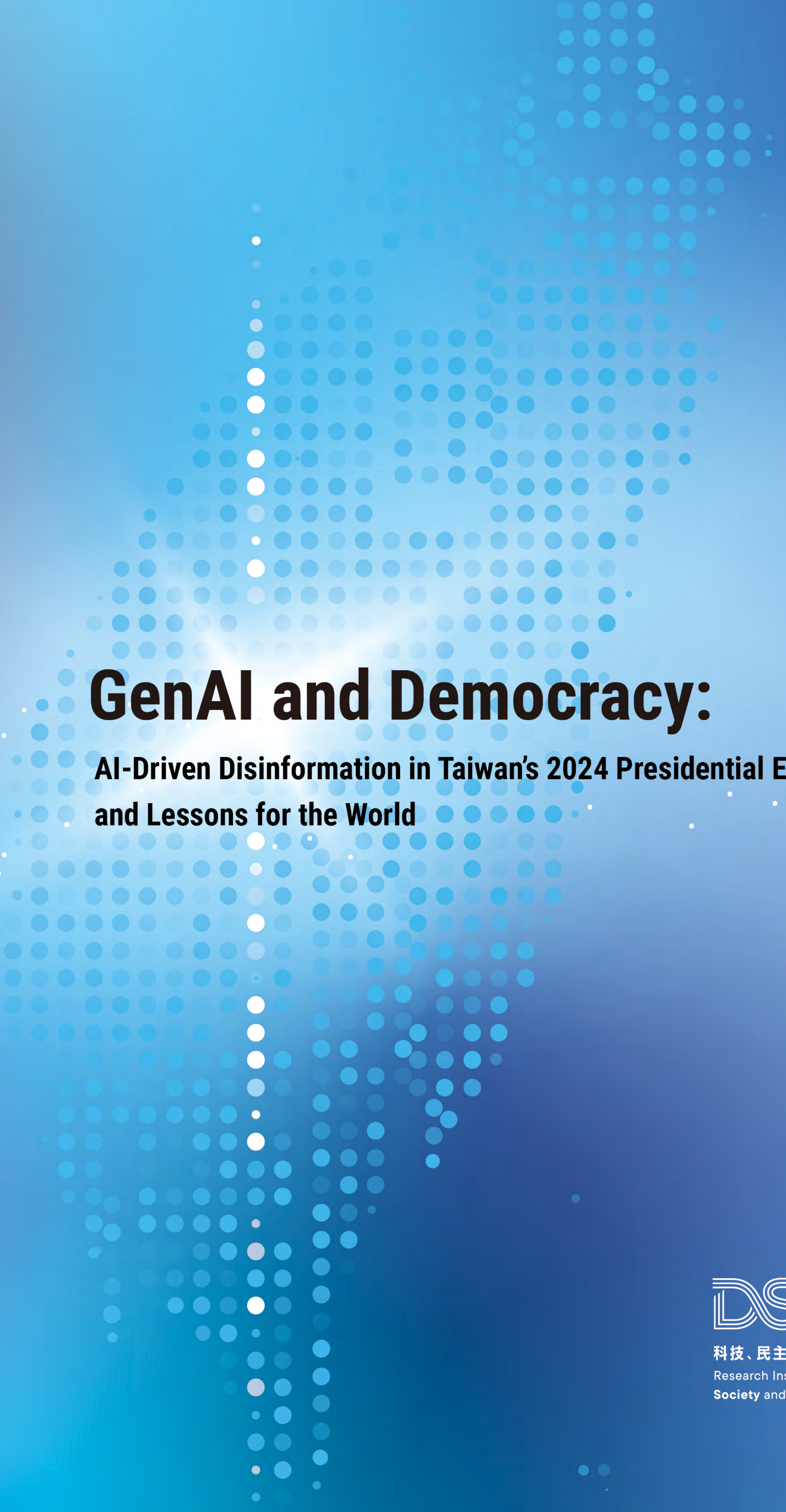


GenAI and Democracy:

AI-Driven Disinformation in Taiwan's 2024 Presidential Election and Lessons for the World

DSET
科技、民主與社會研究中心
Research Institute for Democracy,
Society and Emerging Technology

Research Institute for Democracy, Society and Emerging Technology (DSET)

DSET stands at the forefront of Taiwan's democratic values, weaving these principles into the very fabric of our policy research. As emerging technologies reshape the global landscape, they bring a mix of opportunities and challenges that touch every aspect of political and social spheres. We are committed to crafting governance frameworks for technology that not only protect security but also uphold freedom and sustainability. Our mission is to steer through this new terrain, ensuring that innovation serves democracy, and freedom remains at the heart of technological advancement.

Emerging Technologies and Democracy Research Unit at DSET

The Emerging Technologies and Democracy Research Unit investigates the impact of critical technologies on democratic process and integrity, both in Taiwan and globally. Our key focus areas include artificial intelligence, information manipulation, and supply chain resilience. We explore their integration within democratic frameworks, emphasizing Taiwan's unique position within geopolitical contexts and its role in navigating the complex dynamics of global technological freedom. Our objective is to provoke critical reflection on maintaining democratic practices amid the rapid changes of technological development and international order.

Lead Author

Kai-Shen HUANG (DPhil, Oxford University)

Huang is currently a research fellow at DSET, where he is leading the Emerging Technology and Democracy Research Unit. His research focuses on critical technologies, democratic practices, and supply chain resilience. Previously, he researched AI applications in dispute resolution and public administration in Shanghai. He earned his DPhil in Anthropology from Oxford University in 2020.

Co-Authors

Muyi CHOU (PhD, Humboldt University of Berlin)

Chou is a deputy director at DSET and an assistant professor at National Taiwan Ocean University. She obtained her PhD in Political Science from Humboldt University of Berlin. Her research interests include deliberative democracy, the social and solidarity economy, and social innovation, with a more recent focus on AI and social defense.

Wei-Lin CHEN (PhD, UC San Diego)

Chen is an associated researcher at DSET and a postdoctoral research fellow at the New Zealand Policy Research Institute (NZWRI) at Auckland University of Technology. He completed his PhD in Economics at the University of California, San Diego, in 2023. His research interests are in public economics, development economics, and political economy, with a particular focus on group disparities in public policy practices.

Kyle Yulun KUO (PhD Candidate, NCCU)

Kuo is an associated researcher at DSET and a PhD candidate in Public Administration at National Chengchi University. In 2023-24, he conducted research into disinformation and information manipulation at the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University.

Acknowledgements

This publication is based on a series of research projects initiated in April 2024, under the guidance of DSET's advisory board. We extend our deepest gratitude to all the board members, especially to Ching-Yi LIU, Zsehong TSAI, and Jieh-min WU, for their invaluable insights and steadfast support throughout the duration of this research. Special thanks are also due to numerous experts and organizations, and to all the associated researchers contributing to the publication, including Vakau from Access Now and Isabel Hou from Taiwan AI Academy.

Published in 2024 by the Research Institute for Democracy, Society and Emerging Technology (DSET).

Table of Contents

Summary

Since OpenAI launched ChatGPT in 2022, generative artificial intelligence (GenAI) has revolutionized, if not redefined, the landscape of digital content creation. This technology now enables the efficient and cost-effective production of text, audio, and video, reshaping how information is generated and consumed. Such advancements inevitably raise concerns about the potential for misuse, particularly in sensitive areas such as political processes. In this context, Taiwan's 2024 presidential election provides a critical case study. This report explores and analyzes how GenAI was deployed during the election, presenting specific instances to illustrate its impact on political communication and electoral integrity.

Objectives

This report is structured around five objectives:

1. Defining information manipulation and evaluating its analytical frameworks:

Establish a clear definition of information manipulation and assess the merits and limitations of various analytical frameworks, including DISARM, BEND, and RICHDATA.

2. Presenting and analyzing case studies from Taiwan's 2024 presidential election:

Provide a detailed analysis of multiple instances in which GenAI was utilized for information manipulation in Taiwan, covering four cases: the 58-second audio recording of Ko Wen-je, allegations of Lai Ching-te as an informant of the Kuomintang, fabricated audiovisual content attributed to US Representative Wittman, and fabricated documents purporting to disclose the "secret history" of Tsai Ing-wen.

3. Comparing similar international cases:

Examine four international instances of information manipulation using GenAI, including the use of deepfake technology in India's 2024 elections, the fabricated Pentagon explosion incident in 2023 and the falsified 2024 Biden telephone recording (both in the US), and the counterfeit France 24 TV news broadcast in France in February 2024.

4. Assessing GenAI's effectiveness in manipulating information:

Evaluate the current usage and impact of GenAI in information manipulation, exploring the conditions under which it could have substantial societal effects.

5. Preliminary recommendations for mitigating risks:

Propose strategies for strengthening regulatory measures, enhancing public awareness, and fostering international cooperation to mitigate the potential impacts of GenAI in information manipulation.

Findings

This report identifies two major findings:

1. Moderate impact of GenAI on information manipulation:

The Taiwan Communication Association, Microsoft Threat Analysis Center, AI Labs, and our observations indicate that the impact of GenAI on information manipulation during Taiwan's 2024 presidential election has remained limited and did not significantly alter public opinion.

2. Challenges in content creation and detection:

GenAI has reduced the time and labor costs associated with content creation, making coordinated inauthentic behavior (CIB) harder to detect and placing a significant burden on fact-checking organizations.

Policy Recommendations

This report proposes three initial policy recommendations for further action:

1. Mandate responsible behavior for platforms and GenAI companies:

Platforms should systematically release data associated with the deletion of accounts and posts to aid research efforts. Companies developing GenAI should be required to conduct stringent due diligence to prevent their products from being exploited for information manipulation.

2. Implement preventive measures and address future challenges:

Enhance regulatory oversight of GenAI technologies, raise public awareness, develop AI detection tools, and foster international cooperation for cross-border regulation. Sharing experiences in these areas is vital for preserving the integrity of the information environment and ensuring the fairness of political processes such as elections.

3. Mobilize civil society and enhance tripartite cooperation:

Civil society must be engaged to play an active role in countering information manipulation. Governments, non-governmental organizations, and media outlets should collaborate effectively to mitigate the effects of information manipulation.

Introduction

Since its launch by OpenAI at the end of 2022, ChatGPT has brought GenAI into public view, proving to be a useful tool across various sectors. However, the capabilities of GenAI, including processing human language and generating insights, have also presented new challenges, one of which is its use for disinformation. Lina M. Khan, Chairwoman of the US Federal Trade Commission, highlighted in the New York Times the potential for GenAI to enable rapid and cost-effective production of fraudulent content.[1] Similarly, Sam Altman, CEO of OpenAI, expressed his concerns at a US congressional hearing about the potential for GenAI to facilitate widespread manipulation and fabrication of interactive messages.[2]

The consensus is clear: GenAI significantly lowers the barriers to information manipulation. This development has profound implications for democratic processes, particularly elections, where the integrity of information is crucial. The ability of GenAI to create persuasive, seemingly authentic content can influence public opinion and electoral outcomes, posing a serious threat to the maintenance of fair and free democracies. To address these issues, this report uses Taiwan's 2024 presidential election as a case study, illustrating how GenAI may have been used for information manipulation in democratic processes.

This report is structured to provide an in-depth analysis of information manipulation within the context of GenAI and its impact on democratic elections. It comprises four main sections:

- **Information manipulation and its frameworks:**

  This section starts by defining information manipulation and introduces various analytical frameworks—DISARM, BEND, and RICHDATA—that help elucidate what constitutes information manipulation. These frameworks are compared to highlight their unique perspectives and methodologies, ultimately aiding in a deeper understanding of the dynamics and scope of information manipulation.

- **Cases of AI use in Taiwan's 2024 presidential election:**

  This section explores specific instances where AI was used during the 2024 presidential election in Taiwan. It examines four cases: the 58-second audio recording of Ko Wen-je, the allegations against Lai Ching-te as an informant in the Chunfeng Project, a deepfake video linked to US Representative Wittman, and fabricated documents purporting to reveal "the secret history of Tsai Ing-wen." Each case study highlights how these AI technologies were strategically employed to sway public opinion and manipulate electoral outcomes.

- **Echoes of influence: Tracing AI in elections globally:**

  This section broadens the discussion to the global use of AI in manipulating elections, featuring recent cases taking place in India, the US, and France. It demonstrates the extensive reach and diverse applications of GenAI across different electoral contexts, underscoring its misuse within the wider global democratic processes.

- **Discussion and analysis:**

  This section investigates the role of GenAI in information manipulation by starting with an analysis of overarching tactical objectives. It then explores the specific purposes for which GenAI has been used, employing the DISARM framework to examine the techniques enabled by GenAI. The discussion concludes by assessing the current impact of GenAI and explores potential advancements in its application that could further amplify its societal influence in the future. This analysis provides a comprehensive understanding of the strategic implementation and potential escalation of GenAI-driven information manipulation.

I. Information Manipulation and its Frameworks

1. Defining Information Manipulation

The concept of information manipulation does not have a universally accepted definition across the world. This report synthesizes definitions from multiple sources, including the European External Action Service (EEAS), Taiwan's Information Environment Research Center (IORG), the governments of Canada and New Zealand, the DISARM Foundation, and the American Psychological Association, to offer a comprehensive definition. Information manipulation is hereby defined as a series of intentional acts aimed at adversely affecting the political environment of specific nations or social groups by distorting their informational context. By "political environment," this report means the internal political processes of a country as well as the political values embraced by its populace.[3]

According to this definition, information manipulation should not be viewed as isolated single messages or individual acts of dissemination but, rather, as a series of coordinated actions. For example, the New Zealand government specifically defines "disinformation" as the intentional spread of false or modified information.[4] Reflecting perspectives from the information security community on cyber-attacks,[5] this report advocates for understanding information manipulation as a strategic, interconnected sequence of actions. This approach aligns with that of the EEAS, which characterizes a pattern of disinformation activities as "a pattern of behavior." [6]

Manipulating informational contexts primarily involves actions that influence people's thoughts and decisions, going beyond merely disseminating specific information or messages. It includes a wide range of strategic activities and evaluations. This includes the preparatory steps before information is spread, such as topic selection, targeting specific audience groups, crafting messages, and choosing dissemination channels. Integral to the definition of information environment manipulation are actions such as encouraging targeted audiences to engage in offline activities, assessing the effectiveness of tactics, and concealing traces after information has been disseminated.

This report, therefore, sees information manipulation as a series of deliberate actions that alter the information environment, potentially causing adverse effects on the political environment of targeted countries or social groups. An individual who unknowingly shares false information as a result of information manipulation should not be viewed as an initiator of such an attack but, rather, as a victim of it.

Table 1 | Definitions of Information Manipulation

European External Action Service (EEAS)

European Union EXTERNAL ACTION

"Foreign Information Manipulation and Interference (FIMI) describes a mostly non-illegal pattern of behaviour that threatens or has the potential to negatively impact values, procedures and political processes. Such activity is manipulative in character, conducted in an intentional and coordinated manner, by state or non-state actors, including their proxies inside and outside of their own territory." [7]

Taiwan Information Environment Research Center (IORG)

IO RG 台灣資訊環境研究中心

"Information manipulation" refers to actions involving the dissemination of information that meets at least one of the following criteria:

1. **Source manipulation**: the sources of the information presented are confirmed to be false or cannot be verified as true.
2. **Fact manipulation**: the facts mentioned are proven to be false or partially false or cannot be verified as true.
3. **Inference manipulation**: the inferences drawn lack sufficient factual basis to support the conclusions.

Information manipulation may also involve "coordinated inauthentic behavior," characterized by the repeated dissemination of identical, or highly similar, content or the same links within a short period across news media, social media, and instant messaging platforms.[8]

Executive Yuan, Republic of China (Taiwan)

行政院 Executive Yuan

Fake news: maliciously fabricated and false information that causes harm.[9]

Government of Canada

Canada

"Information manipulation: the act of purposely changing, distorting, or controlling information to change the information environment. This can include partial or full omission of facts, doctored audio or visual content, inauthentic amplification of narratives, trolling, and efforts to censor or coerce self-censorship of information."

"Foreign disinformation: false information that is deliberately created and spread to mislead people, organizations and countries. It is often a part of broader information operations aimed at manipulating audiences. Disinformation is sometimes referred to as 'fake news,' though it can take many forms." [10]

Government of New Zealand

New Zealand Government

"Disinformation is false or modified information knowingly and deliberately shared to cause harm or achieve a broader aim." [11]

DISARM Foundation

DISARM FOUNDATION

"Disinformation: the deliberate promotion of false, misleading or mis-attributed information." [12]

American Psychological Association (APA)

AMERICAN PSYCHOLOGICAL ASSOCIATION

"Disinformation is false information which is deliberately intended to mislead—intentionally misstating the facts." [13]

2. Frameworks for Understanding Information Manipulation

Defining information manipulation goes beyond conceptual analysis. It involves how its practical manifestations are understood within specific social contexts. Therefore, determining what constitutes information manipulation must be based on objective evaluative judgments rather than subjective preferences shaped by various political communities or value systems.

This report adopts fraudulent tactics as a case study to demonstrate our approach to analyzing information manipulation. The analysis entails a comprehensive review of tactics extracted from multiple fraud cases, including an examination of the data types exploited by fraud syndicates, their methods of contacting victims, and the scripts they use.[14] This detailed scrutiny allows us to summarize prevalent fraudulent strategies effectively. By disseminating this knowledge, government agencies and civil organizations can better educate the public about these methods, increase vigilance, and inform key stakeholders, such as bank employees, about when to escalate their awareness.[15]

The process of identifying and preventing information manipulation closely resembles the steps involved in analyzing fraud. Initially, it entails collecting case studies that are suspected of containing false information. These cases are then subjected to a thorough analysis. Ideally, this examination of key cases helps deduce common methods of information manipulation, thereby informing potential strategies and methods for its prevention.

This report examines information manipulation through the framework of offense and defense. The strategic approaches of the two sides differ significantly. Information manipulators employ a top-down strategy: they establish a specific goal, pinpoint vulnerabilities in their target audience, and craft tactics accordingly to meet their objectives. In contrast, defenders begin with observable techniques, analyzing the attacker's tactical goals from the bottom up in an effort to accurately reconstruct the overall strategy of the attack.

Identifying practices of information manipulation is a challenging task, especially in today's politically polarized global environment. The lack of objective interpretation can deepen communication divides, obstructing dialog and consensus-building. In such a context, the need for a clear framework to discern information manipulation is crucial. This framework would provide objective standards for analyzing and summarizing specific cases, which would aid further analysis. It is important to note, as discussed later in this report, that various analytical frameworks may categorize and interpret similar manipulation tactics differently. This variation typically reflects different focuses on attack strategies and stages of manipulation by the creators of these frameworks, rather than mere subjective preferences.

When crafting a prevention strategy, it is crucial for defenders to first recognize their own vulnerabilities through a comprehensive analytical framework. This begins with understanding the attackers' most common tactics and identifying the targeted weaknesses they exploit. Armed with this knowledge, defenders can more effectively develop and deploy tailored detection tools and defenses. These tools might include systems for identifying false content—such as deceptive texts, recordings, and images—or mechanisms to uncover manipulative actions, such as bot accounts or websites spreading disinformation. In addition, defenders can employ long-term strategies, such as holding online platforms accountable and enhancing public education to foster a culture of critical media consumption and information literacy.

Employing a macro-analytical framework enhances defenders' ability to proactively counter information manipulation efforts. By identifying a common tactic employed by attackers, the defense can use the framework not only to detect concurrent strategies but also to predict the attackers' future actions and necessary resources. This proactive approach enables defenders to actively seek evidence of emerging tactics and anticipate the evolution of a given attack. Such strategic foresight is essential for timely response, especially against large-scale, organized disinformation campaigns, where the effectiveness of the defense often depends on the speed of its response. In Taiwan, for example, success in countering information manipulation often results from swift and effective clarifications provided by both government and civil society, as demonstrated by the case studies examined in this report.

This report examines three leading analytical frameworks: DISARM, developed by the DISARM Foundation; BEND, from the Center for Computational Analysis of Social and Organizational Systems (CACOS) at Carnegie Mellon University; and RICHDATA, by the Center for Security and Emerging Technology (CSET) think tank. All three organizations are based in the United States.

