, between 1 January and 23 May 2021, there had been only one COVID-19 death in Australia but 210 'COVID-19 vaccine deaths'. Fact-checking reports confirmed that the data misrepresented a Therapeutic Goods Association report that had found 210 'deaths following vaccination'. The report clarified that there was only a causal link connecting one of these deaths to the vaccines, with the rest most likely due to co-morbidities and old age, but this information was omitted from the infographic. Despite the warning label having a clear intention to persuade users not to open or share content, the post enjoyed 1022 reactions, 740 shares and 350 comments, with many comments debating or condemning the labelling decision, driving a long tail of critical engagement after it had been applied.

By thematically analysing the comments, we gain some insight into the tactics used by the community who demonstrated a high motivation to stay engaged with this content, and for whom the labelling was not a disincentive but a challenge. Specifically, 'social steganography' (Duffy and Meisner, 2023; Marwick and boyd, 2014), and conspiracy 'seeding' or dropping 'breadcrumbs' (links to content in archive bins or less moderated sites) in the comments (Baker, 2022; Moran et al., 2021; Rogers, 2020; Zeng and Schafer, 2021) were tactics that were frequently used to re-distribute labelled content and avoid detection. This agrees with Zeng & Schafer's study, where the use of 'archiving portals or pastebin sites' were identified as tactics that allowed users to access content that was likely to be moderated by the mainstream platforms (Zeng and Schafer, 2021: 1334). Others took advantage of the strong engagement to 'seed' further misinformation and conspiracy content in the comments. One example shared a link to IMOP's website and a page that had been set up to list alleged conflicts of interest of Australian immunisation specialists involved in creating vaccination policy. On the page several specialists are cited as being 'in direct partnership with the ID2020 agenda'. The use of the language 'agenda' is a code language commonly used to reference conspiracy theories that circulated during the pandemic regarding ID2020[4], particularly the claim that ID2020 and one of its main philanthropic funders, Bill Gates, were using mandated vaccines to insert microchips into patients' bodies for population control (Huddleston Jnr., 2020).

Another commenter shared a link to a Canadian podcast series and an interview with a Professor who referenced discredited data showing harmful side-effects from COVID-19 vaccines for children. In other cases, users added links to misinformation and conspiracy theory hosted on sites that champion themselves as ‘moderation lite’, like bitchute, a video hosting site launched in 2017 which describes itself as offering ‘freedom of expression’ and which has become known as a site where de-platformed anti-vaccination influencers and conspiracy-related content are hosted (Rogers, 2020). In one example in our analysis, a user added a link to a bitchute video in the comments and contextualised it by using ‘social steganography’ that dog-whistled to the community, showing support for QAnon style conspiracies and cover-ups of vaccine deaths. As key terms such as ‘vaccine’ and ‘death’ were believed by community members to trigger algorithmic detection, emoji or other code names were often used in their place, as this example shows:

> A friend sent me this link, it’s [sic.] refers to over 4000 deaths of individuals after getting 💉 The true number will not come out, it’s not in the public’s interest to disclose the amount of people that have died within day’s [sic.] of jab.

The analysis revealed multiple uses of encoded terms and language. In another example, the vaccine was referred to as ‘experimental’ rather than a clinically tested vaccine by five commenters, with two of the comments receiving 15 and 40 reactions respectively, mostly supportive. By connecting the thematic analysis to the topic modelling and quantitative content analysis, we can see that frequently used words, such as ‘experimental’ or ‘bitchute’ are not just what the community is discussing, but that they are connected to tactics to avoid algorithmic detection and stay engaged with content the platform is labelling or suppressing.

