

Adaptation and Innovation

MARCH 2025

The Civic Space Response to AI-Infused Elections



Editor
Isabel Linzer

With case studies by:
Laura Zommer
Kian Vesteinsson
Isabel Linzer



The **Center for Democracy & Technology (CDT)** is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

References in this report include original links as well as links archived and shortened by the Perma.cc service. The Perma.cc links also contain information on the date of retrieval and archive.



This report is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Adaptation and Innovation

The Civic Space Response to AI-Infused Elections

Isabel Linzer

This report features case studies by multiple authors.

Laura Zommer wrote the Mexico case study, *Kian Vesteinsson* wrote the Taiwan case study, and *Isabel Linzer* was the report editor and wrote the introduction and case study on AI incident tracking.

Interviews with external experts contributed substantially to the research presented here. Thank you to Daniel Moreno, Dulce Ramos, Tania Montalvo, Liliana Elósegui, Daniela Mendoza for their time and insight for the Mexico case study. We would also like to thank the following people for sharing their time and expertise regarding their respective AI incident tracking projects: Taylor Barkley, Emerson T. Brooking, Cristina de la Puerta, Meredith Furbish, Adrienne Goldstein, Sean McGregor, Kevin Paeth, Max Rizzuto, Dina Sadek, Zeve Sanderson, Daniel Schiff, Kaylyn Jackson Schiff, Christina Walker, Logan Whitehair, and Michael Zelenko.

Additional thanks to Natalia Colombo for support on the Mexico case study and to Allie Funk, Lulu Keng, and Kevin Slaten for reviewing the Taiwan case study. Finally, the report would not have been possible without the support, feedback, and contributions of Tim Harper, Becca Branum, Aliya Bhatia, Amy Winecoff, Nathalie, Maréchal, Samir Jain, Tim Hoagland, and Drew Courtney.

This work is made possible through a grant from the Open Society Foundations.

Contents

Introduction	6
Case Study: Fact-Checking Institutionalized Disinformation in Mexico	10
I. Introduction	10
II. Country context	12
III. Fact-checking and resilience	14
Case Study: Foreign Interference and Decentralized Resilience in Taiwan	24
I. Introduction	24
II. Country context	26
III. Assessment of three interventions	28
Case Study: AI Incident Tracking by Global Civic Space Actors	39
I. Introduction	39
II. Survey of incident tracking	41
III. Challenges, tradeoffs, and answers about the impact of AI	44



Contents

Recommendations	59
Fostering collaboration among civil society	59
Building company policies and information resilience	60
Strengthening information access for research and informed policymaking	61
Endnotes	63



01

Introduction

By Isabel Linzer

AI avatars delivered independent news about Venezuela’s contested election, allowing journalists to protect their identity and avoid politically motivated arrest.¹ Voters in the United Kingdom could cast their ballots for an AI avatar to hold a seat in Parliament.² A deepfake video showed United States President Joe Biden threatening to impose sanctions on South Africa if the incumbent African National Congress won.³

These are a few of the hundreds of ways generative AI was used during elections in 2024, a year that was touted as “the year of elections” and described as the moment in which newly widespread AI tools could do lasting damage to human rights and democracy worldwide. Though technology and security experts have described deepfakes as a threat to elections since at least the mid to late 2010s,⁴ the concentrated attention in 2024 was a reaction to the AI boom in the preceding year. In September 2023, a leading parliamentary candidate in Slovakia lost after a fake audio smearing him was released two days before the election, prompting speculation that the deepfake had changed the election results.⁵ At the beginning of the year, OpenAI’s ChatGPT set a record as the “fastest-growing consumer application in history.”⁶

Though 2024 ended with debates over the extent to which the risks AI posed to elections were overstated, in one way the consequences were clear: the technology changed the way stakeholders around the world did their work. Governments from Brazil to the Philippines passed new laws and regulations to govern the use of generative AI in elections.⁷ The European Commission published guidelines for how large companies should protect the information environment ahead of the June 2024 elections, including by labeling AI-generated content.⁸ US election administrators adopted new communication tactics that were tailored to an AI-infused information environment.⁹

Political campaigns and candidates adopted AI tools to create advertisements and help with voter outreach.¹⁰ Candidates in Indonesia paid for a service that used ChatGPT to write speeches and develop campaign strategies.¹¹ In India, candidates used deepfake audio and video of themselves to enable more personalized outreach to voters.¹² Germany's far right AfD party ran anti-immigrant ads on Meta platforms, some of which incorporated AI-altered images.¹³

Social media platforms and AI developers implemented some election integrity programs, despite recent cuts to trust and safety teams.¹⁴ Twenty-seven technology companies signed the AI Elections Accord, a one-year commitment to addressing "deceptive AI election content" through improved detection, provenance, and other efforts.¹⁵ Google restricted the Gemini chatbot's responses to election-related queries,¹⁶ and OpenAI announced that ChatGPT would redirect users to external sources when users asked about voting ahead of certain elections.¹⁷ Google and Jigsaw worked with media, civil society, and government partners on public media literacy ahead of the European Union elections, including about generative AI.¹⁸

In anticipation of AI tools accelerating or increasing threats to the information environment, civic space actors changed their work, too. This report looks at their contributions to a resilient information environment during the 2024 electoral periods through three case studies: (I) fact-checking collectives in Mexico, (II) decentralization and coordination among civil society in Taiwan, and (III) AI incident tracking projects by media, academics, and civil society organizations.

The case studies highlight a range of approaches to building resilient information environments. They show the ways artificial intelligence complicates that work, as well as how it can be used to support resilience building efforts. The mix of approaches — from fact-checking bots on WhatsApp to cataloging hundreds of deepfakes — tap into information resilience from different angles.

The Mexico case study focuses on the development and tactics of fact-checking collectives, especially in the context of a hostile media environment. The case study also considers the role of AI-generated content in the 2024 election and how WhatsApp and AI are used in fact-checking work.

The Taiwan case study also examines a collaborative but decentralized civil society model. Unlike the Mexican case, however, Taiwan is subject to a prolific amount of Chinese government-linked disinformation campaigns. The case study considers the roles of research into influence operations, fact-checking or information literacy programming, and government policies to counter misinformation.

The third case study looks at how civil society organizations, journalists, and academics tracked the use of generative AI in elections throughout 2024, both in the US and globally. Their work was an important contribution to the current public understanding of how AI was used and offers lessons for improving research and policy in the future, including challenges in data collection and how to conduct well-balanced research on such a high-profile subject. The interviews CDT conducted for this case study also give a snapshot of expert thinking on the impact that generative AI had on elections in 2024.

Though the case studies span different political contexts and types of interventions, common themes emerged. Organizations benefited from complementary or collaborative work with peer groups. They also used AI to bolster their own work. Civic space actors contended with funding and capacity constraints, insufficient access to information from companies, difficulty detecting and verifying AI-generated content, and the politicization of media resilience work, including fact-checking.

Finally, the case studies emphasize that the issue of AI in elections is not temporary. Civic space actors have been addressing the risks and exploring the opportunities AI presents for years — long before the media and policy attention of 2024. These groups will continue to be invaluable resources and partners for public and private actors in 2025 and beyond.

And their work will be urgently needed. The end of companies' commitments under the AI Elections Accord and a global political environment that is increasingly hostile to work relating to elections, fact-checking, and disinformation research mark an absence of leadership on the most pressing threats to information resilience.¹⁹ To that end, the report concludes with recommendations for how companies and civic space actors can continue to support information resilience by fostering collaboration, developing company policies, and strengthening transparency and data access.



02

Case Study: Fact-Checking Institutionalized Disinformation in Mexico

By Laura Zommer

Co-founder and CEO of Factchequeado.

Laura is a recipient of the Maria Moors Cabot Prize Special Citation 2024, as well as a Knight Fellow with the International Center for Journalists (ICFJ) and an Ashoka Fellow. For more than a decade, Laura was the executive director of Chequeado and created LatamChequea, a regional fact-checking network in Latin America. Laura is a member of CDT's Advisory Council and has received numerous other honors for her work in journalism and fact-checking.

I. Introduction

"I have other data (tengo otros datos)," was a frequent refrain from former Mexican President Andrés Manuel López Obrador during his daily press conference, "La Mañanera."²⁰ He used it last year to refute Univision's Jorge Ramos when he was confronted by the journalist with official government statistics highlighting the failure of the "hugs, not bullets" ("abrazos, no balazos") policy to control violence and reduce homicides in Mexico.²¹ "Other data" became the headline of hundreds of journalistic articles during López Obrador's tenure, inspired songs, and was turned into a meme, all of which highlighted why fact-checking political leaders is crucial regardless of political ideology.²² It also underscored the importance of grounding political or public policy debate in reliable information: otherwise, opportunistic actors are free to make claims without evidence or accountability.

Even before the popularization of generative artificial intelligence, social networks and their opaque algorithms shaped the information environment. People around the world could share information more easily than ever, but often the content that was most contentious or provocative spread most widely,²³ even as companies introduced policies to reduce mis- and disinformation on their platforms.

Fact-checking is a means to building informational resilience, protecting elections, and strengthening democracy. Research shows that it does not cause people to change their overarching beliefs, but can correct an individual's factual knowledge²⁴ and reduce the spread of inaccurate information.²⁵ However, for fact-checking to have an impact, it needs speed and scale,²⁶ making AI and technological innovation crucial strategic allies.

Generative AI tools in particular present a serious challenge to the information environment, but they may also have a role to play in strengthening fact-checking efforts, particularly in languages other than English, where information gaps are often substantial. This case study assesses fact-checking efforts during Mexico's 2024 election in light of AI, and discusses the following findings:

- **The use of AI in the 2024 electoral campaign was not as extensive or revolutionary as many had predicted.** However, AI-generated images and audio did impact candidates, particularly in local elections. Generative AI tools have the potential to create further disruptions in future elections, including in 2025 when for the first time the people in Mexico will vote for judges and magistrates.²⁷
- **WhatsApp creates an opportunity for low-friction, user-driven fact-checking.** Fact-checking initiatives have created chatbots on WhatsApp, which is widely used in Mexico, to disseminate content and interact with users. The upsides of this model are especially notable because fact-checking on encrypted platforms requires different strategies than traditional methods, such as labeling.
- **Collaboration is a key component of successful fact-checking in Mexico.** Collective efforts allow for increased efficiency and better engagement with local communities. Close communication between practitioners and developers is crucial to efficient deployment of new technology, including AI, in fact-checking efforts.
- **Artificial intelligence can make fact-checking faster and more efficient.** Fact-checkers in Mexico have used AI for several years to alleviate bottlenecks in their work. This is especially important because speed is a critical component of effective fact-checking.

- **English-dominated technology complicates fact-checking.** Because English is the main language of technological development, fact-checking Spanish language deepfakes is more difficult, and Spanish-speaking developers have reported that they face additional barriers, like the need to re-train models, when working on automated fact-checking projects.

II. Country context

On June 2, 2024, Mexicans went to the polls to elect over 20,000 public officials, including the president and both chambers of the national legislature.²⁸ For the first time in Mexico's history, the 2024 presidential elections featured two leading female candidates. Claudia Sheinbaum Pardo, former mayor of Mexico City and the governing party's candidate, won the presidency, becoming the first woman to hold the office. Sheinbaum secured nearly 60% of the presidential vote to defeat Xóchitl Gálvez, a senator and tech entrepreneur who represented a coalition of historically divergent parties united by their opposition to López Obrador, and Jorge Álvarez Máynez, a relatively unknown former federal congressman from the Movimiento Ciudadano.²⁹

Social media was an important campaign tool for all of the candidates,³⁰ though the Digital Media Observatory (OMD) of the Tecnológico de Monterrey found that Sheinbaum was the dominant candidate online, based on a review of Facebook, Instagram, YouTube, and X data.³¹ Mexicans increasingly rely on social media for news, while the importance of print and television is declining.³²

Though the election of Mexico's first female president was a historic marker of social progress, Sheinbaum's election also promised continuity with the populist policies of López Obrador, founder of the Morena Party (Partido Morena) and mentor to Sheinbaum, who governed from 2018 to 2024. Political analysts and academics have described how the López Obrador administration created a communication system that includes disinformation as a central element.³³ "La Mañanera" became institutionalized as the platform where, from Monday to Friday, the former president—and now

Sheinbaum—could guide the focus of the media and public debate in press conferences lasting between one and three hours. López Obrador reportedly made over 100,000 false, misleading, and unprovable claims during La Mañanera during his first four years in office.³⁴ During these events, López Obrador and Sheinbaum not only promote the achievements of their administrations, but also criticize and even denounce their perceived opponents—including journalists, academics, and civil society leaders.