This report examines the application of these frameworks to help better grasp the practical processes of information manipulation in Taiwan's 2024 presidential election. It begins by detailing how these frameworks categorize various techniques of information manipulation and then discusses the rationale for selecting the DISARM framework for analyzing individual cases.

3. DISARM Framework

The DISARM Foundation, established in 2021, focuses on identifying patterns of information manipulation and devising effective countermeasures.[16] This initiative stems from the Credibility Coalition's misinformation working group, which compiled 63 international cases of information manipulation from 2012 to 2018.[17] These cases were initially analyzed in reference to the ATT&CK (Adversarial Tactics, Techniques & Common Knowledge) framework developed by MITRE, a US non-profit organization specializing in cybersecurity that was established in 2013. This led to the creation of the Adversarial Misinformation and Influence Tactics and Techniques (AMITT) framework.

In 2020, MITRE, in collaboration with Florida International University, adapted the AMITT framework to develop the Structured Process for Information Campaign Enhancement (SP!CE). A year later, the DISARM Foundation was formed to oversee and further develop what is now known as the DISARM framework.[18] By 2022, both the AMITT and SP!CE frameworks were integrated into DISARM, which quickly gained traction, being adopted by the EEAS and MITRE for cybersecurity applications. Presently, the data-sharing systems related to information manipulation that are used by US and EU authorities incorporate the DISARM framework.[19]

The DISARM framework is rooted in the field of information security and builds on the foundational ATT&CK framework developed by MITRE. The ATT&CK framework draws on military concepts, specifically, tactics, techniques, and procedures (TTPs). It explains the tactics that define the objectives of attackers; the techniques they employ; and the procedures they execute to achieve these goals.

The ATT&CK framework integrates TTPs along with essential knowledge and technologies needed to execute information security attacks. It includes principles such as firewall utilization and the psychological underpinnings of human deception. This comprehensive approach enables a broader understanding of the entire attack lifecycle. The creators of the DISARM framework draw parallels between information security and information manipulation, noting that, while information security attacks focus on computers and network systems, information manipulation targets individual minds and social networks. This similarity underscores the analytical approaches common to both domains, emphasizing the strategic targeting of systems, whether technological or human.[20]

The DISARM adopts a structured approach similar to the ATT&CK framework, segmenting the process of information manipulation into four distinct phases: planning, preparation, execution, and evaluation. At each stage, attackers have specific tactical objectives that are achieved through various methods. For example, during the preparation phase, a tactical goal might be "establishing the legitimacy of the attacker," which could involve creating a fake news website. For an attack to be successful, all actions within the so-called kill chain must be flawlessly executed. Consequently, implementing defensive strategies at multiple points in the kill chain, for example, "reducing the credibility of the fake news website," significantly lowers the likelihood of the attacker's success.

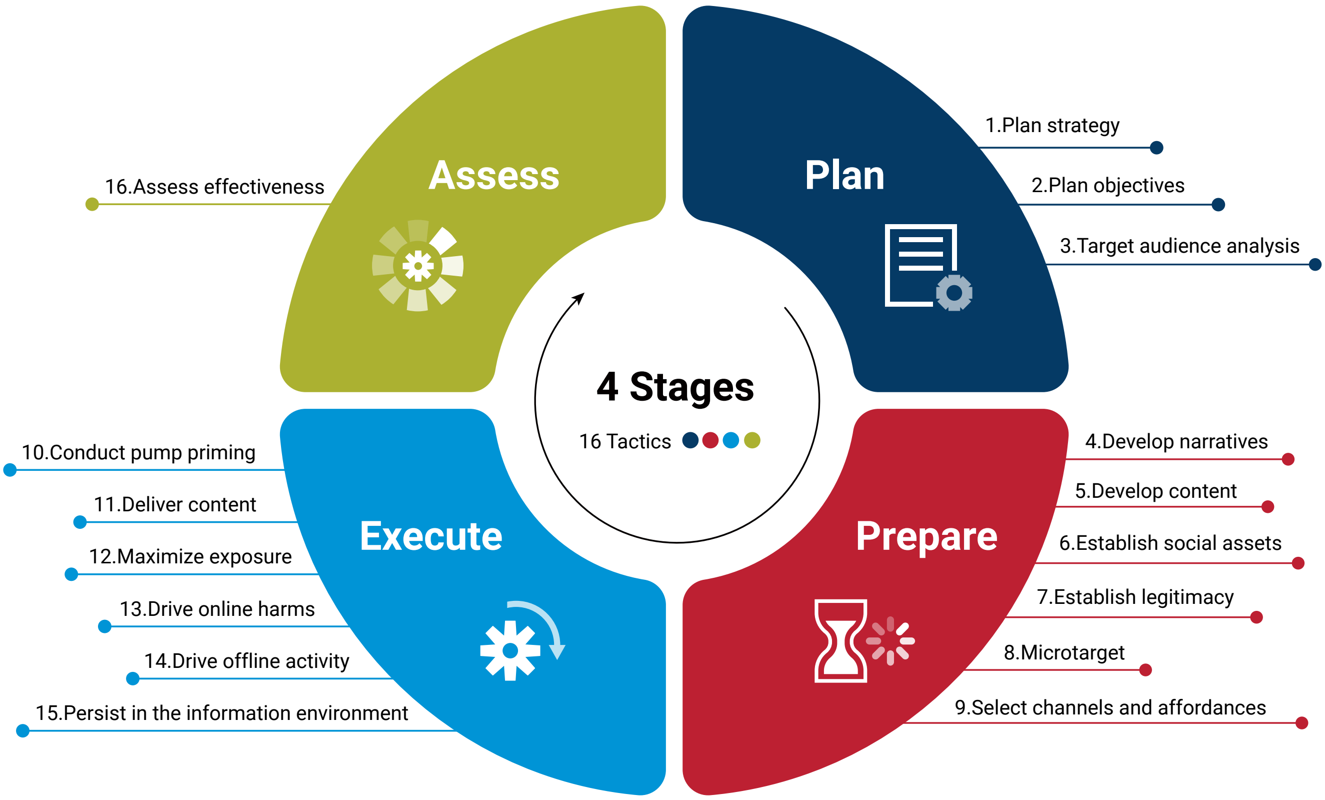
To describe the framework's four stages in more detail: the planning stage focuses on strategic planning, objective setting, and analyzing the target audience; the preparation stage involves developing narrative angles, creating content, establishing legitimacy, building foundational assets within social networks, targeting the audience, and selecting appropriate channels for information dissemination; the execution stage is designed to test the initial dissemination, spread content, proliferate extreme content, create online trauma, and instigate offline activities; and finally, the evaluation stage centers on assessing the effectiveness of the campaign. Each stage is critical in shaping the comprehensive approach required to manipulate information effectively.

The DISARM framework details 244 techniques across its 16 tactical objectives, drawing a parallel to the ATT&CK framework, which differentiates techniques based on their use by "red teams" (attackers) and "blue teams" (defenders). This structured approach not only delineates potential threats but also strategizes proactive defense mechanisms. Therefore, this categorization helps equip a democratic country such as Taiwan, which acts primarily in a defensive role, to effectively identify and implement countermeasures against information manipulation attacks.

For instance, in the context of AI-generated false information—for example, the development of AI-generated images and videos classified as deepfakes (T0086.002 and T0087.001) —proactive strategies such as "prebunking" (C00125) can significantly raise public awareness about AI-generated misinformation. In addition, disinformation campaigns often seek to exploit traditional media to broaden their influence (T0117: Attract traditional media). In response, enhancing media literacy (C00073: Inoculate populations through media literacy training) and urging media outlets to avoid disseminating false information (C00154: Ask media not to report false information) can serve as effective countermeasures to curb the spread and impact of such manipulative practices.

The EEAS has effectively adopted the DISARM framework. This system is used for detecting, analyzing, and documenting information manipulation. It helps identify threats and develop appropriate responses.[21] Building on this, the department has created a response framework specifically aimed at combating foreign information manipulation and interference (FIMI).[22]

Table 2 | DISARM Framework



4. BEND Framework

The BEND framework was developed by the Computational Analysis of Social and Organizational Systems (CASOS) research institute at Carnegie Mellon University (CMU).[23] Located in Pittsburgh, Pennsylvania, CMU has a strong emphasis on computer science, particularly AI and data science. The university conducts numerous interdisciplinary studies and has established collaborative courses and research centers on campus. These centers apply information science across various fields including social policy, finance, management, and decision science, contributing to the development of professionals who utilize computer science techniques to address complex issues.

CASOS is part of CMU's School of Computer Science. It is supported by faculty from five schools within the university. The center focuses on integrating computer science with social science, developing algorithms to analyze human social behavior, and employing information science techniques to predict and assess socio-cultural changes. These efforts support evidence-based public policy formulation. Key projects at CASOS have included using social network analysis to tackle complex social problems and developing predictive models through big data and machine learning.

The BEND framework is an analytical tool created by Kathleen M. Carley and developed by CASOS. It is primarily used to analyze community network security.[24] The framework addresses information manipulation and includes two main components: narrative manipulation and community network manipulation. Each component divides into eight categories, split between positive and negative intentions. This structure results in a total of 16 distinct practice patterns (See Table 3 for more details).

The BEND framework is designed to systematically evaluate community network security by analyzing behavioral patterns within networks and executing strategic influence measures. It serves to identify, analyze, and counteract information manipulation activities on social media platforms, aiming to prevent the dissemination of false information and to maintain the integrity of social structures. Specific case studies employing the BEND framework include the monitoring of social media group dynamics in Ukraine, the assessment of foreign influence campaigns in Syria, and the tracking of COVID-19 conspiracy theories. This framework provides a robust set of tools for understanding and mitigating the impacts of misinformation, thereby helping to safeguard public discourse from the spread of misleading content.[25]

Table 3 | BEND Framework

	Manipulating the narrative	Manipulating the social network
+ Positive	**Engage** Messages that bring up a related but relevant topic	**Back** Actions that increase the importance of the opinion leader or create a new opinion leader
	Explain Messages that provide details on or elaborate the topic	**Build** Actions that create a group or the appearance of a group
	Excite Messages that elicit a positive emotion such as joy or excitement	**Bridge** Actions that build a connection between two or more groups
	Enhance Messages that encourage the topic group to continue with the topic	**Boost** Actions that increase the size of the group or make it appear that it has grown
- Negative	**Dismiss** Messages about why the topic is not important	**Neutralize** Actions decrease the importance of the opinion leader
	Distort Messages that alter the main message of the topic	**Negate** Actions that lead to a group being dismantled or breaking up, or appearing to be broken up
	Dismay Messages that elicit a negative emotion such as sadness or anger	**Narrow** Actions that lead to a group becoming sequestered from other groups or marginalized
	Distract Discussion about a totally different and irrelevant topic	**Neglect** Actions that reduce the size of the group or make it appear that the group has shrunk

5. RICHDATA Framework

The RICHDATA framework, developed by CSET at Georgetown University in the United States, focuses on several key areas: the foundations of AI, including researchers, data, and computing resources; AI applications in defense; AI-related policy; and biotechnology. Established in 2019, CSET also explores the interaction between AI and cybersecurity. This includes the influence of cybersecurity on AI development, international competition in AI, and AI's role in information manipulation—all of which are pertinent to the scope of this report.

The RICHDATA framework systematically organizes methods of information manipulation into seven primary categories, each with specific subcategories that total 28 distinct actions. These are outlined as follows: [26]

1. Reconnaissance involves preliminary activities such as monitoring the environment, identifying intra-group conflicts, adapting narratives into targeted messages, and classifying both the target audience and the information channels they frequent.

2. Infrastructure focuses on the setup required for carrying out information campaigns, including the formation of digital armies, development of digital propaganda methods, construction of information dissemination channels, website creation, and the establishment of security systems for operational teams.

3. Content creation and hijacking pertains to the production of various types of content—textual, visual, and emotional—as well as the dissemination of fabricated information.

4. Deployment, uniquely, does not further break down into subcategories but involves the strategic release and management of the content created in the earlier stage.

5. Amplification seeks to extend the reach of content through sharing, automating propaganda, outsourcing tasks to private sector vendors, engaging influencers or super-spreaders, provoking discussions within the target audience, and embedding populist rhetoric.

6. Troll patrol involves controlled social media operations to manage discourse. This includes launching new accounts to steer conversations, fostering consensus or controversy, and manipulating platform algorithms to favor certain viewpoints.

7. Actualization transitions online efforts into the physical realm, which includes deactivating control over certain social media narratives, organizing events, live-streaming these events, and promoting active offline participation.

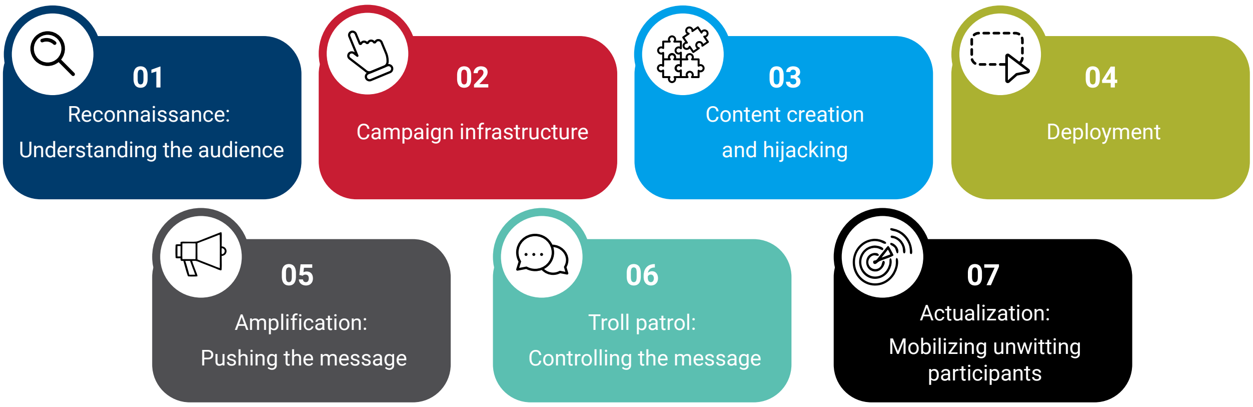
It should be noted that these seven categories defined by RICHDATA align with the four stages of the DISARM framework, from planning to evaluation. Many of the sub-items correspond directly to the tactical objectives or methods outlined within the DISARM framework.

6. Comparing DISARM, RICHDATA, and BEND

The examination of the DISARM, RICHDATA, and BEND frameworks reveals some major differences in their methodological approaches. DISARM and RICHDATA primarily organize manipulation tactics chronologically, focusing on the sequence of actions. In contrast, the BEND framework categorizes tactics based on their strategic intent, employing classifications such as "dismiss," "distort," "dismay," and "distract" to align tactics with the manipulator's objectives.

While the BEND framework excels at scrutinizing the characteristics of fraudulent messages and identifying the intent behind attacks, it is generally tailored for simpler scenarios of misinformation.

Table 4 | RICHDATA Framework



01 Reconnaissance: Understanding the audience

02 Campaign infrastructure

03 Content creation and hijacking

04 Deployment

05 Amplification: Pushing the message

06 Troll patrol: Controlling the message

07 Actualization: Mobilizing unwitting participants

This specificity is detailed in this report in Section 2 ("Defining Information Manipulation"), which emphasizes that understanding information manipulation requires recognizing it as a series of connected actions rather than isolated incidents. Consequently, the BEND framework, with its focus on discrete tactics, may not adequately address the complexities of election misinformation, which often involves nuanced and multifaceted strategies.

Given this context, the BEND framework has not been adopted for our detailed analysis of election misinformation in this report. Instead, the emphasis will be on frameworks that provide a more holistic view of information manipulation, suitable for addressing the intricacies of such cases.

Choosing between the RICHDATA and DISARM frameworks is more challenging due to their similarities. Both frameworks share objectives across DISARM's tactical categories and RICHDATA's techniques, indicating a high degree of overlap. Consequently, the principal distinction between the two does not stem from their classification methods. Instead, this report opts for the DISARM framework as the primary analytical tool for the following four reasons:

First, the DISARM framework, unlike RICHDATA, offers potential countermeasures for each manipulation tactic identified, providing a proactive defense strategy. Although it does not yet have countermeasures for every tactic, this proactive design proves invaluable for government and civil organizations committed to thwarting information manipulation efforts.

Second, the DISARM framework utilizes the STIX data format, widely recognized in cybersecurity for data storage and exchange. This compatibility facilitates the sharing of discovered tactics among analysts, enhancing collaborative efforts to address information manipulation across various contexts.

Third, the open-source nature of the DISARM framework underscores its adaptability and community-driven development. This flexibility allows users to refine and expand the framework in response to evolving manipulation methods, ensuring it remains relevant and effective.

Finally, the dynamic landscape of information manipulation necessitates continual updates to any effective framework. Unlike the RICHDATA framework, which has seen limited updates since its inception, the DISARM framework benefits from ongoing maintenance by the DISARM Foundation. Regular updates, such as the recent upgrade to Version 1.4 in March 2024, ensure it stays current with the latest trends in information manipulation. While the BEND framework is continuously maintained by CMU's CACOS and the RICHDATA framework is updated less frequently, the DISARM framework's commitment to timely updates and community involvement provides a compelling argument for its adoption in dynamic environments where up-to-date, actionable information is crucial.

International exchange and sharing of case studies significantly enhances the analysis of information manipulation. This practice not only consolidates diverse strategic approaches but also improves the analytical capabilities of all participating entities. For Taiwan, incorporating international case studies into discussions about information manipulation offers valuable insights. To facilitate this integration, it is essential to adopt a common language and analytical framework recognized by global cybersecurity communities and authoritative institutions. This approach provides two benefits: it allows for the use of internationally recognized data standards to share instances of information manipulation from Taiwan with the global community, and it enables learning from similar cases worldwide.

In fact, MITRE has effectively established numerous national databases of hacker attack methodologies using the ATT&CK framework, and these serve as an indispensable tool for the cybersecurity community to analyze attack and defense strategies and to exchange case studies.[27]

In summary, there is no global consensus on standards or norms for identifying information manipulation; nevertheless, the DISARM framework's standardized data format, open-source accessibility, and consistent updates position it as a strong candidate for adoption by international organizations.[28] Consequently, this report employs the DISARM framework to analyze the specific cases presented hereafter.

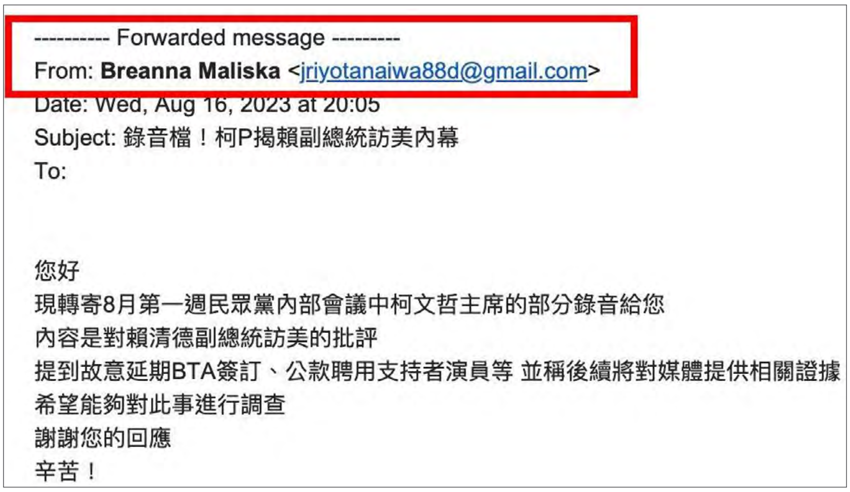
II. Cases of AI Use in Taiwan's 2024 Presidential Election

The 2024 Taiwanese presidential election took place on January 13, with candidates from three major parties: Lai Ching-te and Hsiao Bi-khim from the Democratic Progressive Party, Hou Yu-ih and Jaw Shaw-kong from the Kuomintang, and Ko Wen-je and Cynthia Wu from the Taiwan People's Party. This election witnessed several instances of information manipulation involving AI technology. This report examines four specific cases: a 58-second recording allegedly featuring Ko Wen-je, allegations that Lai Ching-te acted as an informant in the Chunfeng Project, a deepfake video depicting US Representative Wittman, and a fake document concerning undisclosed details of Tsai Ing-wen's past. These cases will be analyzed using the DISARM framework to identify and summarize the manipulation tactics employed.

