Source: *The Straits Times*, pB3
Date: 3 November 2026



A crocodile at Sungei Buloh Wetland Reserve in January. An incident in August concerning an AI-generated image of a crocodile at Pandan Reservoir illustrates the challenges faced by the authorities in determining what is real. ST PHOTO: LIM YAOHUI

Why policing truth is no longer enough

A new global survey shows that as fact-checking and AI labels reach their limits, governments are increasingly turning their attention to the systems that make harmful information spread.



Simon Chesterman

In August, someone sent a photograph apparently showing a crocodile in Pandan Reservoir. PUB ran the image through several AI-detection tools. The results were inconclusive.

Faced with the possibility that there really was a crocodile in the reservoir, the authorities understandably took precautions. Water activities were suspended and a search was conducted. The image was subsequently established to be AI-generated and a man was charged over the incident.

It was, in the scheme of things, a relatively minor falsehood. But it illustrates a much larger problem.

For years, the debate over misinformation has revolved around a deceptively simple question: Is it true?

That question is becoming harder to answer – and less useful as the organising principle for regulation.

Technology is supposed to help. In October, OpenAI announced the introduction of invisible watermarks into eligible ChatGPT text in the European Union, partly in response to the EU AI Act. Yet its own testing demonstrates the limits. In one experiment, replacing just one in 10 words with synonyms reduced detection of the watermark from around 92% to 66%. Replacing a quarter of them reduced detection to 17%.

More fundamentally, a watermark can tell you something about where content came from. It cannot tell you whether that content is true. OpenAI itself emphasises that detecting its watermark does not establish whether a passage is accurate, misleading or harmful.

Meanwhile, another traditional line of defence – fact-checking – is also being rethought. Meta ended its third-party fact-checking programme in the United States and is experimenting with Community Notes elsewhere. Its own Oversight Board has cautioned that crowdsourced notes perform a different function from professional fact-checking and may, in high-risk situations, be vulnerable to manipulation or arrive too late.

These are different examples of the same underlying difficulty. We are producing information faster than we can authenticate, label or correct it. As the volume of information grows, governments around the world are responding in very different ways. Whether those approaches reduce harm without creating new problems remains uncertain.

REGULATION IS BOOMING – AND CHANGING

To provide a better understanding of the challenges we face in this rapidly evolving environment, the National University of Singapore's Information Gyroscope, or iGyro, project is publishing Beyond Falsehoods 2026, the first edition of an annual global index and report on the regulation of misinformation and disinformation.

iGyro is an interdisciplinary research programme examining how misinformation and disinformation is created, disseminated and consumed. Beyond Falsehoods grows out of its regulatory and policy work: mapping what governments are doing and comparing their approaches to build a better evidence base for effective regulation.

The index maps more than 400 legal and policy entries. They range from Singapore's Protection from Online Falsehoods and Manipulation Act (POFMA) and China's criminal-law and AI-labelling rules to the EU's Digital Services Act and South Korea's laws against sexually explicit deepfakes.

What stands out is not simply how much regulation has increased, but how its character is changing. Annual regulatory activity rose from around six instruments a year on average between 2013 and 2017 to 46 in 2023, 67 in 2024 and 71 in 2025.

The purpose is not to rank countries by the number or toughness of their laws, but to make approaches comparable: what harms they target, who bears responsibility, what powers and safeguards they create, and what happens when the laws are used.

Singapore was an early mover. When POFMA was enacted in 2019, much international criticism focused on the danger of allowing governments to decide what was true and false.

That concern has not disappeared. Our study identifies recent examples in India and Mongolia where courts invalidated measures giving public authorities broad powers to determine the truth or falsity of online information.

But the debate has broadened. Around the world, the question is increasingly not whether governments should respond to harmful information, but what they should regulate, who should bear responsibility, and – hardest of all – what actually works.

THE LIMITS OF POLICING TRUTH

The traditional regulatory response targets the individual item of information: identify a false statement, correct it, remove it or punish the person responsible for spreading it.

That model still has a place. Fraud remains fraud if AI makes it easier. Deepfake pornography does not become less harmful because the images are synthetic. Election interference, impersonation and defamation may all justify targeted legal remedies.

But treating the internet as an endless succession of statements to be classified as true or false faces three increasingly obvious problems.