“The Mexican government over the past six years is very different from the one we were fact-checking when we launched [fact-checking efforts] El Sabueso [in 2015] or during Verificado 2018,” said Tania Montalvo, former managing editor at Animal Político and Verificado 2018, and now Programme Manager at the Reuters Institute at the University of Oxford.³⁵ “The level of propaganda in Mexico is on par with Bukelismo in El Salvador.”

Daniel Moreno, who founded Animal Político and has faced attacks from the government, echoed this sentiment: “We spent six years calling the president a liar — with evidence — and yet his party won this year with two out of three votes. There wasn’t a single Mañanera where Andrés Manuel didn’t lie, but he created a system of ‘alternative media or dissemination channels’ to amplify those lies and harass anyone who debunked them, and it proved highly effective.”³⁶ Pro-government accounts, including a coordinated network of YouTube channels, amplify official narratives and other pro-government content, though direct links with the government have not been established.³⁷ Meanwhile, there is also evidence of coordinated online activity and disinformation in opposition to the López Obrador government, including during the 2024 election period.³⁸

Since López Obrador took office in 2018, polling indicates that trust in news has decreased by about 15 percentage points.³⁹ During his tenure, López Obrador regularly attacked the media and unaligned journalists. Mexican fact-checking organizations and journalists have been the target of repeated harassment and threats.⁴⁰ López Obrador came under greater international scrutiny in February 2024 after doxxing a New York Times journalist who was investigating alleged links between López Obrador and drug cartels.⁴¹

The hostility and violence towards the media make accountability efforts — including fact-checking — more urgent, more difficult, and more dangerous.

According to the Electoral Laboratory’s Report on Political Violence, the 2024 elections were the most violent elections in the history of Mexico.⁴² The murders of 13 candidates, 17 pre-candidates and 11 aspirants were recorded during the electoral process. One of the most important ways this has affected the electoral system

is participation: according to the aforementioned report, “in the 17 municipalities where we recorded the greatest number of attacks, there was an average decrease of 7% in participation compared to the 2018 election.”

Freedom House classifies Mexico as “partly free” in its global freedom index; other international civil society organizations, including Reporters Without Borders and the Committee to Protect Journalists, have documented violence against journalists in Mexico.⁴³ In November 2024 the Inter-American Commission on Human Rights’ (IACHR) Special Rapporteur for Freedom of Expression of the Organization of American States condemned the violent media environment, including the murder of at least seven members of the media.^{1 44} The hostility and violence towards the media make accountability efforts — including fact-checking — more urgent, more difficult, and more dangerous.

III. Fact-checking and resilience

Amid a challenging political environment, Mexican fact-checkers have built a collaborative international community and continue to adapt to technological changes. This section analyzes the development of the collective model and considers recent approaches related to WhatsApp, AI chatbots, and deepfake audio.

1 The Special Rapporteur expressed deep concern over the murder of at least seven media professionals, including journalists Enrique Hernández Avilez, Roberto Figueroa, Víctor Morales, Alejandro Martínez Noguez, Mauricio Cruz Solís, and Patricia Ramírez González, as well as radio host Martín Antonio Olivier Rodríguez.

A. Origins of fact-checking in Mexico

Fact-checking in Mexico is characterized by collective efforts as well as innovative responses to, and uses of, technology. **These efforts predated López Obrador's presidency, but rose to greater prominence during his time in office.** In January 2015, the independent digital media outlet Animal Político established El Sabueso. It was inspired by Chequeado in Argentina and Politifact in the United States, as recalled in separate interviews conducted in December 2024 with its founder, Daniel Moreno, and its first editor-in-chief, Dulce Ramos. "From the beginning, the initiative was characterized by innovation: it personified the fact-checker as a dog, used caricatures for ratings, and fully embraced social media to disseminate verified content," explained Ramos.⁴⁵ El Sabueso remains the only Mexican fact-checking initiative certified by the International Fact-Checking Network (IFCN)⁴⁶ and is part of Meta's third-party fact-checking program.⁴⁷

During the 2018 presidential elections, Mexican fact-checking gained national and international prominence with the launch of the alliance Verificado 2018 (known on social media as @VerificadoMX).⁴⁸ This initiative involved a team of 90 people, including professionals and volunteers, who worked collaboratively for four months. This pioneering regional initiative was a collaboration between Animal Político, Pop-Up Newsroom, and AJ+ Español. The network they set up, which included over 60 media outlets, universities, and civil society organizations, republished fact-checks in text, video and other formats to reach a wider audience. Additionally, the team established a WhatsApp number where people could send dubious chain messages and other suspicious content and used a hashtag, #QuieroQueVerifiquen (#IWantYouToVerify), to be fact-checked.⁴⁹

"Verificado 2018 was the first initiative to formally establish a high-quality multimedia journalism desk. We combined formats that broke predefined fact-checking models, presenting information in an engaging and concise manner, with elements that encouraged audience interaction and participation. We didn't limit ourselves to videos; we explored and adapted other formats such as loops, graphics, GIFs, WhatsApp statuses, live streams, and social

media cards,” recalled Alba Mora, editor of AJ+ Español, in a 2018 interview.⁵⁰

“Verificado 2018 positioned itself as the only non-political actor in the election, as citizens don’t fully trust electoral bodies and see them as aligned with political parties. This was very important because it brought the issue of disinformation and fake news into public conversations,” said journalist and professor Daniela Mendoza in a 2020 interview published by Heinrich Böll Stiftung.⁵¹ Mendoza founded Verificado MX, a separate fact-checking organization, in July 2017. Based in Monterrey, the organization is staffed entirely by women and focuses on daily fact-checking, including statements made by the president during La Mañanera, and media literacy.⁵²

Escenario Tlaxcala is another locally focused, small-fact-checking initiative. Originally an internet radio platform, it has since incorporated fact-checking into its work.⁵³ Like El Sabueso and Verificado MX, it is part of the LatamChequea network, which currently brings together over 40 fact-checking organizations from 21 countries.⁵⁴ **International and domestic fact-checking collectives are an opportunity to expand reach and share best practices among partners.** Members can capitalize on shared resources while leveraging their role as trusted local sources to reach different audiences. Expansive partnerships are also helpful protection in the Mexican context, where critical journalism can be dangerous.

B. Assessing AI audio

Although generative AI tools can make deepfakes quickly and easily, false and misleading content is not an AI-specific problem.

According to Arturo Daen, editor of Animal Político’s El Sabueso, most election-related disinformation was not AI-generated.⁵⁵ There were some exceptions, however, and examples of audio content that were potentially generated by AI emerged as an important test of fact-checking efforts in Mexico.

In November 2024, an alleged audio deepfake featured Martí Batres, head of government in Mexico City, planning to interfere with the race to choose

his successor.⁵⁶ Batres denied the veracity of the audio, but the ensuing debate over its origin highlighted **the particular challenge of verifying audio**. “We analysed the material and different tools could not identify with precision if this technology was actually used,” Daen told Reuters Institute.⁵⁷ “The case generated alarm for the possible proliferation of this kind of audio and the difficulty of verifying them. Since then, however, we have not encountered more of this type of material, at least at that scale or level of impact.” In response to the Batres deepfake incident, VerificadoMX published an article about how to identify when content was generated by AI and warning about the challenges in verifying this content, especially audio.⁵⁸

Sheinbaum was the target of several AI-generated audio and video clips, including an audio clip in which she said her campaign was failing in a key state.⁵⁹ One audio clip of then-candidate Sheinbaum went viral on WhatsApp during the 2024 campaign and was widely cited as an example of the challenges that artificial intelligence creates for election integrity. In the 35-second audio clip, Claudia Sheinbaum is heard saying that “the president [López Obrador] represents the past.” It was later determined this clip was created from audio extracted from an old interview.⁶⁰

This example underscores the fact that, although generative AI tools can make deepfakes quickly and easily, false and misleading content is not an AI-specific problem.

The instances of fake or allegedly fake audio recordings in Mexico emphasized the limitations of fact-checking in light of new technological advancements — and specifically in the absence of better provenance and detection techniques. It also

highlighted disparities in technological development; **non-English languages have long been understudied and insufficiently accounted for in technological development**, including in automated content moderation and training LLMs.⁶¹ That English is the primary language of detection development makes identifying Spanish-language deepfakes even more challenging.

That English is the primary language of detection development makes identifying Spanish-language deepfakes even more challenging.

C. The challenge and opportunity of WhatsApp and AI

WhatsApp is a widely used messaging platform in Latin America and among Spanish-speaking communities and is the third most popular social media and messaging platform in Mexico, following Facebook and YouTube.⁶² It offers inexpensive, real-time communication that makes it a tool for sharing news and misinformation alike. People tend to trust information from peers, friends, and family,⁶³ which amplifies the platform's role in the rapid spread of both accurate and misleading content. WhatsApp's end-to-end encryption — which prevents outside access to the content of users' communications — maintains privacy and limits en-masse interventions. Common anti-disinformation tactics, like content moderation and traditional fact-checking strategies, are not possible on encrypted messaging platforms,⁶⁴ requiring creative solutions to improve information resilience while respecting user privacy. For example, WhatsApp has implemented other efforts such as forwarding limits and user reporting to support users' ability to parse and interpret potentially false information, as well as limit its spread.⁶⁵

Nonetheless, **WhatsApp offers significant benefits as a community building tool for fact-checkers**, primarily due to its widespread use and accessibility across diverse populations. In theory, fact-checking on WhatsApp is an immediate, targeted, and personal approach, which presents an opportunity to build trust, correct misinformation, and empower users with the tools to better assess the content they encounter. Operating within trusted personal networks gives fact-checkers a way to reach individuals directly and serve as a trusted source. Additionally, the ability to facilitate real-time communication, outside of a traditional media timeline, enables fact-checkers to address misinformation quickly and efficiently — a key component of reducing the spread of falsehoods. In other words, Mexican fact-checkers have turned the private, person-to-person nature of WhatsApp that is often seen as a weakness in the information environment into a strategic opportunity.

Responsive, user-initiated fact-checking on WhatsApp also stands in contrast with traditional media, which offers less flexibility and personalization.

During the 2024 Mexican election, different initiatives used WhatsApp to combat misinformation. A commonality among these WhatsApp-based initiatives was the bottom-up approach to fact-checking. Users asked questions through trusted channels, which is distinct from more traditional top-down interventions like content-labeling or news alerts. Responsive, user-initiated fact-checking on WhatsApp also stands in contrast with traditional media, which offers less flexibility and personalization.

The National Electoral Institute (INE), Agence France-Presse (AFP), Animal Político, Telemundo, and Meedan collaborated to launch a fact-checking chatbot on April 17, 2024 to debunk false claims about the electoral process. Users could send their questions, including multimedia and links, to the INE channel, called Inés, and receive immediate answers based on fact-checks and explainers that the media organizations submitted to a shared database. Questions about how the election worked (such as how and where to vote) were by far the most popular, followed by requests for fact-checking news. Uptake and fact-checking rates were low; Meedan reported that 6,940 queries were submitted in 54 days. Inés responded to 18% of queries, which meant that those questions were about electoral processes and could be answered using content that was already in the database. Of the remaining 82% of unanswered questions, approximately 4 in 10 could have been answered if Inés had the technical capability to respond belatedly, after a media outlet added relevant information to the database.⁶⁶

Similarly, El Sabueso deployed VerifiChat, a WhatsApp-based service that enabled users to easily verify the accuracy of content by sending links, videos, or images.⁶⁷ Fact-checks were communicated through private messages, often with detailed explanations and links to full articles.

Some initiatives not only focused on debunking misinformation but also built information resilience in other ways, including through prebunking. For example, El Sabueso's newsletter highlighted common misinformation stories, enabling users to identify

falsehoods early on. By responding to questions about voting procedures in addition to fact-checking potential misinformation, Inés may have also built resilience to misinformation by equipping users with reliable information. Both initiatives relied on user questions to guide the publication of fact-checking content, helping to ensure that widely shared concerns were addressed most urgently.

That Mexican voters could access verified election information by asking questions through WhatsApp chats at no cost was, without a doubt, a positive thing. In addition to helping users confirm reliable information, they also helped fact-checking organizations and the electoral authorities to identify information gaps and produce new information.

These efforts also came with challenges, including the need to expand uptake while managing costs. As the Inés data indicates, **usage and response rates would need to increase for the tool to have wide impact**, but it nevertheless demonstrated the potential of WhatsApp-based fact-checking models. That Inés operated for fewer than two months before election day may have contributed to low uptake. In the future, consistent availability, or at least longer pre-election availability, could promote informational chatbot usage before the election cycle begins.