7. Ko Wen-je's 58-Second Audio Recording

On August 16, 2023, at 20:05, an email was sent from a Gmail account with the username "Breanna Maliska" (jriyotanaiwa88d@gmail.com) to multiple major media outlets in Taiwan. The email was titled "Audio Record! Mayor Ko reveals secrets of Vice President Lai's visit to the US." It included an attachment named "Mayor Ko Audio.mp3," purportedly containing a recording from an internal meeting of the Taiwan People's Party in August, chaired by Mayor Ko Wen-je. See Figure 2 for more details.

At 21:24, ETToday News released a story with the headline "Breaking News/Blackmail Attack! A '58-Second Mysterious Audio File' Allegedly Featuring Ko Wen-je Criticizing Lai Ching-te, Prompting Immediate Clarification from Ko's Office." [29] Just five minutes later, at 21:29, the article was shared on PTT, a prominent Taiwanese online forum.[30] By 21:44, the story had spread to Facebook, where the fan page Rational Neutral Voter (李姓中壢選民) posted screenshots of some PTT comments for further



---------- Forwarded message ---------
From: **Breanna Maliska** <jriyotanaiwa88d@gmail.com>
Date: Wed, Aug 16, 2023 at 20:05
Subject: 錄音檔！柯P揭賴副總統訪美內幕
To:

您好
現轉寄8月第一週民眾黨內部會議中柯文哲主席的部分錄音給您
內容是對賴清德副總統訪美的批評
提到故意延期BTA簽訂、公款聘用支持者演員等 並稱後續將對媒體提供相關證據
希望能夠對此事進行調查
謝謝您的回應
辛苦！

Figure 2: Anonymous letter sent to media outlets. Photo sourced from Chen Chih-han's Facebook page

dissemination.[31] Finally, at 21:58, the article found its way to Dcard, a popular forum among Taiwanese students, under the provocative title "Democratic Progressive Party, Don't Go Too Far!" [32]

In response to the audio file, the Taiwan People's Party immediately clarified that it was a fake recording aiming to deceive and cause confusion, and that the matter had been reported for investigation. Spokesperson Chen Chih-han stated on Facebook that the recording's voice, speech speed, and terminology significantly differed from Ko Wen-je's usual manner. The Democratic Progressive Party has urged Ko to file a lawsuit to determine the truth and prevent misinformation from affecting the Taiwan elections.[33]

Table 5 | Chronology

16 AUG 2023	20:05	The Gmail account jriyotanaiwa88d@gmail.com dispatches an email to several major media outlets titled "Audio Record! Mayor Ko reveals secrets of Vice President Lai's visit to the US," which includes a 58-second audio file.
	21:24	ETtoday publishes a news article titled "Breaking News/Blackmail Attack! '58-Second Mysterious Audio File' Mimics Ko Wen-je's Voice Criticizing Lai Ching-te; Ko's Office Urgently Clarifies."
	21:29	A post titled "Breaking News/Blackmail Attack! '58-Second Mysterious Audio File' Mimics Ko Wen-je's Voice Criticizing Lai Ching-te; Ko's Office Issues Urgent Clarification" appears on PTT, sharing an article from ETtoday.
	21:44	The Facebook fan page Rational Neutral Voter (李姓中壢選民) shares a PTT post and includes selected comments from the discussion for further dissemination.
	21:50	Chen Chih-han, the spokesperson for the Taiwan People's Party, clarifies on Facebook that the voice, pace, and phrasing in the audio recording show significant differences from Ko Wen-je's own speech.
	21:58	An article titled "Democratic Progressive Party, don't go too far!" is posted on Dcard, reposting news from ETtoday.
25 AUG 2023		The Investigation Bureau issues a press release stating that the Taipei City Field Division, using deepfake detection software, initially determined that the content of the audio file was very likely to be a deepfake.

This report uses QSearch to track mentions of "Ko Wen-je," "audio file," "Lai Ching-te," "visit US," and "official visit" across different media platforms. The issue garnered attention primarily on August 16–18 and again on the 25th following the Investigation Bureau's response to the case. The distribution of articles across platforms suggests that the impact of the incident was moderate, likely mitigated by prompt clarifications.

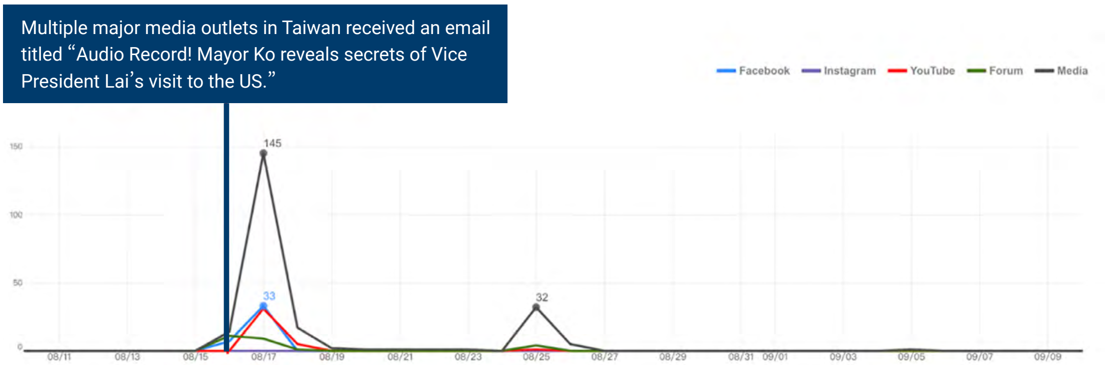

Multiple major media outlets in Taiwan received an email titled "Audio Record! Mayor Ko reveals secrets of Vice President Lai's visit to the US."

Figure 1: The FIMI Response Framework proposed in the EEAS report. Image source: 2nd EEAS Report on Foreign Information Manipulation and Interference Threats (see Footnote 22)

Verification Efforts

The Taiwan FactCheck Center and MyGoPen have not performed forensic analyses on the recording. However, the Taiwan FactCheck Center noted in an article titled "2024 Election Fact-Checking Notes Episode 1: Taiwan's First Pre-Election AI Fake Audio and Tips for Identifying Forged Audio and Video" that the recording appears disjointed, incomplete, and illogical.[35] On August 25, the Investigation Bureau announced that the Taipei City Investigation Office had used deepfake detection software to assess the recording, concluding that it was highly likely to have been significantly manipulated.[36]

Analysis of the Techniques

The tactics identified in the incident involving the 58-second forged audio file of Ko Wen-je, analyzed using the DISARM framework, include several key components. First, the use of an anonymous Gmail account is classified under Tactics T0090 (create inauthentic accounts) and T0112 (email dissemination). Second, the creation of a forged audio recording of Ko Wen-je fits Tactic T0088.001 (develop AI-Generated audio (deepfakes)). Finally, distributing the forged audio to various media outlets to garner news coverage corresponds to Tactic T0117 (attract traditional media).

8. Lai Ching-te Was an Informant of the Chunfeng Project

On December 22, 2023, a video entitled "Explosive! Lai Ching-te's Informant Status Exposed, Chunfeng Files and Audio Leaks" was posted on the YouTube account TrueTJL. The video featured an AI-generated voice purported to be that of retired investigator Lin Zhao-lun. It claimed that presidential candidate Lai Ching-te had been recruited as an informant for the Chunfeng Project, which conducted political surveillance, by Lin Zhao-lun, who was an investigator at the Taipei City Field Division in 1981. It alleged that Jin Guo-biao, the former director of the Bureau's Fifth Department, had confirmed this information.[37]

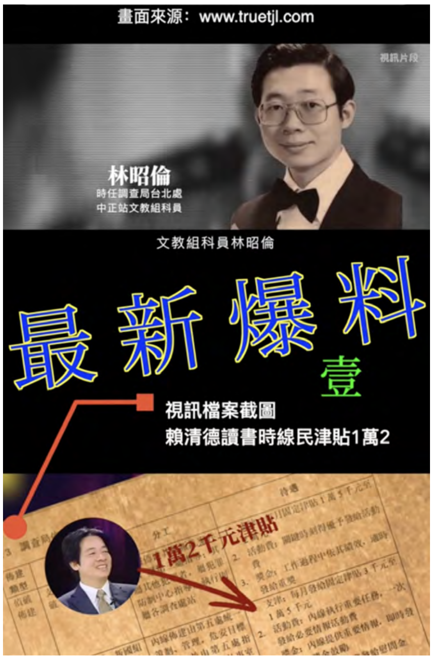


Figure 4: Screenshot from the video titled "Explosive! Lai Ching-te's Informant Status Exposed, Chunfeng Files and Audio Leaks." Source: TikTok

On December 22, following the video's release, an article titled "Explosive! Lai Ching-te's Informant Status Exposed, Chunfeng Files and Audio Leaks" appeared on PTT.[38] Subsequently, on December 26, Chiu Yi posted on Facebook, referring to Lai Ching-te as "actually the biggest informant," [39] and on December 27, he posted an article on PTT's Politics Blackboard titled "Is Lai Ching-te Also an Undercover Agent Sent by the KMT?" This post suggested that Lai Ching-te was an informant for the Investigation Bureau.[40] On December 29, 2023, Dennis Peng's YouTube channel, True Voice of Taiwan, featured an interview with Chang Yu-hua. During the interview, they discussed the content of the TrueTJL video, and Peng claimed that Lai Ching-te was an informant for the Bureau of Investigation.[41]

Table 6 | Chronology

04 DEC 2023	The True Transitional Justice League website was launched, hosted on the IP address 123.253.32.123, which is located in China. Concurrently, the YouTube account TrueTJL was created.
22 DEC 2023	The YouTube account TrueTJL uploaded a video titled "Explosive! Lai Ching-te's Informant Status Exposed, Chunfeng Files and Audio Leaks."
23 DEC 2023	An article appeared on PTT titled "Lai Ching-te's Informant Status Exposed, Chunfeng Files and Audio Leaks."
25 DEC 2023	The Investigation Bureau's Cyber Security Division issued a press release to clarify the situation.
27 DEC 2023	An article titled "Is Lai Ching-te a KMT Undercover Agent?" was posted on the PTT Politics Forum, alleging that Lai Ching-te is an informant for the Investigation Bureau.

This report used QSearch to retrieve articles related to the keywords "Lai Ching-te" OR "Informant Lai Ching-te (線民德)" AND "Chunfeng File OR Chunfeng Project" across various platforms. Before the release of a controversial video on December 22, the YouTube channel Ou Chung-ching had frequently discussed the issue.[42] In response to the video, the Investigation Bureau issued a clarifying press release on December 25.[43] Subsequently, a total of 29 articles was published on this topic on December 25–26, after which the volume of discussion rapidly declined. Given the swift official clarification and the number of articles referencing the topic across platforms, the credibility of the original anonymous source was perceived as relatively low, resulting in the incident having a limited dissemination impact.

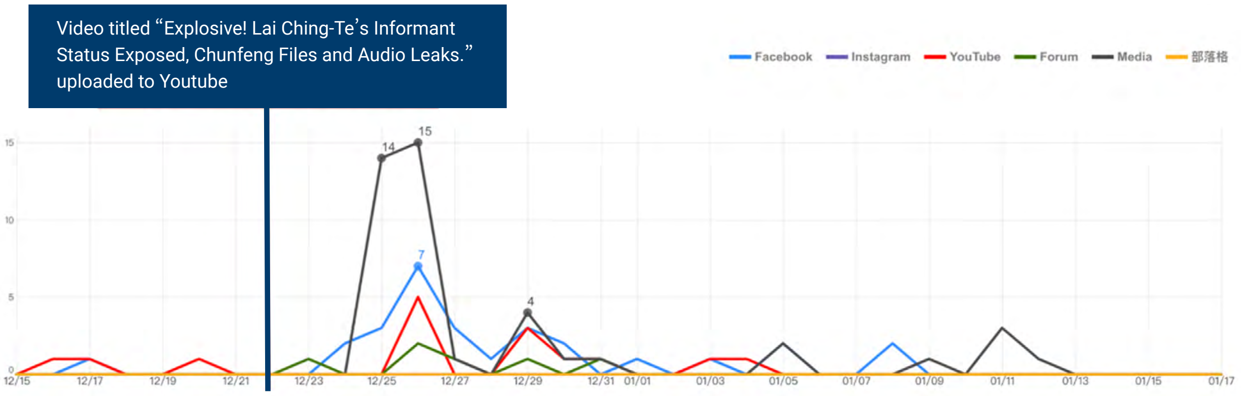

Video titled "Explosive! Lai Ching-Te's Informant Status Exposed, Chunfeng Files and Audio Leaks." uploaded to Youtube

Figure 5: Number of cross-platform articles on the QSearch platform regarding the Lai Ching-te Chunfeng Project incident

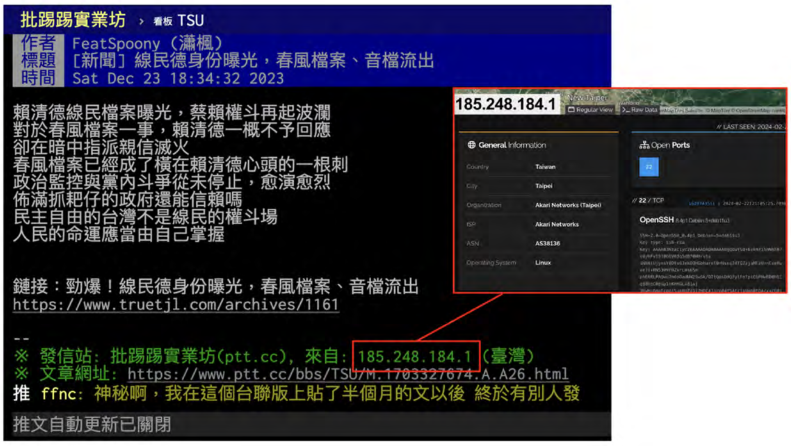
Fact-Checking Processes

After the release of the video claiming Lai Ching-te was an informant for the Chunfeng Project, the Information Security Department of the Investigation Bureau issued a press release on December 25, 2023. The release highlighted that retired investigator Lin Zhao-lun had contacted the Taipei City Field Division to clarify that the voice attributed to him in the video was not his. The release also addressed inaccuracies regarding the former head of the Investigation Bureau, Jin Guo-biao, noting discrepancies with factual accounts, particularly as Jin had been deceased for some time.[44]

The fact-checking platform MyGoPen has discovered that the video posted by the TrueTJL account included a link to truetjl.com, billed as the "True Transformation Justice Alliance." This website's domain was registered through the American registrar Namecheap on December 4, 2023. The server hosting the site, with the IP address 123.253.32.123, is based in China.[45]

Shortly after the video's release on December 22, a post surfaced on the PTT forum on December 23 titled "Lai Ching-te's Informant Status Exposed, Chunfeng Files and Audio Leaks." [46] The post's IP address, 185.248.184.1, points to a Linux server with an open SSH Port—uncommon for a personal computer. This unusual setup suggests that the post may have been published from a Linux server, potentially serving as a springboard for its dissemination.

After the Investigation Bureau issued its press release, the "Truth Transitional Justice Alliance" quickly released another video, titled "From Shameless Shack (賴皮



批踢踢實業坊 › 看板 TSU

作者 FeatSpoony (瀟楓)
標題 [新聞] 線民德身份曝光，春風檔案、音檔流出
時間 Sat Dec 23 18:34:32 2023

賴清德線民檔案曝光，蔡賴權斗再起波瀾
對於春風檔案一事，賴清德一概不予回應
卻在暗中指派親信滅火
春風檔案已經成了橫在賴清德心頭的一根刺
政治監控與黨內斗爭從未停止，愈演愈烈
佈滿抓耙仔的政府還能信賴嗎
民主自由的台灣不是線民的權斗場
人民的命運應當由自己掌握

鏈接：勁爆！線民德身份曝光，春風檔案、音檔流出
https://www.truetjl.com/archives/1161

--
※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 185.248.184.1 (臺灣)
※ 文章網址: https://www.ptt.cc/bbs/TSU/M.1703327674.A.A26.html
推 ffnc: 神秘啊，我在這個台聯版上貼了半個月的文以後 終於有別人發
推文自動更新已關閉

185.248.184.1

// LAST SEEN: 2024-02-

General Information

Country Taiwan
City Taipei
Organization Akari Networks (Taipei)
ISP Akari Networks
ASN AS38136
Operating System Linux

Open Ports

22

// 22 / TCP

OpenSSH 8.4p1 Debian 5+deb11u3

Figure 6: The IP of the post published in real time on PTT is from a Linux server

寮）to Chunfeng Files: The True Face of Lai Ching-te Revealed, Strains Between Tsai and Lai Exposed." The website then issued a warning: "Unless Lai publicly acknowledges his informant role, we will continue to release Chunfeng Files data." Subsequently, updates ceased, and the website was shut down, becoming inaccessible.

The dissemination of the video titled "Lai Ching-te Was an Informant of the Chunfeng Project" raises several red flags. This video was simultaneously shared by multiple YouTube accounts under dubious circumstances. On December 25, eight YouTube channels uploaded the video. Three of these accounts were created on the very same day, while two pre-existing channels had either never posted before or only had a single previous upload. Two other channels, which had historically posted content unrelated to Taiwanese politics, also distributed the video. The Taiwan FactCheck Center suggested, through an account investigation platform, that the channels likely belong to a foreign entity, indicating a strategy involving the use of newly created or disposable channels to spread the video.[47]

Analysis of the Techniques

Since August 2023, media outlets had circulated claims that "Lai Ching-te was an informant of the Chunfeng Project network," suggesting that this was likely a longstanding piece of misinformation. The attackers employed a specific strategy, involving the creation of new websites, YouTube channels, Facebook fan pages, and other digital platforms. On these platforms, they uploaded audio files and videos fabricated using AI technology. This approach was designed to enhance the reach and effect of the misinformation, with the ultimate goal of discrediting a specific political candidate.

This report applies the DISARM framework to deconstruct the methodologies employed in this incident, identifying multiple attack tactics used by the operator. First, the tactic involved referring to Lai Ching-te as an informant of the Chunfeng Project, accompanied by leaks of audio and video from the TrueTJL YouTube account. This method aligns with Strategy T0088.001 of the framework (develop AI-generated audio). Second, the creation of the "True Transitional Justice League" website and the establishment of the TrueTJL YouTube account are covered under Strategies T0098.001 and T0090.001, which cover the creation of inauthentic news sites and the creation of anonymous accounts, respectively.

Furthermore, the operation leveraged pre-existing narratives already circulating on media platforms prior to the release of the materials, such as Temple Talk and statements by Ou Chung-ching labeling Lai Ching-te as a pivotal figure in the Chunfeng Project. These actions are categorized under Tactic T0022.001 (amplify existing conspiracy theory narratives). Finally, the incident involved impersonations of Lin Zhao-lun and Jin Guo-biao in the audio files, a maneuver falling under Strategy T0009, whereby pseudo-experts are employed to lend false credibility to the disinformation.

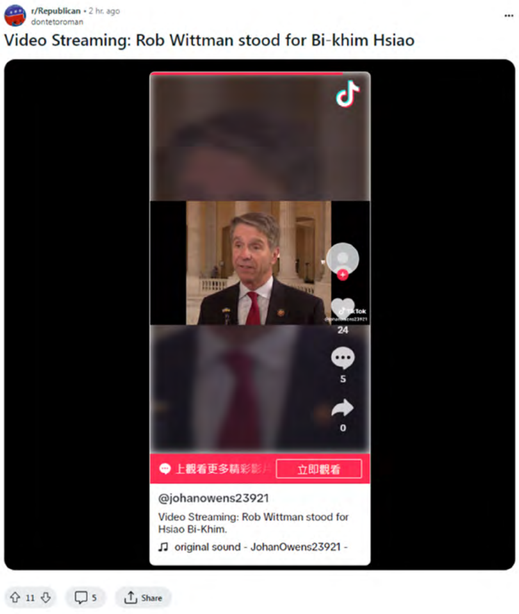
In the later stages of the operation, the account actively uploaded videos to platforms such as Facebook and YouTube (T0105.002: Video sharing) and propagated related news through its websites (T0049.007: Inauthentic sites amplify news and narratives), aiming to magnify the message's impact. Furthermore, the Taiwan FactCheck Center noted the use of newly established or acquired disposable accounts (T0090: Create inauthentic accounts) to simultaneously upload videos on YouTube, thereby attempting to broaden the message's dissemination.

