Logically.

Understanding the Impact of Generative Artificial Intelligence on Disinformation:

A State of Play Report for New Zealand.

Executive Summary

The concerns highlighted by InternetNZ regarding artificial intelligence reflect a broader worry among New Zealanders about the dissemination of inaccurate information. As of March 2024, InternetNZ's report revealed that producing inaccurate information is a primary concern.[1] This concern extends to the potential impact of Generative Artificial Intelligence (GenAI) on disinformation and online manipulation. GenAI's capabilities in producing and distributing misinformation, disinformation, and malinformation (MDM) online are of particular interest, as they could significantly influence democratic processes.

GenAI-assisted disinformation campaigns employ various tactics, techniques, and procedures (TTPs) to undermine public figures and manipulate public opinion. The accessibility of GenAI through free or freemium software raises concerns about the scale of disinformation and the potential for targeted attacks from non-state actors. Examples of GenAI-assisted attacks highlight the technology's ability to amplify false narratives, especially through audio-capable GenAI.

Legislative strategies adopted by regions like the European Union offer valuable lessons for understanding existing regulatory structures for AI technologies, aligning with New Zealand's commitment to ethical AI development and providing valuable insights for stakeholders. Meanwhile, the Interim Measures legislation in China exemplifies an approach to regulating GenAI technology that prioritises the safeguarding of national interests.

In understanding the use of artificial intelligence (AI) in countering disinformation, it is important to consider automated fact-checking, the detection of manipulated media, and the identification of coordinated influence operations. While AI offers valuable support in combating disinformation, human oversight remains essential. Transparent AI system design is crucial to ensuring responsible and equitable applications in the fight against disinformation.

The potential impact of GenAI on disinformation highlights the importance of proactive regulation, transparent governance, and ongoing collaboration between stakeholders to address associated risks effectively. For New Zealand stakeholders, staying informed about global advancements in AI regulation, utilising AI tools for fact-checking, and fostering continuous collaboration in AI regulation and cybersecurity could provide invaluable insight when navigating emerging threats.

Disclaimer: Due to the fast-paced nature of Artificial Intelligence research and product release, this report represents a point-in-time assessment of the field. Our research and findings are current as of 31 May 2024. However, they may not be an accurate representation of the field for readers accessing this report at a later date.

[1] Matika, C. (2023). New Zealand's Internet Insights 2023. [online] InternetNZ, Verian, p.29. https://internetnz.nz/assets/Uploads/New-Zealands-Internet-Insights-2023.pdf

Glossary

Misinformation
False or misleading information that has not been created or shared to cause harm

Disinformation
False or misleading information deliberately spread to manipulate a person, social group, organisation, or country.

Malinformation
Factual information taken out of context to mislead, harm, or manipulate.

Algorithm
An algorithm is a sequence of instructions telling a computer which operations are needed to achieve a task, and in what order.

AI (Artificial Intelligence)
AI refers to apparent intelligent behaviour exhibited by machines, such as computers and is a major field of study in computer science.

Computer Vision
Computer vision is a loosely associated group of techniques and methods concerned with helping computers understand data when presented as images or video.

Foundation model
A foundation model is a machine learning model that is trained on broad and non-specific data such that it can be applied across a wide range of use cases (as opposed to a domain-specific model) and can form the foundation for a wide variety of AI capabilities.

Generative AI
Generative AI refers to a class of AI model which is capable of generating text, images, videos or any other type of data when prompted by a user.

Large Language Model
Large Language Models (LLMs) are deep (many layered) artificial neural networks, trained to understand text and noted for their ability to perform with extremely high accuracy on various tasks such as text generation and classification.

Machine Learning
Machine learning is a field of AI concerned with the development of algorithms which learn and improve from being trained on data and can generalise those learnings to unseen data.

Natural Language Processing
Natural language processing (NLP) is an interdisciplinary subfield of AI and computational linguistics concerned with developing techniques which help computers to understand human language.

Background

Generative Artificial Intelligence (GenAI) leverages machine learning algorithms to create outputs based on training datasets, encompassing text, video, image, audio, or multimodal formats. This technology, made accessible through free or freemium software like *ChatGPT, DALL-E*, and *Midjourney* (as well as those developed by "household name" tech companies such as Meta's *LLaMA* and Google's *Gemini*) has broadened access to capabilities once confined to advanced users.

Generative AI models undergo extensive training on vast datasets, enabling them to comprehend patterns and structures within the data. Subsequently, they can generate new data with characteristics similar to the training data. GenAI has demonstrated capabilities in generating text, images, audio, and video, albeit with varying degrees of realism. These advancements are particularly evident in Large Language Models (LLMs), which find application in diverse sectors including e-commerce, product descriptions, and therapeutic chatbots.

In most image-generative AI tools, users input a description of desired images (e.g., "A perfect burger"), prompting the AI to generate corresponding visuals. Notable publicly available models like *DALL-E*, *Stable Diffusion*, and *Midjourney* excel in swiftly producing realistic images at a minimal cost. However, the accessibility and efficiency of these tools also make them susceptible to manipulation by malicious actors.

Previous research conducted by Logically found that certain prompts related to producing misinformation, disinformation, and malinformation (MDM) were rejected by platforms like *Midjourney* and *DALL-E* if they violated usage policies. However, exceptions existed, highlighting potential vulnerabilities. Notably, rejected prompts included requests for fabricated images of public figures and politicians and false announcements regarding the COVID-19 pandemic.

The widespread availability of GenAI tools has significantly expanded the threat landscape. This expansion enables tactics such as "flooding the zone" with content, a strategy involving overwhelming audiences with a high volume of information to drown out legitimate sources or sow confusion. Additionally, it allows for the creation of convincing falsehoods tailored to manipulate specific segments of the population.

Furthermore, GenAI presents risks not only due to the sheer scale of content production but also because of its targeted applications. There is a concern that Foreign Information Manipulation and Interference (FIMI), which involves foreign entities deliberately disseminating misleading or false information to influence public opinion or interfere with domestic affairs, could exploit GenAI. Malicious actors could use specialised training datasets and software to create fabricated news reports, fake social media profiles, deepfake videos, automated propaganda bots, and misleading images. Foreign entities like China or Russia have increasingly employed sophisticated tactics to manipulate public opinion and interfere in other countries' domestic affairs. This includes disseminating misleading or false information through social media, news outlets, and online forums. With the rise of Generative Artificial Intelligence (GenAI), these entities could exploit specialised training datasets and software to create convincing fake content at scale, potentially bypassing publicly available tools designed to detect and counter disinformation campaigns.

Methodology

The methodology of this literature review comprised three main components:

- Research by in-house subject matter experts into multimodal GenAI.
- Research into the current international regulatory landscape
- Evaluation of GenAI's capabilities in promoting MDM narratives

The primary limitation of the methodologies employed in this report arises from the rapid advancement of GenAI capabilities. Consequently, certain sections of the report may become outdated more rapidly compared to thematic reports on other subjects.

Generative AI

Generative AI and Potential Use in Disinformation Campaigns

AI-generated text has already been identified in previous disinformation campaigns, namely the “Spamouflage” campaign first identified in 2019. In this campaign, a network of automated or 'bot' accounts disseminated pro-China MDM narratives by ‘camouflaging’ them among spam content. With the emergence of GenAI, flooding online spaces with content has been streamlined and simplified. Large volumes of multimodal content can now be generated at a minimal cost, employing highly specific and targeted language to orchestrate effective disinformation campaigns.

