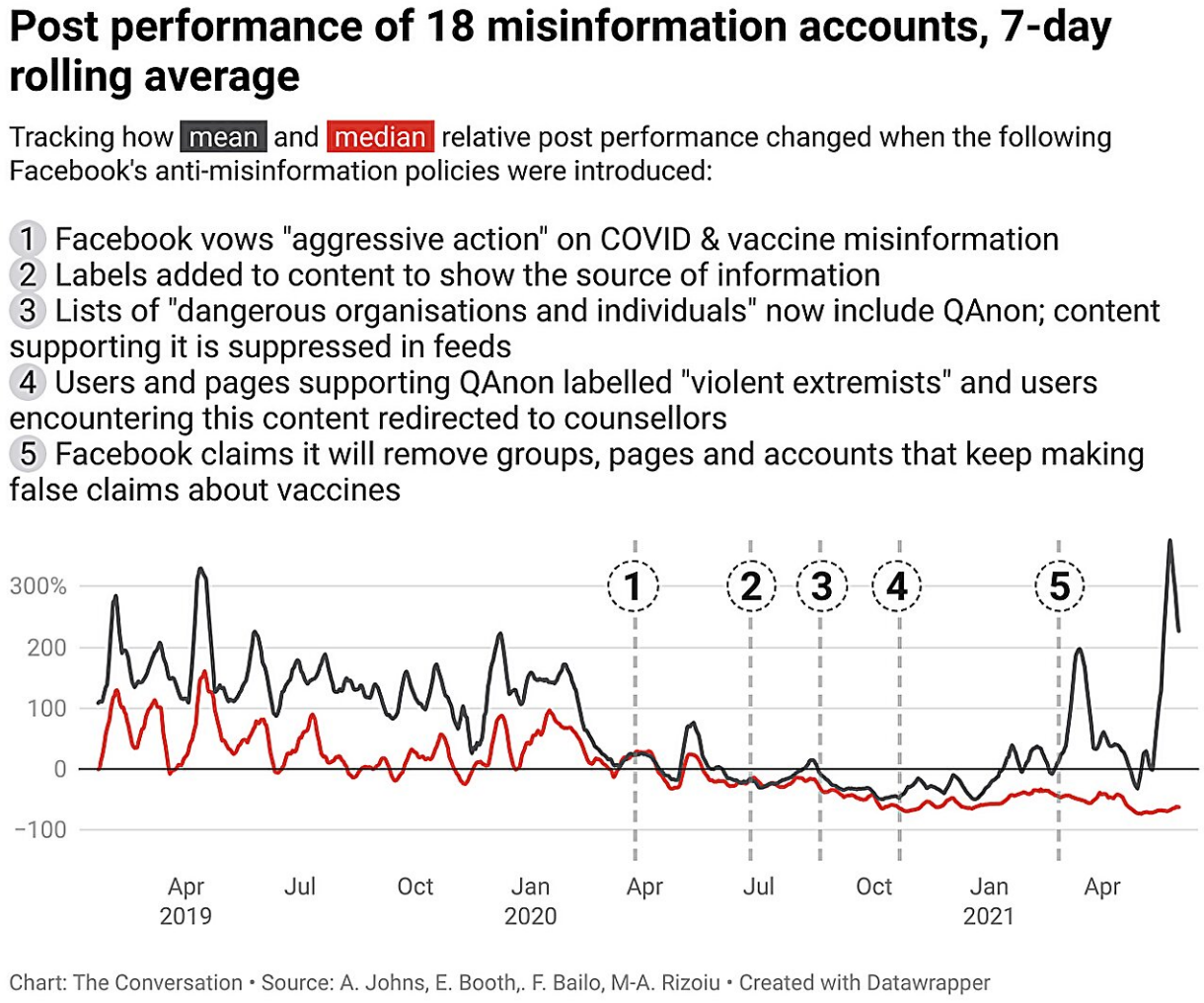
Conspiracy theorist tactics show it's too easy to get around Facebook's content policies

March 23 2024, by Amelia Johns, Emily Booth, Francesco Bailo and Marian-Andrei Rizoiu



Credit: The Conversation

During the COVID pandemic, social media platforms were swarmed by far-right and anti-vaccination communities that spread dangerous conspiracy theories.