9. Deepfake Video Linked to US Representative Wittman

Chronology

On December 29, 2022, at 13:02 Taipei time, a post appeared on Reddit's Republican and Taiwan forums, titled "Rob Wittman Stood for Bi-khim Hsiao." It featured a video clip that was a montage of an interview with Rob Wittman, the vice chairman of the United States House Armed Services Committee, conducted on March 2, 2022 by Washington, D.C.'s WUSA9 television station. The video included an audio segment that seemed to have been created using deepfake technology, purporting to show Wittman expressing support for Bi-khim Hsiao.[48]

On the same day at 17:56, a post appeared on the PTT platform under the username Godisme73, claiming, "The US has taken sides, releasing videos supporting Lai-Hsiao's election," accompanied by a link to the video.[49] Shortly after, Mobile01 featured an article with the headline "PTT Breaking News: the US sides with DPP" by user Qm671006.[50] Later that evening at 23:51, the Facebook fan page I Am Taiwanese, Taiwan Is Our Country uploaded the same video, stating, "Vice chairman of the United States House Armed Services Committee endorses Bi-khim Hsiao." Screenshots shared by users indicate that the video was widely circulated across various Facebook communities.[51]


r/Republican • 2 hr. ago
dontetoroman

Video Streaming: Rob Wittman stood for Bi-khim Hsiao

@johanowens23921
Video Streaming: Rob Wittman stood for Hsiao Bi-Khim.
♫ original sound - JohanOwens23921 -

Figure 7: Screenshot of the original post appearing on Reddit.
Image Source: PTT

Table 7 | Chronology

29
DEC
2023

13:02	An article titled "Rob Wittman Stood for Bi-khim Hsiao" appeared on the Republican and Taiwan subreddits on Reddit.
17:56	An article titled "The US Takes Sides: Video Released Supporting Lai-Hsiao Election" appeared on PTT.
19:08	An article titled "PTT Reveals: US Sides with and Supports the Democratic Progressive Party" appeared on Mobile01.
23:51	A video was uploaded to the Facebook fan page I Am Taiwanese, Taiwan Is Our Country, claiming, "US Military Committee Vice Chairman Appears on Camera in Support of Bi-khim Hsiao."

On December 30, the Taiwan FactCheck Center confirmed that the video was fabricated.[52] Meanwhile, users on PTT posted analyses under titles such as "Lobbying the US for Election Intervention or Flank Groups Fabricating Videos to Deceive Supporters," scrutinizing the video for signs of forgery to challenge the narrative of "the US taking sides." [53] Following this, several users circulated the Taiwan FactCheck Center's findings for further clarification.[54]

Our team utilized QSearch to investigate the presence of content related to the keywords "Vice Chairman of the United States House Armed Services Committee," "Rob Wittman," AND "United States & Endorse" across various platforms from December 27 to January 6. We identified only one fan page within QSearch's monitoring scope that had shared this video. Other reposts, made by individuals in various groups, fell outside QSearch's monitoring capabilities.[55] Despite this, traffic data indicates that major news websites primarily covered the clarifications issued by the Taiwan FactCheck Center, suggesting that the misinformation incident did not achieve widespread dissemination.

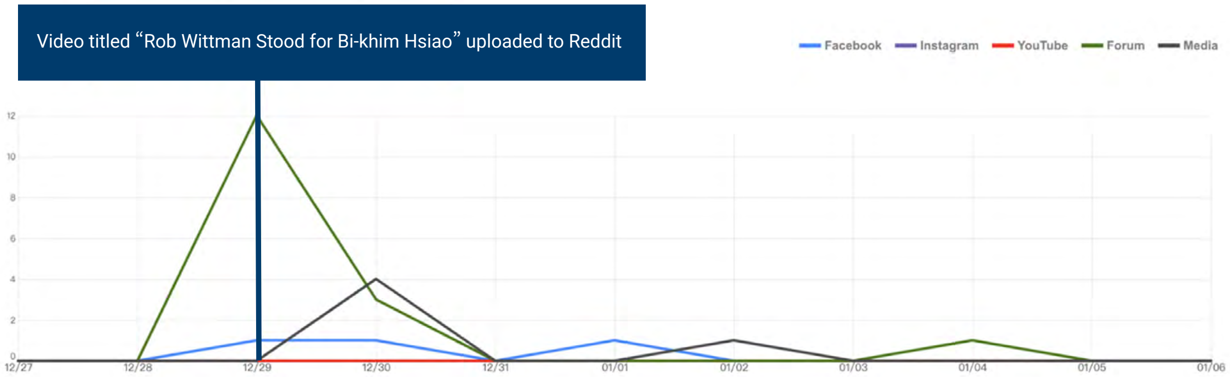


Figure 9: Number of cross-platform articles on the QSearch platform for this case study

Fact-Checking Processes

The Taiwan FactCheck Center has confirmed that the video in question was altered from an interview with Wittman conducted by WUSA9, a television station in Washington, D.C., on March 2, 2022. In the original interview, Wittman discussed the Ukrainian crisis, the American economy, and recovery from COVID-19, without any mention of the presidential election in Taiwan.

Taiwan FactCheck Center's analysis revealed that the authentic video was only 20 seconds long and included a brief appearance of a man in a black suit, indicating that the altered video was spliced from two different segments of the original broadcast. Discrepancies in Wittman's voice and lip movements between the altered and original footage suggest that his lip movements and voice were manipulated in the edited version.

Furthermore, the Taiwan FactCheck Center conducted a search for Wittman's public statements and news coverage on December 29 and found no evidence that he publicly campaigned for the Democratic Progressive Party. Consequently, the Center concluded that the information presented in the edited video was false.[56]

Analysis of the Techniques

In the Wittman case, the attacker employed three primary manipulation tactics. First, the "reuse existing content" tactic, coded as T0084 under the DISARM framework, involved using sourced interview clips from TV stations; second, they downloaded Wittman's audio and altered his mouth movements in the video using AI deepfake technology, corresponding to the "develop AI-generated videos (deepfakes)" tactic, coded as T0087.001; third, the uploader disseminated Wittman's video on Reddit, which was subsequently cross-posted to Taiwan's PTT and Facebook platforms. This action aligns with the "cross-posting" tactic, coded as T0119, which aims to amplify the video's reach and impact.

10. "The Secret History of Tsai Ing-wen"

Chronology

On January 2, 2024, an anonymous individual uploaded a 318-page PDF file titled "The Secret History of Tsai Ing-wen" to the document publishing platform Zenodo. Shortly thereafter, numerous accounts began disseminating the document across social media platforms, including Twitter and Facebook. Several anonymous accounts shared the file with opinion leaders in the Chinese-speaking world.[57]

Simultaneously, online media platforms such as Mirror Fiction began experiencing multiple account invasions, leading to the uploading of documents titled "The Secret History of Tsai Ing-wen." [58] Other platforms, including Wikipedia, Pixnet, Breaking News Commune, TikTok, PTT, and Vocus also saw consecutive postings of articles containing content from "The Secret History of Tsai Ing-wen." [59]

Following the dissemination of these documents, an influx of new users joined YouTube between January 4 and January 10, 2024. These accounts posted a series of videos citing the "Secret History of Tsai Ing-wen." According to a survey by the Australian Strategic Policy Institute (ASPI), up to 490 videos were uploaded by these accounts.[60] The content used GenAI virtual broadcasters, mimicking news reporting styles to recount scenarios from the "Secret History of Tsai Ing-wen" documents. YouTube has since suspended these channels.

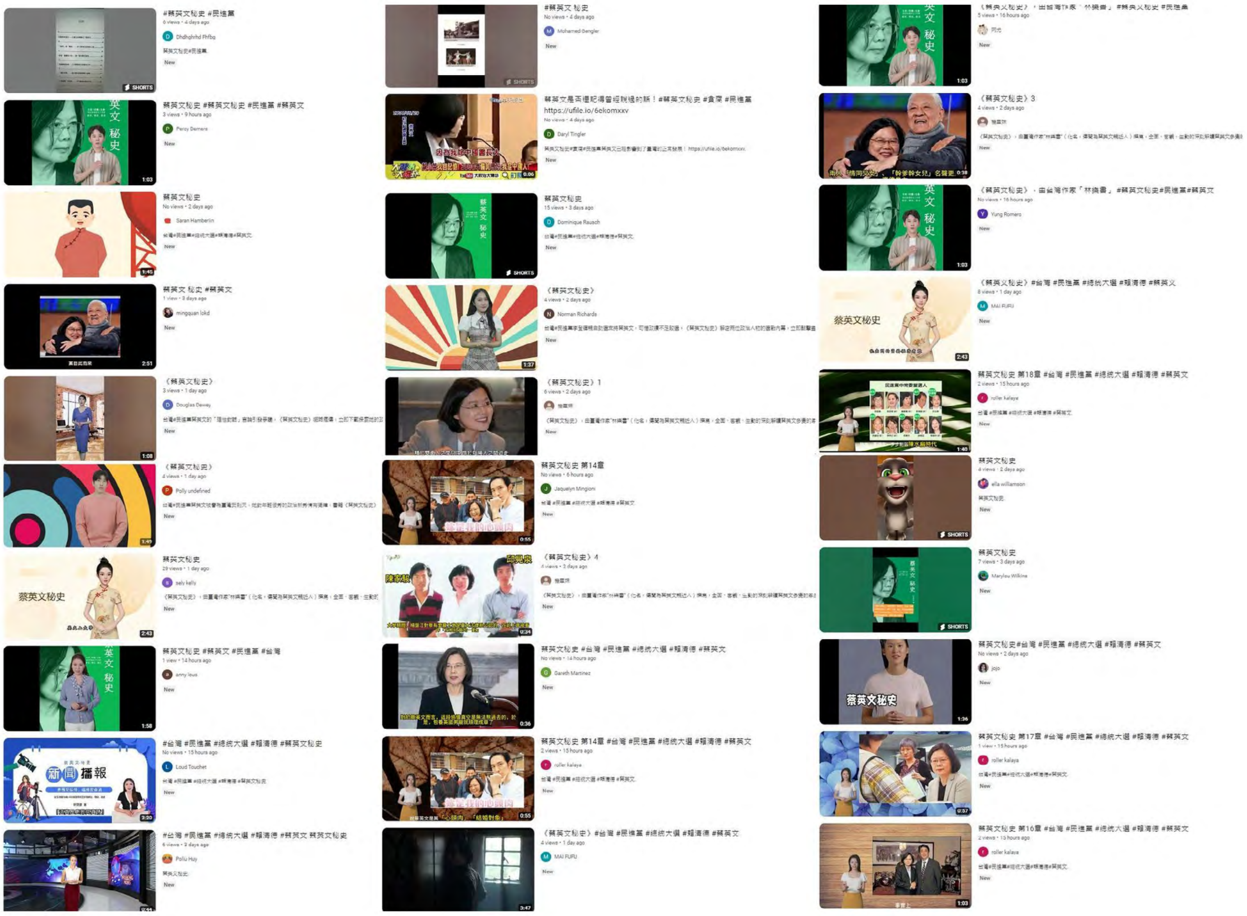
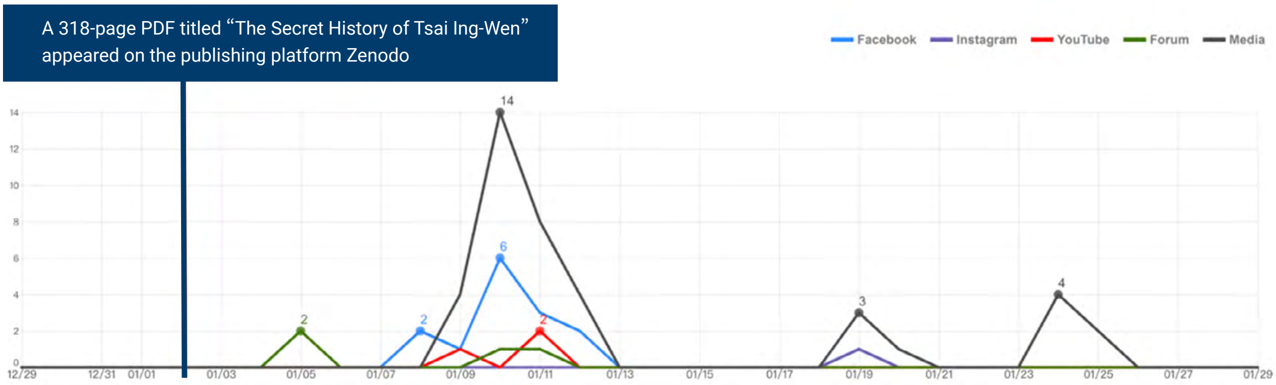


Figure 10: The video titled "The Secret History of Tsai Ing-wen" uploaded on the YouTube platform. Image source: Austin Horng-En Wang's Facebook (王宏恩臉書) 61

Table 8 | Chronology

02 JAN 2024	A 318-page PDF titled "The Secret History of Tsai Ing-wen" appeared on the publishing platform Zenodo.
04 JAN 2024	A large number of fake accounts began uploading videos titled "The Secret History of Tsai Ing-wen" to YouTube.
05 JAN 2024	On the Politics Blackboard section of PTT, a post appeared titled "FB Sparks a 'Secret History' Craze!" that shared a link to the eBook "The Secret History of Tsai Ing-wen."

This report used QSearch to search for the keyword "The Secret History of Tsai Ing-wen" and review the number of related articles across different platforms at the time. Most accounts posting related texts and videos were newly registered and thus were not on the QSearch monitoring list. Data collected by QSearch primarily consisted of news reports and clarifications. The peak of the topic occurred on January 10, 2024, when the incident was reported in the news and clarified by national security units. Based on the article sharing status, the incident did not result in widespread dissemination.



A 318-page PDF titled "The Secret History of Tsai Ing-Wen" appeared on the publishing platform Zenodo

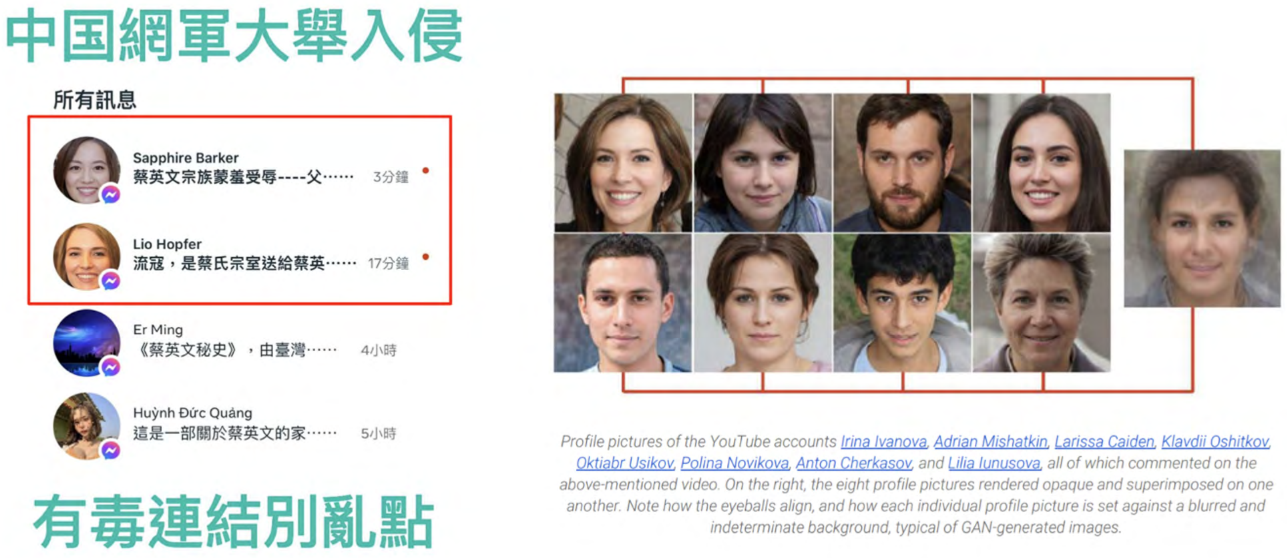
Figure 11: Number of cross-platform articles in the case study of "The Secret History of Tsai Ing-wen" on the QSearch platform

News Verification Processes

According to an analysis by ASPI, the metadata within the PDF of "The Secret History of Tsai Ing-wen" confirms that the document was edited using WPS Word, a piece of word processing software developed by China's Kingsoft Corporation. This software is predominantly used in China.[62]

The content and format of the "Secret History of Tsai Ing-wen" video uploaded on YouTube is complex. According to a report by Radio Taiwan International, Taiwan's national security unit highlighted that the video uses numerous Chinese idioms, such as "Tsai's regime," "island residents," "Trump," and "wild behavior." These terms differ from common expressions in Taiwan.[63] Some videos were edited using "CapCut," a piece of software developed by China's ByteDance, and feature a broadcaster created by D-ID, an American software company, using artificial intelligence.[64] However, since these videos indicate that the D-ID platform was used without payment, they suggest no partnership between the attacker and D-ID. Other videos utilized "Qimiao Meta," a platform by the Chinese company Chumen Wenwen, to create an AI anchor. The ASPI noted that case studies on the platform's website show that it is used by China's police and public security system to produce training videos.[65]

The dissemination of the "Secret History of Tsai Ing-wen" video involved numerous fake accounts. Keelung City Councilor Jiho Chang revealed on Facebook that several accounts sent him messages about the video, with two of these accounts appearing to use AI-generated profile pictures, characterized by identical eye and mouth positions.[66] Taiwan AI Labs analyzed the YouTube accounts posting the video and found that most were created between January 4 and January 8.[67]



中国網軍大舉入侵

所有訊息

Sapphire Barker
蔡英文宗族蒙羞受辱----父······ 3分鐘

Lio Hopfer
流寇，是蔡氏宗室送給蔡英······ 17分鐘

Er Ming
《蔡英文秘史》，由臺灣······ 4小時

Huỳnh Đức Quảng
這是一部關於蔡英文的家······ 5小時

有毒連結別亂點

Profile pictures of the YouTube accounts Irina Ivanova, Adrian Mishatkin, Larissa Caiden, Klavdii Oshitkov, Oktiabr Usikov, Polina Novikova, Anton Cherkasov, and Lilia Iunusova, all of which commented on the above-mentioned video. On the right, the eight profile pictures rendered opaque and superimposed on one another. Note how the eyeballs align, and how each individual profile picture is set against a blurred and indeterminate background, typical of GAN-generated images.

Figure 12: Some content shared on the account of Keelung City Councilor Jiho Chang (2024) appears to be generated by AI, reflecting the Spamouflage pattern previously identified by Graphika (right)

ASPI has identified that the dissemination of the "Secret History of Tsai Ing-wen" on platforms such as Twitter, Facebook, and YouTube is predominantly linked to a fake social network known as "Spamouflage." The Zenodo platform, used to upload the "Secret History of Tsai Ing-wen" PDF, has also been used by Spamouflage to upload a document alleging that COVID-19 originated in the United States, indicating a connection to Spamouflage's operations.[68]

Analysis of the Techniques

The attacker employed five principal techniques during the preparation stage of this incident. First, they fabricated a spurious "secret history of Tsai Ing-wen," attributed to a counterfeit researcher, using Tactics T0009 (Create fake experts) and T0022.002 (develop original conspiracy theory narratives). Second, the attacker uploaded the document to Zenodo and created a series of videos simulating news programs discussing its content, employing tactics T0023.002 (edit open-source content) and T0087.001 (develop AI-generated videos (deepfakes)). Lastly, numerous accounts were registered across various platforms, representing Tactic T0090 (create inauthentic accounts), to facilitate subsequent comments, posts, and message dissemination by the attacker.

During the execution phase of the attack, the attacker sequentially commented on Facebook using the fake accounts they had previously set up. This tactic falls under T0116, which involves commenting on or replying to content. They then reposted the content to platforms such as Vocus and PTT or uploaded videos to YouTube, corresponding with T0115, the posting of content. The attacker manipulated the fake accounts to send content to individuals such as Keelung City Councilor Jiho Chang, aiming to provoke a counterattack in the form of sharing. These actions are categorized as T0039, the tactic of baiting legitimate influencers.