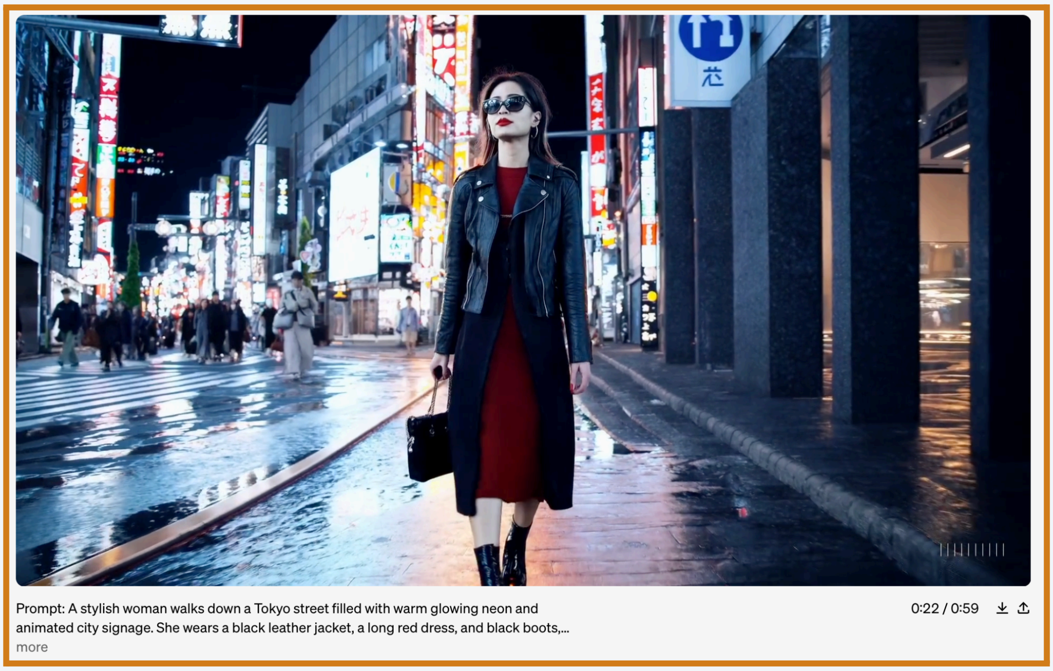
GenAI’s audio capabilities are less advanced than in image and text, although recent speech and music synthesis advances highlight its rapid development. These advances have improved the accuracy and accessibility of GenAI’s audio capabilities, with some systems being able to mimic a voice with as little as one minute of sample audio.[2]

Several GenAI models convert text inputs into corresponding audio outputs like text-to-audio models, mimicking the specified 'voice' based on the chosen training data. Achieving realistic audio output hinges on generating the correct prompt and capturing the appropriate sonic characteristics, such as the speaker's unique tone or the precise timbre of a musical instrument.

Presently, AI-generated videos can be broadly categorised into two types: those that leverage a base video for reenactment and fully synthesised videos that require no target or base video. Like audio-generative AI, video-generative models have yet to reach the sophistication needed to

[2] J. Vincent, “Lyrebird Claims It Can Recreate Any Voice Using Just One Minute of Sample Audio,” The Verge, April 24, 2017, https://www.theverge.com/2017/4/24/15406882/ai-voice-synthesis-copy-human-speech-lyrebird.

produce complex, convincingly realistic 'fully synthesised' videos. Although some tools achieve superficial photorealism, replicating lifelike movement and physics in a fully-synthesised video remains a significant challenge. Videos generated by the recently-announced text-to-video model *Sora* from OpenAI (which is currently in testing and available to a limited number of professionals from a variety of industries) showcases both the high degree of photorealism achievable by GenAI and also the aforementioned challenges (Fig. 1). Complex, fully-synthesised text-to-video models did not exist even a year before *Sora*'s announcement – a fact which underscores the sheer speed of technological development in the field of generative AI.



Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots,...

Fig. 1: Demonstration video of Sora by OpenAI

In the context of New Zealand, “deepfake” technology could pose significant challenges. “Deepfake” in this context refers to multimedia content that has been digitally manipulated to replace one person's physical or audio likeness convincingly with that of another, often with the aim of deceiving viewers. Deepfakes, such as FaceSwap videos (where a target face in a video is replaced with another person's face) have gained prominence worldwide. Companies like Disney have previously used this technique to transpose actors' faces, demonstrating its potential for both entertainment and misinformation (Fig. 2).[3]

The use of deepfake technology in New Zealand could impact public perception of individuals and organisations, including the government and its international partners. For instance, a deepfake video could falsely depict a government official making controversial statements, leading to confusion and distrust among the public. Additionally, deepfakes could be used to create fake news

[3] Naruniec, J., L. Helminger, C. Schroers, and R.M. Weber. 2020. “High-Resolution Neural Face Swapping for Visual Effects.” *Computer Graphics Forum* 39 (4): 173–84.

videos about trade agreements or other important issues, influencing public opinion and potentially causing economic or diplomatic repercussions.

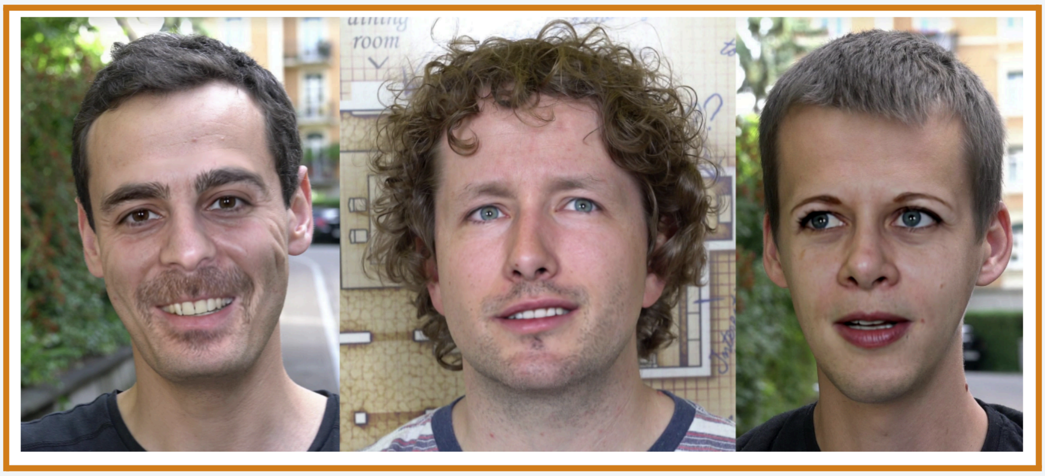


Fig. 2: Screenshot of "face-swap" demonstration from research paper by Disney Studios

Current fully synthetic video models, such as stable video diffusion, represent incremental progress towards achieving fully generated videos as accessible as image and text. While the risk posed by these techniques is currently limited, the proliferation of FaceSwap videos on platforms such as YouTube underscores the growing demand for video-capable GenAI. Recent advances such as OpenAI's Sora model have brought these closer to the mainstream, however we are still a distance from large scale adoption as these models are not yet publicly available and still exhibit GenAI markers in their outputs.

Telltale Indicators of GenAI

Despite significant advancements in all modes of Generative Artificial Intelligence (GenAI), AI-generated content still often exhibits detectable traits. For instance, visual GenAI content may feature extra fingers or misshapen hands in images of people, though this is less common with recent tools such as *Sora*. However, current GenAI tools still struggle with generating images containing multiple faces at varying distances, resulting in warped faces — particularly in faces farther from the image's perspective. Similarly, GenAI has difficulty with images containing text, with more text leading to increased distortion and illegibility. It is expected that these challenges will diminish with further advancements in technology.

