Source: *The Straits Times,* pB3
Date: 15 January 2024

We have weapons against AI-powered deepfakes but fighting truth decay won't be easy

Legal and tech tools to correct, block and slow down false information are being sharpened by governments around the world, but they have their limits.

Simon Chesterman

Though Donald Trump would like credit for coining the term, "fake news" was a headline in The New York Times more than a century ago. The history of propaganda is far older. An early example was Egypt's Ramses the Great, who decorated temples with monuments to his tremendous victory in the 13th century BCE Battle of Kadesh. The outcome was, at best, a stalemate.

But the tools for generating, sharing and consuming dubious information are now very different in the age of AI. This is cause for concern because, over the course of 2024, countries with more than half the world's population will hold national elections.

Elections are foundational to the legitimacy of government. That in turn relies on trust in the process of voting and the determination of the results. And if voters do not accept the outcome, they may be unwilling to accept the authority of the winners.

TRUMPED UP

None of this will surprise those who have followed the Donald Trump show over the past few years. Clumsy election denial is one thing, however. Now imagine authentic-looking CCTV footage of corrupt activity and realistic news accounts of vote-rigging, with AI-empowered agents producing endless variations on these themes, and the prospect of a larger crisis increases.

Analysts are warning that the sophistication of generative AI is only going to increase. This has implications far beyond election outcomes.

Harms that we should worry about include the impact on vulnerable individuals, public institutions, and the bonds of trust that hold society together.

Though much attention focuses on deepfakes of famous people, we should also be concerned about the prospect of getting a phone call from a loved one, or even a video of them, that is faked. I may not be willing to help that Nigerian prince get access to his millions, but who among us would not assist a child or spouse in a panic?

In addition to specific deepfakes gaining traction, a perverse consequence of their proliferation is the "liar's dividend", where people dismiss genuine scandals or allegations of wrongdoing because the basis of truth or falsity has become so muddied and confused. That leads to the broadest type of harm, which is the decline of trust more generally.

If synthetic content ends up flooding the Internet, the consequence may not be that people believe the lies, but that they cease to believe anything at all.

Any uncomfortable or inconvenient information will be dismissed as "fake", while many credulously accept data that reinforces their own world view as "true". Generative AI chatbots like ChatGPT and Bard could exacerbate this if we switch from searching for information on the Internet, which yields multiple possible responses, to asking an intelligent agent that gives us a single answer. As it learns our preferences – and prejudices – it may serve to reinforce them.

Governments are alive to these concerns, motivated also by a measure of regret for failing to regulate social media over the past two decades. Dozens of countries – from the United States to China – are debating policies and legislation.

So what, if anything, should be done?

REGULATE THIS

Up to now, there has been excessive reliance on technical interventions – with limited success. I'm part of a team at the National University of Singapore looking at a broader approach to what we call "digital information resilience". This emphasises the role of consumer behaviour in understanding why people consume fake news and how it affects them, as well as the important role of technology.

My focus is on how regulation and policy can shape supply and demand.

Efforts to regulate any aspect of the digital information pipeline face challenges, particularly if it limits access to information through censorship.

Regulators across the globe are struggling to address perceived harms associated with generative AI while not unduly limiting innovation or driving it elsewhere.

Though there is wariness about unnecessary limits on freedom of speech, even in broadly libertarian jurisdictions like the United States one is not allowed to yell "Fire!" in a crowded theatre.

Generative AI has raised the question of whether the tools that generate content should themselves be regulated. We do not normally regulate private activity – a hateful lie written in my personal diary is not a crime; nor do we punish word processing software for the threats typed on it.

A notable exception is that many jurisdictions make it an offence to create or possess child pornography, including synthetic images in which no actual child was harmed, even if the images are not shared.

For the most part, however, the harm is in the impact the information has on other users and society. In addition to punishing those who intend harms, should the platforms that host and facilitate access also bear responsibility?