Viewed holistically, the attackers aim to inundate social platforms with information about the "secret history of Tsai Ing-wen" (T0049: Flooding the information space) to manipulate the platform's algorithms (T0121:

Manipulate platform algorithm) and increase the visibility of this information to wider audiences.

11. Limited Impact

Upon reviewing the cases from Taiwan's 2024 presidential election, it becomes clear that numerous instances of information manipulation utilized AI. These primarily targeted scandals involving candidates, with only one case falsely representing political statements through fabricated news reports. Methodologically, these instances often employed AI deepfake technology to produce audio or video clips. The dissemination typically began with anonymous accounts, which created breakpoints in information flow. Some of these accounts used AI-generated avatars to enhance their human-like appearance, thereby amplifying their ability to spread disinformation.

The analysis of four cases by QSearch indicates that their impact was minimal and did not result in prolonged, heated discussions. This report suggests that the use of anonymous accounts to disseminate information contributed to the lack of traditional media coverage. Moreover, rapid and timely clarifications through media exposure helped to minimize the impact of these information manipulations.

Table 9 | Cases of AI-Driven Information Manipulation in the 2024 Taiwan Presidential Election

01 Cases	02 Contents	03 Conduits	04 Execution	05 Responses
Ko Wen-je's 58-Second Audio Recording	Using GenAI to create deepfake audio recordings	Registering an anonymous Gmail account	Sending emails to media outlets	Immediate clarification
Lai Ching-te Was an Informant for the Chunfeng Project	Incorporating AI-generated deepfake audio recordings into a video	Creating an anonymous website; registering anonymous YouTube and Facebook accounts	Releasing videos via anonymous websites and accounts	The Investigation Bureau issued a press release to provide clarification
Deepfake Audio-Visual Content Linked to US Representative Wittman	Using GenAI to modify existing news videos	Using existing accounts	Posted on Reddit and subsequently shared on other platforms	The Taiwan FactCheck Center issued a clarification
The Secret History of Tsai Ing-wen	Using GenAI to mass-produce videos	Registering anonymous accounts, some of which use AI-generated avatars	Spreading information using anonymous accounts by sending messages, posting updates, and uploading videos	The media interviewed national security personnel for clarification

III. Echoes of Influence: Tracing AI in Elections Globally

To better understand how GenAI could influence information manipulation during elections, this report includes several notable international cases alongside the Taiwan example. The objective is to expand readers' awareness of election-related information manipulation tactics and to illustrate potential challenges that Taiwan might encounter in future elections.

12. Deepfake Applications in India's Elections

Elections in India have long been intertwined with the use of technology. As early as 2012, the Bharatiya Janata Party (BJP) employed 3D holographic projections of Narendra Modi to enable him to deliver campaign speeches simultaneously across multiple locations. This technology saw extensive use during Modi's 2014 nationwide campaign for prime minister. Importantly, while this technology significantly accelerates the dissemination of information, it does not alter the content of the messages.

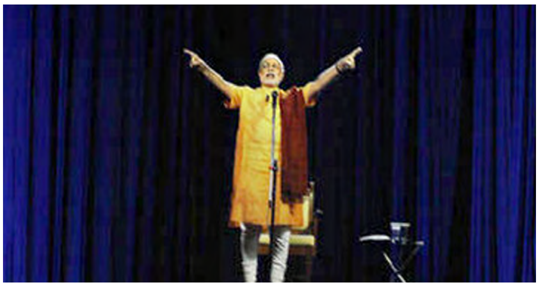


Figure 13: Modi uses 3D holographic technology for simultaneous campaigns across multiple venues. Source: Narendra Modi (2013)

In February 2020, actor-turned-politician and Bharatiya Janata Party legislator Manoj Tiwari employed deepfake technology in his election campaign. Before the local legislative assembly elections, Tiwari engaged Delhi's diverse ethno-social groups with three videos in Hindi, Haryanvi, and English. While the Hindi video was originally recorded, the other two were created using GenAI. This technology altered lip movements and facial expressions to increase the authenticity of the videos.[69]

In October 2020, an AI startup called The Indian Deepfaker was launched.[70] The company leveraged AI technology to replicate the voice of Ashok Gehlot, a chief minister candidate for Rajasthan's state assembly. During the November assembly elections, Gehlot's campaign team used this technology to distribute personalized videos to individual voters through WhatsApp. Each video began by addressing the voter by name, aiming to forge a closer connection between the candidate and the voters and enhance their sense of personal engagement.[71]

The Dravida Munnetra Kazhagam (DMK) in India was led by Muthuvel Karunanidhi from 1969 until his death in 2018. Known as the spiritual leader of the DMK, Karunanidhi posthumously appeared in a video in January 2024, created using AI deepfake technology. In the video, he congratulated a political colleague on the publication of a new book and lavishly praised his son, the incumbent DMK president and Tamil Nadu Chief Minister Muthuvel Karunanidhi Stalin, for his significant political achievements.[72]

In fact, the late Karunanidhi was "revived" multiple times using deepfake technology during 2023–24. Through a series of deepfake speeches, he supported his son's political reputation. The prevalent use of deepfake technology increasingly blurs the distinctions between truth and falsehood, reality and virtuality, perception and fantasy. However, as noted by an unnamed technical consultant in India, "no political party considers manipulating voters through artificial intelligence a crime; it is merely part of campaign strategy." [73]

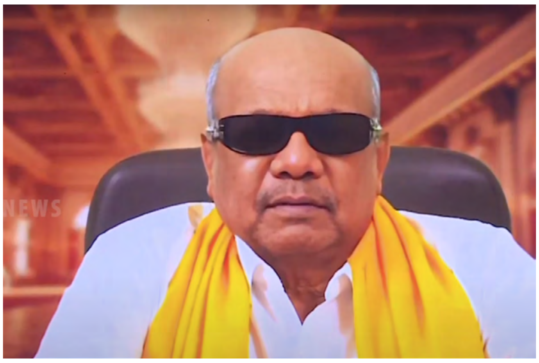


Figure 14: The late political figure Karunanidhi has been resurrected through deepfake technology, making appearances to praise his son's political achievements. Source: Nilesh Christopher (see Footnote 72)

Analysis of the Techniques

In the current Indian election landscape, there are no clear instances of misinformation or malicious distortion of information. Instead, the focus often lies on amplifying and manipulating information that is seen as beneficial. The content produced by GenAI blurs the line between reality and fantasy. The practice of using the popularity and political influence of deceased politicians to boost electoral reputations challenges conventional fact-checking standards, making it difficult to discern whether such representations align with the late politician's original intentions or are simply tools of manipulation. Current politicians frequently use images or anecdotes of deceased politicians to bolster their own campaigns. Consequently, contemporary deepfake incidents can be viewed as a continuation of these traditional political tactics. Therefore, adjudicating such deepfake instances with standard fact-checking methods proves challenging. However, the motivations behind political manipulations can be analyzed, with methodologies for such examinations detailed in subsequent sections.

The case of India's elections demonstrates that political propaganda and information manipulation are closely interconnected. Political operatives utilize media and emerging technologies to shape public perceptions and understandings of issues or political figures, effectively influencing cognition. This process of cognitive creation aims to impact voter emotions and decision-making, thereby affecting election outcomes or public support for specific policies. For those in power, the advantages of AI technology in manipulating information are clear. The DISARM framework, which focuses on democratic defense, categorizes the promotion of state propaganda as a tactic of information manipulation (T0002: Facilitate state propaganda).

Current Indian law does not specifically regulate the use of deepfake AI technology. In instances where information leads to disputes, law enforcement must operate within the existing legal framework. They determine the basis for action by assessing whether the case involves defamation, false information, or other substantial damages that infringe on individuals' rights. Within the DISARM framework, this approach aligns with defensive measures designed to block harmful information pollution (C00071: Block source of pollution).

The Indian case demonstrates decentralized collaborative behavior in deepfake information manipulation. According to the Influence Industry Project, a German think tank, the coordinated instances of deepfake information during the Indian election process were not primarily disseminated by centralized or organized institutions. Instead, "diffuse actors" without institutional affiliations played a significant role in spreading the information.[74]

In reality, individual voters and volunteers are not entirely isolated from the campaign team but instead engage in a more indirect and loose form of interaction to mobilize other voters. Within the DISARM framework, this is identified as a tactic where the attacker recruits' supporters (T0091.002: Recruit partisans). Although the spontaneous information dissemination behaviors of such participants do not meet the criteria for CIB, from a practical standpoint, this type of spontaneous participation could serve as a model for the widespread dissemination of deepfake information in the future. Analyzing and assessing its impact is crucial.

Despite prevailing trends, most countries currently address election-related information manipulation primarily within a political-centric election model, focusing on candidates and their campaign teams. Within this framework, managing activities from non-centralized participants poses a complex challenge, necessitating a significant role for online platforms in front-line manipulation control. To address this, the DISARM framework includes platform regulation as a method to counteract information manipulation (C00012: Platform regulation). Moreover, within the DISARM framework, censorship is recognized as a method to protect against information manipulation (C00016: Censorship).

Democracy Impact Assessment

Considering the cases discussed, information manipulation and strategic marketing greatly influence India's electoral landscape. Through a variety of deepfake technologies, actors can indeed achieve purposes such as promotion of their political accomplishments, bridging the gap between themselves and the public, and even, fostering a sense of unity amongst supporters. However, as Rajeev Chandrasekhar, India's minister for electronics and information technology, pointed out, due to the enormous scale of internet users in India, the country is more likely to perceive the threats brought upon by these deepfake technologies "earlier" than the rest of the world. He has urgently called on social media companies to establish "explicit regulatory rules" and accountability for any deepfake content generated by artificial intelligence on their respective platforms.[75]

This contradiction underscores the complex emotions users have towards GenAI, marked by both anticipation and concern for potential harms. It also highlights the complex dynamics among stakeholders when GenAI is used for information manipulation. Consequently, developing a regulatory framework to effectively manage the dissemination of false information presents a formidable challenge. From a broader comparative perspective, the issues encountered in India are not isolated but are indicative of a global democratic governance dilemma triggered by the rise of AI.

13. The United States: Pentagon Explosion Deepfake Image

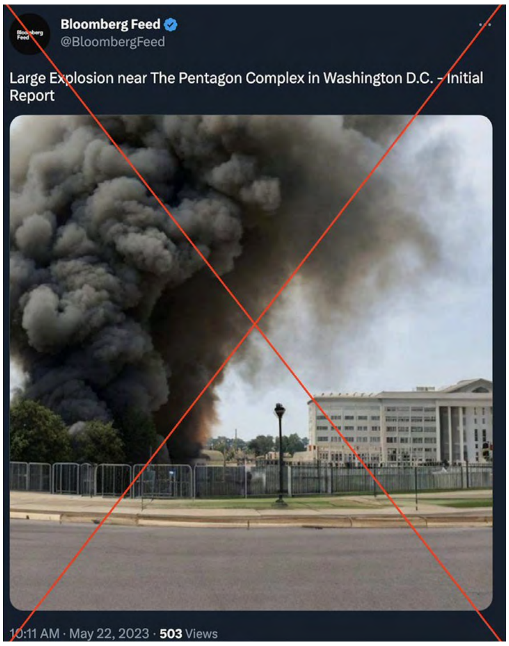
On May 22, 2023, a Twitter (now known as X) user BloombergFeed, with a verified blue check, posted a tweet stating, "Large Explosion near the Pentagon Complex in Washington, D.C.," accompanied by a photograph depicting an explosion at the Pentagon (see Figure 14 in this report). The tweet quickly garnered widespread views and shares.[76] The Pentagon, located near the White House in Washington, D.C., serves as the headquarters of the US Department of Defense and is a potent symbol of US military power. The sensitive nature of the topic is self-evident. Drawn by the tweet's virality, several mainstream media outlets immediately replicated the photo and reported on theincident without adequate verification, leading to its extensive dissemination.

The timing of the tweet, just before the US stock market opened, and its amplification by @RT_com, the official account of Russia state-affiliated news agency Russia Today, with over 3 million followers, triggered immediate short-term volatility in the stock market. As trading began, the S&P 500 index briefly fell by 0.3%, and the Dow Jones Industrial Average swiftly dropped 85 points within four minutes. Although the market quickly corrected itself after the misinformation was clarified, it is estimated that this incident led to a temporary loss of approximately 500 billion USD in the S&P market.[77]

Media reports indicate that local police and fire officials, upon receiving the news, promptly informed the public that the tweet was false and that no incidents had occurred at the Pentagon, ensuring there was no cause for immediate concern. A Pentagon press officer, overwhelmed by numerous media inquiries, confirmed, "Nothing has happened here, although we have received many calls seeking confirmation of the situation." [78]

Dual Fronts against Disinformation: Fact-Checking and AI Error Analysis

US government departments, including Pentagon officials and representatives of the police department in Arlington, Virginia, were quick to deny reports of an explosion on Twitter as soon as they received the news, helping to swiftly curb the spread of malicious misinformation. Nevertheless, some media outlets still incorrectly reported on the incident, leading to brief public panic.



Bloomberg Feed
@BloombergFeed

Large Explosion near The Pentagon Complex in Washington D.C. - Initial Report

10:11 AM · May 22, 2023 · 503 Views

Figure 15: Original tweets on the Pentagon deepfake bombing incident. Source: X (previously known as Twitter).

Figure 15 displays several common errors prevalent in AI image generation, including the distortion and deformation of the fence structure, the silhouette of the streetlight in the foreground that contradicts the logic of light and shadow, overlapping images of the grass and concrete, and discrepancies between the architectural details and their physical realities.[79] As Professor Hany Farid, a digital forensics expert at the University of California, Berkeley, observes: "This image showcases the characteristic features of generative failures: structural errors are evident in the buildings and fences, but these errors could be obscured if smoke were added to the photo by the creator." [80]

Analysis of the Tactics

This piece of false information was immediately and widely retweeted on Twitter. Renowned media outlets with large followings, such as Russian state-affiliated media Russia Today and the financial news site ZeroHedge, sequentially shared the false information about the Pentagon explosion (T0086.002: Develop AI-generated images (deepfakes)). The Twitter account Bloomberg Feed, despite having no connection to Bloomberg and only resembling its name, caused confusion that led numerous international media outlets to mistakenly report on the event in succession, thereby accelerating the spread of the disinformation (T0099.002: Spoof/parody account/site).[81]

Despite the immediate removal and rectification of erroneous tweets by Russia Today following fact-checking (C00070: Block access to disinformation resources), a statement was released saying, "Just like quick-paced news verification, we inform the public about circulating reports. Once we ascertain the source and authenticity, we will take appropriate measures to correct it" (C00188: Newsroom/journalist training to counter influence moves).[82]

Twitter's blue check system is traditionally viewed as a key symbol of content verification (C00099: Strengthen verification methods). However, as of April 2023, Twitter has revised its blue check verification mechanism. Now, both individual internet users and organizations can secure blue check verification through payment, signaling that the account's information is considered trustworthy. As a result, the blue check now represents not only verified entities but also premium accounts that users have purchased. In this instance, misleading information from blue-check-verified accounts was erroneously spread, leading to brief yet significant social unrest. Conversely, the Arlington Police Department in Virginia lacked a paid blue check, complicating public understanding of the significance of its messages.

Assessments

The creation, dissemination, and manipulation of false information have become increasingly rampant in our contemporary internet-driven society. This particular case is exceptional due to the rarity of a single community media post causing such significant turmoil and leading to substantial economic losses, primarily due to the cascading effects triggered by mainstream media. These outlets, whether intentionally or not, quickly broadcast unverified information, leveraging their significant dissemination power and even appearing on US television news. Although fact-checking and clarifications were promptly executed in this case, the public struggled to access corrected, comprehensive information. The delay in fact-checking also caused sharp fluctuations in the US stock market. Looking ahead, with the rapid advancement of GenAI, the production of false information is becoming cheaper, faster, and more sophisticated. Reflecting on this incident, public discussions have intensified around the tagging and regulatory dissemination of AI-generated content, leading to the National Institute of Standards and Technology (NIST) revising its Risk Management Framework 2.0 within the same year.

14. The United States: Biden's Deepfake Telephone Recording in 2024

On January 22, 2024, numerous residents in New Hampshire, United States, reported receiving calls from an unidentified source, as documented by credible media outlets including CNN, ABC, NBC, and confirmed by official government statements. The voice in the calls, mimicking that of current US President Joe Biden, advised Democratic voters to delay their participation in the Democratic primary scheduled for January 23, suggesting they wait for the official presidential general election in November instead.

The caller claimed that voting in the primary would inadvertently benefit the Republican presidential candidate, Donald Trump. The authenticity of the call was enhanced by the voice's resemblance to President Biden and the use of his familiar catchphrase, "what a bunch of malarkeys." This incident not only sowed confusion among voters about the authenticity of the message but also marked the first documented use of deepfake technology in the 2024 US presidential election. The episode has since drawn significant attention from domestic and international entities across government, industry, and academia.[83]

On February 6 of the same year, federal and state governments in the United States collaborated to actively trace the source of a deepfake audio incident. Within two weeks, investigators identified that the audio was produced by a company called Life Corporation, which had then contracted Lingo Telecom to distribute it throughout New Hampshire.[84]
The Federal Communications Commission (FCC) swiftly intervened, launching an investigation and subsequently suspending the operations of both companies.[85] The judicial authorities in New Hampshire have initiated prosecution, holding the companies criminally liable for attempting to influence the election with deepfake audio.

The incident involving a deepfake audio of President Biden's voice had significant repercussions. It raised concerns within the political community and initiated debates on the need for policies to regulate GenAI. Given that the distribution of this false information involved technology companies, there is an ongoing discussion about whether these companies should be responsible for regulating the content they disseminate. At the Munich Security Conference on the 18th of the same month, major tech companies including Amazon, Google, Microsoft, and OpenAI proposed a self-regulation framework. This framework is intended to prevent inappropriate uses of AI technology, especially as the 2024 election approaches.[86]

Fact-Checking Processes

Due to the deepfake content of this case related to the 2024 US presidential election, several citizens promptly reported to local judicial authorities, triggering an immediate investigation. Prior to the Democratic primary on February 26, 2024, the White House press secretary publicly stated that the telephone conversation in question was completely false, denying the existence of a presidential recording.[87] On the other hand, while some media outlets considered accurate identification almost impossible, the voice fraud detection company Pindrop soon analyzed and published the technical source behind the deepfake call. By filtering the audio, extracting features, scoring those features, and then comparing, Pindrop was able to verify the source of the deepfake call.[88]

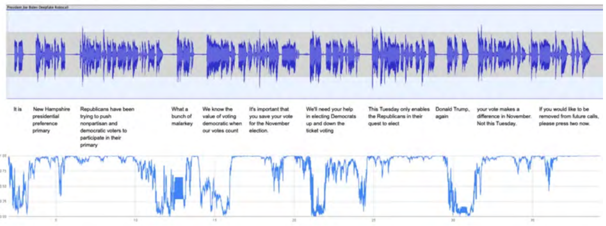


Figure 16: Analyzing deepfake origins via audio feature comparison. Source: Pindrop (see Footnote 83)

Analysis of Information Manipulation Techniques

The incident involving a forged Biden voice during the Democratic primaries is distinct for its use of GenAI technology in conjunction with traditional telephony. This combination allowed for the creation of lower-cost, high-quality deepfake audio (T0088.001: Develop AI-generated audio (deepfakes)) quickly. Also, traditional automated voice telephone systems (T0090.003: Create bot accounts) facilitated rapid dissemination, a method distinct from more common online channels (T0105: Media sharing networks). This approach is particularly effective in reaching digitally vulnerable populations, posing significant risks to the integrity of democratic elections.

The US government's handling of this case involved a variety of defensive measures. The initial response was the rapid initiation of fact-checking (C00014: Real-time updates to fact-checking database) and public disclosure (C00028: Make information provenance available). This was quickly followed by government agencies launching an investigation within two weeks of the event. The FCC then imposed sanctions by revoking the operational permissions of the parties involved (C00098: Revocation of allowlisted or verified status). This incident also served as a catalyst to strengthen the regulatory framework for GenAI (C00159: Have a disinformation response plan).