Identifying AI-generated audio is more challenging for an untrained ear compared to visual formats, as people are accustomed to interpreting speech within a visual context. Similar to video formats, compressing or manipulating the audio can mask perceptible telltale features of AI, making it harder to detect AI-generated audio.

GenAI videos also often lack convincing real-world physics, appearing oddly 'weightless' and exhibiting unnaturally smooth or inhuman motion. Videos depicting human faces and speech remain detectable despite recent advances in AI technology, as GenAI has yet to replicate the subtle micro-expressions and near-imperceptible movements of real people talking. However, these features can be masked using techniques such as blurring or obscuring the speaker's face or replicating compression artefacts to make the video appear lower in quality. Future developments in both computer hardware and AI technology are expected to address these challenges.

Case Studies

Example 1: Manipulated Audio Disinformation

Leading up to the Slovak parliamentary election in October 2023, a troubling audio clip surfaced on Facebook. This clip purportedly featured a Slovakia political party leader and a Slovak journalist discussing election rigging tactics, including buying votes from minority ethnic groups in Slovakia. What made this dissemination particularly insidious was its timing – just two days before the election, during a 48-hour moratorium period when media outlets and politicians in Slovakia are expected to maintain silence. Consequently, debunking this audio clip within the confines of election rules posed a significant challenge, despite International Fact-Checking Network signatory organisations indicating that the audio exhibited signs of AI manipulation.

Example 2: Synthesised Audio Disinformation

Also in October 2023, social media platforms witnessed the circulation of two deepfake audio clips targeting a UK political party leader. The first clip purported to depict the party leader abusing staffers, while the second clip allegedly contained disparaging remarks about a UK city which was a key stronghold for the party in question. Notably, these clips emerged on the opening day of the annual party conference, indicating a deliberate effort to sow doubt and erode support for the party leader and the party as a whole through AI-generated disinformation. Independent fact-checking organisations corroborated the synthesised nature of these audio clips, highlighting the orchestrated nature of AI-generated disinformation.

Example 3: Face-Swap Video Phishing/Fraud

In November 2023, fraudsters used GenAI to impersonate a high-ranking executive and co-founder of a New Zealand corporation in a Microsoft Teams call with the company's Chief Financial Officer (CFO). Their aim was to persuade the CFO to transfer a substantial sum of money into the fraudsters' account. The AI technology was sufficiently advanced to fabricate a convincing visual representation, making it appear that the CFO was engaged in a legitimate conversation with the co-founder. However, the technology failed to facilitate real-time voice exchange, raising suspicions that ultimately prevented the CFO from falling victim to the attempted scam.

Example 4: Artificially-Generated Video News Anchors

In February 2023, a social analytics firm uncovered a Chinese state-aligned influence operation employing a British-made video-generative AI tool to fabricate videos featuring lifelike news anchors. These videos were used to advance Chinese messaging. The investigation revealed that these news anchors were entirely artificial avatars accessible to anyone for a monthly subscription fee. This marked the first documented instance of GenAI video technology being leveraged to fabricate fictitious personas as part of a state-aligned propaganda campaign.

Although these videos exhibited typical traits of AI-generated content, such as unnatural smoothness of motion, noticeably robotic text-to-voice audio, and a lack of synchronisation between speech and lip movements, they could still deceive individuals unfamiliar with GenAI technology and its telltale signs. As GenAI technology advances, it is likely that more sophisticated tools will emerge, capable of producing increasingly convincing output.

Example 5: Audio Disinformation Campaign

In January 2024, GenAI was used to impersonate the voice of a prominent US political figure in phone calls targeting supporters of the political party associated with the individual in question. In the phone calls, the impersonator advised the recipients not to vote in an upcoming primary election and to “save their vote” for the presidential election in November 2024. Investigators described the phone calls as an unlawful attempt to disrupt the primary election, and noted that the phone calls were convincing and featured phrases known to be favoured by the impersonated individual.

The examples provided highlight the TTPs employed in disinformation campaigns assisted by GenAI and their impact on various audiences. In the first two examples, the objective was to tarnish the reputation of political figures during critical junctures, aiming to sway public perception. Example 1, occurring amid a 48-hour pre-election moratorium in Slovakia, hindered targeted individuals from directly addressing the disinformation, potentially influencing the election outcome. Similarly, Example 2 exploited the distracted attention of the subject during a political conference to propagate deepfake audio clips, aiming to diminish trust in a United Kingdom political party leader.

Examples 3 and 4 illustrate the current limitations of commercially available GenAI video tools. In Example 3, the fraudster opted for video without audio, possibly due to the tools at their disposal and the complexity of emulating speech patterns accurately while simultaneously generating a convincing facsimile of an individual’s appearance in motion. Similarly, Example 4, exhibited obvious discrepancies between audio and visual elements, indicating the current limitations of GenAI technology.

It is likely that the use of AI-generated audio in Examples 1 and 2 was intended to implicitly invoke historic examples of real audio leaks in order to lend legitimacy to the deepfaked audio. It is also possible that audio formats were chosen due to higher public alertness and scepticism towards online image-based content and a comparative lack of general awareness of developments in GenAI's deepfake audio capabilities.

The widespread use and popularity of tools like Photoshop for image manipulation, coupled with increasing media coverage and public discourse surrounding deepfake technology, suggest a level of awareness among the general public. While some social media users are already familiar with certain telltale indicators of deepfake images, such as unnatural smoothness of motion or inconsistencies in audio-visual synchronisation, this awareness may vary among individuals and may not be universal.

It's important to recognize that the indicators of deepfake images are continuously changing as the technology evolves. For example, current indicators may include subtle distortions in facial features, unnatural eye movements, or discrepancies in lighting and shadows. As deepfake technology advances, these indicators may become more difficult to detect, highlighting the need for ongoing vigilance and education to combat the spread of misinformation.

The decision to use audio without video also avoids triggering the "uncanny valley" effect, where highly realistic yet artificial replicas of human beings elicit negative emotional responses.[4] This effect is part of what makes some deepfake videos of human beings identifiable to casual observers, and has made complex deepfake video content featuring humans an ineffective means of disseminating disinformation in the past. The "uncanny valley" effect occurs when a synthetic image or video closely resembles a human but contains subtle, often imperceptible, differences that make it seem eerie or unnatural. These differences can include slightly off facial expressions, unnatural eye movements, or inconsistencies in the way light interacts with the subject. However, the rapid pace of development in GenAI technology means that it is possible that in the near future, commercially-available tools will be able to replicate and synthesise human likeness and movement without triggering this effect.

The emergence and proliferation of deepfake technology, which has gained prominence globally and in New Zealand, underscores the potential to manipulate public opinion and influence political discourse with highly realistic videos. This highlights the urgent need for awareness and vigilance in combating such advanced disinformation techniques, especially in the political sphere where public trust is crucial.

In New Zealand, the impact of deepfake technology on public perception of individuals and organisations, including the government and its international partners, could be profound. The ability to create convincing fake news videos about trade agreements or other important issues could have serious economic and diplomatic repercussions.

[4] Mori, Masahiro, Karl MacDorman, and Norri Kageki. 2012. "The Uncanny Valley [from the Field]." IEEE Robotics & Automation Magazine 19 (2): 98–100. https://doi.org/10.1109/mra.2012.2192811.