While many of the conspiracy theories discussed by the community were targeted at government and public health authorities, the move by social media platforms to suppress or reduce the distribution of anti-vaccination content, and label content containing misinformation, fuelled further conspiracies regarding big tech and their complicity with big pharma and governments. Baker’s study on anti-vaccine wellness influencers particularly identifies how community members in her study used encoded language, such as MSM to dog whistle about the existence of QAnon style agendas – without the need for terms such as ‘Deep state’, ‘plandemic’, ‘5G’, the ‘Great Reset’ or others that might trigger platform moderation. Baker concludes that this is an effective way to maintain engagement on mainstream platforms and build up a following before re-directing users to ‘less regulated platforms such as Gab, Parler, MeWe and Telegram’ (Baker, 2022: 16). This language was also frequently used by the account holders and commenters in this study, with msm being understood to be involved in covering up the true number of vaccine deaths:

> MSM are in on this whole thing, only report on what the elites tell them to. Clearly you are not doing any research but listening to msm […] This is a completely experimental ‘vaccine’ […] it was rushed through and given provisional approval with zero long-term safety testing.

Some comments, however, didn’t even try to conceal their support for dangerous conspiracy theories, identifying COVID-19 as a scam to control the world’s population and referencing the ‘New World Order’ (NWO) conspiracy, or what has been referred to as the ‘Great Reset’:

> There are so many Doctors and scientists coming out saying the vaccines are no good but MSM isnt reporting it and fb tries to limit it. Dont you wonder why when the elites of the world now openly talk about the New World order/the great reset and the need to reduce the population are many if the same people who benefit financially from the vaccines? Gates, Fauci etc.

'Dangerous organisation' labelling and community response. The second commenting thread of interest (responding to Figure 6) contained community reactions to Meta's 'Dangerous organisations' policy update, introduced in October 2021, where accounts that regularly shared QAnon-related materials were labelled 'extremist' and users redirected to a counselling service. The labelled content in the post shown in Figure 6 is re-shared by the official account holder with a caption that calls out Facebook for being hypocritical, given that it is a company that 'deals in lies and manipulation'.

Reactions to this post are of a much more emotionally charged tone and show higher levels of engagement with the post receiving 598 comments. More than half of the comments took a straightforward anti-censorship stance against Facebook and big tech. But 28% responded by sharing misinformation and/or conspiracy theory in their comments, including referring to Facebook and big tech as co-conspirators with governments and other global 'elites' in a broader agenda to silence 'the truth'.

As above, several comments referenced 'experimental' vaccines, sometimes referring to them as gene therapy, while MSM/msm and 'the agenda' were also frequently used. One comment read: 'He only speaks the truth about what's real we need more people like this man, we are people humans with rights not Guinea pigs for experimental vaccines 👍' and another claimed 'FAKEBOOK Is a total joke and they too will be held responsible for not allowing the truth to be exposed. MSM are the liars and been paid to follow a very false narrative 😡'.

A small selection of commenters called on community members to leave Facebook and join less moderated platforms, including Telegram and the now defunct Parler.

But, while commenters may direct account holders to less moderated platforms to avoid content moderation, nonetheless, Facebook is an important platform to grow community and followers, so account holders often tried to game the algorithm to increase followers and engagement while also driving users to less moderated platforms for more open discussion.

Shadow banning and bridging to 'dark platforms'. Ethnographic observation of the Facebook community where our last commenting thread was sourced (Figure 7) showed that livestreams were the main content shared with followers. These livestreams were scheduled and promoted through the main page and included frequent interviews with 'experts' or 'insiders' discussing alternative health remedies and negative health consequences from immunisation.

Livestreams themselves could be considered an effort to evade content moderation, as discussion cannot be captured by automated filters for typed text, and the length of the videos (ranging from 30 min to over an hour) requires a greater investment of time to review (although the previous example shows that Meta often labelled this content). The choice to mainly use video content paid off for the page in the early days of the pandemic with engagement being strong. As of late November 2021, the page's posts, which also included paid advertisements for anti-vaccination and anti-lockdown protests, were receiving between 200 and 300 comments per post, with some exceeding 800 comments. These same posts were also receiving likes that ranged between 500–2000 likes. While the account holder may have anticipated that engagement with the page would decrease following the gradual easing of lockdowns in Australia after mid-2021, page engagement actually seemed to increase (most likely coinciding with government-enforced vaccine mandates in some industries). But it was also during this time period that the page owner and star of regular livestreams began to complain that the page was being shadow banned by Facebook, with views of livestreams sharply dropping at the time Figure 7 was posted. Suspecting she had been shadow banned, she calls on her followers to recommend a 'good, censorship free, livestreaming platform'.

