First Draft

claire@firstdraftnews.com

Cite this as: *BMJ* 2021;372:n26

http://dx.doi.org/10.1136/bmj.n26

Published: 21 January 2021

ESSAY

Too little, too late: social media companies' failure to tackle vaccine misinformation poses a real threat

As the world looks to the new covid-19 vaccines with hope, there are major worries about how social media will affect uptake. **Claire Wardle** and **Eric Singerman** ask what the companies in charge should be doing to stem the misinformation tide

Claire Wardle, Eric Singerman

The major social media companies are facing wide criticism for failing to deal with vaccine misinformation on their platforms. In response, the likes of Facebook and Twitter and Google (which owns YouTube) have stated that they will take more action against false and misleading information about covid-19 vaccines.[1]

This is undeniably positive, but these policy updates will not cover many types of posts that have the potential to lead to vaccine hesitancy. Take a mother's recent post on a public Facebook group: "Prior to her 6 week vaccinations, my daughter was perfectly fine," but afterwards "she was having major seizures . . . has anyone else had this happen after their 6 week vaccinations?" Should this post be removed? I doubt many people would think so. But should it be labelled as potentially misleading? Should it have a link to a vaccine information centre? Should it be demoted so that fewer people see it in their newsfeeds?

These questions are only growing more urgent. Global immunisation levels against diseases such as measles remain generally high, but an increasing, loud minority of people around the world have become more reluctant to take vaccines and less willing to listen to long trusted scientific institutions.[2] It's hard to pin down the causal links between misinformation, vaccines, and trust—especially without access to the social media companies' proprietary data—but the effects of misinformation should not be dismissed, especially during a pandemic. Tackling misinformation presents several difficult challenges.

Disinformation and misinformation

Let's start with definitions. Disinformation and misinformation are not the same thing. When someone deliberately creates or shares false or misleading content, and they intend to cause harm, that's disinformation. When they do so unwittingly and don't intend harm, it's misinformation. They both include outright lies and imposter news outlets, but also more benign content like misleading headlines or even satire. A fake Biden campaign website that admitted it was a parody on its front page could count as misinformation, for example.[3] Note that much of this content is neither fake nor news.

Given how multifaceted misinformation and disinformation are, quantifying them is a tall order. Even if they were easy to define and measure, they are hard to explain. People spread disinformation for all kinds of reasons: financial gain, power, or just for fun. People spread misinformation to connect with their communities, to test out ideas, and to showcase their beliefs and identities online.

Compounding these problems is the fact that virtually all of this is legal. Most discussions about misinformation and disinformation, not to mention company policies, start with freedom of speech and the First Amendment.[4] Aside from terrorist content and imagery of child sexual abuse, this approach puts a premium on the marketplace of ideas, founded on the idea that counter speech, not censorship, is the best way to deal with falsehoods. It avoids penalising harmful speech until its harms are clear and imminent. And it struggles to realise that not all legal speech deserves the same "freedom of reach" on the internet.[5] This mindset helps explain the platforms' tepid responses to health misinformation.

Confusion in the age of covid-19

Before the covid-19 pandemic, social media companies had taken a hands-off approach, at least until 2016 when the Brexit referendum, along with elections in the Philippines and US, woke them up to political disinformation. And until recently they had done next to nothing to combat health misinformation. To experts, this oversight was especially worrying.[6]

This laidback approach changed in 2018 when a series of measles outbreaks in the US seemed to be fuelled by vaccine misinformation. This was certainly not the first time that misinformation potentially affected a public health crisis, but because this took place in America, home of Facebook, Google, Twitter and others, it got the companies' attention.[7] For the most part, their first steps were limited. Relying on counter speech, Facebook and Twitter added educational pop-ups for users searching for vaccine content. And focusing on content that might cause "real world harm," they changed their recommendation algorithms to suppress false statements about vaccines. Notably, targeted social media ads escaped scrutiny, and some companies went further than others—Pinterest limited the results of any search for "vaccines" to trusted official sources like the World Health Organization.[8] But social media continued to be a breeding ground for health misinformation[9] as the world entered a shocking pandemic.

It's only now, as pressure on the companies from governments, scientists, doctors, and the public hits breaking point, that they have changed their health misinformation policies all together. Facebook, Twitter, and YouTube all took a more assertive and expansive view of "harm."[10,11] Facebook, for example, targets "false claims about the safety, efficacy, ingredients, or side effects of the vaccines."[12] Previously these types of claims would have been flagged by factcheckers and demoted in people's newsfeeds. Now, they are being removed.