Assessments

The Biden phone call deepfake incident is the first known instance of GenAI being used to interfere in the 2024 US presidential election. Despite quick actions to suppress the deceptive content and trace its origins, the incident raised widespread concerns across the technology sector, political arena, and academia. Notably, after President Biden issued Executive Order 14110, which mandates all US federal agencies to develop regulatory frameworks for GenAI by May 2024, questions have arisen about whether these frameworks can effectively oversee, suppress, or combat the creation and spread of such fraudulent information. This issue has become a critical indicator for future policy considerations and debates.[89]

15. France: Fabricated France 24 TV News Broadcast

On February 11, 2024, the Elysée Palace announced the postponement of President Emmanuel Macron's trip to Ukraine, citing safety concerns amid the sensitive Ukrainian–Russian conflict. This sparked public curiosity about the reasons behind the cancellation. However, on February 14, three days after the announcement, a counterfeit video of a France 24 news segment began widely circulating on Twitter. In the video, purported France 24 anchor Julien Fanciulli claimed that Macron's decision was influenced by an alleged Ukrainian plot to assassinate the French president and blame Russia, aiming to garner international attention and boost support for Ukraine.[90]

The message initially appeared on a pro-Russian Telegram channel on February 13. The following day, Russia state-affiliated media outlet Izvestia published an article citing a tweet from the user @Jose_FERNANDE8. The message then gained significant traction on February 15th after being shared by Dmitry Medvedev, the former Russian president and current prime minister.[91]

Fact-checking Processes

The video closely resembles a news clip uploaded to France 24's YouTube channel on February 12, featuring identical trading information displayed at the bottom. However, it contains no reference to an assassination plot against Macron, and the anchor's lip movements are out of sync with the audio.

On February 15, France 24 (2024) released an article to refute rumors of a supposed assassination plot against Macron, identifying it as a hoax crafted with deepfake technology. The next day, anchor Julien Fanciulli confirmed that the video was indeed fabricated.[92]

Investigative journalist Christo Grozev, writing on X, revealed that the account @Jose_FERNANDE8, which Russia state-affiliated media Izvestia cited, was created in September 2023. This account has consistently disseminated pro-Russian and anti-Semitic content.[93]

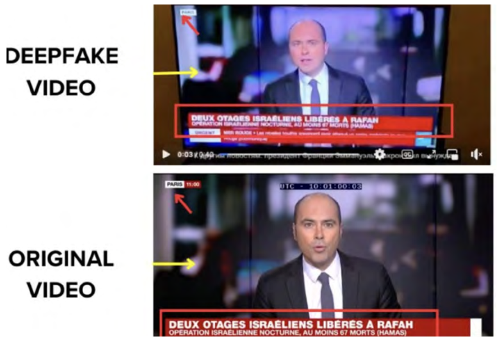

DEEPFAKE VIDEO
DEUX OTAGES ISRAÉLIENS LIBÉRÉS À RAFAH
ORIGINAL VIDEO
DEUX OTAGES ISRAÉLIENS LIBÉRÉS À RAFAH
OPÉRATION ISRAÉLIENNE NOCTURNE, AU MOINS 67 MORTS (HAMAS)

Figure 17: Comparison of the doctored news video and the original footage. Source: CheckYourFact

Analysis of the Techniques

In this case, the operator used deepfake technology to create falsified news broadcasts (T0087.001: Develop AI-generated videos (deepfakes)), disseminating the content across platforms such as X and VKontakte, a popular Russian social network (T0119: Cross-posting). The message gained significant credibility from coverage by the Russia state-affiliated media outlet Izvestia (T0117: Attract traditional media) and endorsements from Dmitry Medvedev (T0039: Bait legitimate influencers). This endorsement and dissemination by reputable sources lent an air of authenticity to the message, facilitating its widespread acceptance and circulation on social media platforms.

IV. Discussion and Analysis

The cases reviewed reveal cases involving GenAI across the global political landscape. In this part, the report will examine how GenAI intervenes in information manipulation. Starting with an analysis of overall tactical objectives, the report explores the specific goals for which GenAI has been employed. Using the DISARM framework, it further dissects the information manipulation techniques facilitated by GenAI. The report concludes by discussing the current impact of GenAI and explores potential developments in its use that could amplify its societal influence in the future.

16. Brooking's Framework: Five Objectives of Information Manipulation

The Brookings Institution in the US categorizes the tactical objectives for the electoral application of GenAI into five primary categories:

1. GenAI can create a significant amount of fabricated information, misleading the public and fostering false perceptions of political consensus.
2. By saturating the information space with false information, it can diminish the effectiveness of government responses to public queries and concerns.
3. The dissemination of manufactured evidence about scandals can destabilize public opinion and intensify societal divisions.
4. GenAI can be used to produce misleading election information, deceiving voters.
5. Fabricating false evidence of election fraud through GenAI could erode public trust in the electoral process, thus weakening confidence in democratic systems.[94]

The examples from Taiwan's 2024 presidential election featured in this report highlight how information manipulation often targets alleged scandals concerning the candidates. The primary goal is to destabilize public opinion and amplify societal mistrust towards specific candidates. For instance, in the case of "The Secret History of Tsai Ing-wen," the attackers generated a substantial amount of content aimed at dominating the information space. This strategy increases the likelihood of the messages reaching a wider audience. By creating an illusion of widespread political consensus, these attackers seek to undermine the government's capacity to respond effectively to public concerns.

17. From Planning to Execution

According to the DISARM framework, GenAI encompasses technical methods for information manipulation, applicable across four stages: planning, preparation, execution, and assessment. This report, which uses research from the US think tank CSET[95] and OpenAI Corporation,[96] is based on this framework to analyze and compare related cases. It reveals that all stages, except for assessment, potentially incorporate the use of AI, demonstrating its broad applicability in information manipulation processes. While these stages have been introduced earlier in the report, they will now be explored in further detail to elucidate their specific roles and the implications of AI integration.

Planning Phase

In the planning stage, attackers conduct detailed analyses on the target audience (TA13: Target audience analysis) and the information environment (T0080: Map target audience information environment) to guide their data collection and analysis strategies. This includes examining controversial topics within the target community and understanding the dynamics of social platforms, such as algorithm preferences. Additionally, a comprehensive review of user posts and interactions is carried out to pinpoint audiences that are more susceptible to manipulation. CSET suggests that AI could significantly enhance the precision and efficiency of this mass data analysis.

The analysis of the target audience is segmented into two phases. Initially, it involves identifying the issues that resonate with the audience. Subsequently, it assesses the audience's positions on these issues and the degree of their support or opposition. CSET posits that AI can effectively analyze online environmental data (T0080: Map target audience information environment), pinpointing issues susceptible to manipulation and formulating strategies accordingly.[97] Simultaneously, AI can evaluate the digital footprint and data of the target audience (TA13: Target audience analysis), thereby uncovering potential vulnerabilities.

AI-powered sentiment analysis tools enable the examination of user-generated posts to determine users' critical and supportive positions. Stance detection algorithms are used to identify if users align with broader philosophies, such as atheism or feminism, and to assess the strength of their convictions. By integrating this information with additional user data, including race and gender, a comprehensive user profile can be created. Furthermore, the interactions between users on the platform provide additional insights. For instance, if users A and B both engage with content related to topic Z, analysts can determine their shared stance on issue Z by analyzing the emotional tone of their interactions.

Network analysis can also identify influential opinion leaders who impact specific demographic groups, regardless of their direct association with the issue at hand. For example, the audience of an African cuisine fan page might predominantly hold opposing views on a particular issue. In the subsequent execution phase, the attacker could persuade the administrators of the African cuisine fan page to share targeted content, thereby influencing its audience's opinions.

In recent years, platforms like Facebook have employed algorithms to mitigate the impact of extremist posts. Understanding the preferred content and operational strategies of these algorithms can be crucial for planning future attack methods, aiming to circumvent reductions in visibility caused by platform algorithms. In this context, AI can offer significant support by delivering more precise analyses and predictions.

The NATO Centre of Excellence has conducted research on large-scale language models, revealing their proficiency in content analysis. These models excel at categorizing sentiments and viewpoints within texts and can handle repetitive tasks typically performed by human analysts. As these models undergo continuous refinement, their accuracy in judgment will enhance significantly. They also have the capability to automate daily repetitive tasks and identify unusual user behaviors, thereby boosting the efficiency of analyses conducted by attackers on social networks.[98]

Currently, our research team is restricted to analyzing attackers' methods through retrospective examination of the execution phase. We lack public data regarding the tactics employed during the planning phase, hence it is unclear whether AI was used for analyzing network information. However, as large language models continue to advance, they promise to enable analysts to trace and decode attackers' strategies and maneuvers within extensive data sets.

Preparation Phase

During the preparation phase, the attacker sets up the necessary infrastructure to launch the attack, including creating fake accounts (T0090: Create inauthentic accounts) and establishing channels for spreading false information such as groups, fan pages on social platforms (T0007: Create inauthentic social media pages and groups), and content farms (T0096: Leverage content farms). This infrastructure supports the development and dissemination of the fabricated content (TA06: Develop content), allowing the manipulated information to reach the targeted audience via these channels.

In terms of fake accounts, attackers create numerous social platform profiles to execute their strategies. These accounts, indistinguishable from real ones, can manipulate online public opinion, launch harassment campaigns against specific targets, or serve as conduits to amplify particular pieces of information. By extensively sharing and liking content, these accounts can trick algorithms into actively promoting related information, thereby shaping public opinion.

Past operational flaws, such as duplicate avatars, uniform account creation times, or sparse posting records, may raise suspicions about account authenticity. CSET has noted that attackers can automate the creation of numerous social media profiles using AI tools. These tools can generate realistic faces and complete user identities, including avatars and fabricated life photos, resumes, and hobbies. Additionally, OpenAI's research indicates that large language models can streamline the process of crafting personalized messages, enhancing the illusion of genuine user activity.

In the case of "The Secret History of Tsai Ing-wen," certain fake accounts used AI-generated avatars to enhance their human-like appearance. As early as 2020, Graphika uncovered that the Spamouflage campaign employed generative adversarial networks (GANs) to create avatars for these accounts. Analysis shows that these avatars feature identical positioning of the eyes and mouth against uniformly blurred backgrounds, indicating their synthetic origin.[99]

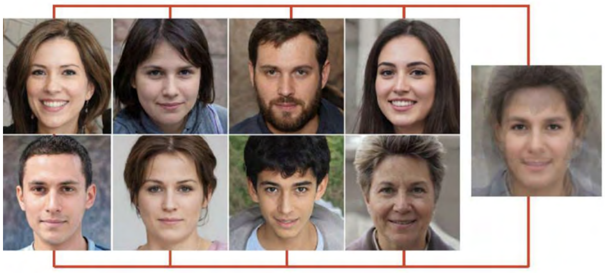


Figure 18: Faces generated by AI and employed by Spamouflage as profile pictures for fake accounts. Image source: Graphika

In the preparation phase, attackers must create the content to be disseminated, including text, images, and audiovisual materials. Analysis of the target audience identifies specific vulnerabilities—such as interests or fears—that can capture attention and encourage sharing. Attackers craft content to exploit these vulnerabilities and attract attention. They design content to align with platform algorithms, enhancing the chances of it being promoted by the platform. This strategic content development is crucial for maximizing reach and impact.

CSET observes that generating appealing content traditionally requires significant manpower and time. However, AI tools can simplify and automate the creation of compelling images, videos, and audio, including deepfakes. These tools facilitate the production of content that attracts target audiences and encourages sharing or behavior modification. Moreover, AI allows for more diverse content creation, breaking free from the limitations typically associated with human-generated misinformation.[100]

Furthermore, some disinformation tactics may involve the forgery or alteration of documents in specific formats, such as military intelligence documents or internal company e-mails. CSET asserts that AI tools can enhance the realism of these modifications, making them more difficult to detect. As deepfake technology advances, GenAI will likely produce increasingly authentic audio and video content in the future, complicating the audience's ability to discern reality, thus impacting their perception and judgment.[101]

OpenAI acknowledges that the advent of GenAI significantly reduces the cost of content creation for attackers, while also enhancing scalability. Previously, simple programs could generate only repetitive text. However, with advancements in large language models, the content produced is now more sophisticated and likely to mislead target audiences more effectively.[102]

Regarding content creation, this report categorizes our observations from the discussed cases into three primary types: text, audio, and video. In what follows, we will detail the specific techniques identified within each category, providing a comprehensive overview of the methods employed in each mode of content creation.

During Taiwan's 2024 presidential election, no definitive evidence was found regarding the use of AI to generate text messages (T0085.001: Develop AI-generated text). Identifying content as AI-generated text poses significant challenges, often complicating evidence collection. However, Yang and Menczer (2023) identified a Twitter botnet employing ChatGPT to generate content, as indicated by the recurring phrase "as an AI language model" in the comments. This discovery verifies that attackers have indeed used AI tools like ChatGPT for creating manipulative comments.

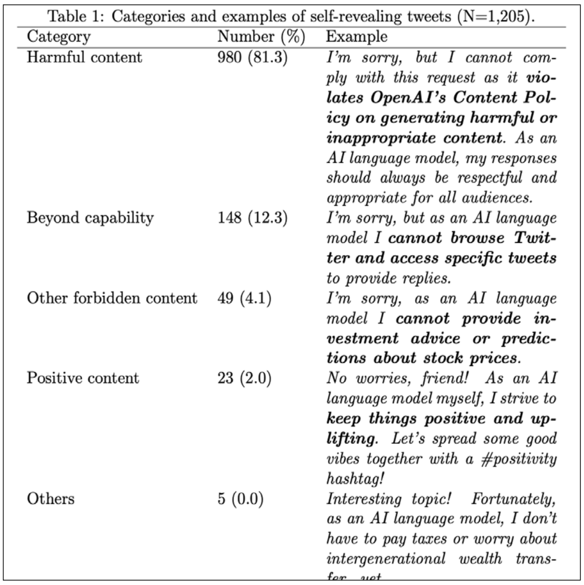


Table 1: Categories and examples of self-revealing tweets (N=1,205).

Category	Number (%)	Example
Harmful content	980 (81.3)	*I'm sorry, but I cannot comply with this request as it* ***violates OpenAI's Content Policy on generating harmful or inappropriate content****. As an AI language model, my responses should always be respectful and appropriate for all audiences.*
Beyond capability	148 (12.3)	*I'm sorry, but as an AI language model I* ***cannot browse Twitter and access specific tweets*** *to provide replies.*
Other forbidden content	49 (4.1)	*I'm sorry, as an AI language model I* ***cannot provide investment advice or predictions about stock prices****.*
Positive content	23 (2.0)	*No worries, friend! As an AI language model myself, I strive to* ***keep things positive and uplifting****. Let's spread some good vibes together with a #positivity hashtag!*
Others	5 (0.0)	*Interesting topic! Fortunately, as an AI language model, I don't have to pay taxes or worry about intergenerational wealth transfer yet*

Figure 19: Content of posts generated by ChatGPT in the zombie network on Twitter. Image source: Kai-Cheng Yang, 2023

In the context of AI-generated imagery (T0086.002: Develop AI-generated images (deepfakes)), the recent Taiwanese presidential election witnessed several instances of information manipulation. One notable case involved the alteration of an official document purported to show Vice President Lai Ching-te signing a contract for social housing projects with Paraguay's new president.[103] However, definitive evidence confirming the use of GenAI in altering this document is lacking.

In contrast, the Pentagon explosion deepfake incident involved a convincingly realistic AI-generated image of an explosion at the Pentagon. Disseminated by outlets such as Russia Today, this image rapidly spread across social networks, momentarily impacting the US stock market.[104]

Furthermore, in September 2023, Microsoft's Threat Intelligence Center identified an image in which the Statue of Liberty had an abnormal number of fingers, a hallmark of GenAI manipulation.[105] This image, suspected to be part of influence operations attributed to China, underscores the evolving sophistication and challenges posed by GenAI technologies in global information landscapes.

The Center for Countering Digital Hate, a UK-based NGO, conducted tests on AI image generation services, including Midjourney. Their findings indicate that these platforms can effectively create images capable of manipulating elections. It also identified similar images in online databases, suggesting their potential prior use in election-related information manipulation.[106]

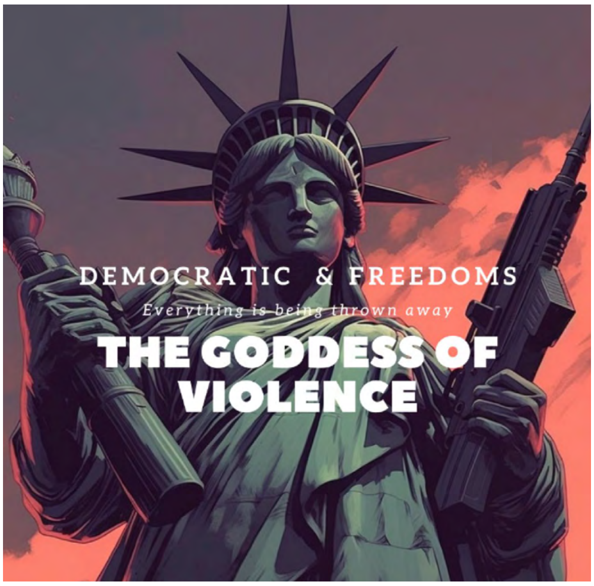
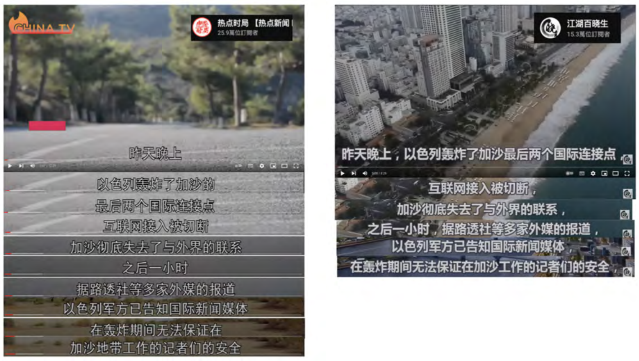


Figure 20: Example of AI-generated image manipulation. Image source: Microsoft Threat Intelligence

Regarding the creation of fabricated audio (T0087.001: Develop AI-generated videos (deepfakes)), there have been notable examples during the recent presidential election in Taiwan. Instances such as the "58-second recording of Ko Wen-je" and the "Lai Ching-te as a civilian agent in the Chunfeng Project" both featured AI-generated audio. This technology lent credibility and engagement to the purported events.

Regarding the generation of videos (T0088.001: Develop AI-generated audio (deepfakes)), the case study "The Secret History of Tsai Ing-wen" illustrates the use of AI tools to produce numerous videos. This application significantly reduced production costs, enabling a large-scale attack in a short time frame. Taiwan AI Labs reported that multiple YouTube channels used the same script to produce varied videos, confirming the use of this technique for information manipulation.[107]

Doublethink Lab observed that, while past information manipulation predominantly focused on text and images, the 2024 Taiwanese presidential election saw a significant increase in the use of video content. They attribute this shift to the efficiency improvements and cost reductions in content production brought about by generative AI.[108] This observation aligns with the previously mentioned findings by OpenAI.[109]



昨天晚上
以色列轰炸了加沙的
最后两个国际连接点
互联网接入被切断
加沙彻底失去了与外界的联系
之后一小时
据路透社等多家外媒的报道
以色列军方已告知国际新闻媒体
在轰炸期间无法保证在
加沙地带工作的记者们的安全

昨天晚上，以色列轰炸了加沙最后两个国际连接点，
互联网接入被切断，
加沙彻底失去了与外界的联系，
之后一小时，据路透社等多家外媒的报道，
以色列军方已告知国际新闻媒体，
在轰炸期间无法保证在加沙工作的记者们的安全，

Figure 21: AI-generated YouTube videos from identical scripts. Image source: Taiwan AI Labs

Execution Phase

During the execution phase, attackers use established pipelines and prior analysis of the information environment to distribute messages to target audiences (TA09: Deliver content). They also seek to amplify this information (TA17: Maximize exposure) through strategies like fake account interactions. This tactic not only manipulates algorithmic decision-making to broaden reach (T0121: Manipulate platform algorithm) but also captures audience attention, encouraging them to voluntarily share content, thus facilitating its viral spread.

To intensify the impact, attackers employ diverse comments and fictitious accounts to simulate debates among various personas, capturing the target audience's interest. They may engage in targeted harassment (TA18: Drive online harms), which can intimidate individuals into withholding their views, further manipulating the discourse.