Not surprisingly, the comments replying to this post suggest a range of 'dark platforms' like those mentioned by Rogers (2020) and Zeng and Schafer (2021). Unlike IMOP, the comments were fewer in number (N = 95), perhaps because of shadow banning. This was confirmed by one user who commented ‘you are becoming very hard to find here on facebook’, and another user saying ‘Instagram has blocked you. I cannot find you’. Among comments responding to the call, livestreaming sites such as Rumble were recommended, with the platform becoming home for an Australian far-right influencer de-platformed from the mainstream social media platforms, Avi Yemeni.

Similar recommendations were made of Twitch, a livestreaming site popular with gamers but which has since attracted political influencers, including alt-right influencers, owing to the ease with which users can get paid for content through the site or link to external websites and payment services (Russonello, 2021): ‘I know so many people who get censored on so many apps especially Facebook and twitch seems to work for them’. Although one user suggested that Twitch had performed some recent ‘purges’ that might make it less appealing: ‘I didn’t even think of twitch!! But I wonder about censorship though, I think it’s already done some purges over this past year’.

But even despite claims of censorship and ‘shadow banning’, leading the account holder to consider alternative platforms, the page continued to register stable engagement in the timeframe and had a well-stocked shop on Facebook, through which they sold books on the ‘plandemic’, arguments against mask-wearing, and even t-shirts which identify the wearer as an anti-vaxxer. This demonstrates that Facebook’s design allowed the page owner to profit from users’ purchases of content through the store, including even more extreme content than what was able to be published on the page.

Conclusion

By analysing page performance of 18 Facebook accounts found to share COVID-19 misinformation and conspiracies against key content policy announcements, and through a quantitative and qualitative analysis of comments on overperforming pages, the paper asks whether Meta’s content moderation policies to ‘reduce’ rather than remove offending content have effectively slowed the spread of vaccine misinformation and conspiracy theories on Facebook (RQ1). It also asks whether the communities impacted behave in the way Meta’s modelling and assumptions (underlying these techniques) predict? (RQ2), and, finally, we investigated what tactics these communities adopted to mitigate or evade these forms of platform governance, and with what success (RQ3)?

The findings addressing RQ1 revealed two trends. In the first stage of analysis, the performance scores of selected accounts showed that strategies to reduce distribution of content were partially effective in slowing the spread. Our analysis showed that median account performance clearly dropped after February 2020, when Meta announced they would be taking a more aggressive approach to moderating COVID-19 and vaccine misinformation, and it kept declining as Meta announced new measures. Nonetheless, the mean performance showed a different trend, with median and mean scores diverging around three policy announcements between March and October 2020: (a) the decision to add more labelling to COVID-19-related misinformation, (b) adding QAnon to the Dangerous Organisations list and reducing content showing support for QAnon conspiracy theories and (c) adding labels to content and accounts found to continually share QAnon-related content, indicating they were ‘extremist’. After these announcements, the mean performance score began to trend upwards based on a few accounts which grew their followers and performance despite these more targeted measures. This trend continued after 26 February 2021, when Facebook vowed to remove rather than reduce false vaccine claims. Our findings showed that the two accounts not only maintained their performance score but overperformed during this time period and increased their followers (see Figure 3 and 4), while continuing to share vaccine misinformation.

Though it is impossible to identify a strong causal explanation for why this divergence occurred without access to Meta's internal data and decision-making, a content and thematic analysis of comments parsed from three posts across the two overperforming accounts provide some insight into how the user community responded to shadow banning and labelling by Facebook, which, we argue, helps to explain how these accounts maintained strong engagement with COVID-19 conspiracy theory and misinformation, on the platform and beyond it.

Addressing RQ3, across all three posts, there was evidence of the community more actively seeking workarounds to continue to access content affected by labelling or demotion techniques, rather than relying on Facebook's algorithm. Of the tactics adopted by followers to avoid algorithmic detection and moderation, a common tactic was 'conspiracy seeding' or sharing misinformation and conspiracy-related content in the comments, using URL archiving tools to 'hide' this content, or linking to content on less moderated platforms, such as Bitchute. There was also evidence of the use of social steganography or encoded language, typos and emojis to replace terms that may be subject to algorithmic detection and intervention, a tactic which the scholarship tells us has been used by left and right leaning communities to bypass 'shadow bans'.