Unlike other WhatsApp-based tools discussed above, Inés raises another critical aspect of the fact-checking ecosystem: government collaboration. While the partnership between INE media offered advantages in terms of efficiency and potential scale, it also raised a challenge for trust in the media regarding their impartiality and independence. **There are risks inherent to collaborating with government agencies which merit serious discussion.** While improving access to authoritative information from the government about time, place, and manner of voting is a positive use for this type of partnership, normalizing the government as an arbiter of truth in other contexts, and regarding mis- and disinformation more broadly, has the potential to set a dangerous precedent.

Government partnership models have been employed elsewhere as well, including in the last presidential elections in Brazil with Confirma 2022, which has been lauded as a success story.⁶⁸ In 2024, an attempt by the Indian government to gain greater fact-

checking authority was struck down in court as unconstitutional and a risk to free expression and press freedom.⁶⁹ Partnerships have been refused in other cases too. For instance, Meedan wanted to implement this model in the 2023 Argentine elections, but Chequeado and the National Electoral Chamber did not agree to do so. The López Obrador government previously launched a fact-checking initiative that did not hold the president accountable and created confusion by adopting the name “Verificado,” demonstrating the risk of official government involvement in fact-checking.⁷⁰ Concern about government involvement in fact-checking in Mexico is therefore warranted and should be approached with caution in the future.

Nevertheless, Meedan’s report highlighted **several suggestions for how to improve the Inés chatbot in the future, which could also guide similar efforts**. Recommendations include proactively producing and sharing explainers; activating earlier in the election season; and providing shorter, more conversational messages. The possibility of shortening messages raises a potential tradeoff in automated fact-checking: balancing detail and user appeal. The report also notes that for Inés, in addition to verifications, explainers about the electoral process (e.g., how to vote, how to register, where to find out where to vote) could have been produced to respond to basic and repeated queries. VerifiChat and VerificadoMX’s, which are managed by people instead of being automated like Inés, did just that.⁷¹

D. AI used to support fact-checking: Increased speed without losing quality

Timely intervention matters to successfully countering misinformation.⁷² WhatsApp-based chatbots discussed in the previous section used AI to produce responses more quickly than in previous elections. This work in 2024 was the continuation of a longstanding trend of innovation among fact-checking initiatives. One of Inés’s key strengths was its ability to manage a high volume of inquiries simultaneously, particularly during peak times such as on Election Day. A user question would query the API, which allowed the bot to respond in real time, drawing from the project’s database and providing immediate answers to users’ questions.

Inés was not the first fact-checking initiative to use AI to accelerate fact-checking; Chequeado was a global pioneer in doing so.⁷³ Its work during 2024 stemmed from development of Chequeabot in 2016, which initially served to separate facts from opinions in speeches and prevent fact-checkers in Argentina from spending long hours listening to radio, TV or video interviews to identify checkable phrases.⁷⁴ Later, new features enabled Chequeabot to alert editors when a false claim was repeated, to convert audios and videos to text in real time, and to monitor disinformation in Spanish more efficiently.⁷⁵

As of 2024, Chequeabot had several more functionalities and is used by fact-checking organizations in seven countries in the region, including VerificadoMX.⁷⁶ Mendoza said that, for a small team like the one she leads at Verificado, this technology is a great help in gaining speed and saving resources. “We use all of them: El Desgrabador, Qué se checka and El Periscopio,” Mendoza added in an exchange on WhatsApp in January 2025.⁷⁷

For its part, Chequeabot’s ability to monitor media and identify content to fact-check made it possible to **reduce the hours of repetitive work** by members of the VerificadoMX team and other organizations in Latin America. It also contributed to faster content distribution. This nearly decade-long experience shows how organizations can experiment with AI despite limited resources. Organizations engaging in this type of work should identify problems and bottlenecks, such as time spent on duplicative fact-checks or transcribing live events, and work collaboratively with developers and partners to find the best possible solutions for a given budget.

But incorporating AI into fact-checking also presents challenges. One major obstacle in using AI and other tech-based tools to support fact-checking is the lack of technical development in languages other than English, as discussed in the section on AI audio. When Chequeado worked on an automation project in collaboration with Full Fact, for example, the Argentine development team noticed that their English colleagues could advance faster without having to re-train models or adjust functionalities.

Implementing AI tools to alleviate certain bottlenecks can leave humans more time to focus on the tasks for which they are still better, such as understanding context, humor, slang, and satire.

Another challenge is the ephemeral nature of third party tools or functionalities. For instance, El Periscopio, which seeks to optimize disinformation monitoring in Spanish on social networks in a single dashboard, suffers every time a social network makes a change to its API or access permissions.⁷⁸ This makes it necessary

to continuously invest resources in adjusting the tool so it can continue to function. An extreme example of how **access to data** affects the work of organizations like El Periscopio was Meta's shutdown of CrowdTangle, a tool that helped journalists and civil society actors identify and respond to mis- and disinformation.⁷⁹

An additional consideration is the circumstances in which people trust AI tools to provide information. This is an area for further research. For instance, *Radio Fórmula* introduced NAT in 2023, an AI-generated news presenter,⁸⁰ but as of 2024 only 37% of Mexicans reported feeling comfortable with news produced mostly by human journalists with the assistance of AI, and only 26% said they were comfortable with news produced mostly by AI.⁸¹ Trust in AI-based sources may change over time as people become accustomed to the technology, and organizations should consider how those views may impact their perceived reliability.

Despite these challenges, AI presents important opportunities for fact-checking organizations — both large and, especially, small — that need to produce content faster, more cheaply, in innovative formats, and for diverse audiences. Experience in Mexico shows that collaboration between organizations with capacity constraints and developers is vital for these tools to be useful and adopted and used to gain efficiency. Implementing AI tools to alleviate certain bottlenecks can leave humans more time to focus on the tasks for which they are still better, such as understanding context, humor, slang, and satire.



03

Case Study: Foreign Interference and Decentralized Resilience in Taiwan

By Kian Vesteinsson

Senior research analyst for technology and democracy at Freedom House.

Kian co-authored Freedom on the Net 2024: The Struggle for Trust Online, along with several assessments about internet freedom in Taiwan. This case study owes a great deal to the researchers, fact-checkers, and practitioners who joined Freedom House at a May 2024 workshop and participated in interviews throughout 2023 and 2024.

I. Introduction

Taiwan's online environment faces a deluge of disinformation campaigns, many of them credibly linked to the People's Republic of China and the ruling Chinese Communist Party. These influence operations deploy cutting-edge and old-fashioned techniques to reduce trust in institutions and sway Taiwanese people on sensitive issues, intensifying around elections, including ahead of Taiwan's consequential general election in January 2024. Campaigning in the 2024 election coincided with the boom of generative artificial intelligence services; Taiwanese voters encountered false and misleading information about the candidates created by AI tools — an early harbinger of how generative AI is reshaping influence operations. Meanwhile, false and misleading content flourished among Taiwanese junk news sites, bulletin board sites, and content creators. The persistence of influence operations and false and misleading information about politics is corrosive for the Taiwanese information environment, with considerable societal consequences. In the words of Eve Chiu, editor-in-chief of Taiwan Fact Check Center, "the biggest damage of disinformation is that people don't trust: they don't trust institutions and they don't trust democracy."⁸²

In the face of the Chinese government's efforts to develop and deploy influence over Taiwan, stakeholders across the country have developed innovative methods of resilience. Taiwanese civil society organizations are highly networked and collaborative, in part a legacy of the 2014 Sunrise Movement, when protestors leveraged civil disobedience and online activism to postpone passage of a trade bill that favored China. They are also closely linked with an active community of public interest technologists, loosely organized in part under the umbrella of the civic tech community called g0v ("gov-zero"). Taiwan's government engages robustly with civil society organizations, particularly as the ruling Democratic Progressive Party (DPP) has sought to bolster Taiwan's international reputation as a model for digital democracy.

These conditions have produced an engaged, innovative, and diverse network of groups working to safeguard reliable online information during elections by countering influence operations and building societal resilience to disinformation. Technologists are embedded within the community of civil society organizations, creating room for innovative technical research into influence operations and boutique technical tools to disseminate accurate and reliable information. These efforts have a strong commitment to transparency and collective participation, a reflection of the g0v community, which serve to cultivate trust from the broader Taiwanese community. While Taiwan faces the same problems as many democracies around the world — sharpening political polarization, declining trust in democratic institutions — its community of stakeholders is uniquely positioned to build a more trustworthy online information environment.

During the 2024 elections, Taiwanese civil society leveraged strong, diverse networks and digital innovation to monitor Chinese influence campaigns, bolster access to reliable information, and build broader societal resilience. Key findings include:

- Taiwan's disinformation researchers observed the **increased use of generative AI as a part of more conventional campaigns**. In influence operations attributed to China-based actors, for example, AI-generated video avatars bolstered the reach of rumors smearing Taiwanese government officials.

- While researchers found widespread evidence of false and misleading information campaigns, the **campaigns focused on longer-term issues**, like skepticism of positive relations with the United States, and may have had less impact on electoral participation.
- **Fact-checkers socialized accurate information to diverse groups** across Taiwanese society, whether through transparent crowd-sourcing platforms or tailored programming at the community level.
- Fact-checking **chatbots helped people access accurate information quickly and easily**. Their integration into popular social media and messaging apps facilitates dissemination, and offer approaches that respect end-to-end encryption.
- Policymakers enacted **legislative responses to electoral misinformation with transparency and proportionality as underlying values**. The measures reflect the influence of the civic tech community and stand in stark contrast to previous years, when authorities ordered such content removed.

II. Country context

On January 13, 2024, Taiwanese voters went to the polls to select a new President and all 113 members of the Legislative Yuan, the unicameral legislature. Voters elected Vice President Lai Ching-te of the Democratic Progressive Party (DPP) to assume the presidency, marking the first time in Taiwanese history that a party has held the presidency for three consecutive terms. The opposition Kuomintang (KMT) secured a slim plurality of seats in the Legislative Yuan and has since legislated in coalition with the Taiwan People's Party (TPP), an rising third party.⁸³ The DPP has considered Taiwan a de facto independent nation and advocated for a separate national identity. The KMT has favored closer relations with China; prior to Taiwan's democratic transition, the KMT ruled Taiwan as a one-party authoritarian state for decades.

The election was competitive, particularly because former president Tsai Ing-wen of the DPP could not run again due to term limits. The DPP had controlled the Executive Yuan and Legislative Yuan since 2016 but fared poorly during countrywide local elections in 2022. Lai faced New Taipei mayor Hou Yu-ih of the KMT and former Taipei mayor Ko Wen-je, who represented the four-year-old TPP as its first presidential candidate; Terry Gou of Hon Hai Precision Industry (Foxconn) ran a brief independent campaign. Lai led in most election-cycle polling, though polls indicated that voters would support a proposed KMT-TPP presidential coalition. The coalition ultimately failed to materialize over a dispute between Hou and Ko on who would lead the ticket.⁸⁴ Dissatisfaction over the Taiwanese economy dominated domestic policy debates during the election. Foreign policy issues, especially regarding relations with China and the United States, were also significant.

Influence operations were prominent, many of them linked to actors based in China.⁸⁵ Prominent narratives exploited key election-related issues. For example, some disinformation campaigns sought to spread rumors about scandals involving Tsai, Lai, or other DPP officials to undermine public perception of their fitness for office.⁸⁶ Others tried to spread or amplify distrust in the Taiwanese government's public service provision, including on contentious economic issues. For example, one prominent China-linked influence operation disseminated false stories that a government-sponsored effort to import eggs to alleviate a shortage resulted in the entry of spoiled, carcinogenic, or poisoned eggs. A third category of narratives sought to amplify distrust in US-Taiwan relations, for which Taiwanese researchers coined the phrase "US skepticism,"⁸⁷ such as with false stories about contaminated pork imports from the United States.⁸⁸ In some cases, the disinformation campaigns deployed AI-generated avatars to amplify false claims or featured AI-generated content.⁸⁹

Taiwan's civil society organizations have been instrumental in uncovering and attributing influence operations and building societal resilience to disinformation campaigns. Their success is informed by strong organizational and informal networks, in part a legacy of the 2014 Sunflower Movement, a prodemocracy protest movement that saw mobilization across Taiwanese

Formal and informal connections create opportunities for exchange, helping Taiwan’s relatively small community of civil society organizations collaborate without duplicating work.

society. Networking is incentivized by registration requirements: Taiwanese law requires that new nongovernmental organization secure 30 members in order to incorporate;⁹⁰ civil society groups often engage their peers and collaborators to meet this requirement. These formal and informal connections create opportunities for exchange, helping Taiwan’s relatively small community of civil society organizations collaborate without duplicating work.

