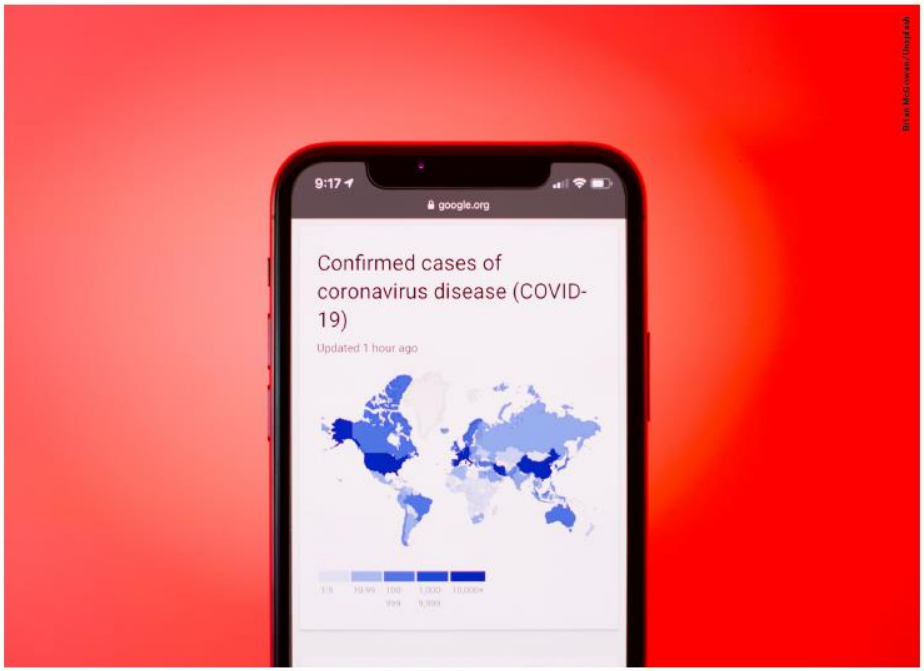

Brian McGowan/Unsplash

SOCIAL MEDIA

Deploying AI to tackle misinformation online

The rise of fake news and hate speech online has required tech giants to use AI to moderate content. But with the rapid spread of misinformation surrounding the coronavirus pandemic, are we at a stage where we can fully trust tech to moderate content?

Marianne Eloise

As the novel coronavirus (COVID-19) pandemic has progressed, information surrounding it changes by the minute. As quickly as new advice on keeping ourselves safe is dispersed online, it's often dispelled. Think, for example, of how many times you've heard in recent weeks that wearing face masks is either useless or necessary to reduce viral load?

With the uncertain nature of this pandemic, nobody knows exactly what we're supposed to do, which means even well-meaning people are distributing misinformation in an effort to promote safety.

As much of the information surrounding COVID-19 is being circulated on the internet, it's partly down to the platforms themselves to crack down on false information. At the offices of tech giants, in particular Facebook, content moderation is an extremely high-pressure job.

Moderators have strict rules of entry, working in secure rooms that they cannot even bring their own phones into. Moderation is highly sensitive and often traumatic, and Facebook outsources the work to contractors who are not allowed to work from home due to "safety, privacy and legal reasons", according to a Facebook spokesperson quoted recently in *The Intercept*.

While staff have now been sent home, there remains the important question of how content will be moderated at a critical time when misinformation could be a matter of life or death. Although employees are still able to work on helping to train machine-learning systems from home, without human moderators, tech giants such as Facebook and YouTube will be leaning heavily on AI content moderation. This is a controversial move that has its drawbacks as AI often doesn't have the capabilities or nuance to identify misinformation.

Professor Andy Pardoe, AI expert and author, believes the pandemic has brought into focus the ways that the future of work is already with us. "We need a flexible and robust digital workforce able to adapt to dynamic changes in environment and business priorities," he says.

Pardoe adds that part of this journey towards a robust workforce will require the use of various tools. "From automation to artificial intelligence, these tools can provide many benefits and efficiencies if implemented in a considered and controlled way," he says.