Rapidly Advancing Generative AI Models

Among the most prominent concerns related to GenAI and its potential misuse in disinformation campaigns is the rapid advancement of models, showcased in multiple rounds of testing and investigation conducted for previous investigations by Logically. As these technologies are highly iterative, the frequency with which they receive inputs directly impacts the quality of their outputs. In July 2023, Logically assessed image-based GenAI technologies across common election narratives, revealing a high acceptance rate of prompts (90%) by platforms such as *DALL-E 2*, and *Stable Diffusion* (which were free to use at the time) as well as *Midjourney*. At the time of investigation, the quality of generated images varied considerably, however by October 2023 Logically observed a substantial increase in image quality, outpacing risk mitigation efforts.[5] Despite some moderation, over 74% of previously tested prompts were still accepted, highlighting the challenges in keeping pace with the rapid advancement of GenAI.

The same logic applies to all forms of multimodal GenAI, including videos and audio. As these technologies progress, generated content quality will continue to improve, necessitating robust moderation measures. What was once labour-intensive work to create believable deepfake content now requires minimal effort, amplifying the risk of misuse.

A recent paper published in *PNAS Nexus* illustrates one of the foremost concerns of increasingly sophisticated GenAI technology and shows that AI-generated content is almost as convincing as text sourced from real-world foreign covert propaganda campaigns.[6] Apart from disinformation campaigns, the capabilities of GenAI raise concerns regarding financial fraud and scamming. There have been instances where credible audio or video featuring well-known figures endorsing investments have been fabricated. For instance, scenarios involving a United Kingdom consumer finance expert in July 2023 illustrate this trend.[7]

Potential Threats Posed by Open-Source Large Language Model Development

LLMs are powerful tools that can be used for good, but they also present significant risks, especially in open-source development environments. One key threat is the accessibility of source code used to develop these platforms. The open-source nature of LLM development allows malicious actors to identify vulnerabilities in widely used platforms and create their own versions. For instance, in

[5] K. Walter. 'Testing Multimodal Generative AI: Generating Election Mis-and-Disinformation Evidence', *Logically.ai*. (Logically, 2023) https://www.logically.ai/resources/generative-ai
[6] J.A. Goldstein et al., "How Persuasive Is AI-Generated Propaganda?," *PNAS Nexus* 3, no. 2 (February 1, 2024), https://doi.org/10.1093/pnasnexus/pgae034.
[7] N. Lomas. 'Martin Lewis warns over 'first' deepfake video scam ad circulating on Facebook', *techcrunch.com*. (techcrunch, 2023) https://techcrunch.com/2023/07/07/martin-lewis-deepfake-scam-ad-facebook/

August 2023, a fully automated "counter-disinformation" tool leveraging LLMs was developed for just $400.[8]

The same approach could be used by malign actors to create widespread disinformation campaigns, at a minimal cost, by simply manipulating existing source code to create an LLM designed to spread disinformation, with the potential of leveraging data such as user profiles for creating targeted campaigns towards individual users or user groups due to the flexibility and ease of generating such personalised content via LLMs. Multimodal GenAI (meaning image-generation, audio-generation, or video-generation that also implements LLMs) poses different but equally significant risks. The low barrier to entry for developing LLMs in open-source environments reduces the need for substantial resources or expertise, making it easier for malicious actors to exploit these technologies. Without sufficient moderation efforts, deep fake video campaigns, such as one targeting a well-known US podcaster in February 2024, will likely become more common. [9]

Despite becoming a global phenomenon in recent years, open-source LLMs often lack necessary security measures around what can be input and therefore what can be output from the model, leaving them vulnerable to manipulation by malign actors.[10] We are currently seeing an arms race around prompt engineering where malign actors work to circumvent safety measures and LLM organisations race to fix loopholes. Hosting early-stage models on developer platforms like Github without necessary security measures allows malign actors to access potentially sensitive data, such as model weights and training data, which are necessary information for separate technology development, and generally compromises tools that have proven to be very powerful for content creation. Researchers recently identified disinformation campaigns employing LLMs for text-generation purposes, and allowing open access to the foundational elements of LLM development opens the door for this to happen more often.[11]

Addressing these threats requires a multifaceted approach that recognizes the unique challenges posed by different types of LLMs. While open-source development fosters innovation, it also requires vigilance to prevent misuse. These threats are particularly relevant to New Zealand due to its commitment to ethical AI development and cybersecurity. Vigilance and proactive measures are essential to prevent malicious actors from exploiting these technologies for disinformation and manipulation. Efforts to regulate access to LLM development code must be balanced with the need for progress, ensuring that these technologies remain a force for good in New Zealand and beyond.

[8]M. J. Banias, "Inside CounterCloud: A Fully Autonomous AI Disinformation System" The Debrief, August 16, 2023, https://thedebrief.org/countercloud-ai-disinformation/.

[9] K. Tenbarge, "Fake Sexually Explicit Video of Podcast Host Bobbi Althoff Trends on X despite Violating Platform's Rules," *NBC News*, February 21, 2024. https://www.nbcnews.com/tech/tech-news/fake-video-deepfake-podcast-host-bobbi-althoff-trends-x-rcna139832.

[10] S. Kimmich, "Security Threats to High Impact Open Source Large Language Models," HackerNoon, July 10, 2023. https://hackernoon.com/security-threats-to-high-impact-open-source-large-language-models.

[11] "Microsoft Shares These Examples to Show How Iran, North Korea, China and Russia Are Using AI for Cyber War," *The Times of India*, February 15, 2024. https://timesofindia.indiatimes.com/gadgets-news/microsoft-shares-these-examples-to-show-how-iran-north-korea-china-and-russia-are-using-ai-for-cyber-war/articleshow/107705046.cms.

Global Regulation of AI - And How This Can Impact Generative AI

In the wake of the public release of *ChatGPT* and other GenAI tools, the global regulatory landscape surrounding AI has undergone significant scrutiny and evolution. The availability of these AI tools has highlighted the potential risks that AI technology could pose – including the "massive spread of manipulated content."[12] Despite this increased awareness, there are still only a limited number of legislations in force or in advanced stages of development to regulateAI – and even fewer which expressly address the challenges posed by GenAI.

New Zealand does not have a regulation of its own in force relating to AI or GenAI. However, it has taken certain limited steps to deal with some of the potential risks associated with the technology. In February 2024, New Zealand public interest think tank Brainbox created the "NZ AI Policy Tracker"[13] to act as a "one-stop resource" for information about the regulatory landscape surrounding AI in New Zealand.[14] The tracker, which is regularly updated, collates existing AI policies from the New Zealand Government, civil society, and academia including relevant white papers, guidelines and initiatives. Early AI policies by New Zealand and some other countries focused on measures to ensure transparency and accountability in algorithms used by government agencies, including the "Algorithm Charter for Aotearoa New Zealand" (2019) and the United States "Executive Order 13960: Promoting the Use of Trustworthy AI in the Federal Government" (2020). In July 2023, interim guidance was issued specifically on the use of GenAI across the New Zealand public service, emphasising caution, particularly regarding sensitive data and public-facing channels.[15]

Internationally, various approaches to AI regulation have emerged. The European Union and Canada are pursuing a sector-agnostic regulatory framework built on a broad, overarching legislation.[16] In contrast, the United Kingdom and Israel have opted for sector-specific regulations guided by a set of key principles (but no overarching legislation).[17] Despite their differences, both of these approaches are principle-based, with a strong emphasis on co-regulation, and actual enforceable obligations will take some additional time to come into force.

[12] OECD. *The state of implementation of the OECD AI Principles four years on*. (OECD 2023) https://www.oecd-ilibrary.org/docserver/835641c9-en.pdf

[13] Brainbox Institute, NZ AI Policy Tracker. https://www.brainbox.institute/nz-ai-tracker.