Again, while the findings in this study do not show clear causality between tactics employed by users to evade or mitigate content moderation techniques and the page's overperformance on key metrics such as followers and shares of posts, it does provide some insight into possible reasons why performance does not dim on some accounts. It also presents some interesting counter-arguments to Meta's underlying assumptions regarding labelling and shadow banning (RQ2). With regard to labelling, the underlying assumption – that by informing users that content contains misinformation or dangerous content they will be persuaded not to share this content – does not hold for community's invested in this contents' promotion. In this study and others, where communities are invested in particular narratives, the findings show that they will go to great lengths to engage with this content and share it. In relation to shadow banning or suppression, the assumption that this approach is better than removing content because it avoids attention being brought to removed content, which encourages users to find workarounds, this logic also didn't hold up in this study. Instead the decision to shadow ban is often noticeable to users in a community. Rather than allowing such content to be suppressed, users in this study instead mobilised to find work arounds – in essence, they came together as a community to game the algorithm rather than allowing the algorithm to determine what content they accessed and how.

This demonstrates that while Meta's content policies during the COVID-19 pandemic were partially effective in reducing the spread of COVID and vaccine misinformation, such approaches are not effective for communities that are strongly invested in spreading COVID and vaccine misinformation, and more research targeting susceptible communities and tactics of resistance to content moderation policies is required if the platforms are serious about addressing some of the gaps identified in this study.

Acknowledgements

Previous versions of this article were presented at the ADM+S Misinformation Workshop (2022). We acknowledge the Ethics Committee of the University of Technology Sydney for their approval of our research application (UTS HREC REF NO. ETH19-3877).

Data availability statement

The data underlying this article, the R code to replicate the analysis, and the online supplemental material are available in the Harvard Dataverse at (Johns et al., 2022).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author disclosed receipt of the following financial support for the research, authorship and/or publication of this article: The authors declare that the work produced in this paper was enabled by funding received from the Facebook (Meta) Content Policy Award 2019.

Supplemental material

Supplemental material for this article is available online.

ORCID iD

Amelia Johns https://orcid.org/0000-0002-3946-7869

Notes

1. More details on sampling and on the quantitative data analysis are available in the Online appendix and in the replication materials.
2. The filtering regex was ‘covid|vaccin|pandemic|plandemic|lockdown|mandatory|coronavirus|virus|jab|mask|CV-19|informed choice|bill gates|vaxeen’.
3. See https://help.crowdtangle.com/en/articles/2013937-how-do-you-calculate-overperforming-scores for details.
4. ID2020 is a public-private consortium established to address the United Nations 2030 Sustainable Development Goal of providing digital identity for all people, including undocumented refugees, to access healthcare and other needs.

References

Are C (2021) The shadowban cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies* 22(8): 2002–2019.

Baker SA (2022) Alt. health influencers: how wellness culture and web culture have been weaponised to promote conspiracy theories and far-right extremism during the COVID-19 pandemic. *European Journal of Cultural Studies* 25(1): 3–24.

Blei DM, Ng AY and Jordan MI (2003) Latent dirichlet allocation. *The Journal of Machine Learning Research* 3: 993–1022.

Clegg N (2020) Combating COVID-19 misinformation across our apps. Available at: https://about.fb.com/news/2020/03/combating-covid-19-misinformation/ (accessed 1 March 2024).

de Keulenaar E, Burton AG and Kisjes I (2021) Deplatforming, demotion and folk theories of big tech persecution. *Fronteiras—Estudos Midiáticos* 23(2): 118–139.

Díaz Á and Hecht-Felella L (2021) Double standards in social media content moderation. Report, Brennan Center for Justice. New York, USA. Available at: https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation (accessed 1 March 2024).