The g0v community, a civic tech movement that participated actively in the Sunflower Movement, has served as an incubator and coordination space for activists and civil society writ large. The g0v manifesto emphasizes open collaboration on issues of importance to the public and in support of speech and information transparency, fueled by open-source and decentralized contributions — one famous g0v motto runs “don’t ask why nobody did this, admit you are the nobody first.”⁹¹

III. Assessment of three interventions

This case study considers three distinct interventions: civil society research into influence operations, fact-checking or information literacy programming conducted by independent organizations, and policymaking tailored to countering misinformation without infringing on free expression. Particular attention is paid to how interventions have shifted over time, accounting for the now-widespread availability of generative AI tools.

This analysis considers independent disinformation research and fact-checking practices as separate but related interventions. Disinformation researchers set out to identify, monitor, and catalogue influence operations in order to set the stage for action; fact-checkers identify and correct specific false and misleading claims, and deliver corrections to the broader public. Taiwan’s disinformation researchers and fact-checkers work closely together, and some organizations produce both research and fact-checking.

For example, the Taiwan Information Environment Research Center (IORG, formerly the Information Operations Research Group) assesses whether specific claims are credible or manipulated, and produces rigorous research and analysis relating to influence operations more broadly.

Though the report is primarily focused on civic space actors, this analysis includes policymaking because of the unique relationship between the civic tech community and Taiwanese government, driven in large part by the g0v movement.

A. Independent research on influence operations

Taiwanese civil society organizations, academics, and journalists have invested heavily in the effort to expose and catalogue influence operations on the Taiwanese internet, including around electoral periods. Research products include investigations analyzing influence operations, short-form updates that build on previous investigations, and meta-analysis that produces insights on trends. Such research informs targeted interventions from other stakeholders, including Taiwanese civil society, policymakers, regional and global social media platforms, and the international community.

Civil society groups produce cutting-edge research, several focusing on election-related research specifically. Doublethink Lab (DTL) analyzes influence operations and false and misleading content on the Taiwanese internet, including during elections.⁹² The IORG covers the Mandarin-language information environment, including efforts to spread false and misleading content during elections. DTL and IORG, both founded in 2019, have produced research on the role of false and misleading content in several recent elections, including the 2020 general election, 2022 local elections, and the 2024 general election.⁹³ Taiwanese government and academic researchers also produce research on false and misleading content and influence operations.

Freedom House research has found that **Taiwan faces an intense degree of Chinese influence efforts**, offline and online.⁹⁴ In studying China-linked influence operations over the course of several electoral periods, researchers have built a nuanced understanding of how such efforts shape Taiwan's information environment.

Taiwanese researchers have documented the innovation of new trends and tactics by Chinese actors, shaping the collective understanding of how online influence campaigns work in practice. For example, they have found that influence operations targeting Taiwanese voters are transnational, featuring considerable cross-strait exchange — including Taiwanese commentators who engage with China-backed state media accounts⁹⁵ — and engaging participants across East and Southeast Asia.⁹⁶ They involve a diverse set of actors and platforms, including junk news sites that churn out content to artificially inflate advertising revenue and content creators who boost CCP talking points, wittingly or not.⁹⁷ Jason Liu, senior policy analyst at Access Now, observed that influence operations increasingly exploit the profit motive of the internet: “You click and earn money. It doesn't really matter if [content creators] believe in the fuff or a political party — it's more important to understand how they can make money.”⁹⁸ Researchers have also identified variation between influence operations targeting English-speaking and Mandarin-speaking audiences, indicating that perpetrators of influence operations may deploy different tactics dependent on context.⁹⁹

DTL found that **AI-generated content featured in influence operations during the 2024 election**, usually as one component of a wider campaign featuring more conventional tactics. For example, AI-generated virtual news anchors featured prominently in a campaign to smear former president Tsai; the campaign revolved around an ebook titled “Secret History of Tsai Ing-wen” that was amplified by a network of inauthentic accounts on a range of platforms. In DTL's post-election summary of influence operations, the researchers noted that “while text and meme-based content have traditionally been prevalent in information manipulation, the increase in video content as a primary technique has signaled a shift in strategy.”¹⁰⁰

The efficacy of research as an intervention depends on stakeholders in the private or public sector putting it to use. DTL and IORG pair their research efforts with advocacy and recommendations for public and private sector stakeholders. The close networks of Taiwan's civil society community may also offer a path to distribute and disseminate research findings further. Notably, the networks include policymakers: DTL co-founder Puma Shen was elected as a DPP legislator-at-large in the 2024 election.¹⁰¹

Some elements of the success of these groups may be unique to Taiwan. Some 88 to 90 percent of Taiwan's 23 million people use the internet,¹⁰² a vast majority of whom write with the same script,¹⁰³ allowing for more nuanced research. The deep collaboration with technical experts, often fueled by the g0v community, is also noteworthy, as discussed throughout this study.

Taiwan's researchers have sought to shape the field of disinformation research beyond the island, deploying their unique expertise to strengthen interventions in other contexts. DTL has built on its research to establish theoretical and practical contributions to the study of influence operations and developed collaborative relationships in other countries, including through exchange with Ukrainian civil society.¹⁰⁴ Taiwanese organizations have worked with collaborators across South and Southeast Asia and the Pacific to cultivate research and resilience to disinformation in those contexts.¹⁰⁵ These efforts have been facilitated in part by greater collaboration between Taiwanese and international civil society organizations, aided by an effort by the Tsai administration to relax registration requirements for international nongovernmental organizations.¹⁰⁶

B. Collaborative and community-minded fact-checking

Fact-checking online content is now an established field. Taiwanese fact-checkers have developed innovative methods of conducting fact-checks and delivering fact-checked content to a wide range of people.

Organizations like Taiwan Fact Check Center and MyGoPen conduct fact-checking of claims on the Taiwanese internet by engaging journalists and researchers to examine rumors and falsehoods.¹⁰⁷ For example, Taiwan Fact Check Center sought to assess a pre-election claim that Lai had agreed to give hundreds of millions of dollars to Paraguay for public housing during his August 2023 visit to the country.¹⁰⁸ Fact-checkers with the center found that the rumor appeared to have originated in a manipulated photo of a Spanish-language memo posted on PTT (Taiwan's largest bulletin board system, a type of open forum comparable to Reddit), and spread widely as frustration with Taiwan's public infrastructure became an issue in the general election campaign.¹⁰⁹

The Cofacts platform offers a unique model for fact-checking. People who visit the Cofacts platform or engage with chatbots on messaging apps like LINE — Taiwan's most popular social media and messaging platform¹¹⁰ — can request that an article or claim be fact-checked. Cofacts introduced its LINE-compatible chatbot in 2018, making it a leader in using emergent technology to develop information resilience.¹¹¹ Cofacts fact-checkers, who may be professional or non-professional contributors, can submit responses. The platforms' users can read through all submitted fact-checks, endorse those they find to be accurate, or submit their own responses. The platform is open-source and Cofacts releases its analytics data to the public.

Several organizations lead **programs at the community level that build information literacy** among less online or highly connected populations, including familiarity with fact-checking. Civil society group Fake News Cleaner conducts trainings at the community level, including programming for seniors, that equip participants with information literacy concepts.¹¹² IORG has created resources for middle and high school teachers to engage with students on information literacy and developed programs with community organizations across the country.¹¹³

Reflecting on Taiwan's innovative approaches to fact-checking, Chihhao Yu, co-director of IORG, noted that "openness and transparency are still our greatest strength."¹¹⁴ These values

“Openness and transparency are still our greatest strength.”

– Chihhao Yu, co-director of the Taiwan Information Environment Research Center

undergird Taiwan’s innovative approaches to fact-checking, whether through transparency as a design choice on the Cofacts platform or in the community-driven approaches of the information literacy programs.

Taiwan’s fact-checking groups developed their efforts with the aim of being complementary and not duplicative. Scholars Chiaoning Su and Wei-Ping Li attribute this benefit as a product of the relative recency of the field in Taiwan. Su and Li noted that the different approaches of Taiwanese fact-checking groups — including diverging perspectives on what kinds of content should be covered and how fact-checks are supplied — create a “collaborative safety net” that serve a variety of audiences.¹¹⁵

Taiwan’s fact-checkers have deployed cutting-edge tools to support their work, including AI and machine learning, another indicator of the high degree of technical capacity brought by g0v contributors. For many organizations, these technologies increase the pace and reach of fact-checking. Taiwan Fact Check Center has experimented with natural language processing tools to cluster requests for fact-checks and filter them against previous corrections.¹¹⁶ Cofacts has deployed machine learning to manage the high volume of requests for fact-checks and organize them to facilitate corrections from editors.¹¹⁷ MyGoPen also uses bots to respond to the thousands of fact-checking requests they receive. The organization’s founder, Charles Yeh, told The Guardian last year that AI is helpful beyond its role enabling a user interface: “[AI] speeds up the checking process — helps with comparison, identification and translation — and we use it for some situations.”¹¹⁸ A Carnegie Endowment meta-analysis of studies about countering disinformation found that such technical innovations are essential in overcoming one fundamental challenge for fact-checkers: it is much less time-consuming to create false or misleading information than it is to debunk it.¹¹⁹

In-app integration minimizes the friction of seeking out fact-checks and enables sharing among social groups.

Fact-checkers around the world face another **fundamental challenge: reaching wider audiences** beyond those who already seek out fact-checks. Users can integrate chatbots operated by fact-checking groups and independent developers into LINE. The developer of a chatbot called Auntie Meiyu suggests that people add the bot, which pulls from MyGoPen data and other sources, to group chats to supply friendly fact-checks and fraud detection.¹²⁰

These types of in-app integration minimize the friction of seeking out fact-checks and enables sharing among social groups, widening the audience for fact-checks, though automated review of messages raises privacy concerns. LINE has also experimented with chatbots that offer fact-checking capability without changing the platform's encryption protocols.¹²¹ Instead, a user can share a message they want to fact-check by forwarding it or copying the text of a message and sending it to the chatbot. Although the individual message that a person chooses to fact-check is no longer private, the original conversation remains protected by end-to-end encryption.

While traditional and tech-enabled fact-checking can debunk and provide in-the-moment information, the information literacy programming conducted by Taiwanese fact-checking groups serves to engage people in more holistic and long-term efforts. According to IORG codirector Yu, "sustained engagement over time" is essential for the success of programming that aims to cultivate community-based information literacy programs.¹²² The Carnegie Endowment meta-analysis found that the most successful media literacy programs cultivate "confidence and a sense of responsibility alongside skills development."

Taiwanese fact-checkers report that their work is increasingly politicized, similar to those around the world whose work came under attack during 2024 elections.¹²³ A Taiwanese fact-checker shared that their organization debunked a disproportionate number of election-related rumors about the DPP ahead of the 2024 vote, simply because rumors about the DPP were more common than those about other parties. This imbalance in output prompted accusations of bias, many of which the fact-checker suspected were made in bad faith in order to delegitimize the organization's

Harassment remains a threat that limits the ability of fact-checkers to do their work safely.

work.¹²⁴ Other fact-checkers reported facing online harassment — in some cases originating from anonymous trolls, in others by content creators who were the subject of fact checks.¹²⁵ While transparent and participatory models of fact-checking like Cofacts may mitigate allegations of bias to a certain extent, harassment remains a threat that limits the ability of fact-checkers to do their work safely.

Funding also remains a critical challenge for Taiwanese fact-checking groups. Fact-checking organizations in Taiwan struggle with many of the challenges that their global peers face: how to remain solvent on grant funding, whether to take funding from sources that may carry reputational risk, and so on. The g0v community provides room for software developers interested in public participation to incubate projects, offsetting some costs that fact-checking organizations in other countries might face when hiring programmers to create new tools. In return, anthropologist Aaron Su notes that g0v projects offer software engineers, often highly paid, a unique opportunity for “targeting [their labor] toward the direct and immediate betterment of the larger community at hand.”¹²⁶ However, sustainability remains a persistent issue for Taiwanese groups.

C. Rights-respecting policymaking alongside civic tech innovation

Ahead of the 2024 election, Taiwanese policymakers passed laws that took a more tailored and rights-respecting approach to false and misleading election-related information. These mark an improvement compared with previous efforts that introduced criminal penalties for false or misleading online content, or sought to create new avenues for authorities to order the removal of such content from online content hosts.