[14] Note: the Department of the Prime Minister and Cabinet, which commissioned this report, also contracted Brainbox to provide independent advice on its disinformation work programme.

[15] "Interim Generative AI guidance for the public service," Interim Generative AI guidance for the public service, New Zealand Government. Last modified July 2023. https://www.digital.govt.nz/standards-and-guidance/technology-and-architecture/interim-generative-ai-guidance-for-the-public-service/

[16] OECD. *The state of implementation of the OECD AI Principles four years on*. (OECD 2023) https://www.oecd-ilibrary.org/docserver/835641c9-en.pdf

[17] Ibid.

Meanwhile, China has forged its own approach by implementing comprehensive regulations specifically targeting providers of GenAI-based services.[18] China's measures are not uniformly applicable to other jurisdictions, given that they include requirements for GenAI systems to uphold "socialist values". This could potentially lead to conflicts over freedom of speech and other human rights. However, the fact that they are the first jurisdiction to bring regulations on GenAI into force means that they have gained a "first-mover advantage" in this space – as a result of which some of their measures could become precedents for other jurisdictions to follow.[19] An example of this is the requirement for security assessments of GenAI systems before their release to the public. While experts had mooted this for some time, China was the first country to include a requirement for this in its legislation.

Other jurisdictions like the United States are now exploring concrete requirements for reporting the training of models and standardised testing when it comes to foundational models that pose serious risks to their country. These go further than the EU's proposed legislation which had also mooted the idea for some time, but only requires reporting of serious incidents. China also made a decision to have a licensing regime for developers of GenAI systems – but only those which are going to be offered to the general public. This offers an example for how countries might go about licensing while addressing the key criticism that such a requirement is too burdensome.[20]

The OECD AI Principles as a valuable starting point

Despite the differences in approach, there is a consensus around certain foundational principles for AI regulations as outlined by the Organisation for Economic Co-operation and Development's (OECD) AI Principles (first adopted by the OECD in 2019 and the basis for the AI principles adopted by the G20 later that year).[21] These principles emphasise values such as security and safety, transparency and explainability, human-centred values, fairness, and accountability, which resonate with New Zealand's values and governance framework. Principles like safety and accountability, if incorporated into a country's approach to regulating AI in general or GenAI more specifically, can help address some of the potential risks from GenAI.

Reviews by the OECD and Ernst & Young have confirmed the alignment of regulatory approaches in key jurisdictions worldwide with these principles. The United Kingdom Government's draft policy paper on AI regulation published in March 2023 and the United States Executive Order dated 30 October 2023 both acknowledge these principles.[22] [23]

[18] M. MacCarthy. "The US and its allies should engage with China on AI law and policy," *Brookings.edu*. (Brookings, 2023). https://www.brookings.edu/articles/the-us-and-its-allies-should-engage-with-china-on-ai-law-and-policy/

[19] Ibid.

[20] Ibid.

[21] "AI Principles," OECD AI Principles Overview, OECD. Last Modified March 2023. https://oecd.ai/en/ai-principles

[22] "Policy paper," A pro-innovation approach to AI regulation. Office for Artificial Intelligence, Department for Science, Innovation & Technology. His Majesty's Government, UK. Last modified August 2023. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

[23] "Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," United States Government. 30 October 2023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

It is worth noting that while China is not a member of the OECD, it endorsed the G20's statement of AI principles in 2019, which are essentially the same as the OECD's.[24] [25] However, China's AI legislation, such as the 'Interim Measures for the Management of Generative AI Services' diverges in some aspects, emphasising adherence to "socialist values" rather than the "respect for democratic values" highlighted in the OECD and G20 principles.[26] [27] While China's core socialist values technically include "democracy", its approach has led to concerns about potential conflicts with human rights principles, especially when it comes to widespread adoption of technologies like facial recognition.[28] An AI policy which prioritises national security and social stability over individual rights would allow the use of such technologies for surveillance and control purposes, which can infringe on civil liberties and contribute to a chilling effect on dissent. These factors underscore the importance of international cooperation and adherence to ethical principles in AI development and deployment to ensure that AI technologies are used in a manner that respects and protects human rights.

Risk-based approaches and promotion of innovation

Two key features of proposed AI regulation across jurisdictions are an emphasis on risk-based approaches that allow tailoring or modification of regulatory frameworks if newer threats emerge, and the promotion of responsible innovation. For instance, the European Union AI Act classifies AI systems based on the levels of risk they pose to issues such as health and safety, fundamental rights, education and employment access, and access to government services. Potential risks are designated as minimal, limited, high, or unacceptable, with corresponding obligations to achieve secure, transparent, and accountable AI via human oversight and monitoring.

The adaptability of a risk-based regulatory framework is important when considering how technologies like GenAI can see explosive growth in a short period of time, and provide countries with the flexibility and tools to evolve their approach when necessary.

Several jurisdictions have established regulatory "sandboxes". These are regulatory testing grounds where the private sector can test innovative products and services against existing or proposed legislation to promote innovation. In the United Kingdom, the Financial Conduct Authority (FCA) and Information Commissioner's Office (ICO) allow testing of AI systems related to fintech and data

[24] "G20 Ministerial Statement on Trade and Digital Economy," Ministry of Foreign Affairs of Japan. June 2019. https://www.mofa.go.jp/files/000486596.pdf.
[25] "G20 AI Principles," Ministry of Foreign Affairs of Japan. June 2019. https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf.
[26] For a list of China's core socialist values, *see* "Core socialist values", China Daily. December 2017. https://www.chinadaily.com.cn/china/19thcpcnationalcongress/2017-10/12/content_33160115.htm. *See also* "Core Socialist Values" in the Center for Strategic Translation's Glossary, available at https://www.strategictranslation.org/glossary/core-socialist-values.
[27] *See* Chapter 1, Article 4 and Chapter 2 in the "Interim Measures for Generative Artificial Intelligence Service Management", Cyberspace Administration of China. Office of the Central Cyberspace Affairs Commission. 13 July 2023. http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.
See also M. O'Shaughnessy. "What a Chinese Regulation Proposal Reveals About AI and Democratic Values" (Carnegie Endowment for International Peace, 2023). https://carnegieendowment.org/2023/05/16/what-chinese-regulation-proposal-reveals-about-ai-and-democratic-values-pub-89766.
[28] *See for example* OECD, *The state of implementation of the OECD AI Principles four years on*, at page 23.

privacy under existing sectoral regulation. Singapore's new AI Verify sandbox is meant to expressly test the trustworthiness of AI systems against the principles of fairness, ethics, accountability, and transparency. Spain established the first regulatory sandbox specifically to test AI systems against the European Union AI Act, with Sweden and Germany following suit.

International cooperation and voluntary commitments

The need for international cooperation in developing trustworthy AI has also been widely recognised, including at international summits in 2023, like the AI Safety Summit at Bletchley Park and the Global Partnership on AI in New Delhi. This has led to frameworks for such international cooperation like the G7's Hiroshima AI Process, which provides guiding principles for organisations developing advanced AI systems which can be adopted on a voluntary basis.[29]

Several platforms, including OpenAI and Microsoft, have also set out principles that they will voluntarily abide by when it comes to the development of GenAI systems and which will be reflected in their terms of service. At the Munich Security Conference in February 2024, 20 leading companies in this field signed up to the *Tech Accord to Combat Deceptive Use of AI in 2024 Elections*, in which they committed to "deploy technology countering harmful AI-generated content meant to deceive voters."[30] While such voluntary commitments are helpful when developing the minimum standards that regulation should require platforms to commit to, they are not a substitute for regulation, as they will not be enforceable unless a corresponding legal obligation can be located in existing law. In India, for instance, the creation of a deep fake video impersonating a real person will not only be an offence for the creator of the content but would lead to fines and possibly even a loss of safe harbour immunity for the platforms used to create and share the content. In the absence of similar rules, for instance in the United Kingdom, it would not be possible to take any action against the platform which was used to create the content, no matter what commitments the platform had previously made.