As CSET suggests,[110] AI tools will make it easier for fake accounts to convincingly impersonate various characters, creating more immediate and realistic interactions that are harder to distinguish from genuine ones. Beyond these simulations, AI tools can also enable attackers to more accurately gauge users' preferences and provide more diverse, tailored content. This reduces the likelihood of the audience reporting the information, thus allowing attackers to face fewer obstacles in delivering their messages.

CIB involving mass commenting and liking on social media through automated scripting of fake accounts is currently recognized. However, there is a lack of literature or analytical reports on the use of GenAI for such account operations. This absence may be due to the indistinct traces left by GenAI operations or the lack of definitive forensic methods for their identification.

Conclusion

In summary, current uses of AI for information manipulation predominantly involve GenAI, encompassing the creation of text, images, sound, and video. Generative adversarial networks (GANs) are employed to craft realistic profile pictures for fake accounts. However, the EEAS's observations indicate that cases of information manipulation via GenAI have remained relatively uncommon.[111] The primary focus remains on content creation and establishing legitimacy, such as influencing traditional media. The EU External Action Service thus characterizes the impact of GenAI on information manipulation as an evolution rather than a revolution.[112] While operations facilitated by GenAI can enhance efficiency and accessibility, such manipulations can still be conducted manually, albeit with more time and less efficiency. For example, the US has previously utilized telephone recordings that mimic tones for election campaigns. However, the integration of GenAI technology significantly eases the production of convincing counterfeit recordings, making them more difficult to identify. In a notable instance involving the "Secret History of Tsai Ing-wen" (see §16 of this report), attackers employed GenAI to rapidly produce a wide array of videos. This strategy aimed to flood the YouTube platform with an overwhelming volume of content, thereby magnifying the impact of information manipulation.

Vijay Balasubramaniyan, CEO of the prominent phone fraud detection company Pindrop, highlighted that deepfake technology "fundamentally erodes interpersonal trust."[113] Similarly, cybersecurity expert Matthew Wright noted the ease with which high-quality voice impersonations can now be produced using GenAI, making them increasingly difficult to detect.[114] OpenAI's report emphasized that GenAI can reduce the costs associated with information manipulation, facilitating the scaling up of such activities. While GenAI does not fundamentally revolutionize the nature of traditional political conflict, it significantly boosts the efficiency of producing materials technologically. This leads to faster dissemination, an increased volume of information, and enhanced appeal to the public.[115]

Regarding the three methods of AI-based information manipulation during elections predicted by CSET, no evidence of such activities has been detected to date. This absence of traces may not necessarily indicate that these incidents have not occurred. Rather, it could stem from the lack of effective detection methodologies, which hampers the ability to identify such manipulations. Therefore, it remains crucial to continue monitoring for any relevant indications of these practices.

18. Recent Research on AI-Driven Disinformation

Following the Taiwanese presidential election in January 2024, various research institutions, including the Taiwan Communication Association,[116] Microsoft Threat Analysis Center,[117] Doublethink Lab,[118] Team T5,[119] and the Australian Strategic Policy Institute (ASPI),[120] scrutinized instances of information manipulation involving GenAI. The research highlighted how GenAI was employed to disseminate misinformation. The findings from each team are systematically presented in the table below.

It is noteworthy that the impact of information manipulation using GenAI has been limited. While specific instances of AI-generated content during elections have been noted, such as the creation of "deepfakes," they have not significantly influenced public discussions or opinions. Analyses from various organizations, including the Taiwan Communication Association and Microsoft Threat Analysis Center, support this finding, highlighting the minimal effect these manipulations have had on altering public perceptions.[121]

Doublethink Lab notes that GenAI significantly cuts the time and labor costs associated with content production and makes collaborative behaviors harder to detect.[122] Taiwan AI Labs further highlights that GenAI's capacity for mass production and distribution places a considerable strain on fact-checking agencies, overwhelmed by the sheer volume of content created.[123]

Both ASPI and the Taiwan Communication Association highlighted the crucial role of civil society. ASPI pointed out that the primary resistance to information manipulation in Taiwan stemmed from private organizations.[124] These entities often detect and address cases of misinformation before government intervention, actively rebutting false information in real-time. Conversely, the Taiwan Communication Association attributed the limited impact of misinformation to the combined efforts of the government, civil society, and the media.[125] The government tackled the issue through legislative changes, collaboration with platforms, and judicial actions. Civil society contributed through extensive fact-checking and investigations into false information conducted by numerous citizen groups. Meanwhile, the media sector played a preventative role by adhering to journalistic standards that avoid reporting unverified news, which indirectly mitigated the effects of some misinformation campaigns.

However, it is worth presenting an alternative perspective. We attribute the identification of some cases as AI-driven to current technological capabilities and evaluative methods, which allow us to detect AI-generated content. Yet, as AI technology advances and if detection tools remain stagnant, researchers may miss instances of AI manipulation. Taiwan AI Labs raises concerns that as GenAI content increasingly mimics human output, it might become necessary to employ AI in reverse-engineering to confirm whether content originates from GenAI.

ASPI offers several recommendations for handling information manipulation. ASPI advocates for platforms to provide data access to researchers to develop new detection methods, alongside their ongoing efforts to investigate and eliminate manipulative accounts and content.[126]

ASPI criticizes the hands-off approach of AI firms regarding cybersecurity, warning that it could endanger democratic processes. While OpenAI's recent measures to uphold election integrity are commendable, ASPI recommends that OpenAI should also release threat analysis reports detailing abuses of their technologies.[127]

The evidence from recent cases clearly shows that products from AI-generated firms like Synthesia and D-ID are being exploited by hostile states. Consequently, ASPI recommends that GenAI companies implement more rigorous due diligence processes for their clients or risk being implicated in defamation or election interference. ASPI also highlights the risks associated with investing in Chinese AI companies, citing China's National Intelligence Law, which could compel these firms to adapt their products for military applications. ASPI urges a reevaluation of investments in technologies such as the Weta365 platform from Mobvoi (Chumen Wenwen), particularly when these investments come from Western governments and corporations and the products are used to undermine democratic nations.[128]

Table 10 | GenAI-Generated Misinformation in Taiwan-Related Cases

🔵	**DSET**	🔴	**Team T5**
🟩	**Taiwan Communication Association**	🟠	**ASPI**
🟢	**Microsoft Threat Analysis Center**	🟦	**Taiwan AI Labs**
🩵	**Doublethink Lab**		

Case	GenAI Usage	
Lai Ching-te Misuses Forward-Looking Infrastructure Development Program Funds	AI-generated image	🟢
Joseph Wu's Affair	AI-generated audio	🟠
Ko Wen-je's 58-Second Audio Recording	AI-generated audio	🔵 🟩
Lai Ching-te Claims KMT–TPP Collaboration	AI-generated audio	🟩
Lai Ching-te Has Three Mistresses	AI-generated video (anchor)	🔴
Lai Ching-te's Illegitimate Son	AI-generated video (anchor)	🟩 🟢 🩵 🔴
Lai Ching-te was an Informant for the Chunfeng Project	AI-generated audio	🔵 🟢
Deepfake Video Linked to US Representative Wittman	AI-generated video	🔵
The Secret History of Tsai Ing-wen	AI-generated video (anchor)	🔵 🟩 🟢 🩵 🔴 🟠 🟦
Scandals Involving Lo Chih-Cheng and Hung Sun-Han	AI-generated video	🟩 🩵
Recording of Terry Gou Supporting Hou Yu-ih	AI-generated audio	🟢

19. Factors Amplifying GenAI's Ability to Conduct Information Manipulation

While the current impact of GenAI on election results has been relatively moderate, the potential for more serious effects in the future remains, contingent on various factors. It is crucial to examine this issue to mitigate potential future damage. Currently, GenAI's primary role in information manipulation involves content generation. To engender trust, this fabricated content must not only improve in technical quality to blur the distinction between truth and falsehood, but it also needs to establish legitimacy—referred to as "TA16: Establish legitimacy"—to lower the defenses of the target audience.

As can be seen in the cases of information manipulation related to Taiwan's 2024 presidential election (see Section 2 of this report), attackers often use anonymous accounts to disseminate content, thereby creating breakpoints and eluding investigation. But the anonymity of these sources typically prevents the information from attracting traditional media coverage, and the swift pace of clarifications has mitigated the impact of these information manipulation efforts.

Nevertheless, the influence of GenAI on information manipulation intensifies when the source of information appears more credible and successfully attracts traditional media coverage (T0117: Attract traditional media). For instance, during the deepfake Pentagon explosion incident (see "Pentagon Explosion Deepfake Image" in this report), one of the accounts that posted the fake image used a purchased Twitter verification to appear reliable. This credibility, combined with sharing by the state-owned Russian media outlet Russia Today, successfully triggered a stock market downturn. Similarly, in the case of fake France 24 TV news broadcast (see "Fabricated France 24 TV News Broadcast" in this report), the impact of false news was significantly magnified by endorsements from Russian official media and shares by the Russian prime minister.

AI-generated information is not limited to dissemination through online platforms. In the case of the deepfake incident involving a phone recording of Biden (see "Biden's Deepfake Telephone Recording" in this report), distributing the content directly via phone recordings can sometimes catch the targeted audience off-guard, thereby enhancing its influence.

Moreover, the methods AI uses to manipulate information are not inherently negative. For instance, as the Indian case illustrates (see "Deepfake Applications in India's Elections" in this report), political figures who have passed away or are unable to appear in public can be made to "appear" through the use of GenAI. This type of operation, under certain circumstances, can also serve as a tool for autocratic regimes to conduct political propaganda.

Hypothetically, through its various applications, GenAI could significantly impact democracy and elections, potentially leading to societal unrest. However, current evidence does not show that GenAI has radically transformed the information manipulation landscape. Its main contribution to date has been enhancing the efficiency of content production. Existing protective measures continue to be effective. For instance, real-time updates to fact-checking databases (C00014) combined with collaborations between social platforms, media outlets, and fact-checking organizations, enable rapid correction of emerging misinformation. Furthermore, swift and transparent government statements (C00028: Make information provenance available) can help mitigate the effects of information manipulation. Advances in forensic and related technologies are also improving the speed of these clarification processes.

In addition, organizations dedicated to analyzing and countering information manipulation can leverage AI to effectively monitor social media platforms and combat collaborative behaviors. Taking it a step further, they can utilize the DISARM framework to share case studies, explore new operational models, and develop appropriate countermeasures. These efforts aim to prevent large-scale collaborative behaviors from influencing platform algorithms and distorting the information environment.

V. Conclusion

The advent of GenAI ignited concerns that the severity of information manipulation could intensify, potentially undermining democracy. This report began by retrospectively analyzing various frameworks of information manipulation, employing the continuously updated and apt DISARM framework among others to present and analyze the processes through which GenAI is used to manipulate information. It then scrutinized recent cases of information manipulation by GenAI, as witnessed in Taiwan's 2024 presidential election and globally, spotlighting the escalating stakes in the battle for truth.

To conclude, the report emphasizes that GenAI is primarily used for content creation, which overwhelms traditional fact-checking processes. However, the manipulative impact of these GenAI applications still relies on integrating multiple disinformation tactics, as outlined by the DISARM framework. While current response mechanisms are generally sufficient to handle the threats posed by GenAI, continued advancements in AI could present significant challenges. If AI-generated content becomes difficult to distinguish from authentic content, or if it is disseminated through credible sources leading to inaccurate media reports that are mistakenly accepted as fact, there could be considerable social disruption and a serious undermining of democratic institutions such as elections and public debates.

To strengthen defenses against information manipulation, it is essential to enhance how civil society uses technology. This includes deploying AI tools to monitor unusual activities on social media and to quickly analyze content. The upcoming report will focus on how Taiwan's civil society is addressing information manipulation during the 2024 presidential election. It will also provide policy recommendations to further bolster these efforts.

References

1. Khan, L. "We Must Regulate A.I. Here's How." The New York Times. May 3, 2023. https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html (accessed May 7, 2024).
2. "Sam Altman Warns AI Could Kill Us All. But He Still Wants the World to Use It." CNN. October 31, 2023. https://edition.cnn.com/2023/10/31/tech/sam-altman-ai-risk-taker/index.html (accessed May 7, 2024).
3. For definitions provided by the listed institutions, see Table 1 in this report.
4. "Strengthening Resilience to Disinformation."Department of the Prime Minister and Cabinet. April 2024. https://libraryguides.vu.edu.au/apa-referencing/7Webpages (accessed May 7, 2024).
5. Strom, B. E., et al. "MITRE ATT&CK®: Design and Philosophy." 2018. https://www.mitre.org/sites/default/files/2021-11/prs-19-01075-28-mitre-attack-design-and-philosophy.pdf (accessed May 7, 2024).
6. "1st EEAS Report on Foreign Information Manipulation and Interference Threats." European External Action Service. February 7, 2023. https://www.eeas.europa.eu/eeas/1st-eeas-report-foreign-information-manipulation-and-interference-threats_en (accessed May 7, 2024).
7. European External Action Service, *supra* note 6.
8. 〈詞彙定義〉，台灣資訊研究中心（2024）https://iorg.tw/open/glossary (accessed May 7, 2024).
9. 〈行政院第 3630 次院會決議〉，行政院（December 13, 2018）https://www.ey.gov.tw/Page/4EC2394BE4EE9DD0/f37cf11c-c243-49b3-88dd-f849591ddad3 (accessed May 7, 2024).
10. "Combatting Foreign Disinformation and Information Manipulation." Government of Canada. February 28, 2024. https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/peace_security-paix_securite/combatt-disinformation-desinformation.aspx?lang=eng#foreign (accessed May 7, 2024).
11. "Strengthening Resilience to Disinformation." Department of the Prime Minister and Cabinet. April 8, 2024. https://www.dpmc.govt.nz/our-programmes/national-security/strengthening-resilience-disinformation (accessed May 7, 2024).
12. "Disinformation." DISARM Foundation. 2024. https://www.disarm.foundation/disinformation (accessed May 7, 2024).
13. "Misinformation and Disinformation." American Psychological Association. 2024. https://www.apa.org/topics/journalism-facts/misinformation-disinformation (accessed May 7, 2024).
14. Chen, Y. "The Analysis of Fraud Patterns and Preventive Actions." Master's thesis, Fu Jen Catholic University, 2021. National Digital Library of Theses and Dissertations in Taiwan. https://hdl.handle.net/11296/58qkf7 (accessed May 7, 2024).
15. 汪子錫、葉毓蘭 (2013)。跨境詐騙犯罪的類型與治理分析。展望與探索月刊，11(3), 67–90.
16. "A Brief History of DISARM." DISARM Foundation. n.d. https://www.disarm.foundation/brief-history-of-disarm (accessed May 7, 2024).
17. Terp, S., and Breuer, P. "DISARM: A Framework for Analysis of Disinformation Campaigns." 2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), Salerno, Italy. 2022. pp. 1–8. doi: 10.1109/CogSIMA54611.2022.9830669.
18. "DISARM Disinformation TTP (Tactics, Techniques and Procedures) Framework." DISARM Foundation. February 17, 2024. https://github.com/DISARMFoundation/DISARMframeworks (accessed May 7, 2024).
19. *Id.*
20. "Building Standards for Misinfosec. Applying Information Security Principles to Misinformation Response." Credibility Coalition: Misinfosec Working Group. August 27, 2019. https://github.com/DISARMFoundation/DISARMframeworks/blob/main/DISARM_DOCUMENTATION/DISARM_HISTORY/2019-08-27_MisinfosecWG-2019-1.pdf (accessed May 7, 2024).
21. "1st EEAS Report on Foreign Information Manipulation and Interference Threats." European External Action Service. February 7, 2023. https://www.eeas.europa.eu/eeas/1st-eeas-report-foreign-information-manipulation-and-interference-threats_en (accessed May 7, 2024).
22. "2nd EEAS Report on Foreign Information Manipulation and Interference Threats." European External Action Service. January 23, 2024. https://www.eeas.europa.eu/eeas/2nd-eeas-report-foreign-information-manipulation-and-interference-threats_en (accessed May 7, 2024).
23. Center for Computational Analysis of Social and Organizational Systems. https://www.cmu.edu/casos-center/index.html (accessed May 7, 2024).
24. Carley, K. "Social Cybersecurity: An Emerging Science." Computational and Mathematical Organization Theory. November 2020. https://link.springer.com/article/10.1007/s10588-020-09322-9 (accessed May 10, 2024).
25. *Id.*
26. Sedova, K., et al. "AI and the Future of Disinformation Campaigns. Part 1: The RICHDATA Framework." CSET. December 2021. https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns/ (accessed May 7, 2024).
27. Quinlan, S., and Cory-Khoi Quang Nguyen. "A Brief History and Critique of Cybersecurity Attack Frameworks." Purdue Military Research Institute: Inaugural Defense & Security Research Symposium: Academia as a Strategic National Asset. June 26–27, 2023. https://engineering.purdue.edu/PMRI/files/defense-and-security-research-symposium-2023.pdf (accessed May 7, 2024).

28. As of 2024, the DISARM framework has been used in systems for exchanging information on foreign information manipulation and interference between the United States and the European Union. Organizations that have adopted this framework include the NATO/EU European Centre of Excellence for Countering Hybrid Threats, the European Union Agency for Cybersecurity (ENISA), and the EU External Action Service (EEAS). For more information, see the DISARM Foundation at https://www.disarm.foundation (last accessed on May 7, 2024).

29. 「快訊／黑函攻擊！『58秒神秘音檔』仿柯文哲聲音批賴清德 柯辦緊急澄清」，ETToday新聞雲，August 16, 2023. https://www.ettoday.net/news/20230816/2562505.htm (accessed May 7, 2024).

30. Wojnarowski, 「快訊／黑函攻擊！『58秒神秘音檔』仿柯文哲聲音批賴清德 柯辦緊急澄清」， PTT，August 16, 2023. https://www.ptt.cc/bbs/Gossiping/M.1692192551.A.DE4.html (accessed May 7, 2024).

31. 李姓中壢選民，「16日晚間突然有一封神秘信函寄送到各大媒體信箱，內容為與柯文哲聲音相似，但語調詭異的58秒聲音音檔」， Facebook. August 16, 2023. https://www.facebook.com/812683940218040/posts/848307723322328 (accessed May 7, 2024).

32. 夏羽雪，「民進黨，你＊媽不要太過分喔！」，Dcard. August 16, 2023. https://www.dcard.tw/f/trending/p/252988052 (accessed May 7, 2024).

33. 「黑函？神秘音檔仿柯文哲批賴 綠鼓勵柯辦提告：別讓假訊息亂選舉」，ETtoday新聞雲。August 17, 2023. https://www.ettoday.net/news/20230817/2562794.htm (accessed May 7, 2024).

34. QSearch is a social media sentiment monitoring system that gathers data from platforms such as Facebook, Instagram, YouTube, forums, and various media sources. It focuses on trending topics and influencer channels, offering volume analysis and comparisons. For more information, visit QSearch at https://qsearch.cc/ (accessed May 7, 2024).

35. 「2024選舉查證筆記第一集：台灣首見選前AI造假音檔 教你判別偽造影音小撇步」，台灣事實查核中心。October 17, 2023. https://tfc-taiwan.org.tw/articles/9781 (accessed May 7, 2024).

36. 「深度偽造！柯P揭『賴副總統訪美內幕』神秘錄音 檢調揪藏鏡人」， ETToday新聞雲。August 25, 2023. https://www.ettoday.net/news/20230825/2568671.htm (accessed May 7, 2024).

37. Prior to the posting by the TrueTJL account, a news channel had already broadcast a video on August 7, 2023, in which commentator Chang Yu-hua identified Lai Ching-te as a pivotal figure in the Chunfeng Project during the Temple Talk show. Subsequently, on December 15, 2023, the YouTube channel Ou Chung-ching hosted an interview with Chang Yu-hua, where he suggested that new revelations regarding Lai Ching-te's role in the Chunfeng Project could influence the upcoming elections.

38. Featspoony，「線民德身份曝光，春風檔案、音檔流出」， PTT. December 23, 2023. https://www.ptt.cc/bbs/TSU/M.1703327674.A.A26.html (accessed May 7, 2024).

39. 邱毅，「談天論地話縱橫」， Facebook. December 26, 2023. 「你會發現賴清德原來是最大尾的抓耙子，也是蔡賴之間長期矛盾心結的根源」，https://www.facebook.com/284075159749398/posts/905173817639526 (accessed May 7, 2024).