> **We weren't even close to this capacity before the pandemic and nothing has changed. Humans are still essential**

Shannon Vallor, an expert in the ethics of data and AI, believes the use of AI will move beyond content moderation throughout the duration of the outbreak. She says: "There will be strong pressures to use AI to boost public health efforts such as contact tracing and outbreak detection, as well as to amplify the human capacity to serve people in isolation, whether that's through remote symptom and risk assessment at scale or linking people to delivery of vital goods and services they need."

However, the urgent and sudden need to automate areas of the workforce could prove to be problematic. "Rushing to implement AI tools in this time of crisis, without expert knowledge and advice, may only deliver limited value and functionality that may not fully solve the challenges faced with a disrupted and remote workforce," she says.

This is something experts are in near-total agreement on. Jean-Claude Goldenstein, founder and chief executive of CREOpoint, labels the spread of misinformation an "infodemic" and believes an over-reliance on AI is now leading to harmful disinformation spreading too fast about COVID-19 treatment and possible cures. "There are hard tradeoffs between the health of content moderators, consistent content moderation and privacy," says Goldenstein.

YouTube has conceded that implementing AI tools at this stage might require "suppressing the good" content as a payoff. This means users might see seemingly unfair or randomised blocks imposed on non-insidious content as it gets swept up with genuinely dangerous content that needs suppressing. It's better, in their opinion, to be too safe than sorry.

"Users and creators may see increased video removals, including some videos that may not violate policies," YouTube warns, which is something of a sore spot for creators as their AI content moderation tools have seen innocuous content by, say, LGBT creators unfairly suppressed in the past.

This confusion plays into the issue at hand. "As we have seen with using AI to tackle hate speech, where those calling out racism or homophobia are often flagged for hateful speech because some of the same language must be used to describe it, AI can't easily tell the difference between a post that is spreading a hoax theory about COVID-19 and a post thoughtfully discussing the same hoax to debunk it. So you will get a lot of false-positive flags," says Vallor. She believes human review is necessary to ensure platforms don't suppress helpful and corrective information.

The pandemic has forced tech giants into a position where they may have to roll out AI moderation technology more widely than it's ready to be used. But Vallor believes we aren't even close to this being a reasonable possibility.

"We weren't even close to this capacity before the pandemic and nothing has changed. Humans are still essential to understanding context, assessing risk, and appropriately balancing and making tradeoffs between competing values in content moderation," she says. "A combination of AI and human moderation is the only way to be even modestly effective in this battle."

Vallor believes that those spreading and creating misinformation need to be found and penalised, not just on a post-by-post basis but also personally, while removing incentives to spread misinformation. "As long as people can profit and face no lasting social consequences from wilfully generating and spreading dangerous falsehoods online, then we are likely to be fighting a losing battle," she says.

However, there have been positive moves, such as Google pulling Alex Jones's Infowars app after he used it to spread coronavirus falsehoods. "It's not a technology problem. It's a problem of public integrity and there isn't yet the public will to fix it. What is needed is for us to gain the collective social courage to recognise and confront the entire phenomenon, namely the wilful or reckless spreading of falsehoods about matters of vital concern," Vallor adds.

Although CREOpoint's Goldenstein believes we will see more misinformation spreading over the coming months, there are ways to combat it that do not put employees in danger. "Technology giants like Twitter, Facebook, Google are Microsoft are likely to get together again, at the intersection of AI and trust, with tech pioneers such as CREOpoint and others from academia, agencies, brands, fact checkers, experts and regulators, as well as traditional media," he says.

But will this change? Will AI ever be fully capable of moderating content on these platforms? "We will need human moderation for the foreseeable future," says Vallor, adding that barring a leap forward in the methodologies, progress will continue to be slow.

However, while the capabilities of AI in accurately moderating content remain limited, it's potentially the only way forward for tech giants if they are to keep their employees safe. ●

60%

global citizens agree that it is acceptable to temporarily cut off social media platforms in times of crisis to stop the spread of false information

54%

believe people are capable of separating fact from fiction, so would not support a social media ban to stop the flow of fake news

49%

say they do not trust social media companies to ensure the factuality of the content on their platforms

45%

say social media platforms are the best, most accurate sources of news and information

Ipsos 2019