Duffy BE and Meisner C (2023) Platform governance at the margins: social media creators’ experiences with algorithmic (in)visibility. *Media, Culture & Society* 45(2): 285–304.

Gerrard Y (2018) Beyond the hashtag: circumventing content moderation on social media. *New Media & Society* 20(12): 4492–4511.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, London: Yale University Press.

Gillespie T (2022) Do not recommend? Reduction as a form of content moderation. *Social Media + Society* 8(3): 20563051221117.

Goldman E (2021) Content moderation remedies, 28 Michigan of Technology Law Review. 1. Available at: https://repository.law.umich.edu/mtlr/vol28/iss1/2 (accessed 1 March 2024).

Grondahl T, Pajola L, Juuti M, et al. (2018) Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (AISec) Available at: https://arxiv.org/abs/1808.09115 (accessed 1 March 2024).

Guo C, Zheng N and Guo C(John) (2023) Seeing is not believing: a nuanced view of misinformation warning efficacy on video-sharing social media platforms. *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW2): 1–35.

Johns A, Bailo F, Booth E, et al. (2022) Replication data for: You've been shadowbanned: Has Facebook's strategy to suppress rather than remove COVID-19 vaccine misinformation actually slowed the spread?, Harvard Dataverse, V2, DOI: 10.7910/DVN/A9RNBS.

Huddleston Jnr T (2020) Bill Gates is top target for coronavirus conspiracy theories. *CNBC*, April 17. Available at: https://www.cnbc.com/2020/04/17/bill-gates-is-top-target-for-coronavirus-conspiracy-theories-report.html (accessed 1 March 2024).

Karppi T and Nieborg DB (2021) Facebook confessions: corporate abdication and Silicon Valley dystopianism. *New Media & Society* 23(9): 2634–2649.

Kong Q, Booth E, Bailo F, et al. (2022) Slipping to the extreme: A mixed method to explain how extreme opinions infiltrate online discussions. *Proceedings of the International AAAI Conference on Web and Social Media* 16(1): 524–535.

Martel C and Rand DG (2023) Misinformation warning labels are widely effective: a review of warning effects and their moderating features. *Current Opinion in Psychology* 54: 101710.

Marwick AE and Boyd D (2014) Networked privacy: how teenagers negotiate context in social media. *New Media & Society* 16(7): 1051–1067.

Matamoros-Fernández A (2017) Platformed racism: the mediation and circulation of an Australian race-based controversy on twitter, Facebook and YouTube. *Information, Communication & Society* 20(6): 930–946.

Meta (2022) Meta integrity timeline. Available at: https://transparency.fb.com/en-gb/policies/improving/timeline/ (accessed 1 March 2024).

Moran RE, Koltai K, Grasso I, et al. (2021) ViralityProject.Org. Available at: https://www.viralityproject.org/rapid-response/content-moderation-avoidance-strategies-used-to-promote-vaccine-hesitant-content (accessed 1 March 2024).

Myers West S (2018) Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. *New Media & Society* 20(11): 4366–4383.

Nicholas G (2022) *Shedding light on shadowbanning*. Washington D.C: Report, Center for Democracy and Technology. https://cdt.org/insights/sheddinglight-on-shadowbanning/.

Pennycook G, Cannon TD and Rand DG (2018) Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General* 147(12): 1865–1880.

Rogers R (2020) Deplatforming: following extreme internet celebrities to telegram and alternative social media. *European Journal of Communication* 35(3): 213–229.

Russonello G (2021) Twitch, Where Far-Right Influencers Feel at Home: On Politics. *New York Times (Online)*.

Savolainen L (2022) The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society* 44(6): 1091–1109.

Suzor NP, Myers-West S, Quodling A, et al. (2019) What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 13: 18.

Tiidenberg K and van der Nagel E (2020) *Sex and Social Media*. Melbourne: Emerald Publishing.

Zeng J and Schäfer MS (2021) Conceptualizing "dark platforms". COVID-19-related conspiracy theories on 8kun and gab. *Digital Journalism* 9(9): 1321–1343.

Zuckerberg M (2018) Blueprint for Content Governance. Available at: https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/?hc_location=ufi (accessed 1 March 2024).