The first is scale. Generative AI has radically lowered the cost of producing convincing text, images, audio and video. A CNA investigation this year found a factory-like system behind hundreds of videos using synthetic or manipulated female presenters to push false or misleading narratives about Singapore and Malaysia. Another found hundreds of fake videos using Jack Ma's likeness, many apparently produced through coordinated or automated workflows.

There is simply too much content for fact-checkers, platforms or governments to adjudicate every claim before it spreads.

The second problem is ambiguity. Some of the most damaging information online is not simply "false". A genuine photograph can be presented with a misleading caption. A fact can be twisted when stripped of context. Authentic information can be selectively disclosed to intimidate or manipulate. Conversely, AI-generated material is not necessarily false: a synthetic image illustrating a news story may be entirely benign.

The third is authentication. Provenance technologies such as watermarks and content credentials are useful and worth developing. But they answer a different question. They can provide evidence about the origin or editing history of content; they cannot certify truth.

This is why the regulatory landscape is beginning to move beyond falsehoods.

FOLLOW THE AMPLIFICATION

Perhaps the most significant finding in our data is an upstream shift in responsibility.

Before 2016, online platforms – the social-media, video-sharing and messaging services typified by Facebook, Instagram, YouTube, X, TikTok and WhatsApp – were targeted in 38% of the measures we examined. Among instruments adopted between 2023 and 2025, that figure had risen to 69%. Over the same period, the proportion targeting individual creators fell from 89% to 65%, while the proportion targeting people who knowingly disseminated falsehoods fell from 64% to 35%.

The traditional regulatory response targets the individual item of information: identify a false statement, correct it, remove it or punish the person responsible for spreading it. That model still has a place. Fraud remains fraud if AI makes it easier. Deepfake pornography does not become less harmful because the images are synthetic. Election interference, impersonation and defamation may all justify targeted legal remedies. But treating the internet as an endless succession of statements to be classified as true or false faces three increasingly obvious problems.

This does not mean letting users off the hook. It means recognising that the architecture of an information system matters at least as much as the individual pieces of content travelling through it.

A lie told to five people is different from one algorithmically recommended to five million. A deepfake sitting unnoticed on a website poses a different risk from one promoted because outrage drives engagement. If a platform's design determines what goes viral, who sees it, how quickly it spreads and whether corrections ever catch up, regulation cannot sensibly focus only on the person who pressed "post".

This is the logic behind the EU's Digital Services Act. Its important innovations are not instructions to Brussels bureaucrats to adjudicate individual statements as true or false. Instead, large platforms must assess systemic risks, increase transparency, provide access to researchers and take responsibility for aspects of how their systems operate.

That approach is neither simple nor uncontroversial. It took almost two years from the European Commission opening proceedings against X to its first non-compliance decision in December 2025, imposing a €120 million (S$172 million) fine over verification design, advertising transparency and researcher access. X is challenging the decision.

But the direction of travel matters. The regulatory question becomes less "Is this post false?" and more "What systems caused potentially harmful material to reach so many people, and what reasonable responsibilities should attach to those systems?"

That also creates space for interventions short of censorship. Platforms can introduce friction before material is forwarded, reduce artificial amplification, make paid influence transparent, give researchers meaningful access to data, preserve provenance information where possible, and provide effective mechanisms for reporting and appeal.

None of this eliminates the need for fact-checking, media literacy or targeted laws against harmful content. Nor does it eliminate the danger of regulatory overreach. Platform regulation can itself become excessive, expensive or politically abused.

The challenge is to regulate processes and risks without turning governments into editors of the internet.

Singapore entered this debate early. POFMA confronted directly the uncomfortable question of who gets to say that something is false. Seven years later, that question remains important – but it is no longer enough.

The next phase of regulation must ask not only whether a piece of information is true, but what harm it causes, how it reaches people, and who designed the systems that made that reach possible.

That is also where an index such as Beyond Falsehoods can help. Each annual edition will track not only how many laws are passed, but what they regulate, where responsibility lies, what safeguards they include, and how they are implemented. The aim is not to prescribe a single model, but to help distinguish promising approaches from those that are ineffective, disproportionate or overtaken by technology.

That is what it means to move beyond falsehoods.

• Simon Chesterman is David Marshall Professor of Law at the National University of Singapore. Beyond Falsehoods 2026 was written with Fakhar Abbas and Eka Nugraha Putra as part of the Information Gyroscope (iGyro) project at NUS. The report is available at BeyondFalsehoods.com.

Source: The Straits Times © SPH Media Limited. Permission required for reproduction.