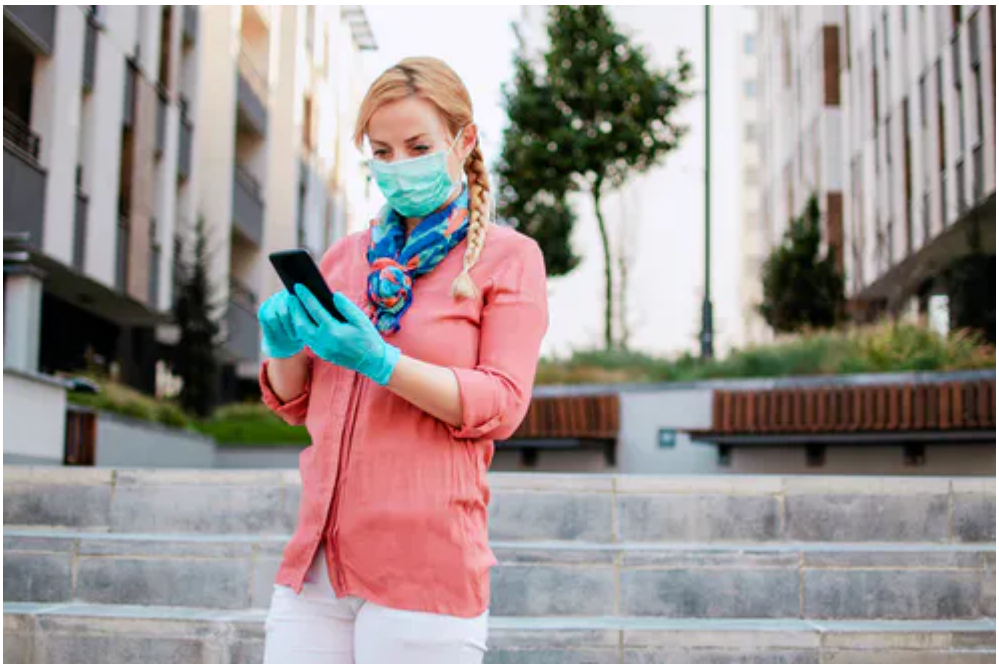
THE CONVERSATION
Academic rigour, journalistic flair



Pearl PhotoPix/Shutterstock

Misinformation: tech companies are removing 'harmful' coronavirus content – but who decides what that means?

August 27, 2020 4.29pm BST

The "infodemic" of misinformation about coronavirus has made it difficult to distinguish accurate information from false and misleading advice. The major technology companies have responded to this challenge by taking the unprecedented move of working together to combat misinformation about COVID-19.

Part of this initiative involves promoting content from government healthcare agencies and other authoritative sources, and introducing measures to identify and remove content that could cause harm. For example, Twitter has broadened its definition of harm to address content that contradicts guidance from authoritative sources of public health information.

Facebook has hired extra fact-checking services to remove misinformation that could lead to imminent physical harm. YouTube has published a COVID-19 Medical Misinformation Policy that disallows "content about COVID-19 that poses a serious risk of egregious harm".

The problem with this approach is that there is no common understanding of what constitutes harm. The different ways these companies define harm can produce very different results, which undermines public trust in the capacity for tech firms to

Authors

Stephanie Alice Baker
Senior Lecturer in Sociology, City, University of London

Matthew Wade
Lecturer in Social Inquiry, La Trobe University

Michael James Walsh
Associate Professor, University of Canberra

moderate health information. As we argue in a recent research paper, to address this problem these companies need to be more consistent in how they define harm and more transparent in how they respond to it.

Science is subject to change

A key problem with evaluating health misinformation during the pandemic has been the novelty of the virus. There's still much we don't know about COVID-19, and much of what we think we know is likely to change based on emerging findings and new discoveries. This has a direct impact on what content is considered harmful.

The pressure for scientists to produce and share their findings during the pandemic can also undermine the quality of scientific research. Pre-print servers allow scientists to rapidly publish research before it is reviewed. High-quality randomised controlled trials take time. Several articles in peer-reviewed journals have been retracted due to unreliable data sources.

Even the World Health Organization (WHO) has changed its position on the transmission and prevention of the disease. For example, it didn't begin recommending that healthy people wear face masks in public until June 5, "based on new scientific findings".

The World Health Organization has updated its advice as new evidence has emerged. FABRICE COFFRINI/EPA

Yet the major social media companies have pledged to remove claims that contradict guidance from the WHO. As a result, they could remove content that later turns out to be accurate.

This highlights the limits of basing harm policies on a single authoritative source. Change is intrinsic to the scientific method. Even authoritative advice is subject to debate, modification and revision.

Harm is political

Assessing harm in this way also fails to account for inconsistencies in public health messaging in different countries. For example, Sweden and New Zealand's initial responses to COVID-19 were diametrically opposed, the former based on "herd immunity" and the latter aiming to eliminate the virus. Yet both were based on authoritative, scientific advice. Even within countries, public health policies differ at the state and national level and there is disagreement between scientific experts.

Exactly what is considered harmful can become politicised, as debates over the use of malaria drug hydroxychloroquine and ibuprofen as potential treatments for COVID-19 exemplify. What's more, there are some questions that science cannot solely answer. For example, whether to prioritise public health or the economy. These are ethical considerations that remain highly contested.

Moderating online content inevitably involves arbitrating between competing interests and values. To respond to the speed and scale of user-generated content, social media moderation mostly relies on computer algorithms. Users are also able to flag or report potentially harmful content.

Despite being designed to reduce harm, these systems can be gamed by savvy users to generate publicity and distrust. This is particularly the case with disinformation campaigns, which seek to provoke fear, uncertainty and doubt.

Users can take advantage of the nuanced language around disease prevention and treatments. For example, personal anecdotes about "immune-boosting" diets and supplements can be misleading but difficult to verify. As a result, these claims don't always fall under the definition of harm.

Similarly, the use of humour and taking content out of context ("the weaponisation of context") are strategies commonly used to bypass content moderation. Internet memes, images and questions have also played a crucial role in generating distrust of mainstream science and politics during the pandemic and helped fuel conspiracy theories.

Transparency and trust

The vagueness and inconsistency of technology companies' content moderation mean that some content and user accounts are demoted or removed while other arguably harmful content remains online. The "transparency reports" published by Twitter and Facebook only contain general statistics about country requests for content removal and little detail of what is removed and why.

This lack of transparency means these companies can't be adequately held to account for the problems with their attempts to tackle misinformation, and the situation is unlikely to improve. For this reason, we believe tech companies should be required to publish details of their moderation algorithms and a record of the health misinformation removed. This would increase accountability and enable public debate where content or accounts appear to have been removed unfairly.

In addition, these companies should highlight claims that might not be overtly harmful but are potentially misleading or at odds with official advice. This kind of labelling would provide users with credible information with which to interpret these claims without suppressing debate.

Through greater consistency and transparency in their moderation, technology companies will provide more reliable content and increase public trust – something that has never been more important.

Social media **Coronavirus** **Misinformation** **COVID-19**