These included the false claims that vaccines are a form of population control, and that the virus was a "deep state" plot. Governments and the World Health Organization redirected precious resources from vaccination campaigns to debunk these falsehoods.

As the tide of misinformation grew, platforms were accused of not doing enough to stop the spread. To address these concerns, Meta, the parent company of Facebook, made several policy announcements in 2020–21. However, it hesitated to remove "borderline" content, or content that didn't cause direct physical harm, save for one policy change in February 2021 that expanded the content removal lists.

To stem the tide, Meta continued to rely more heavily on algorithmic moderation techniques to reduce the visibility of misinformation in users' feeds, search and recommendations—known as shadowbanning. They also used fact-checkers to label misinformation.

While shadowbanning is widely seen as a concerningly opaque technique, our new research, published in the journal Media International Australia, instead asks: was it effective?

What did we investigate?

We used two measures to answer this question. First, after identifying 18 Australian far-right and anti-vaccination accounts that consistently shared misinformation between January 2019 and July 2021, we analyzed the performance of these accounts using key metrics.

Second, we mapped this performance against five content moderation

policy announcements for Meta's flagship platform, Facebook.

The findings revealed two divergent trends. After March 2020 the overall performance of the accounts—that is, their median performance—suffered a decline. And yet their mean performance shows increasing levels after October 2020.

This is because, while the majority of the monitored accounts underperformed, a few accounts overperformed instead, and strongly so. In fact, they continued to overperform and attract new followers even after the alleged policy change in February 2021.

Shadowbanning as a badge of pride

To examine why, we scraped and thematically analyzed comments and user reactions from posts on these accounts. We found users had a high motivation to stay engaged with problematic content. Labeling and shadowbanning were viewed as motivating challenges.

Specifically, users frequently used "social steganography"—using deliberate typos or code words for key terms—to evade algorithmic detection. We also saw conspiracy "seeding" where users add links to archiving sites or less moderated sites in comments to re-distribute content Facebook labeled as misinformation, and to avoid detection.

In one example, a user added a link to a BitChute video with keywords that dog-whistled support for QAnon style conspiracies. As terms such as "vaccine" were believed to trigger algorithmic detection, emoji or other code names were used in their place:

"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."

While many conspiracy theories were targeted at government and public health authorities, platform suppression of content fueled further conspiracies regarding big tech and their complicity with "Big Pharma" and governments.

This was evident in the use of keywords such as MSM ("mainstream media") to reference QAnon style agendas:

"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.'"

Another comment thread showed reactions to Meta's dangerous organizations policy update, where accounts that regularly shared QAnon-content were labeled "extremist". In the reactions, MSM and "the agenda" appeared frequently.

Some users recommended that sensitive content be moved to alternative platforms. We observed one anti-vaccination influencer complain that their page was being shadowbanned by Facebook, and calling on their followers to recommend a "good, censorship free, livestreaming platform".

The replies suggested moderation-lite sites such as Rumble. Similar recommendations were made for Twitch, a livestreaming site popular with gamers which has since attracted far-right political influencers.

As one user said,

"I know so many people who get censored on so many apps especially Facebook and Twitch seems to work for them."

How can content moderation fix the problem?

These tactics of coordination to detect shadowbans, resist labeling and fight the algorithm provide some insight into why engagement didn't dim on some of these "overperforming" accounts despite all the policies Meta put in place.

This shows that Meta's suppression techniques, while partially effective in containing the spread, do nothing to prevent those invested in sharing (and finding) misinformation from doing so.

Firmer policies on content removal and user banning would help address the problem. However, Meta's announcement last year suggests the company has little appetite for this. Any loosening of policy changes will all but ensure this misinformation playground will continue to thrive.

Provided by The Conversation

Citation: Conspiracy theorist tactics show it's too easy to get around Facebook's content policies (2024, March 23) retrieved 27 April 2024 from https://techxplore.com/news/2024-03-conspiracy-theorist-tactics-easy-facebook.html