40. Missshark，「賴皮德也是KMT派出的臥底？」，PTT. December 27, 2023. https://www.ptt.cc/bbs/HatePolitics/M.1703614267.A.BD1.html (accessed May 7, 2024).

41. 政經關不了，「張友驊驚爆春風專案內幕！賴清德被吸收當線民 人事時地大揭露」，Facebook. December 29, 2023. https://www.facebook.com/watch/?v=896824548837283 (accessed May 7, 2024).

42. 歐崇敬，「賴清德為家連哭5次！歐崇敬：趙少康繼續追打安全嗎？歐崇敬：為何台名嘴對賴清德地下黨員身分及春風專案不敢評論？」， YouTube. December 20, 2023. https://www.youtube.com/watch?v=k7SXR1FY9Ps (accessed May 7, 2024).

43. 法務部調查局，「法務部調查局針對網傳『賴清德為調查局春風專案線民』不實訊息澄清說明」， December 27, 2023. https://www.mjib.gov.tw/news/Details/29/957 (accessed May 7, 2024).

44. 法務部調查局，「法務部調查局針對網傳『賴清德為調查局春風專案線民』不實訊息澄清說明」，December 27, 2023. https://www.mjib.gov.tw/news/Details/29/957 (accessed May 7, 2024).

45. 「【缺乏背景】賴清德與春風專案的影片？錄音檔曝光？無實質證據佐證」，MyGoPen. December 26, 2023. https://www.mygopen.com/2023/12/informer.html (accessed May 7, 2024).

46. Featspoony，「線民德身份曝光，春風檔案、音檔流出」，PTT. December 23, 2023. https://www.ptt.cc/bbs/TSU/M.1703327674.A.A26.html (accessed May 7, 2024).

47. 「【謠言風向球】選舉傳言新招！ 用AI生成的政治影音 再用假帳號擴散」，台灣事實查核中心。January 5, 2024. https://tfc-taiwan.org.tw/articles/10119 (accessed May 7, 2024).

48. 「【影音變造】網傳影片『美國聯邦眾議員軍事委員會副主席魏特曼12月29日受訪公開為台灣某黨總統候選人拉票』？」，台灣事實查核中心。December 30, 2023. https://tfc-taiwan.org.tw/articles/10066 (accessed May 7, 2024).

49. Godisme73，「美國站邊了 發影片力挺賴蕭當選」，PTT. December 29, 2023. https://www.ptt.cc/bbs/Gossiping/M.1703843819.A.817.html (accessed May 7, 2024).

50. Qm671006，「PTT爆卦美國站邊支持民進黨」，Mobile01. December 29, 2023. https://www.mobile01.com/topicdetail.php?f=638&t=6897449 (accessed May 7, 2024).

51. 一百五，「綠營瘋傳，美政府已公開力挺賴蕭配，稱蕭是最能守護台灣、與美國配合的人」，Facebook. December 31, 2023. https://www.facebook.com/YiBaiWu/posts/pfbid02AY3xoqaPxRtz7onb8ddAcHGuaLXjBAMP4RaKq868vbEjpbgSKc8XvDaYkzgDB7tal (accessed May 7, 2024).

52. 「【影音變造】網傳影片『美國聯邦眾議員軍事委員會副主席魏特曼12月29日受訪公開為台灣某黨總統候選人拉票』？」，台灣事實查核中心。December 30, 2023. https://tfc-taiwan.org.tw/articles/10066 (accessed May 7, 2024).

53. Fuh101878，「遊說美國介選 or 側翼變造影片騙支持」，PTT. December 30, 2023. https://www.ptt.cc/bbs/HatePolitics/M.1703905871.A.7B3.html (accessed May 7, 2024).

54. Naruto，「Re: 遊說美國介選 or 側翼變造影片騙支持」，PTT. December 30, 2023. https://www.ptt.cc/bbs/HatePolitics/M.1703936255.A.1C7.html (accessed May 7, 2024).

55. QSearch only monitors public Facebook pages and does not track personal Facebook accounts.

56. 「【影音變造】網傳影片『美國聯邦眾議員軍事委員會副主席魏特曼 12 月 29 日受訪公開為台灣某黨總統候選人拉票』？」， 台灣事實查核中心。December 30, 2023. https://tfc-taiwan.org.tw/articles/10066 (accessed May 7, 2024).

57. 張之豪，「中国網軍現在四處散佈一本不知道是哪個白癡寫的《蔡英文秘史》，不外乎就是『數典忘祖』、『出身不純』、『媚日親美』」，Facebook. January 8, 2024. https://www.facebook.com/JihoTiun/posts/pfbid02Ey5UEVVDfqhFeYHqGhfbvZXc2C1xrde7ge5VRpFSYAWYG2ijtkRew9XNpMxh5wqsl (accessed May 7, 2024).

58. 「假的！大量 AI 生成『蔡英文秘史』入侵網路平台 鏡文學急處置封鎖」，鏡傳媒。January 10, 2024. https://www.mirrormedia.mg/story/20240110edi027 (accessed May 7, 2024).

59. 「大量 AI 生成影音攻擊蔡總統 國安人士研判中國網路作戰演練」，中央廣播電台。January 10, 2024. https://www.rti.org.tw/news/view/id/2192329 (accessed May 7, 2024).

60. Zhang, A. "As Taiwan Voted, Beijing Spammed AI Avatars, Faked Paternity Tests and 'Leaked' Documents." ASPI. January 18, 2024. https://www.aspistrategist.org.au/as-taiwan-voted-beijing-spammed-ai-avatars-faked-paternity-tests-and-leaked-fake-documents/ (accessed May 7, 2024).

61. Austin Horng-en Wang is an assistant professor in the Department of Political Science at the University of Nevada, Las Vegas, with research focuses including electoral behavior, East Asian politics, and political psychology. Profile page: https://www.unlv.edu/people/austin-wang (accessed May 7, 2024).

62. Zhang, A., *supra* note 60.

63. 中央廣播電台，「大量 AI 生成影音攻擊蔡總統 國安人士研判中國網路作戰演練」。January 10, 2024. https://www.rti.org.tw/news/view/id/2192329 (accessed May 7, 2024).

64. Zhang, A., *supra* note 60.

65. D-ID is a company that creates platforms for producing interactive audiovisual content using AI. Its products, such as Creative Reality™ Studio and AI Agents, can generate virtual characters and produce videos similar to news anchor broadcasts.

66. 張之豪，「中国網軍現在四處散佈一本不知道是哪個白癡寫的《蔡英文秘史》，不外乎就是『數典忘祖』、『出身不純』、『媚日親美』」，Facebook. January 8, 2024. https://www.facebook.com/JihoTiun/posts/pfbid02Ey5UEVVDfqhFeYHqGhfbvZXc2C1xrde7ge5VRpFSYAWYG2ijtkRew9XNpMxh5wqsl (accessed May 7, 2024).

67. "2024 Taiwan Presidential Election Information Manipulation AI Observation Report." Taiwan AI Labs. January 31, 2024. https://ailabs.tw/uncategorized/2024-taiwan-presidential-election-information-manipulation-ai-observation-report/ (accessed May 7, 2024).

68. Spamouflage was first exposed in 2019 by the social media research institute Graphika. This network used fake accounts on YouTube, Facebook, and Twitter to post videos in Chinese. For reference, see Nimmo, B., et al. "Cross-Platform Spam Network Targeted Hong Kong Protests." Graphika. September 25, 2019. https://graphika.com/reports/spamouflage (accessed May 7, 2024).

69. Jee, C. "An Indian Politician Is Using Deepfake Technology to Win New Voters." MIT Technology Review. February 19, 2020. https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/ (accessed May 7, 2024).

70. The Indian Deepfaker is renowned for its deepfake technology, particularly in the areas of image manipulation and voice cloning. Beyond aiding in election-related activities, the company also collaborates with prominent institutions like Netflix and New York University, integrating deepfake technology with marketing, education, and surveillance technologies. For more information, visit https://theindiandeepfaker.com/.

71. Sharma, Y. "Deepfake Democracy: Behind the AI Trickery Shaping India's 2024 Election." Al Jazeera. February 20, 2024. https://www.aljazeera.com/news/2024/2/20/deepfake-democracy-behind-the-ai-trickery-shaping-indias-2024-elections (accessed May 7, 2024).

72. Christopher, N. "How AI Is Resurrecting Dead Indian Politicians as Election Looms." Al Jazeera. February 12, 2024. https://www.aljazeera.com/economy/2024/2/12/how-ai-is-used-to-resurrect-dead-indian-politicians-as-elections-loom (accessed May 7, 2024); Purtill, J., " 從深度偽政治背書到聊天機器人競選者：人工智慧正在改變選舉方式 ." ABC News. February 22, 2024. https://www.abc.net.au/chinese/2024-02-22/ai-eletions-deepfakes-generative-campaign-endorsement-democracy/103500274 (accessed May 7, 2024).

73. Sharma, Y., *supra* note 71.

74. The Influence Industry Project, "2024 Indian Elections Series: Volunteers, Government Data, and Regulation." 2024. https://influenceindustry.org/en/highlights/highlight-india-case-studies/ (accessed May 7, 2024).

75. Firstpost. "IT Min Rajeev Chandrasekhar Warns Social Platforms to Act Against Deepfakes Ahead of 2024 General Election." January 29, 2024. https://www.firstpost.com/tech/it-min-rajeev-chandrasekhar-warns-social-platforms-to-act-against-deepfakes-ahead-of-2024-election-13664432.html (accessed May 7, 2024).

76. Edwards, B. "Fake Pentagon 'Explosion' Photo Sows Confusion on Twitter." Ars Technica. May 23, 2023. https://arstechnica.com/information-technology/2023/05/ai-generated-image-of-explosion-near-pentagon-goes-viral-sparks-brief-panic/ (accessed May 2, 2024).

77. Oremus, W., Harwell, D., & Armus, T. "A Tweet About a Pentagon Explosion Was Fake. It Still Went Viral." Washington Post. May 22, 2023. https://www.washingtonpost.com/technology/2023/05/22/pentagon-explosion-ai-image-hoax/ (accessed May 7, 2024).

78. Marcelo, P. "FACT FOCUS: Fake Image of Pentagon Explosion Briefly Sends Jitters through Stock Market." AP News. May 23, 2023. https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4 (accessed May 7, 2024).

79. Passantino, D. O. "'Verified' Twitter Accounts Share Fake Image of 'Explosion' Near Pentagon, Causing Confusion." CNN. May 22, 2023. https://www.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html (accessed May 7, 2024).

80. *Id.*

81. Marcelo, P., *supra* note 78.

82. *Id.*

83. See, e.g., Tolan, C., O'Sullivan, D., & Winter, J. "How a Biden AI Robocall in New Hampshire Allegedly Links Back to a Texas Strip Mall." CNN Politics. February 8, 2024. https://www.cnn.com/2024/02/07/politics/biden-robocall-texas-strip-mall-invs/index.html (accessed May 7, 2024); Pezenik, S., & Shepherd, B. "Fake Biden Robocall Urges New Hampshire Voters to Skip Their Primary." ABC News. January 22, 2024. https://abcnews.go.com/Politics/fake-biden-robocall-urges-new-hampshire-voters-skip/story?id=106580926 (accessed May 7, 2024); Seitz-Wald, A. "Democratic Operative Admits to Commissioning Fake Biden Robocall That Used AI." NBC News. February 26, 2024. https://www.nbcnews.com/politics/2024-election/democratic-operative-admits-commissioning-fake-biden-robocall-used-ai-rcna140402 (accessed May 7, 2024); Ali, S., & Will, W. "AI-Generated Robocall Impersonates Biden in Apparent Attempt to Suppress Votes in New Hampshire." The Globe and Mail. January 22, 2024. https://www.theglobeandmail.com/world/us-politics/article-ai-generated-robocall-impersonates-biden-in-apparent-attempt-to/ (accessed May 7, 2024).

84. Balasubramaniyan, V. "Pindrop Reveals TTS Engine Behind Biden AI Robocall." Pindrop. January 25, 2024. https://www.pindrop.com/blog/pindrop-reveals-tts-engine-behind-biden-ai-robocall (accessed May 2, 2024).

85. FCC. "FCC Makes AI-Generated Voices in Robocalls Illegal." Federal Communications Commission. February 8, 2024. https://www.fcc.gov/document/fcc-makes-ai-generated-voices-robocalls-illegal (accessed May 7, 2024).

86. O'Brien, M., & Swenson, A. "Tech Companies Sign Accord to Combat AI-Generated Election Trickery." AP News. February 16, 2024. https://apnews.com/article/ai-generated-election-deepfakes-munich-accord-meta-google-microsoft-tiktok-x-c40924ffc68c94fac74fa994c520fc06 (accessed May 2, 2024).

87. Ali, S., & Will, W., *supra* note 83.

88. Balasubramaniyan, V., *supra* note 84.

89. The White House. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." The White House. October 30, 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/ (accessed May 7, 2024).

90. Khatsenkova, S. "Did France 24 Air a Segment Claiming Ukraine Ordered Emmanuel Macron's Assassination?" Euro News. February 20, 2024. https://www.euronews.com/my-europe/2024/02/20/did-france-24-air-a-segment-claiming-ukraine-ordered-emmanuel-macrons-assassination (accessed May 7, 2024).

91. "Fact Check: Viral Video Purporting to Show France24 Report about Alleged Macron Assassination Attempt is a Deepfake." CheckYourFact. February 19, 2024. https://checkyourfact.com/2024/02/19/fact-check-france24-macron-assassination-attempt-deepfake/ (accessed May 7, 2024).

92. Khatsenkova, S., *supra* note 90.

93. *Id.*

94. Wirtschafter, V. "The Impact of Generative AI in a Global Election Year." Technical report. The Brookings Institution. January 30, 2024. https://www.brookings.edu/articles/the-impact-of-generative-ai-in-a-global-election-year/ (accessed May 7, 2024).

95. Sedova, K., et al. "AI and the Future of Disinformation Campaigns. Part 2: A Threat Model." CSET. December 2021. https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns/ (accessed May 7, 2024).

96. "Forecasting Potential Misuses of Language Models for Disinformation Campaigns and How to Reduce Risk." OpenAI. January 11, 2023. https://openai.com/research/forecasting-misuse (accessed May 7, 2024).

97. Sedova, K., et al., *supra* note 95.

98. Fredheim, R. "Virtual Manipulation Brief 2023/1: Generative AI and Its Implications for Social Media Analysis." NATO Strategic Communications Centre of Excellence. June 5, 2023. https://stratcomcoe.org/publications/virtual-manipulation-brief-20231-generative-ai-and-its-implications-for-social-media-analysis/287 (accessed May 7, 2024).

99. Nimmo, B., et al. "Spamouflage Goes to America." Graphika. August 12, 2020. https://graphika.com/reports/spamouflage-dragon-goes-to-america (accessed May 7, 2024).

100. Sedova, K., et al., *supra* note 95.

101. *Id.*

102. "Forecasting Potential Misuses of Language Models for Disinformation Campaigns and How to Reduce Risk." OpenAI. January 11, 2023. https://openai.com/research/forecasting-misuse (accessed May 7, 2024).

103. 陳培煌、陳慧敏，「【謠言風向球】外交假公文事件簿 查核中心帶你一起洞察造謠手法」， Edited by 陳偉婷，台灣事實查核中心。August 31, 2023. https://tfc-taiwan.org.tw/articles/9533 (accessed May 7, 2024).

104. “Fake Pentagon ‘Explosion’ Photo Sows Confusion on Twitter.” Edwards, B. May 23, 2023. https://arstechnica.com/information-technology/2023/05/ai-generated-image-of-explosion-near-pentagon-goes-viral-sparks-brief-panic/ (accessed May 7, 2024).

105. “Digital Threats from East Asia Increase in Breadth and Effectiveness.” Microsoft Threat Intelligence. September 7, 2023. https://www.microsoft.com/en-us/security/business/security-insider/reports/nation-state-reports/digital-threats-from-east-asia-increase-in-breadth-and-effectiveness/ (accessed May 7, 2024).

106. “Fake Image Factories. How AI Image Generators Threaten Election Integrity and Democracy.” Center for Countering Digital Hate. March 6, 2024. https://counterhate.com/research/fake-image-factories/ (accessed May 7, 2024).

107. “2024 Taiwan Presidential Election Information Manipulation AI Observation Report.” Taiwan AI Labs. January 31, 2024. https://ailabs.tw/uncategorized/2024-taiwan-presidential-election-information-manipulation-ai-observation-report/ (accessed May 7, 2024).

108. 「人造多重宇宙：2024台灣大選境外資訊操作與影響觀察報告」，台灣民主實驗室。June 5, 2024. https://medium.com/doublethinklab-tw/493423f9bba8 (accessed June 6, 2024).

109. OpenAI, *supra* note 102.

110. Sedova, K. et al., *supra* note 95.

111. “2nd EEAS Report on Foreign Information Manipulation and Interference Threats.” European External Action Service. January 23, 2024. https://www.eeas.europa.eu/eeas/2nd-eeas-report-foreign-information-manipulation-and-interference-threats_en (accessed May 7, 2024).

112. *Id.*

113. “Scammers Made a Biden Deepfake. Here’s Why It Wasn’t Very Good.” POLITICO Tech. January 26, 2024. https://politico-tech.simplecast.com/episodes/scammers-made-a-biden-deepfake-heres-why-it-wasnt-very-good (accessed May 7, 2024).

114. “Audio Deepfake Scams: Criminals Are Using AI to Sound like Family and People Are Falling for It.” Khatsenkova, S. March 25, 2023. https://www.euronews.com/next/2023/03/25/audio-deepfake-scams-criminals-are-using-ai-to-sound-like-family-and-people-are-falling-fo (accessed May 7, 2024).

115. “Forecasting Potential Misuses of Language Models for Disinformation Campaigns and How to Reduce Risk.” OpenAI. January 11, 2023. https://openai.com/research/forecasting-misuse (accessed May 7, 2024).

116. Hung, C., Fu, W., Liu, C., & Tsai, H. “AI Disinformation Attacks and Taiwan’s Responses during the 2024 Presidential Election.” Taiwan Communication Association. April 12, 2024.

117. “China Tests US Voter Fault Lines and Ramps AI Content to Boost Its Geopolitical Interests.” Microsoft Threat Analysis Center. April 4, 2024. https://blogs.microsoft.com/on-the-issues/2024/04/04/china-ai-influence-elections-mtac-cybersecurity/ (accessed May 7, 2024).

118. 「2024台灣選舉：境外資訊影響觀測報告初步分析」，台灣民主實驗室。January 19, 2024. https://medium.com/doublethinklab-tw/fe7f819aeabd (accessed May 7, 2024); 「人造多重宇宙：2024台灣大選境外資訊操作與影響觀察報告」，台灣民主實驗室。June 5, 2024. https://medium.com/doublethinklab-tw/493423f9bba8 (accessed June 6, 2024).

119. “Cyber Threats against Taiwan’s 2024 Presidential Election.” Team T5. March 4, 2024. https://teamt5.org/en/posts/whitepaper-cyber-threats-against-taiwan-s-2024-presidential-election/ (accessed May 7, 2024).

120. Zhang, A. “As Taiwan Voted, Beijing Spammed AI Avatars, Faked Paternity Tests and ‘Leaked’ Documents.” ASPI. January 18, 2024. https://www.aspistrategist.org.au/as-taiwan-voted-beijing-spammed-ai-avatars-faked-paternity-tests-and-leaked-fake-documents/ (accessed May 7, 2024).

121. Hung, C., et al., *supra* note 115; Microsoft Threat Analysis Center, *supra* note 117.

122. 「2024台灣選舉：境外資訊影響觀測報告初步分析」， 台灣民主實驗室。January 19, 2024. https://medium.com/doublethinklab-tw/fe7f819aeabd (accessed May 7, 2024); 「人造多重宇宙：2024台灣大選境外資訊操作與影響觀察報告」，台灣民主實驗室。June 5, 2024. https://medium.com/doublethinklab-tw/493423f9bba8 (accessed June 6, 2024).

123. Taiwan AI Labs, *supra* note 107.

124. Zhang, A., *supra* note 120.

125. Hung, C. et al., *supra* note 116.

126. Zhang, A., *supra* note 120.

127. *Id.*

128. *Id.*

科技、民主與社會研究中心

Research Institute for **Democracy**, **Society** and **Emerging Technology**