Disinformation wars: The fight against fake news in the age of AI

New Scientist 12 September 2023 By Graham Lawton

Researchers and governments are finally battling back against the deluge of false information online, just as artificial intelligence threatens to supercharge the problem

IN OCTOBER 2021, Phil Howard, an internet researcher at the University of Oxford, was alerted to a preposterous story on social media. It alleged that the covid-19 pandemic was started by a shipment of Maine lobsters that arrived in Wuhan, China, days before the first outbreak. He and his colleagues spent months trying to track down the source and didn't get to the bottom of it – except that it probably originated in China, possibly through the state-owned TV channel CGTN.

"I felt my career had hit a new low," says Howard. "What was so ridiculous was the enormous effort that we needed to expose a ridiculous attempt to manipulate public opinion. I realised that I didn't want to do that work myself, so I decided to try and come up with an initiative that would do something about the problem in a systematic way."

Today, Howard is chair of a new organisation called the International Panel on the Information Environment, one of many initiatives pushing back against the pollution of the information ecosystem. Regulators, too, are finally lacing up their own boots after spending years sitting on their hands.

The stakes couldn't be higher, with the recent rise of generative artificial intelligence and its capacity to produce persuasive disinformation on an industrial scale. Many researchers are saying that the next two years are make or break in the information wars, as deep-pocketed bad actors escalate their disinformation campaigns, while the good guys fight back. Which side prevails will determine how the information environment – and everything it shapes, from people's beliefs about vaccines to the outcomes of elections – will operate for the foreseeable future.

Misinformation and its nefarious cousin disinformation, defined as misleading information that is seeded deliberately, have been around for thousands of years. But the advent of social media was a watershed in the sense that it put the tools of disinformation into the hands of the masses, with seismic consequences.

Why fake news spreads fast

Research shows that fake news spreads six times as fast as true news, says former Google employee Tristan Harris, who now runs the Center for Humane Technology in San Francisco. "Even though there's a very small number of extreme voices out there, social media takes that 5 per cent of the population and then stretches it out over the whole movie screen of humanity."

Why false information spreads so far and wide is well known. The algorithms that serve up content on platforms such as Facebook, Instagram and X (formerly Twitter) have two overriding goals: to capture and keep people's attention, and to motivate them to share content by rewarding them for doing so. The more provocative the content, the more successful that strategy. "People are getting social rewards for sharing very emotionally provoking information or information that's not the most accurate," says Gizem Ceylan at Yale University. Moreover, the act of sharing becomes habitual. "Over time, you become like a pigeon pecking at a button with the hope of getting food, but you don't realise it," she says.

Ceylan's research demonstrates how the problem manifests. In an experiment, she and two colleagues showed Facebook users true and fake headlines and asked them whether they would share each of them. The heaviest users of the site said they would share 37 per cent of the fake headlines and 43 per cent of the true ones. "They were completely insensitive to the truthfulness of the information," says Ceylan. Most people want an unpolluted information environment, she adds. The trouble is that the reward structure of social media thwarts the will of the people.

There are real world consequences, personified in many people's minds by the attack on the US Capitol building on 6 January 2021, seemingly inspired by conspiracy theory about electoral fraud. Disinformation has also been blamed for other societal phenomena, including political polarisation, climate and election denial, and anti-vaccination sentiment. But the evidence joining the dots is lacking. "What's hard is to know how many people change their minds or change their votes because of something said," says Howard.

Indeed, recent research suggests that social media isn't to blame for political polarisation, though the results are controversial. In early 2020, a team of researchers from Meta approached two social scientists, Talia Stroud at the University of Texas at Austin and Joshua Tucker at New York University, to propose a research project on how its social media platforms Facebook and Instagram influenced users in the period around the US presidential election. Meta offered the researchers unprecedented access to its data and freedom to choose the research questions and study designs. The result was the 2020 Facebook and Instagram Election Study, which published its first results in July this year.

Social media under scrutiny

The research showed that Meta's algorithms have a massive influence on the content people see, and that people with different ideologies see very different political content. All of which is unsurprising. It was in addressing the other questions on the agenda – primarily, whether tweaking the algorithms that determine what people see had any effects on their political positions – where it got interesting.

Recall that the election was on 3 November. Between 23 September and 23 December 2020, the academics made three major changes to Meta's algorithms, tested them on three separate groups of around 23,000 consenting users and measured the effects. The first change was to remove posts shared by users' friends, or "reshares". The second was to switch off personalised feeds, which prioritise content the user is expected to engage with more, and instead present content chronologically. The final change was to deprioritise content from politically like-minded sources.

Meta CEO Mark Zuckerberg prepares to testify to Congress about the role of Facebook in spreading misinformation, among other things, in 2018

Then came the big reveal: none of these changes had any influence on the users' political attitudes or voting intentions. None. "We now know changing the algorithm for even a few months isn't likely to change people's political attitudes," says Stroud.