In advance of the elections, policymakers deployed interventions **focused on transparency and more limited, proportionate provisions** for the removal of election-related falsehoods. The Legislative Yuan passed amendments to the Public Officials Election and Recall Act in 2023 that require political advertisements to disclose their funder.¹²⁷ The 2023 amendments, which cover both

online and offline political ads, also prohibit political advertisements funded by foreigners, particularly those from mainland China, Hong Kong, or Macau, and require that internet service providers or other companies verify the nationality of the advertiser.¹²⁸ The 2023 amendments also establish new safeguards against AI-generated misinformation. Under the new rules, political candidates may report nonconsensual and misleading content of themselves created by generative AI services to the police. Subject to confirmation from law enforcement technologists that the content is AI-generated, the candidate may submit a request to platforms to remove the content within two days.¹²⁹

Previous Taiwanese lawmaking took a much more restrictive approach to false and misleading content about sensitive issues. The Social Order Maintenance Act (SOMA) bars the spread of false information online, and authorities have invoked it to penalize political speech, while the 2019 Anti-Infiltration Act criminalizes the spread of election-related disinformation sponsored by “foreign hostile forces.” During the Covid-19 pandemic, the DPP government passed new rules to criminalize false statements about the epidemic; many people faced fines or short-term detentions for false or misleading comments about Covid-19 cases or vaccines.¹³⁰

The Taiwanese government’s approach in 2024 centered transparency and proportionality as underlying values while enacting a much-needed response to electoral misinformation. It stands in contrast to that of other democracies around the world — Freedom House research has found that interventions related to election information in many countries infringed on human rights.¹³¹ The more narrow and rights-respecting approach also contrasts with the Taiwanese government’s actions in previous elections and in response to other forms of harmful online content (for example, a controversy over a nontransparent website blocking mechanism deployed against fraud and scam websites).¹³²

The unique approach to false and misleading election-related content in 2024 reflects a range of factors. The civic tech community has a long history of collaboration with the DPP government, and prominent members have advanced **an agenda of transparency and openness**, including around the question of how government should respond to mis- and disinformation. At the

same time, the DDP government faced several recent controversies and criticism around previous efforts to create rules for the online environment, including election-related misinformation. Finally, civil society organizations deployed the interventions described above, serving as a societal response beyond government action.

The civic tech community has a long history of collaboration with the DPP government. g0v members have joined the ranks of government, most notably Audrey Tang, who served as Taiwan's first minister of digital affairs and is now cyber ambassador-at-large.¹³³ In the aftermath of the Sunflower Movement, Taiwanese policymakers, Tang among them, began to embed the g0v values of openness and collaboration into its daily operations, including in the effort to counter influence operations. For example, Tang told TIME in 2017 that “we take freedom of speech much more seriously than most of the other Asian countries,” informing her commitment to media literacy education, fact-checking, and transparency as a counter-balance to influence operations.¹³⁴ During Tang's tenure at the Ministry of Digital Affairs, government agencies experimented in how to effectively counter false information in a timely way without resorting to censorship.¹³⁵ In one such initiative, the Taiwanese government adopted the “222 principle” in 2019, which set out a requirement for government agencies to issue social media-friendly corrections (with a title no longer than 20 words, no more than 200 words of explanation, and with two images) within one hour.¹³⁶

Available evidence suggests that the earlier era of anti-misinformation policies did not considerably shape the online environment ahead of 2024 election. SOMA convictions for election-related speech were sparse during the 2024 election. Indeed, Taiwanese courts acquitted at least two people who were arrested under SOMA charges for false claims about election fraud made on social media.¹³⁷ The Anti-Infiltration Act, meanwhile, was invoked sparingly and only in relation to campaign-related corruption.¹³⁸

One factor in this change may have been the **desire to avoid allegations that the DPP government deployed restrictions on misinformation to its own benefit** during campaigning. Such criticism emerged when the DPP took harsh measures in response to false and misleading information in other contexts, including around the 2020 election.¹³⁹ In one prominent controversy, the

KMT criticized the DPP government for de-licensing CTi News, online news outlet and satellite channel, over concerns of Chinese influence.¹⁴⁰

Another factor may be the **backlash against the DPP government's effort to pass technology regulation**. In 2022, the DPP government unveiled a package of technology regulation known as the Digital Intermediary Services Act (DISA). Modeled in part on the European Union's Digital Services Act, DISA would have obligated covered platforms to establish significant transparency and notice-and-appeal mechanisms. It would also have required platforms to remove content under court orders or label it subject to government orders. These provisions prompted concern about censorship from free expression advocates and harsh criticism from the opposition KMT; some civil society organization supported the attempt at platform regulation while criticizing the legislation itself. The proposed DISA lapsed, and was not revived.¹⁴¹ The DPP government has not proposed tech regulation in the aftermath, instead focusing on passing legislation in 2024 aimed at online fraud and scams.

Taiwan's shift towards more narrow, rights-respecting policymaking around false and misleading election-related content in 2023 could serve as a model for lawmakers around the world. So should the past decade of democratic deliberation about which responses to misinformation work and which do not, buttressed by civil society organizations and civic tech activists. The g0v community and the global civic tech movement have reflected extensively on what elements of g0v might be replicated in other contexts.¹⁴² In any circumstance, governments seeking to lay the groundwork for a similar approach should look for opportunities for long-term collaboration with civic tech activists and ensure that civil society organizations have the funding they need to operate.



04

Case Study: AI Incident Tracking by Global Civic Space Actors

By Isabel Linzer

Elections and Democracy Fellow at the Center for Democracy & Technology.

Isabel focuses on research and policy pertaining to elections and technology, including the role of generative AI and mis- and disinformation. She has worked on issues at the intersection of human rights, democracy, and technology throughout her career, including with Freedom House, the UN Office of the High Commissioner for Human Rights, and the National Democratic Institute.

I. Introduction

Throughout 2024, numerous civic space actors tracked the use of generative AI in politics and elections, both within the US and globally. These civil society organizations, journalists, and academics all sought to answer some version of a question that has dominated the past year: what consequences does widespread access to generative AI tools have for elections? By tracking incidents, civic space actors tried to answer this question with evidence.

Amid a fragmented ecosystem and disjointed policymaking by governments and technology companies, civic space actors compiled thousands of data points in an organized, transparent, and centralized way. Their work facilitates access to information, enabling informed policy advocacy and accountability for governments and technology companies. Reflecting this point, Kaylyn Jackson Schiff, an assistant professor at Purdue University and co-creator of the Political Deepfakes Incident Database (PDID), said, "We're really trying to think about these broader impacts on the information environment. Impacts on peoples' trust in social media platforms, trust in traditional media, trust in political institutions."¹⁴³ In addition to providing in-the-moment documentation, incident tracking provides a contemporaneous record that can be used for

future research and limit post hoc attempts to manipulate or rewrite facts. Persistent knowledge creation efforts like the PDID are more important than ever, as companies limit researcher access to data¹⁴⁴ and governments and activists politicize disinformation research.¹⁴⁵

This case study reviews eight efforts to track political- and election-related AI incidents and assesses how they contribute to public knowledge and information resilience: The AI and Elections Tracker (Abundance Institute), the AI Elections Project (WIRED), the AI Incidents Database, the AI Political Archive (New York University, University of North Carolina, and American Association of Political Consultants Foundation), the Political Deepfakes Incident Database (Purdue University), Spitting Images: Tracking Deepfakes and Generative AI in Elections (German Marshall Fund), the 2024 AI Elections Tracker (Rest of World), and the 2024 Foreign Interference Attribution Tracker (DFRLab). This is not an exhaustive list of tracking work, but includes a set of prominent projects that vary in scope, providing opportunities to compare the work across sectors, methodologies, and geographies. Several of the projects were continuations of work that preceded the “year of elections,” but all gained greater relevance and urgency in light of the focus on AI and elections in 2024.

CDT conducted desk research and spoke with the researchers behind seven of the eight tracking efforts to gain a deeper understanding of their goals, process, and the evolution of their work.² The teams also provided insight into the challenges they encountered. The conversations offered lessons for future work, as well as a snapshot of expert thinking on the consequences of AI in elections as of December 2024. Key findings include:

- The choice between **prioritizing breadth of incidents** (such as geography, topic, or relevant sector) **or the depth** (quantity of data) is a key divergence in incident tracking methodologies. This resourcing decision not only has implications for data collection

2 CDT interviewed experts affiliated with each of the eight listed projects except for the WIRED AI Elections Project. For all projects, CDT’s desk research included reviewing publicly available materials about the projects, the data they published, and related publications, including articles from academic, journalistic, and civil society sources.

methods and ability to draw qualitative or quantitative findings, but for policy. Breadth can help guide policy by stress-testing edge cases, while depth in a more specific area can support statistically valid “evidence-based” policymaking. The resource choice is therefore also a prioritization of how to approach policy advocacy.

- **A lack of transparency and independent access to data** inhibits incident tracking. Better access could alleviate capacity constraints and facilitate research that would support more sophisticated policy improvements.
- **AI should be considered in concert with pressing, pre-existing mis- and disinformation issues**, and altered content more broadly. While researchers tended to agree that the risks AI posed to elections last year were overstated, they also cautioned against premature triumphalism. Companies, governments and other stakeholders should continue to take risks seriously.
- **Incident tracking does not capture all harms**, so anyone using tracking research should consider how to account for cumulative effects.

II. Survey of incident tracking

The eight trackers we reviewed varied in scope. Some, like DFRLab’s Foreign Interference Attribution Tracker, were US-specific, while others captured global data. Six dealt with election-specific content and two were broader in scope. Civil society organizations produced four of the trackers we studied, media outlets were behind two, and academics organizations created the remaining pair. They also captured different types of incidents, ranging from an expansive collection of cross-sector AI-related harms (AI Incidents Database), to a catalogue exclusively of deepfakes (German Marshall Fund), to a database of media coverage of AI incidents (Abundance Institute). The variety of definitions, geographic scope, and other methodological approaches among trackers creates, as several researchers we spoke with said, an “ecosystem” of information.

Despite significant differences in the scope of each database, researchers behind the projects were broadly aligned on the goals and intended audience of their work. Specifically, they were unified in the hope that a mix of civil society, journalists, researchers, media, and policymakers would find their work useful. Researchers from NYU, for example, hoped that their work could support evidence-based policymaking on generative AI and political advertising.¹⁴⁶ Similarly, the Abundance Institute has used its tracking research to inform advocacy towards policymakers, including in a comment to the Federal Communications Commission on AI disclosure in political ads.¹⁴⁷

Preparing a non-expert or public audience was also an explicit aim, in line with inoculation theory.¹⁴⁸ One of DFRLab's goals was to "increase public resilience against future foreign influence and interference efforts, especially in online spaces," in addition to providing a public record of incidents, building public attribution standards, and serving as a resource on the topic.¹⁴⁹ The Purdue team's 2024 publication about the PDID lays out the landscape of goals:

The fundamental goal of the database is to provide a publicly available resource to advance research, practice, and governance efforts surrounding deepfakes. Journalists, fact-checking organizations, media literacy educators, and members of the public can utilize the database to evaluate the veracity of specific instances or identify broader trends of concern. In addition, the PDID can assist in understanding the effectiveness of watermarking and detection techniques, media literacy efforts, and other policy initiatives on deepfake dissemination and impact.¹⁵⁰

Though specifically referring to the PDID work, the use cases laid out above align closely with other projects. The Purdue team is also developing a website designed to make public access to their data even easier.¹⁵¹

Several researchers noted that project goals evolved over time, however. The AI Incident Database (AIID) has been public since 2020.¹⁵² An early goal was to raise basic awareness that it was possible to track AI incidents. That point is now well-established —

Diverse research approaches combine to identify edge cases and outstanding questions that a single approach alone would not.

as is clearly demonstrated by the many trackers that exist, including those outside the scope of this report — and the current focus is to educate different stakeholders, including policymakers, at a deeper level.¹⁵³ Others noted a shift in goals and stakeholders during the

election period and going forward; while journalists and fact-checkers were important users during the pre-election period, policy advocacy and research are a greater focus in the medium and long term.

So far, engagement with these different stakeholders has largely been successful. Teams reported outreach from media and policy groups, engagement with academia and government offices, and examples of their work in education, including in syllabi. Some

directly shared their findings with social media platforms and other technology companies.