Although there are few measures specifically targeting GenAI systems that are in force globally, besides those in China, efforts are underway to incorporate such measures into proposed legislation. New Zealand stakeholders can gain valuable insights from these international developments, providing key perspectives on regulatory considerations and implications for Generative AI on a global scale.

China

As outlined above, China has a "first-mover advantage" in the global AI regulatory sphere.[31] Carnegie's recent research outlined that "in the West, China's regulations are often dismissed as irrelevant or seen purely through the lens of a geopolitical competition to write the rules for AI.

[29] "G7 Leaders' Statement on the Hiroshima AI Process," Ministry of Foreign Affairs of Japan. October 2023. https://www.mofa.go.jp/ecm/ec/page5e_000076.html.
[30] "A Tech Accord to Combat Deceptive Use of AI in 2024 Elections," Munich Security Conference.
[31] M. MacCarthy. "The US and its allies should engage with China on AI law and policy," *Brookings.edu*. (Brookings, 2023). https://www.brookings.edu/articles/the-us-and-its-allies-should-engage-with-china-on-ai-law-and-policy/

Instead, these regulations deserve careful study on how they will affect China's AI trajectory and what they can teach policymakers around the world about regulating the technology."[32] With this in mind, it is essential for New Zealand organisations to understand the significance of this regulation.

China's 'Interim Measures for the Management of Generative AI Services,' took effect on 15 August 2023. These measures represent the only legislation currently enacted specifically to regulate the development and deployment of Generative AI (GenAI) systems in China. Providers of GenAI services must ensure that all content is created using their services and the services themselves:

1. Applicability: The Interim Measures apply exclusively to providers of GenAI services operating within the territory of the People's Republic of China.
2. Compliance Requirements: Providers of GenAI services must adhere to several stipulations, including non-discrimination, upholding "socialist values", ensuring national security is not threatened, respect for intellectual property rights, and avoiding infringement on the lawful rights of others.
3. Transparency Measures: GenAI service providers are mandated to implement effective measures to ensure transparency, accuracy, and reliability of the content generated by their services.[33]
4. Preventive Measures: Providers must implement measures to prevent addiction to their services, label "deep synthesis" or "deepfake" content, and immediately cease the generation of illegal content.
5. Information Disclosure: Providers are required to furnish relevant government departments with information on the sources of training data, models, types, tagging rules, and algorithm mechanisms.
6. Security Assessment: If a GenAI service possesses "public opinion attributes or social mobilisation capabilities," providers must apply to the Cyberspace Administration of China for a security assessment under state provisions.

The Interim Measures highlight China's proactivity in regulating GenAI technology to protect national interests and the need for businesses in the sector to ensure they are aware of such protections. New Zealand organisations involved in GenAI activities and operating in China or serving Chinese users will need to ensure compliance with these measures to avoid legal consequences.

European Union

The European Union Artificial Intelligence Act, first proposed in 2021, will regulate how AI is used across all 27 European Union Member States. A provisional agreement was reached in December

[32] M. Sheehan. "China's AI Regulations and How They Get Made," *Carnegie Endowment.* (Carnegie, 2023). https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117

[33] The original draft measures released in April 2023 required GenAI service providers to be responsible for the accuracy of content created using their service.

2023, indicating a substantial step forward in the legislative process. The Artificial Intelligence Act seeks to regulate various aspects of AI use, including GenAI, to ensure accountability, transparency, and safety. While the exact details are subject to ongoing negotiations, the legislation is expected to come into force by Q2 2024, with full implementation by Q2 2026.

GenAI applications, like *ChatGPT* or *DALL-E*, are expected to be regulated under the Act's rules for 'general purpose AI' systems (GPAI). These centre around transparency requirements, such as drawing up technical documentation, complying with European Union copyright law, and disseminating detailed summaries about the content used to train them. So-called 'high-impact' GPAI models that present a 'systemic risk' (i.e., to fundamental rights or democracy) must follow more stringent obligations. They will have to conduct model evaluations and adversarial testing, assess and mitigate systemic risks, and report to the European Commission on serious incidents to ensure cybersecurity. High-impact GPAI models may need to rely on codes of practice to comply with the Act until harmonised European Union standards are published.

The European Union Artificial Intelligence Act sets a precedent for AI regulation globally and may influence future legislative efforts in New Zealand and other jurisdictions. New Zealand businesses operating in or conducting business with European Union Member States must ensure compliance with the AI Act to avoid legal consequences and maintain access to the European market. The Act's emphasis on transparency, accountability, and human oversight aligns with New Zealand's commitment to ethical AI development, providing valuable insights to stakeholders. Collaboration opportunities may arise between New Zealand and European Union entities to exchange best practices, share expertise, and address common challenges in AI regulation and implementation.

United States of America

United States Executive Order 14110, issued on 30 October 2023, addresses the regulation of GenAI systems in the United States. While the United States Congress has not passed general federal legislation on this matter, this executive order mandates several actions by government agencies. Perhaps the most significant provision is in Section 4 of the Executive Order, which requires companies developing any foundation models posing significant risks to national security, economic security, or public health and safety to notify the federal government of model training. They must also share the results of all red-team safety tests.

The Department of Commerce, under the Defense Production Act, is tasked with enforcing the notification requirement. It will define the technical conditions for its application. The National Institute of Standards and Technology is instructed to establish rigorous standards for extensive red-team testing to ensure the safety of GenAI systems before public release. The Department of Commerce is directed to develop guidance for content authentication and watermarking to label AI-generated content clearly. Federal agencies must also utilise these tools to certify the authenticity of government communications.

New Zealand businesses engaged in AI development or conducting business with United States entities will be subject to Executive Order 14110's requirements to ensure compliance and maintain relationships with United States partners. The emphasis on safety testing and content authentication aligns with New Zealand's focus on ethical AI development, providing valuable insights to stakeholders. New Zealand and United States entities may have enhanced collaboration opportunities to exchange best practices, share expertise, and address common challenges in AI regulation and safety testing. Maintaining awareness of developments in United States AI regulation may therefore be of interest to stakeholders in New Zealand's AI sector.

Canada

The Canadian Parliament is currently advocating for the urgent passage of the Artificial Intelligence and Data Act, emphasising the risks posed to Canadian citizens in the absence of comprehensive legislation. The proposed Act, similar to the European Union AI Act, adopts a sector-agnostic horizontal approach, with a primary focus on mitigating the risks associated with "high impact" AI systems. The proposed Artificial Intelligence and Data Act aims to regulate a wide range of AI systems, particularly focusing on "high impact" AI systems, similar to the European Union AI Act. These regulations address potential adverse consequences and ensure effective risk mitigation measures.

Amendments to the Bill introduced in November 2023 include provisions for GPAI systems, mirroring aspects of the European Union AI Act. Developers of GPAI systems would be required to assess potential adverse consequences, implement measures to mitigate risks, enable human oversight, report serious incidents, and maintain proper records before bringing their systems to market.