Meta lapped up the results. In a lengthy press release, Nick Clegg, president of global affairs, said that the studies "add to a growing body of research showing there is little evidence that key features of Meta’s platforms alone cause harmful ‘affective’ polarization, or have meaningful effects on key political attitudes, beliefs or behaviors”. Academics say this is a rush to judgement. “I think their finding is that there are no short-term noticeable effects,” says Howard. “The long-term effects are still a bit of an unknown.”

It was a similarly frustrating lack of clarity regarding the impact of social media on political outcomes that partly convinced Howard that a systematic scientific effort was urgently needed to track, understand and counter disinformation. He teamed up with veteran activist Sheldon Himelfarb, who runs a non-profit organisation called PeaceTech Lab in Washington DC. In April 2021, they proposed the foundation of an international body on disinformation modelled on the Intergovernmental Panel on Climate Change. The International Panel on the Information Environment, or IPIE, was formally launched at the US National Academy of Sciences in Washington DC in May.

In its first piece of work, IPIE systematically reviewed hundreds of papers on strategies to tackle disinformation on social media. It found that of multiple plausible ways to do so, only two are proven to work: content labelling, where posts are tagged with fact-checks and disclosures about who paid for them; and corrective information, where somebody with authority debunks disinformation on social media in a separate, unlinked piece of content. “For now, these are the only two solutions we feel confident about,” says Howard.

How to combat misinformation

Moderation, where platforms take down content or suspend and block accounts, works somewhat, as does media literacy training for users. But everything else that has been tried – allowing users to report disinformation, inhibiting the sharing of dodgy content, telling users they have consumed misinformation, redirecting users to reliable sources and so on – are as-yet unproven.

The fundamental problem is that none of this can compel the platforms to act, says Howard. “Experience has shown us that we cannot leave it to the social media platforms or the other tech platforms alone,” says Melissa Fleming, the UN’s under-secretary-general for global communications. “We do feel like we are in an information war.”

That is where governments enter the fray, and they are finally taking action. Leading the charge is the European Union’s Digital Services Act. This legislation, which will be applied from 17 February 2024, will compel larger social media platforms to limit the spread of illegal content, be more transparent about their moderation decisions and allow users to challenge them when they believe their content has been unfairly removed or down-rated.

“Most importantly, it obliges [the platforms] to have risk assessments on how the design and functionality of their platforms influences elections, civic discourse and how illegal content is disseminated,” says Agne Kaarlep, a former policy officer for digital services and platforms at the European Commission and now head of policy at content-moderation company Tremau.

The Digital Services Act also compels social media companies to be more open with their data. That is potentially a big win for scientists, says Howard. “So many of the effects that we were uncertain about are going to take deep, deep data access [to identify]” he says.

Similar legislation is afoot in the UK, Australia, Brazil and Canada. The UN is also drawing up a code of conduct. But the US is conspicuously absent from the regulatory party. This is largely because of the US Constitution’s first amendment of 1791, which says that congress will make no law abridging the freedom of speech.

However, free speech may not be a trump card. A recent study led by Stephan Lewandowsky at the University of Bristol, UK, polled 2564 people in the US on their attitudes towards content moderation. “The overarching finding was that people are OK with content moderation,” he says.

“They do support the removal of misinformation that is demonstrably causing harm.”

Perhaps the biggest problem, however, is that all this is happening in the face of another looming threat: generative AI tools that can produce convincing text, pictures and video in response to human prompts. “When we think about the way in which disinformation is produced and distributed at the moment, generative AI is going to supercharge that,” says Stephen King, CEO of the foundation Luminate, which works to address digital threats to democracy. “What we will be looking at in a year’s time, two years’ time, is going to be of a scale and magnitude which we can’t really imagine. The problem is going to get worse before it gets better.”

Bad actors may use publicly available generative AI, such as ChatGPT, or even build their own propaganda algorithms trained on disinformation, says Howard (see “What ChatGPT knows about AI disinformation”, below). “You can now generate 10,000 different lies and figure out what goes viral,” says Frances Haugen, formerly of Facebook, who in 2021 disclosed documents detailing the company’s failure to curb misinformation, and is now at Beyond the Screen, a non-governmental organisation that campaigns for greater checks and balances on social media platforms. “We’re going to be playing a very, very different game.”

Recent research suggests it may be a losing one. “When you train generative AI on human-written propaganda, the algorithm’s propaganda is viewed as more persuasive, at least a little bit, because it’s more succinct, it’s shorter, it’s punchier,” says Sander van der Linden, a psychologist at the University of Cambridge.

In July, the disinformation-fighting tech company Logically released the results of a chilling experiment with generative AI. Its researchers asked three image-generating AIs – Midjourney, DALL-E 2 and Stable Diffusion – to produce fake images relating to hot-button political issues in the US, India and the UK. Sometimes AI’s can refuse to answer such queries because their developers have decided it steps on dicey territory. However, most of their prompts were accepted and some produced extremely plausible disinformation images, including those of electoral fraud in the US, protesters toppling a statue of Isaac Newton in Cambridge, UK, and Muslim women wearing saffron scarves in support of the Hindu nationalist Bharatiya Janata party in Delhi, India.