Meanwhile, teams took **a mix of approaches to collaboration.**

The researchers we spoke with largely felt that they had found a niche among the array of incident tracking efforts, though some acknowledged that redundancies existed and were inefficient. Several sought out collaboration and were in touch with each other or with researchers working on other tracking projects not covered by this report. For example, aiming to “improve interoperability and enhance rigor towards greater impact, [the PDID team] consulted existing and forthcoming coding taxonomies developed by the AIID, Center for Security and Emerging Technology, and AI Vulnerabilities Database.”¹⁵⁴ Rest of World’s methodology says the organization is “keen to work with other organizations and researchers who are monitoring tech’s influence on elections.”¹⁵⁵

At the most ambitious end of the collaborative model was the AIID. The AIID aims to create a collective or federation model that “empower[s] AI incident research in different domains,” while also avoiding “safety data and safety insight being totally fragmented, which is both an organizational, cooperation, and technical problem.”¹⁵⁶ One way they work toward this goal is by integrating existing taxonomies, rather than proposing their own. The database currently uses the Center for Security and Emerging Technology AI Harm Taxonomy and the Goals, Methods, and Failures taxonomy.¹⁵⁷

A mix of definitions and project scopes was a benefit in many ways: diverse research approaches combine to identify edge cases and outstanding questions that a single approach alone would not. Some researchers we spoke with, however, expressed frustration that the **language around generative AI is imprecise and imbued with value judgements**. The term “deepfake,” for example, may even stigmatize AI-generated content in a way that similar content created through other means, including “cheapfakes,” would not be subject to. This distinction recalls the “behavior”-based approach that companies often take when combatting disinformation, where treatment is based on technical attributes, rather than content, impact, or harm.¹⁵⁸ Imprecise definitions may inadvertently reproduce flaws seen in relevant policy areas in the past, including content moderation, and add confusion to current debates between researchers and policymakers.

III. Challenges, tradeoffs, and answers about the impact of AI

This section considers the tradeoffs and challenges in collecting data, developing sufficient capacity and expertise, and responding to the predominant narrative that generative AI, especially deepfakes, would cause chaos in elections. Each offers lessons for how future research can be improved, considerations for policymakers and advocates who use this research, and guidance for how funders can support civic space actors’ work. The final part of this section concerns expert opinions on the animating question of the ways or extent to which AI had an impact on elections in 2024. Rather than offering a clear answer, this analysis points to more unresolved questions and knowledge gaps that researchers and policymakers working on AI in elections should consider going forward.

Description of AI Incident Trackers

Civil society

- The **AI and Elections Tracker**, created by the Abundance Institute, tracks news articles about AI in the US election. A primary goal of the research was to keep a contemporaneous record, so as to assess findings against popular narratives about the importance of AI and deepfakes in the election.
- The **AI Incident Database (AIID)** is a collaborative effort by several organizations, including the Digital Safety Research Institute, Georgetown's Center for Security and Emerging Technology, and the Partnership on AI. The AIID has the broadest scope of any tracker we reviewed, indexing all harmful uses of artificial intelligence.¹⁵⁹ While not elections-specific, the database is searchable and, moreover, the researchers we spoke with described the AIID as a starting place for more specialized work, ranging from AI in healthcare to elections.¹⁶⁰
- **Spitting Images: Tracking Deepfakes and Generative AI in Elections** is a German Marshall Fund (GMF) project that is global in scope and attempts to understand how AI is being used, in both deceptive and benign ways. The tracker categorizes the type of media and records the timeline, description, and media coverage of the incident. The project aims to help voters, policymakers, and researchers understand the impact of AI and allow comparison across countries.¹⁶¹
- DFRLab's **2024 Foreign Interference Tracker** is specific to the 2024 US election. Because the focus is on foreign interference, it only captures the use of AI to the extent that it is known to be used in interference efforts. Generative AI was listed as a method used in 39 of the 90 incidents included in the dataset at the time of writing. The tracker scores each incident's overall reach and impact as well as how credible the attribution is.

Academia

- **The AI Political Archive**, launched in July 2024 as partnership between New York University, the American Association of Political Consultants Foundation, and the University of North Carolina,¹⁶² focuses on the use of AI in political communications. Its definition of political communications includes political ads and social media posts, and the Archive aims to capture data about AI usage in national and down-ballot races.
- Scholars at Purdue University developed the **Political Deepfakes Incidents Database (PDID)**. The project's scope extends beyond elections and instead focuses on political incidents, including content shared by or about a politician, or about a politically salient event.¹⁶³ The database, which included over 700 records at the time of writing, includes not just AI-generated materials but also cheapfakes or other manipulated images. It also includes material that is real, but that users claimed was fake.

Media

- **The AI Elections Project** is a collection of global examples of AI's use in elections, published by WIRED. Each documented example includes an image, brief description, date, and linked source. Though the project is described as "tracking every instance of AI's use in and around" the 60+ elections in 2024, as of the time of writing the project included a smaller selection that nevertheless offered broad geographic scope and a range of use cases.
- Rest of World's **2024 AI Elections Tracker** is also global in scope, but designed to offer a selection of examples rather than a comprehensive database. Its stated goal is to "create a database of examples that can be used to understand the many ways in which AI is being deployed around elections." It worked with reporters in countries that held elections to identify any election-related incidents, regardless of creator, intent, or type of media.

A. Data collection

Technology companies already possess, in great detail, the kinds of data that researchers devote huge amounts of resources to imperfectly collect.

The eight projects used different definitions, variables, and methods, but researchers encountered four common problems during data collection. First was **access to data from companies**. As CDT has previously documented, there is extreme information asymmetry

between technology companies and civic space actors.¹⁶⁴ The problem has been made worse in recent years by platforms rolling back data access options. In 2023, Twitter (now X) announced that it would no longer offer free access to its API.¹⁶⁵ In 2024, Meta shut down CrowdTangle, which previously facilitated social media monitoring that researchers used to study the online information environment, including mis- and disinformation. The decision, taken during such a crucial election year, drew widespread criticism

from a range of academics and civil society organizations, including CDT.¹⁶⁶ AI companies, though increasingly relevant and influential, are often opaque. They lack strong transparency practices and tend to be reluctant to share data with researchers due to a range of privacy, reputation, and market-related concerns.¹⁶⁷

“Loss of access has not been a small thing,” one DFRLab researcher said of the CrowdTangle shutdown. Indeed, DFRLab’s 2024 Foreign Interference Attribution Tracker (FIAT) was based on the 2020 version of the project. The 2024 version dropped the “Attribution Impact” score, which aggregated engagement data from Facebook, Reddit, and Twitter, from the 2020 edition. “Come 2024, almost none of that infrastructure was possible to spin up again,” DFRLab told CDT.¹⁶⁸ The 2024 project announcement describes the change in more detail: “Due to the shutdown of Meta’s CrowdTangle tool and increased restrictions on APIs, this measure could not be replicated for the new dataset.”¹⁶⁹

DFRLab’s work is not the only project affected. The AIID team found that the impact of tracking work is limited by the lack of information from companies around AI-related incidents.¹⁷⁰ NYU researchers also expressed frustration with data access. “I wish no one needed to create their own database at all,” said Zeve Sanderson, Executive Director of the NYU Center for Social Media & Politics. “It’s an extremely bad use of anybody’s time, mostly because these data

are there,” referring to the fact that technology companies already possess, in great detail, the kinds of data that researchers devote huge amounts of resources to imperfectly collect.¹⁷¹ The NYU team also noted that when platforms did disclose data, it was not always machine-readable. That meant manually scrolling through an ad library, for example.¹⁷²

CDT found similar problems with ad libraries in research published in 2024.¹⁷³ Meta hosts a web-based repository with keyword search, while Google/YouTube’s web repository does not feature a keyword search. Snapchat and X provide downloadable CSV files without keyword search. The Meta and Google repositories provide a visual of the ad creative, while Snap and X only include a link. Similarly, a 2024 study by Mozilla concluded that among ad repositories offered by 11 large tech companies, “none is a fully-functional ad repository and none will provide researchers and civil society groups with the tools and data they need to effectively monitor the impact of VLOs’ [Very Large Online Platforms and Search Engines] advertisement on Europe’s upcoming elections.”¹⁷⁴ As Mozilla and others have observed, when there is transparency, the type of information shared is inconsistent between companies and therefore is still of limited use for aggregating and drawing comparisons.

The second data collection challenge was the need to **positively identify uses of generative AI**. Detecting AI-generated content is notoriously difficult. Publicly available AI image detectors are unreliable and sometimes reach conflicting conclusions.¹⁷⁵ Detection of other mediums, including audio and text, can be even more difficult, as these types of content offer researchers less information compared with images or videos.¹⁷⁶

The Purdue researchers acknowledged a certain degree of uncertainty in their coding due to the limitations posed by detection technology, and noted it could be a greater problem in the future as the quality of generated content improves. Those same limitations also made it difficult to correctly identify AI-generated content as opposed to cheapfakes or photoshopped images.¹⁷⁷ Rather than verifying authenticity themselves, researchers at the German Marshall Fund (GMF) primarily relied on media coverage, academic articles, fact-checking organizations, or other experts to confirm

that content was a deepfake and identify what type of media was involved in the incident (i.e., image, video, etc.).¹⁷⁸

The third common challenge in data collection was soliciting **public incident submission**. The Rest of World, Abundance Institute, WIRED, AIID, and NYU projects had public submission options. The PDID team reported that it is considering implementing a similar option in the future.¹⁷⁹ Based on the interviews conducted for this report, public submission was minimal, with some projects receiving no public input. While this outcome was disappointing, creating a submission channel is not resource-intensive, so may remain worthwhile even if the rate of quality submissions is low.

A final challenge in collecting incidents was navigating the risk of **publishing potentially harmful information**. The Purdue researchers expressed concern about sharing their current database widely, because it contains examples of mis- and disinformation. The language or images they collected could inadvertently expose users to misleading information or be recycled and used to spread mis- or disinformation further. As part of their plan to improve public access to the data, they are working to watermark the deepfakes in their database to protect against

misuse.¹⁸⁰ This is a responsible but time-consuming process that highlights the ethical challenges in generating research for public use and knowledge, as well as the need for improvements to technical provenance that would reduce the burden on researchers.

Improvements to technical provenance would reduce the burden on researchers.

The GMF team also grappled with **how to responsibly document sensitive material**, including non-consensual sexually explicit content. Part of its process was ensuring that the sources it cited did not leave a “paper trail” to the harmful image. GMF’s methodological choice to rely on media coverage was also partly informed by a decision not to surface potentially harmful content that was not already in the public domain.¹⁸¹ The decision to use media coverage instead of publicizing new instances draws an important contrast between the civil society and media approach to tracking AI incidents during the election, as uncovering new, noteworthy incidents is part of the journalistic mandate. That being

said, Rest of World also addressed this concern in its methodology: “In rare cases, we may opt not to publish the content if doing so could have particularly damaging consequences.”¹⁸²

In addition to these challenges, the comparison between projects highlighted an important difference in how they defined success, related to database size. The Purdue researchers were optimistic about the potential uses and impact of the data they collected, while the NYU team saw its database and other similar databases as not yet large enough the kind of sophisticated empirical research that is increasingly standard. “The ability to infer anything from these databases is extremely limiting,” said Sanderson.¹⁸³

Database size also came up among civil society organizations. The Abundance Institute contained over 46,500 entries at the time of writing,¹⁸⁴ the group described its project as one of the largest media databases tracking AI in the election, and identified scale as one of its successes. GMF’s Spitting Images, on the other hand, included fewer than 200 cases at the time of writing and researchers were less concerned with whether the database included “enough” instances. Instead, they identified the project’s global footprint as their biggest success, highlighting a tradeoff between breadth versus depth of data.¹⁸⁵

The concern about size and the breadth versus depth tradeoff implicates two different models of using research to inform policy. One relies on quantitative evidence, as intended in the NYU case, and the other, following the GMF model, highlights diverse circumstances and edge cases as a way of stress-testing solutions.

B. Capacity and expertise

Incident tracking takes a huge amount of labor, simply in terms of human hours. In all cases CDT looked at, the **small teams doing the work encountered capacity constraints**. One team we spoke with narrowed the scope of its project after beginning the work because the research process was so time consuming. Another had to hire an additional team member. Global projects encountered additional complications, compared with US-focused projects,

as they more frequently dealt with language barriers and needed sufficient expertise to interpret a greater diversity of political and social contexts.

One way to alleviate the burden of human data collection was automation. However, this option required a high degree of technical expertise and introduced other challenges. The Abundance Institute used a subscription service to scrape web content that matched its search parameters.¹⁸⁶ While this helped build a large database, it also collected noise. The team noted that using the data, therefore, required a degree of data literacy and technical skills,¹⁸⁷ which many civic space actors do not have.