The new policies also target claims that are only misleading, as well as those designed to spread confusion, by adding labels and demoting them. Repeat offenders could have their accounts disabled, sometimes permanently.

Consensus and responsibility

Despite this stronger stance, Facebook, Google, and Twitter are still uncomfortable accepting responsibility. They are not, they claim, "arbiters of truth," merely middle men providing a platform to their users, the public. The companies fall back on directives from health organisations to determine what counts as false, misleading, or confusing, whether it's international bodies like WHO or national bodies like the US Centers for Disease Control and Prevention and the NHS.

The decision to rely on expert organisations makes sense in principle, but in practice matters aren't so simple. For one, scientific consensus struggles to keep pace with misinformation. Through the summer of 2020, health agencies flip flopped on guidance concerning masks and airborne transmission, while misinformation on these topics was allowed to fester. By May, it was becoming increasingly clear that health and political misinformation are hard to tease apart. Anti-quarantine and anti-mask groups on Facebook were using political arguments to evade the censure of expert agreement on science. Making matters worse, disinformation campaigns targeted long trusted scientific institutions, with the President of the United States, Donald Trump, playing an influential role in undermining them.[13]

Like scientific consensus, it's hard to find consensus on what counts as misinformation in the first place. Consider a recent article in the *Spectator*, with the headline "Landmark Danish study finds no significant effect for facemask wearers." To much fanfare, the article was labelled misinformation and removed from Facebook. As Kamran Abbasi, *The BMJ*'s executive editor, pointed out, the study in question didn't damn mask wearing so much as reach inconclusive results. Nor did it discuss, much less question, viral spread among mask wearers. Abbasi deemed the *Spectator* article symptomatic of a "disagreement among experts" that came down to interpretation of the study itself, and he criticised its take down.[14]

If misinformation were only a problem of falsehoods, this case would be simpler. But remember that misinformation includes well intentioned but misleading headlines. Whether or not the *Spectator* article offered a reasonable interpretation of the study, the question of whether its headline misled would remain. Focusing on fabricated content and demonstrably false claims misses a diverse range of content that is much harder to define and deal with.

Often, focusing on individual examples like the *Spectator* article—whether they should be removed, flagged, or labelled as false with a factcheck—turns into a pointless game of whack-a-mole. We will never be able to find, let alone tackle, all misinformation and disinformation. In the meantime, individual posts build larger, more problematic narratives. One photo depicting a soldier administering a vaccine, one tweet about a low grade fever after vaccination, or one blogpost claiming that vaccines cause autism would have limited impact individually. But together, they form a deeper story that erodes trust and pushes people to question the safety of vaccines.

Unfortunately, you can't just factcheck, label, or remove a narrative. They shape and sometimes dangerously warp how we make sense of the world.[15,16] No matter how companies tackle these issues, their policies will come up short. On the one hand, even the most clearly written policies have flaws. Bad actors spreading disinformation will find loopholes, like those who posted websites that had been removed, by using new, seemingly harmless, links from the Internet Archive.[17] And benign, well intentioned posts will get caught in the net, like the iconic "napalm girl" photos from the Vietnam war that were removed for violating a ban on nudity. On the other hand, more malleable policies are criticised for putting too much discretion in the hands of the social media companies. When they instituted a ban on "hate speech," for example, critics were left wondering what, exactly, had been banned. The history of content moderation is one of grappling with policies that are either clear but inflexible or adaptable but too vague and discretionary.

It's up to all of us

As we enter 2021 and covid-19 vaccines are at last rolled out, misinformation is undoubtedly going to pose a serious barrier to uptake. The social media companies are at least showing a willingness to intervene. But people wishing to undermine trust in the vaccine won't be using outright lies. Instead, they will be leading campaigns designed to undermine the institutions, companies, and people managing the rollout. They will be posting vaccine injury stories and providing first person videos detailing side effects that are difficult to factcheck. And, when well meaning local radio stations ask on Facebook, "Will you be getting the covid vaccine?" the comments will be flooded with conspiracy ideas and suggestions.

The question for the companies is whether they're prepared to tackle this, even if such posts don't break their current guidelines. This will sit uneasily with people who recognise that changing policies during a public health emergency could lead to a slippery slope that ends up curtailing freedom of speech. What's required is more innovative, agile responses that go beyond the simple questions of whether to simply remove, demote, or label. We need responses that acknowledge the complexity of defining misinformation, of relying on scientific consensus, and of acknowledging the power of narratives. Unfortunately, we don't have time to design them. So while we implore the social media companies to take a more active role, it is us, those who use social media, who need to start taking responsibility for our posting and sharing.