In the United States, this would require a review of Section 230 of the 1996 Communications Decency Act, which absolves Internet platforms of responsibility for the content posted on them.

Singapore adopted the Protection from Online Falsehoods and Manipulation Act (Pofma), which empowers ministers to make correction orders for false statements of fact if it is in the public interest to do so.

GOVERNMENTS VERSUS FAKE NEWS

Though Singapore was criticised when it adopted Pofma in 2019, governments around the world are considering similar legislation to deal with the problem of fake news.

Australia released a draft Bill in 2023 on Combating Misinformation and Disinformation that has been hotly debated – including its fair share of fake news. Around the same time, the European Union's Digital Services Act came into force, while Britain passed a new Online Safety Act.

All struggle with the problem of how to deal with "legal but harmful" content online.

Australia's Bill would have granted its media regulator more power to question platforms on their efforts to combat misinformation. Backlash against perceived threats to free speech led the government to postpone its introduction in Parliament until later in 2024, with promises to "improve the Bill".

Ofcom, the body tasked with enforcing the new British law, states that it is "not responsible for removing online content", but will help ensure that firms have effective systems in place to prevent harm.

Such gentle measures may be contrasted with China's more robust approach, where the "great firewall" is often characterised by over-inclusion. Some years ago, Winnie the Pooh was briefly blocked because of memes comparing him to President Xi Jinping; earlier efforts to limit discussion of the "Jasmine Revolution" unfolding across the Arab world in 2011 led to a real-world impact on online sales of jasmine tea.

Correcting or blocking content is not the only means of addressing the problem, of course. Limiting the speed with which false information can be transmitted is another option, analogous to the circuit breakers that protect stock exchanges from high-frequency trading algorithms sending prices spiralling.

In India in 2018, WhatsApp began limiting the ability to forward messages after lynch mobs killed several people following rumours circulated on the platform. A study based on real data in India, Brazil and Indonesia showed that such methods can delay the spread of information, but are not effective in blocking the propagation of disinformation campaigns in public groups.

Another platform-based approach is to be more transparent about the provenance of information. Several now promise to label content that is synthetic, though the ease of creation now makes this a challenging game of catch-up.

Tellingly, the US tech companies that agreed to voluntary watermarking in 2023 limited those commitments to images and video, echoed in the Biden Administration's October 2023 executive order. Synthetic text is nearly impossible to label consistently; as it becomes easier to generate multimedia, it is likely that images and video will go the same way.

In fact, as synthetic media becomes more common, it may be easier to label content that is human rather than AI.

Trusted organisations may also watermark images so that users can identify where a photo comes from. The problem here is that tracking such data requires effort and many users demonstrate little interest in spending the time to verify whether information is true or not.

Twitter (prior to its acquisition by Mr Elon Musk) introduced a "read before you retweet" prompt, which was intended to stop knee-jerk sharing of news based solely on the headline. It appeared to have a positive impact, but was not enough to stop the slide into toxicity post-Musk.

THE FOURTH ESTATE

The ideal, of course, is for users to take responsibility for what they consume and share. Those of us who grew up watching curated nightly news or scanning a physical newspaper may be mystified by a generation that learns about current events from social media feeds and the next video on TikTok.

Yet concerns about the information diet of the public are as old as democracy itself. Some months before the US Constitution was drafted in 1787, Thomas Jefferson pondered whether it would be better to have a government without newspapers or newspapers without a government. "I should not hesitate a moment to prefer the latter," he concluded, making clear that he meant that all citizens should receive those papers and be capable of reading them.

As voters around the world head to the polls in 2024, no government has solved the problem of fake news. But as you consider your own information diet, exercise common sense. Try to remain critical without becoming cynical.

And if you see something that seems too good to be true, it probably is.

• Simon Chesterman is vice-provost at the National University of Singapore and founding dean of NUS College, as well as senior director of AI governance at AI Singapore. Information about the Information Gyroscope project is available at https://ctic.nus.edu.sg/igyro