Human review was still required when data collection was automated. Initially, Abundance used an automated sentiment analysis tool, but discontinued it because it “found the analysis inaccurate and not particularly informative.”¹⁸⁸ One consideration, the team explained, was that once it determined that the automated sentiment analysis was not reliable or helpful, continuing sentiment analysis manually with such a large dataset would have been too great of a capacity challenge.¹⁸⁹ Purdue’s database also features sentiment analysis, involving human review of each incident.

DFRLab relied on internal technical expertise, including building a custom prompt for the Claude API to code its cases. The automated coding was later manually compared and reconciled with independent human coding of each case. In its assessment, the investment in automation was worthwhile. Even though the project was supported by in-house technical skills, its stability was still to

some extent dependent on the external technology. For example, Anthropic pushed an update during DFRLab’s collection period, which required the team to reconfigure its prompt.¹⁹⁰

Underlying the capacity constraints described above are the financial constraints that most civic space actors encounter. Time-intensive projects like the ones reviewed in this report implicate staff time and hiring decisions, and the subscription services needed

to scrape data cost money. Funding models do not necessarily maximize the effectiveness of data collection projects, which gain value as their scope and scale expand. One group we spoke with

Funding models do not necessarily maximize the effectiveness of data collection projects, which gain value as their scope and scale expand.

explicitly said its project scope was dependent on funding, and the ability to scale the project up in the future would depend on additional funding. As summarized by a researcher from a different project when speaking about the possibility of continuing their work, **“the data becomes exponentially more valuable as it goes on.”**¹⁹¹

C. Calibrating around the AI hype narrative

In the background of these projects, even those that began multiple years ago, was the narrative that 2024 was the headline year for generative AI in elections. Teams we spoke with were trying to calibrate their approach in response to the AI hype narrative throughout their research process. The Abundance Institute likened the discourse around AI in elections to historical hype-cycles and fears around other technological innovation.¹⁹² In the 1930s, for example, people were afraid of telephones.¹⁹³ Other researchers

reflected similar sentiments. At the core of the Purdue team’s work, and part of why it included cheapfakes in addition to AI deepfakes, was a focus on AI as potentially distracting from a wider problem. “We just want to make sure people aren’t directing all their attention to deepfakes at the exclusion of all these other longstanding problematic ways of contributing misinformation.”¹⁹⁴

Paired with the concern about disproportionately focusing on AI in 2024, however, was **the fear of overcorrection**. The standard for AI having an impact on 2024 elections was set so high that it was almost impossible to meet in the absence of a single instance that caused a dramatic upheaval in election outcomes. Much attention was, of course, focused on the US election, where a history of election

denialism and influence operations appeared ripe for deepfake-instigated disruption. While there were several high profile examples of deceptive generated content in the US election,¹⁹⁵ the worst case scenario did not arrive. Analysts should in part consider, however, the counterfactual that one high-risk scenario was avoided

“We just want to make sure people aren’t directing all their attention to deepfakes at the exclusion of all these other longstanding problematic ways of contributing misinformation.”

*- Kaylyn Jackson Schiff,
Political Deepfakes Incident
Database*

because Donald Trump’s swift victory meant that one of the most precarious situations — a lengthy, contested vote counting and certification process after voting ended — did not occur. There are no guarantees that future elections will be equally uneventful, especially in close races.

Companies often deploy vital resources during a short period around election day, even though risks — if not media attention and public pressure — last much longer.

Last January, the NYU team published an article warning that there was a **risk of overestimating the danger posed by AI in elections**, following a pattern previously seen with fears about misinformation.¹⁹⁶ In December 2024, the NYU authors reiterated that point, confirming that AI risks were overstated but that the “fear is it’ll swing too far in the other direction.” One consequence they raised was that guardrails would be removed despite the potential that the worst uses in AI-powered microtargeting are forthcoming.¹⁹⁷ This is a real danger. As seen in past elections, companies often deploy vital resources during

a short period around election day, even though risks — if not media attention and public pressure — last much longer.¹⁹⁸ Meta’s announcement in January 2025 that it would end its third-party fact-checking model in the US, which was implemented following the 2016 election,¹⁹⁹ and stop downranking misinformation²⁰⁰ is a further indication that the pendulum may be swinging too far towards an approach that discounts potential risks to the information space.

The focus on dangerous consequences of AI meant that other, sometimes positive, uses were underappreciated, some of which tracking efforts helped identify. The Rest of World reporting found that it was rare to come across material that “actually intended to deceive.”²⁰¹ Defying its own expectations, the Purdue team found **a surprising amount of satire**, as well generated content that positively boosted candidates. Its research also clarified the prominence of unrealistic deepfakes, in contrast to photorealistic images intended to deceive. Similarly, CDT’s own research found extensive uses of AI by political campaigns in the US that were not readily apparent during the election cycle, including helping with data analysis, writing scripts for canvassers or fundraising outreach, and drafting text in specific dialects (like Mexican Spanish or Cuban Spanish).²⁰²

D. Evaluating the impact of AI on 2024 elections

Quantifying the impact of mis- and disinformation is already extremely difficult, and a far wider range of uses of generative AI in elections needs to be accounted for when considering the technology's overall impact on 2024 elections. Research teams we spoke with had **different perspectives on whether generative AI had an impact on elections** in 2024, or even if there is sufficient evidence to answer that question. At one end of the spectrum, the Abundance Institute's post-election update said that its "analysis identifies no evidence that generative AI negatively affected the process or the outcome of the 2024 US election."²⁰³ The NYU team told CDT that the impact of generative AI on the US election is unknown. GMF's Adrienne Goldstein noted that, from a global perspective, there were specific instances when AI-generated content had a clear impact, even if they did not occur in the US election.²⁰⁴ Zelenko observed, "Misinfo is really cheap, and AI makes

it cheaper. But it doesn't supercharge it the way we thought it might and didn't have the global destructive effect we thought it might."²⁰⁵

Real-time collection and verification fills an information gap among the media and research community.

But a further limitation of incident tracking is that it distills the use of AI into discrete pieces. There is a need for that type of work — DFRLab, for example, pointed out that real-time collection and verification fills an information gap among the media and research community.²⁰⁶ It is also why, as described in the previous section, there is a growing recognition of the diverse and unexpected uses of generative AI in elections.

The tradeoff, however, is that by quantifying individual instances, tracking does not automatically reflect the cumulative effect or ongoing harm of certain uses of generative AI tools. Many of the researchers CDT spoke with were aware of this limitation. The AIID team wrote in September, "While some harms caused by AI systems are the results of discrete events with defined and intuitive event timelines, other AI incidents can span ambiguous amounts

By quantifying individual instances, tracking does not automatically reflect the cumulative effect or ongoing harm of certain uses of generative AI tools.

of time, present as series of events that are difficult to neatly separate and aggregate in a database setting, or obscure when exactly harm can be considered to have occurred.²⁰⁷ Or as Christina Walker of Purdue University summarized, “It’s much harder to think about what the potential harm for a deepfake is when it’s not always as instantaneous as a [crashed self-driving] car.”²⁰⁸ In a similar vein, Zelenko assessed that even though no single piece of content had a decisive impact, generative AI “made the information space louder and noisier and trashier, and in that case has destroyed faith in what they see more broadly.”²⁰⁹

In the end, satire, memes, and other easily identifiable or unrealistic synthetic content comprised a large portion of incidents documented by the trackers CDT reviewed. As described previously, most of the non-journalistic trackers relied on media coverage of AI incidents to build out or fact-check their databases. Most of the “bad” uses of AI that NYU found were ones that had been reported on, while their original research identified more examples that were “anodyne” and not deceptive.²¹⁰ As Kevin Paeth of AIID said, “A lot of AI incident reporting today relies on third party media and journalism and research by these parties. Inherently, the interest of certain kinds of harms and AI incidents will go down and be less newsworthy, despite them posing the same or even greater risks over time as tools become global.” In other words, incident tracking is not designed to reflect subtle technology-related changes that nevertheless amount to meaningful, cumulative shifts in election and information environments.

There are at least three potential harms and directions for future research that are largely not captured by incident tracking: access to information, the liar’s dividend, and the alteration of belief permission structures. The right of **access to information**, including reliable information, is a core component of the right to free expression. To the extent that mis- and disinformation, including AI-generated content, proliferates, it can disrupt access to information which, in some cases, carries implications for a host of other rights, including the right to free and fair elections.²¹¹ Encountering disinformation about time, place, and manner

“Certain kinds of harms and AI incidents will be less newsworthy, despite them posing the same or even greater risks over time as the tools become global.”

- Kevin Paeth, AI Incident Database

of voting is the most straightforward example; a legitimate democratic election requires that voters have accurate information about participation. A second possibility is a severely degraded information space where voters struggle to sift through the “noise” and find accurate information which they can use to form their own opinions. Instead of being actively misled, in this scenario voters simply cannot access the information they need. Searching through low-quality or misleading content imposes a time and knowledge burden on voters; it is not conducive to accessible and inclusive public participation. And though highly plausible, the harm in this scenario is difficult to measure.

While there is general agreement that AI makes content creation easier, one consideration around which experts did not express consensus was the extent to which AI makes information dissemination easier. The dissemination question is critical to evaluating access to information and overall trust in the information environment since information that is created but not seen has little or no effect. One question is how the availability of large quantities of synthetic information will affect the content that social media platforms display to users through algorithms that personalize users’ feeds – in particular, to what degree content made possible by generative AI tools could appear in user feeds. The sources and quality of the synthetic content could inform how the platforms’ algorithm interprets it, depending on the platform’s content moderation policies, but we should anticipate seeing increasingly sophisticated synthetic content over time, which may not be as easy to identify as artificial activity, such as bot networks, has been in the past. On the other hand, companies are working on standards such as C2PA that would help identify the provenance of at least some forms of content. How these developments play out will affect whether and to what extent social media algorithms distinguish authentic and synthetic content for purposes of determining what shows up in a user’s feed, including content intended to be false or misleading election-related information.

The way hashtags have been used in disinformation campaigns is an illustrative analogue. One strategy is to create and consistently use a hashtag, making it trend and therefore gaining visibility.²¹² Another is hashtag hijacking, where a campaign uses an existing hashtag to boost visibility or uses popular hashtags on posts that offer a counternarrative to the movement originally associated with the hashtag. While this can be a method of engaging in online debate, it can also be used strategically to bury the relevant information.²¹³ Demonstrating the way generative AI can support high-volume uses of hashtags, in early 2024, an X account posted AI-generated images related to the newly appointed French Prime Minister at a “high-frequency,” leveraging the #Gabybug hashtag to gain visibility.²¹⁴

The second and third harms stem from how the **proliferation of fake content undermines trust in the information environment**, even in authentic content. The second is the **“liar’s dividend,”** the concept that when people are aware that content they encounter may be fake, bad actors can dismiss even authentic content as synthetic.²¹⁵ During the US election, a North Carolina gubernatorial candidate heavily implied that allegations about his inappropriate and racist online activity were the result of AI-generated deepfakes.²¹⁶ In 2023, an Indian politician denied the authenticity of three damaging audio recordings, blaming them on AI, even though experts concluded that at least two were authentic.²¹⁷ This type of scenario, where someone in power casts doubt on unsavory accusations against them is the most common conceptualization of the liar’s dividend. The Purdue team tried to capture this risk in its database by including real images that sharers presented as deepfakes.²¹⁸ Declining trust in institutions is well-established, and the liar’s dividend functioning as a potential accelerant²¹⁹ is a difficult-to-measure cumulative factor that merits further research. It should not be discounted in assessments of how AI impacted elections in 2024, even if the concrete outcomes are yet unknown.

The third harm is that **degraded trust in the information environment can offer conscious or unconscious permission structures** for individuals to more easily disregard information that flouts their pre-existing beliefs or accept information that conforms to them. Two versions of this are motivated reasoning, where people consciously or subconsciously intake and process information

with a specific end goal in mind,²²⁰ and confirmation bias, the psychological principle stating that people innately discount information that does not conform to their beliefs and opinions.²²¹ Motivated reasoning and confirmation bias are especially important to consider when evaluating the consequences of unrealistic synthetic content, including memes and other creations that are not intended to deceive. For example, future research should consider whether surreal or artistic images that likely do not deceive the consumer — such as Kamala Harris in a communist uniform²²² — can nevertheless fit into existing schemas. A related query is about the impact of high volumes of such content, given that AI may be instrumental in creating it in large quantities. These questions are particularly important given that several of the trackers found frequent uses of memes and unrealistic imagery. Once again, there is potential for profound societal consequences that are not immediately apparent.