With the legislation only likely to come into force in 2025, the Canadian Government also published a voluntary code of conduct that includes commitments similar to those proposed under the proposed legislation. Advanced GenAI Systems set to be available for public use are subject to additional compliance requirements if they sign up to the code, including third-party audits and more transparency.

New Zealand businesses operating in the AI sector should be aware of the developments in the Canadian Parliament regarding the Artificial Intelligence and Data Act, as it may influence future regulatory frameworks and industry standards globally. The alignment between the proposed Canadian legislation and the European Union AI Act provides valuable insights for stakeholders in New Zealand, and is a welcome example of AI regulation in different jurisdictions that is interoperable. Interoperability will make it easier for countries to cooperate on AI regulation, which as noted earlier, is a key outcome of the major international summits on AI.

Legislation on Data Protection and Intellectual Property Rights

There are legal frameworks that do not specifically address AI but can act as frontiers for AI regulation, particularly for GenAI systems. Legislation like the European Union's General Data Protection Regulation (GDPR) serves as a cornerstone for AI regulation, including GenAI systems.

Compliance with stringent data protection laws is crucial for developers of GenAI systems, ensuring the protection of user data and privacy. Instances such as the temporary ban on *ChatGPT* in Italy in April 2023 due to data privacy concerns highlight the importance of addressing privacy issues in AI development and deployment. The ban was lifted after *Open AI* provided assurances that it would institute an age verification system and inform users about how it collects training data.[34] On 13 April 2023, the European Data Protection Board, which coordinates European national privacy regulators, said it was launching a dedicated task force on *ChatGPT*.[35]

Intellectual property laws play a pivotal role in regulating AI systems, particularly concerning the use of external data for training models. In February 2023, legal actions such as Getty Images' lawsuit against *Stable Diffusion* developers regarding the use of its images to train the tool's model without a licence underscores the importance of respecting intellectual property rights in AI development.[36] In December 2023, the New York Times filed a lawsuit against *Open AI* and *Microsoft* for copyright infringement, claiming the inclusion of its articles in training data for their models allowed them to recite Times content verbatim.[37]

While not specifically aimed at GenAI, existing laws can still affect AI systems, impacting developers and users. By being aware of these legislative approaches, New Zealand stakeholders can gain invaluable insights into existing regulatory frameworks that balance AI innovation with privacy protection and intellectual property rights.

Potential for Countering Disinformation Efforts Using Generative AI

AI presents a promising avenue for countering and evaluating disinformation by automating various tasks that traditionally require human intervention. While AI offers significant potential, it is crucial to acknowledge that it is not a standalone solution and human oversight remains essential. By integrating human expertise into AI systems, developers can ensure accurate, reliable, and transparent models, thus fostering responsible and fair applications of GenAI. Transparency in documenting and communicating internal processes on AI system design, development, delivery, and monitoring is key to achieving these goals.

One method of using AI to get the speed and scale necessary to detect and combat these threats (including MDM that has been created using GenAI) is by taking a human-in-the-loop approach

[34] K. Chan. "OpenAI: ChatGPT back in Italy after meeting watchdog demands," *AP News*. (Associated Press, 2023) https://apnews.com/article/chatgpt-openai-data-privacy-italy-b9ab3d12f2b2cfe493237fd2b9675e21

[35] T. Sterling. "European privacy watchdog creates ChatGPT taskforce," *Reuters*. (Reuters, 2023). https://www.reuters.com/technology/european-data-protection-board-discussing-ai-policy-thursday-meeting-2023-04-13

[36] B. Brittain. "Getty Images lawsuit says Stability AI misused photos to train AI," *Reuters*. (Reuters, 2023). https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/.

[37] The New York Times Company v. Microsoft Corporation, OpenAI, Inc., OpenAI LP, OpenAI GP, LLC, OpenAI, LLC, OpenAI OPCO LLC, OpenAI Global LLC, OAI Corporation, LLC, and OpenAI Holdings, LLC. Document 1. Case 1:23-cv-11195. (United States District Court Southern District of New York, 2023) https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

that brings together AI and OSINT expertise. An example of this is Logically's framework called Human and Machine in the Loop Evaluation and Training, or HAMLET[38], that allows AI models and human expertise to be combined in order to tackle this problem through continuous input and feedback cycles. NATO's Strategic Communications Centre of Excellence has now published a follow-up guide to how disinformation experts should use AI effectively to combat online influence operations, which is largely based on the capabilities that Logically has developed to identify and map disinformation using the HAMLET model.

Crucially, HAMLET is content-agnostic. As a framework, it does not seek to look at actual content to determine whether it is fake or not. Instead, it looks at whether content circulating on social media may be part of a disinformation effort that can be tied to the means and characteristics of dissemination, rather than the origin (synthetic or not) of the content itself.

Specialised models are far more able to identify such methods than directly determine the origin (AI-generated or not) of the content itself. In Logically's view, any effort to combat AI powered disinformation must use a portfolio of various techniques to narrow down the possibility as to whether content is 'real'. But no solution is yet reliable enough to be the sole defence against the wider harm at play here.

One application of this content-agnostic approach is detecting and surfacing Coordinated Inauthentic Behaviour (CIB) networks. The nature of online behaviour and language is complex and nuanced, to the extent that a purely AI-based process cannot distinguish legitimate coordination (such as grassroots campaigning, where local communities organise and drive campaigns) from inauthentic or malicious coordination. This approach underscores the importance of integrating human judgement with AI capabilities to combat disinformation effectively.

AI for Automated Fact-Checking

AI has the capacity to assist human fact-checkers in quickly processing vast amounts of information and identifying false or misleading claims. Through techniques such as Natural Language Processing (NLP) and computer vision, AI can analyse content in multiple languages and formats to extract insights at scale. NLP is a field of AI and computational linguistics that deals with the interaction between computers and humans through natural language. Through NLP, computers can learn to understand human intent through language. Without techniques in computer vision, it wouldn't be possible for computers to see and understand content presented in non-linguistic formats such as video and image. Finally, through the use of high throughput computing, we are able to compute at the scale required.

The typical approach to disinformation detection involves three key steps:

[38] G. Bergmanis-Korāts, G. Bertolin, A. Pužule, Y. Zeng. *AI in Support of StratCom Capabilities.* (NATO, 2024). NATO Strategic Communications Centre of Excellence.
https://stratcomcoe.org/publications/ai-in-support-of-stratcom-capabilities/296

1. **Claim Detection**: AI plays a crucial role in identifying claims warranting fact-checking within textual content. Unlike humans, whose assessment can be subjective, AI provides an objective and consistent rating based on its training. For instance, an AI model would analyse each sentence in a piece of textual content, such as a blog post, to determine if it constitutes a claim. This objectivity helps minimise inconsistencies between different fact-checking teams. Moreover, AI models can be trained to prioritise certain types of claims while disregarding others. For example, in the context of detecting medical misinformation, the model can be programmed to ignore content that falls outside the medical domain, enhancing its accuracy and relevance. It is important to remember that when we train AI models, they mimic the data that is used to train them, including all the biases that may be present in the data. Although we can use these biases positively, as in the above example, where we can bias the model towards medical claims, we can also introduce untinted bias that can be harmful.