AI-driven disinformation

“Eighty-six per cent of the images that we tried to create did actually bring through some disinformation in really hyper-realistic images,” says Beth Lambert at Logically, and a former head of counter disinformation in the UK Department for Culture, Media and Sport. When humans were producing disinformation there used to be a trade-off between quality and quantity, she says. Not any more.

The speed at which the technology is moving is also creating concern. At the moment, generative AIs can produce text, images or video, but not in combination. The next development will be multimodal versions capable of generating all three at the same time on the same topic, says Stevie Bergman of the ethics research team at AI company Google DeepMind. “Being able to do multiple of these has particular implications for how to limit misinformation or disinformation, and what misinformation or disinformation even looks like,” she says. Multi-agent AI, where generative AIs talk to one another to solve problems, is also advancing rapidly, she says, with unknown consequences.

A more distant but plausible possibility is what Sam Stockwell at the Alan Turing Institute in London calls autonomous intelligent agents: generative AIs that can be given a task and that then go away and figure out how to do it without further prompts from a human. “You have the prospect of an AI agent running a disinformation strategy,” he says.

Through all the gloom, there are potential upsides. Fact-checkers already use AI to discover disinformation and generative AIs may make that task more efficient. “Without AI, we cannot do content moderation at scale in a timely fashion,” says Sarah Shirazyan, manager of content policy at Meta. “For example, AI can actively identify and detect harmful content, be it hate speech or mass production of fake accounts, and then take it down before thousands of users report it, or even see it.”

But as previous research shows, content moderation won’t be enough. “Fact-checking and debunking are not going to solve our problems,” says Ceylan.

The attack on the US Capitol building on 6 January 2021 followed a conspiracy theory of electoral fraud during the 2020 election

One potentially more effective intervention is to borrow techniques from cryptography, says Paul Romer at Boston College, who shared the 2018 Nobel prize for economics. “What we need to do is to address this fundamental piece of infrastructure of digital authenticity: how do I know who’s behind this? This is a solved problem in cryptography, there’s a system called digital signatures, but we just haven’t taken the trouble to implement it [on social media].”

In cryptography, digital signatures are used to verify the authenticity of the sender of a message or document, and the idea is that they could be adapted to verify that the producer of a social media post is human, says Romer. AI-generated content could also be forcibly digitally watermarked, making it easier for social media platforms to recognise and flag. According to Shirazyan, big companies including Meta, Google and OpenAI have agreed on voluntary commitments to do so.

There are other ways to mitigate information pollution. “A lot of people say the long-term solution is to have digital literacy courses in high schools to teach people to be more sophisticated technology users,” says Howard. “But it’s not always clear what a good digital literacy programme looks like. That will probably be one of the next things we work on at the IPIE.”

We do at least have good models. According to Jussi Toivanen at Finland’s National Cyber Security Centre, his country – which borders Russia and was recently the subject of a Kremlin-backed disinformation campaign alleging that Finnish social workers were taking Russian-born children into custody and selling them to gay couples in the US – has spiked the disinformation gun. “Media literacy has been part of our official school curriculum since the 1970s,” he says. “It starts in kindergarten, it continues to basic education, it continues in adult education.”

Finland is an exemplar, says van der Linden, who, for his part, has been experimenting with what he calls “inoculation”. That is, exposing people to weakened doses of some of the standard psychological manipulation techniques – denial, deflection, conspiracy theories, trolling, gaslighting, fearmongering, inciting outrage and presenting false dilemmas such as “you’re either with me or against me” – and then deconstructing them so people can build up cognitive resistance. A large body of evidence suggests that this inoculation approach works.

So the battle lines are drawn. On one side are the social media giants and the bad actors who flood their services with malign information, now armed with generative AI. On the other are the academics, legislators and the general public. How it plays out is hard to predict, but the outcome will affect us all, as nobody is immune to being played.

“The best-case scenario is that the research community can pull itself together in a coherent way to make reasonable demands about access to data,” says Howard. “The worst outcome is that the technology platforms will simply refuse to comply and regulators won’t have the political will to act.”

What ChatGPT knows about AI disinformation

The rise of chatbots powered by artificial intelligence threatens to supercharge the production of false information (see main story). We asked OpenAI's ChatGPT whether generative AI can produce disinformation. The results were, in their own way, instructive.

"Yes, generative AI can potentially produce disinformation. Generative AI models, like GPT-3, are designed to generate text based on patterns and examples present in the data they were trained on. They do not have an inherent understanding of truth or ethics and are agnostic to the factual accuracy of the information they generate.

"If the training data includes examples of disinformation or misleading information, the AI model may learn to generate similar content when prompted with certain cues. Furthermore, if a user intentionally feeds the model with false or misleading information, it can produce responses that propagate disinformation.

"OpenAI and other responsible AI developers are aware of this risk and encourage ethical use of AI to prevent the spread of disinformation. They may implement moderation systems, provide guidelines to users on ethical AI usage, and try to minimize the likelihood of models generating harmful content. However, it's essential to remember that the potential for generating disinformation exists, and it's the responsibility of users to use AI technology responsibly and consider the consequences of their actions."