Let's hope that, by the next pandemic, these challenges will have been tackled in ways that don't leave us feeling as vulnerable to disinformation and misinformation as we do today.

Biographies

Claire Wardle is a leading expert on user generated content, verification, and misinformation. She is co-founder and US director of First Draft, the world's foremost non-profit organisation focused on research and practice to tackle misinformation and disinformation. In 2017 she co-wrote a report for the Council of Europe called*Information disorder: toward an interdisciplinary framework for research and policymaking*. Previously, she was a research fellow at the Shorenstein Center for Media, Politics, and Public Policy and research director at the Tow Center for Digital Journalism at Columbia Journalism School. She has worked with newsrooms and humanitarian organisations around the world, providing training and consultancy on digital transformation. She earned a PhD in

communications and an MA in political science from the University of Pennsylvania.

Eric Singerman is a research assistant and policy adviser at First Draft. He is currently a law student at the University of Chicago.

Provenance and peer review: Commissioned; not externally peer reviewed

Competing interests: We have read and understood BMJ policy on declaration of interests and have no relevant interests to declare.

1 Sanchez K. Facebook will remove COVID-19 vaccine misinformation. *The Verge* 3 Dec 2020. https://www.theverge.com/2020/12/3/22150425/facebook-covid-19-vaccine-coronavirus-misinformation-ban.

2 Smith R, Cubbon S, Wardle C. Under the surface: Covid-19 vaccine narratives, misinformation and data deficits on social media. *First Draft* 2020. https://firstdraftnews.org/wp-content/uploads/2020/11/FirstDraft_Underthesurface_Fullreport_Final.pdf?x27751.

3 Wardle C. Understanding Information Disorder. *First Draft* Oct 2019. https://firstdraftnews.org/wp-content/uploads/2019/10/Information_Disorder_Digital_AW.pdf?x76701.

4 Klonick K. The new governors: the people, rules, and processes governing online speech. *Harv Law Rev* 2018;131.https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/.

5 DiResta R. Free speech is not the same as free reach. *Wired* 30 Aug 2018. https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/.

6 Larson HJ. The biggest pandemic risk? Viral misinformation. *Nature* 2018;562:309. doi: 10.1038/d41586-018-07034-4 pmid: 30327527

7 Diresta R, Wardle W. Online misinformation about vaccines. *The Sabin-Aspen Vaccine Science & Policy Group* May 2020. https://www.sabin.org/sites/sabin.org/files/sabin-aspen-report-2020_meeting_the_challenge_of_vaccine_hesitancy.pdf.

8 Telford T. Pinterest is blocking search results about vaccines to protect users from misinformation. *Washington Post* 21 Feb 2019. https://www.washingtonpost.com/business/2019/02/21/pinterest-is-blocking-all-vaccine-related-searches-all-or-nothing-approach-policing-health-misinformation/.

9 Cook J. Instagram's search results for vaccines are a public health nightmare. *Huff Post* 2 Feb 2020. https://www.huffpost.com/entry/instagram-promoting-anti-vax-anti-vaccine_n_5e347c50c5b69a19a4aede0c.

10 Coronavirus: staying safe and informed on Twitter. *Twitter* 3 Apr 2020. https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#protecting.

11 COVID-19 Medical Misinformation Policy. *YouTube*. 2020. https://support.google.com/youtube/answer/9891785.

12 Jin K. Keeping people safe and informed about the coronavirus. *Facebook* 11 Dec 2020. https://about.fb.com/news/2020/12/coronavirus/#misinformation-update.

13 Townes D. "Do no harm"—assessing the impact of prioritising US political disinformation over health misinformation in 2020. *First Draft* 7 Dec 2020. https://firstdraftnews.org/latest/do-no-harm/.

14 Abbasi K. The curious case of the Danish mask study. *BMJ* 2020;371:m4586.

15 Wardle C. The drip, drip, drip of misinformation on covid-19 vaccine. *Boston Globe* 12 Nov 2020. https://www.bostonglobe.com/2020/11/12/opinion/drip-drip-drip-misinformation-covid-19-vaccine/.

16 Smith R, Cubbon S, Wardle C. Under the surface: Covid-19 vaccine narratives, misinformation and data deficits on social media. *First Draft*. 2020. https://firstdraftnews.org/wp-content/uploads/2020/11/FirstDraft_Underthesurface_Fullreport_Final.pdf?x27751.

17 Donovan J. Covid hoaxes are using a loophole to stay alive—even after content is deleted. *MIT Technology Review*. 30 Apr 2020. https://www.technologyreview.com/2020/04/30/1000881/covid-hoaxes-zombie-content-wayback-machine-disinformation/.