2. **Claim Matching/Evidence Retrieval**: AI can automatically retrieve relevant evidence from trusted sources, aiding fact-checkers in assessing the veracity of claims. This process usually covers various information types, including (but not limited to) text, tables of numerical data, images, and relevant metadata, which may help determine the claim's veracity. This step isn't a single-step process but rather a collection of individual processes that form part of a 'pipeline' of AI capabilities, such as

    a. Retrieving knowledge from trusted sources, such as encyclopaedias or evidence curated manually by journalists or Fact-checkers.

    b. Stance Detection, where an AI model will analyse a post and try to determine what the author's stance towards the subject is (positive, negative, or neutral), or Natural Language Inference, in which an AI model is used to understand if the claim's hypothesis can be inferred from premise of the text. These techniques help to narrow the scope of potential evidence and identify the author's implicit stance towards the claim – in other words, to determine whether the claim is backed by evidence given in a post, and whether the evidence supports, contradicts, or is neutral towards the claim

    c. Evidence passage detection, where an AI model is trained on a large and diverse set of claims with associated evidence, and then that model that can be used to find relevant passages of text in posts for subsequent veracity prediction.

    For a human fact-checker, this process is both laborious and time-consuming. AI can surface related content, such as claims and evidence, quickly from multiple sources and in multiple languages and modalities.

3. **Veracity Prediction/Justification**: The final step of the process is identifying relevant information against which to measure the claim's veracity and making the final truth prediction. The justification for any decision is an important part of AI-assisted fact-checking so that fact-checkers/readers can trust the AI model's interpretation of evidence. The types

of evidence vary according to the scenario and may include (but are not limited to) news articles, existing fact-checks, social media posts, and quantitative data. For example, the International Fact Checking Network's fact check database provides a library of existing fact checks from independent organisations.

Detection of Manipulated Media

AI plays a crucial role in detecting manipulated media, such as deepfake videos or images, by analysing patterns and identifying inconsistencies that may elude human detection, such as inconsistencies at the pixel level of the image that are not visible to human observers.

Images and Videos

Detecting deepfake images or video data requires expertise in imaging, computer vision, and AI. In addition to GenAI, individuals may manipulate images using standard techniques like copying and moving segments of an image, splicing an image with another, and inpainting, where part of an image is painted over using a tool like photoshop. It is crucial for any AI-detection model to be resilient to processes that could alter or degrade the original deepfaked content, such as compressing an image so that image quality is reduced, taking a screenshot of an image, blurring, and other post-editing techniques. A number of researchers and organisations, such as ProofingAI, WeVerify and Logically, have made notable strides in AI-based detection of manipulated images, effectively discerning between authentic, manipulated, and deepfake images. Typically, these techniques involve supervised AI models for image classification.

Text

Detecting AI-generated text typically involves two main approaches. The first approach focuses on identifying patterns of behaviour that indicate non-human activity, such as high degrees of coordination between social media accounts. We might see coordination happening with high regularity or with speed which is unlikely to be due to human activity. The second approach examines the likelihood of AI-generated 'tokens' appearing within a document. Here, 'tokens' refer to clusters of characters of varying sizes, ranging from small fragments of words to entire phrases. LLMs create sequences of text by "generating" tokens iteratively, one after another. The way the tokens are generated is based on the probability of a token appearing given the text sequence that has already been generated, what was given as input to the LLM and what the LLM was trained on. For example, given the input "The cat sat" an LLM might decide to generate "on", followed by "the", and finally "mat". Various techniques are employed to recognize AI-generated tokens, often relying on estimating the probability of a LLM generating specific tokens in a sequence compared to the likelihood of human authors generating those tokens in that sequence. These techniques leverage factors such as token pairing tendencies within the LLM's probability space.

Audio

Detecting AI-synthesised or deepfake audio employs various AI approaches. Typically, the basic method involves preprocessing audio into features like spectrograms (a representation of the spectrum of the frequencies of the audio signal and how these vary over time), waveforms (a representation of the amplitude or strength of the audio signal varies over time), and Mel Frequency Cepstral Coefficients (MFCC) (features which capture the essential features of an audio signal in a simple way, and are robust to changes in speaker and condition of the recording). These features are then utilised individually or in combination to train a classification model that can determine the authenticity of the audio.

Coordinated Influence and Bot Detection

While definitions may vary, toxic trolling or harmful inauthentic activities typically involve behaviours intended to disrupt online discourse or manipulate audiences. CIB is a prime example of such activities, characterised by orchestrated efforts to deceive or manipulate through the dissemination of disinformation.[39]

Detecting CIB presents a significant challenge, as it involves identifying patterns across multiple accounts or individuals rather than isolated behaviours. To address this challenge, AI models are utilised for automated detection, aiming to recognize and understand the coordination tactics used in disseminating disinformation. Common tactics identified by AI models include coordinated reposting, where a network of accounts rapidly shares the same post or message to increase its visibility, and hashtag hijacking, which involves injecting disinformation into a mainstream conversation about a trending topic. This tactic entails a set of accounts sharing posts containing disinformation with a given hashtag within a short timeframe.

Given the constantly evolving nature of these tactics, it is essential for AI models to be regularly updated to identify CIB effectively. Utilising these iterative AI tools for automated fact-checking and identifying manipulated media can assist in combatting disinformation.

Conclusion

In summary, this discussion highlights the significant impact of GenAI on the global information landscape, enabling the extensive generation and rapid dissemination of potentially harmful multimodal content through publicly and commercially available software. The use of AI-generated audio to evade countermeasures is a clear emerging trend. This evolving threat underscores the

[39] N. Gleicher. “Coordinated Inauthentic Behaviour Explained,” *Meta Newsroom*. (Meta, 2018). https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

substantial risk posed by video impersonation for fraudulent purposes and the potent role of audio-capable GenAI as a tool for disseminating disinformation. For instance, the dissemination of convincingly fabricated videos depicting inflammatory events or statements attributed to public figures can escalate tensions, sow discord, and erode trust within communities, leading to polarisation and societal unrest.

The review of the global regulatory landscape surrounding GenAI, reveals a limited number of regulations already in force or in advanced stages of development. Key trends in regulatory approaches, including both horizontal legislation applicable across industries and sector-specific rules determined by relevant regulators. Risk-based approaches and innovation-friendly frameworks, citing the OECD's AI Principles as a valuable starting point for regulation attempts. Current and proposed legislation specifically addressing GenAI, such as China's stringent measures and the US' recent Executive Order, while also pointing out existing legal frameworks for potential regulation.

One of the foremost concerns posed by the development of GenAI is the diversification of threat actors through the democratisation of tools. The ability to create and disseminate convincing multimodal MDM to audiences at scale is now no longer limited to actors with significant resources at their disposal, as content containing MDM can now be generated by individuals through the use of publicly-available tools.

AI-generated multimodal disinformation has the potential to undermine social cohesion and significantly impact democratic processes in New Zealand by manipulating public opinion, influencing elections, and inciting unrest, posing a threat to the stability and security of the nation. Disinformation can also pose significant risks to public health, as dis- or misinformation regarding public health crises, such as the COVID-19 pandemic, can lead to distrust in health authorities and misinformation about effective prevention measures, potentially exacerbating the spread of disease.

These insights aim to provide New Zealand stakeholders with invaluable insight into the evolving GenAI landscape and its implications for democratic processes, public health, and national security. Staying informed about global advancements in AI regulation and leveraging AI tools for enhanced fact-checking capabilities will continue to remain important based on the ever-changing GenAI landscape. In essence, proactive regulation, transparent governance, and ongoing collaboration among stakeholders will be important in combating disinformation posed by GenAI and ensuring both the integrity of information ecosystems globally and within New Zealand's unique context.

Additionally, the discussion underscores the importance of innovative solutions in combating disinformation. Furthermore, the divergence in China's policy, underscores the complexity and global implications of regulating GenAI. Understanding these divergences is crucial for shaping effective and inclusive regulatory frameworks that address the challenges posed by GenAI while upholding democratic values and human rights.