The outstanding questions about AI's harms are also a reminder that generative AI may supercharge components of mis- and disinformation, but is not entirely transformative. Access to information and confirmation bias were challenging topics to address long before "deepfake" entered common parlance. In Schiff's words, "A bigger question that we're having as a society is 'What do we mean by a manipulated image?'"²²³ Generative AI has simply increased the urgency of that debate. All three risks — disrupted access to information, the liar's dividend and altered permission structures — are heightened by the current global anti-fact-checking trend, including the end of Meta's fact-checking program in the US and Google's decision not to include fact-checking in search results, videos, and algorithmic ranking in the EU.²²⁴

Perspectives on what should be done about AI content are varied. In researching this report, CDT heard calls for technical provenance tools, government-mandated disclosure requirements, a sociotechnical approach to interventions, and deep investment in traditional journalism. Some of these solutions are familiar to those who have followed policy debates around mis- and disinformation or "fake news" laws, as are concerns that regulating the use of generative AI would interfere with freedom of expression. Weighing these concerns, Zelenko said, "I hope we can always stay one

step ahead in terms of actually being able to verify generative AI material, but I struggle with imagining a future in which we can draw a very clear line around AI-generated disinformation material and ban it, for instance.”²²⁵

The lesson to draw from the past year and the 15 experts who CDT spoke with is the urgency of finding balance, so the pendulum does not swing between phases of AI hype that spur policies that interfere with human rights and periods when the potential harms of AI in elections fades from view. Following the relatively smooth US election, a shift in discourse towards proclaiming the over-exaggeration of AI risks, the conclusion of commitments made under the AI Elections Accords,²²⁶ and the end of election-related surge funding and constant media attention that flourished under the “2024 year of elections” banner, 2025 threatens to be the latter. There will be dozens of elections in 2025. Generative AI tools will continue to improve. Balance requires vigilance by civic space actors, attention from policymakers, and deliberate efforts by technology companies to continue to take seriously their role in mitigating harms from their services.



05

Recommendations

By Isabel Linzer

Drawing from the analysis in this report, CDT recommends that companies and civic space actors continue contributing to a resilient information space by fostering collaboration, developing company policies, and strengthening transparency and data access.

Fostering collaboration among civil society

- Civic space actors should **prioritize interoperability and collaboration whenever possible**. Group agenda setting and information sharing can help reduce duplication and alleviate capacity issues.
- Organizations should engage their networks to **improve language use and definitions** around generative AI, AI, and other types of manipulated or misleading content. Similar to the increasingly standard distinction between mis- and disinformation, developing consistent and accurate terminology related to generative AI, deepfakes, cheapfakes, manipulated content, deceptive content, and other related concepts will support clear public communications and advocacy.

- Funders should consider how to make their **grants supportive of civic space collaboration**. Some funding structures incentivize competition between peer organizations, which creates disincentives for information sharing and collaboration. Funders should consider supporting similar projects that can speak to each other, and other collaborative models.

Building company policies and information resilience

- Social media companies should **update and reinforce existing policies to ensure they account for the harms of false and misleading AI-generated content**, including policies on hate speech, election interference, hot-button issues, misinformation, political advertising, and manipulated media. They should audit their existing enforcement mechanisms for these policies, including friction, labeling, and transparency, to best mitigate the potential harms posed by use-cases of generative AI. Companies should consult civil society experts when formulating and implementing these policies, and ensure that changes respect free expression.
- Companies should **invest in fact-checking partnerships and provide funding to independent fact-checking organizations**. Fact-checking can improve access to quality information without infringing on the right to free expression. Before implementing additions or alternatives, including user-generated fact labels, companies should fund independent researchers to evaluate the speed, reach, and biases of those approaches, which can have shortcomings as compared with professional fact-checking partnerships.²²⁷
- Companies should promote and **direct users to external authoritative sources of election-related information**. They should consult with independent civil society organizations to determine what approach and sources are appropriate in a given country's political context.

- Generative AI developers and governance stakeholders should **take a sociotechnical approach to building and governing AI systems.**²²⁸ This is especially important for developing policy related to elections, as it is crucial to take the diversity of global political and social contexts into account.
- Generative AI developers and deployers should create **rights-respecting guardrails for political campaign use cases involving demographic data**, including use to develop targeted advertising or other targeting strategies. These use cases create heightened risks for privacy violations and hyperlocal mis- and disinformation, but also offer opportunities for campaigns to communicate more effectively with constituents and communities that may be difficult to reach.

Strengthening information access for research and informed policymaking

- Companies should continue to **invest in provenance and detection technology**. They should make authentication tools publicly available, and report findings in a privacy-preserving and anonymized way.
- AI developers and deployers should **commit to greater transparency**, including by adopting approaches like those in the Santa Clara Principles. They should disclose what their election integrity policies are, how their policies are enforced and tested, and when they are in place.
- Companies should **improve the usability of the data they disclose**. Ad libraries, for example, should include metadata and be machine-readable, so that researchers can make better use of the data.
- Online services should **test their election integrity policies in languages other than English** including by investing in high-quality benchmarks, red-teaming, and other forms of external engagement and disclosing to users which languages their systems are trained and tested in.

- Civic space actors involved in tracking or documentation efforts should take steps to **limit the potential harm from sharing misleading or harmful content**, including AI-generated material. They can ensure that sources do not lead back to harmful content, such as non-consensual sexual or intimate deepfakes. They should also consider adding watermarks or other provenance indicators, to reduce the possibility of examples in databases or research being misused.



06 Endnotes

Introduction

- 1 "Venezuelans Use AI Avatars and Instagram Live to Fight Back Maduro's Repression," *Global Voices*, August 19, 2024, <https://perma.cc/BKQ8-NB74>.
- 2 Vittoria Elliot, "There's an AI Candidate Running for Parliament in the UK | WIRED," WIRED, accessed February 21, 2025, <https://www.wired.com/story/ai-candidate-running-for-parliament-uk/>.
- 3 Taryn Hourie, "US President Joe Biden Did Not Warn of South African Sanctions If Ruling ANC Wins 2024 Elections – Ignore AI-Generated Video," *Africa Check*, April 30, 2024, <https://perma.cc/KT6F-BJ2X>.
- 4 Jeremy Hsu, "Experts Bet on First Deepfakes Political Scandal - IEEE Spectrum," *IEEE Spectrum*, June 22, 2018, <https://perma.cc/RTZ7-VXXC>; Deb Riechmann, "I Never Said That! High-Tech Deception of 'deepfake' Videos | AP News," *AP News*, July 2, 2018, <https://apnews.com/article/north-america-donald-trump-ap-top-news-elections-artificial-intelligence-21fa207a1254401197fd1e0d7ecd14cb>.
- 5 Lluís de Nadal and Peter Jančárik, "Beyond the Deepfake Hype: AI, Democracy, and 'the Slovak Case,'" *Harvard Kennedy School Misinformation Review*, August 22, 2024, <https://perma.cc/PFQ3-UDXU>.
- 6 Krystal Hu, "ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note," *Reuters*, February 2, 2023, <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>.
- 7 Beatriz Farrugia, "Brazil's Electoral Deepfake Law Tested as AI-Generated Content Targeted Local Elections," *DFRLab*, November 26, 2024, <https://perma.cc/U7P5-YGUM>; "Guidelines on the Use of Social Media, Artificial Intelligence, and Internet Technology, for Digital Election Campaign, and the Prohibition and Punishment of Its Misuse for Disinformation, and Misinformation, in Connection with the 2025 National and Local Elections and the Barmm Parliamentary Elections," Pub. L. No. Resolution No. 11064 (n.d.), https://comelec.gov.ph/php-tpls-attachments/2025NLE/Resolutions/com_res_11064.pdf.



- 8 "Commission Publishes Guidelines under the DSA," March 25, 2024, European Commission - European Commission, accessed February 21, 2025, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1707.
- 9 Cat Zakrzewski et al., "Debunking Misinformation Failed. Welcome to 'Pre-Bunking,'" *The Washington Post*, May 26, 2024, <https://www.washingtonpost.com/technology/2024/05/26/us-election-misinformation-prebunking/>; "AI Toolkit for Election Officials" (U.S. Election Assistance Commission, n.d.); "Securing Election Infrastructure Against the Tactics of Foreign Malign Influence Operations" (Cybersecurity & Infrastructure Security Agency, April 2024).
- 10 Forthcoming CDT research on campaigns' uses of AI.
- 11 Kate Lamb, Potkin, Fanny, and Teresia, Ananda, "Generative AI May Change Elections This Year. Indonesia Shows How," *Reuters*, February 8, 2024, <https://www.reuters.com/technology/generative-ai-faces-major-test-indonesia-holds-largest-election-since-boom-2024-02-08/>.
- 12 Nilesh Christopher and Varsha Bansal, "Indian Voters Are Being Bombarded With Millions of Deepfakes. Political Candidates Approve," *Wired*, May 20, 2024, <https://perma.cc/PYS7-9BE9>.
- 13 Vittoria Elliott, "Germany's Far-Right Party Is Running Hateful Ads on Facebook and Instagram," *Wired*, May 29, 2024, <https://www.wired.com/story/meta-racist-ads-germany-eu-elections/>.
- 14 Ben Goggin, "Big Tech Companies Reveal Trust and Safety Cuts in Disclosures to Senate Judiciary Committee," *NBC News*, March 29, 2024, <https://perma.cc/H87U-X94S>.
- 15 "A Tech Accord to Combat Deceptive Use of AI in 2024 Elections," Munich Security Conference, 2024, <https://perma.cc/D5EQ-Z2W9>.
- 16 Hayden Field, "Google Restricts Election-Related Queries for Its Gemini Chatbot," *CNBC*, March 12, 2024, <https://perma.cc/48ZW-BC2E>; "Supporting the 2024 Indian General Election," Google, March 12, 2024, <https://perma.cc/C8SN-YSDN>.
- 17 "How OpenAI Is Approaching 2024 Worldwide Elections," OpenAI, February 15, 2024, <https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections/>.
- 18 "Countering Online Manipulation in Europe," Prebunking with Google, accessed February 21, 2025, <https://prebunking.withgoogle.com/eu-prebunking/>; *Decontextualization*. Spot It. Stop It., 2024, <https://perma.cc/ES85-M7YZ>; Zakrzewski et al., "Debunking Misinformation Failed. Welcome to 'Pre-Bunking.'"

- 19 Isabel Linzer and Aliya Bhatia, "Center for Democracy & Technology's Submission to the United Nations Special Rapporteur for Freedom of Opinion and Expression," *Center for Democracy and Technology*, January 15, 2025, <https://perma.cc/3SEA-849T>; William T. Adler and Samir Jain, "Seismic Shifts: How Economic, Technological, and Political Trends Are Challenging Independent Counter-Election-Disinformation Initiatives in the United States" (Center for Democracy & Technology, September 2023), <https://perma.cc/29TP-2NTQ>; Alexandra Reeve Givens, "The Munich Security Conference Provides an Opportunity to Improve on the AI Elections Accord," *Just Security*, February 13, 2025, <https://perma.cc/53KY-D5BW>.

Case Study: Fact-Checking Institutionalized Disinformation in Mexico

- 20 *Fin de Sexenio: ¿Qué Nos Dejó La Conferencia Mañanera de AMLO?*, 2024, <https://perma.cc/JCT4-CE93>.
- 21 Jorge Ramos Avalos, "AMLO: 186,380 MURDERS (AND COUNTING)," *Jorgeramos.Com* (blog), September 30, 2024, <https://perma.cc/8LS5-2DV8>; "Tengo Otros Datos," *AMLO Niega Fracaso En Su Estrategia de Seguridad | #ÚltimasNoticias #Shorts*, 2024, <https://perma.cc/3FKK-GJXS>.
- 22 "'Yo Tengo Los Mismos Datos': Chumel Torres Se Burló de AMLO Con Un Meme," *Infobae*, October 19, 2021, <https://perma.cc/L7VX-WMUM>; Masta Quba and P. Jaguar, *Nosotras Tenemos Otros Datos*, 2021, <https://perma.cc/5C5A-8SCT>.
- 23 Kim Peters, Yoshihisa Kashima, and Anna Clark, "Talking about Others: Emotionality and the Dissemination of Social Information," *European Journal of Social Psychology* 39, no. 2 (2009): 207–22, <https://doi.org/10.1002/ejsp.523>; Stefan Stieglitz and Linh Dang-Xuan, "Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior," *Journal of Management Information Systems* 29, no. 4 (2013): 217–48, <https://doi.org/10.2753/MIS0742-1222290408>; William J. Brady et al., "Emotion Shapes the Diffusion of Moralized Content in Social Networks," *Proceedings of the National Academy of Sciences* 114, no. 28 (2017): 7313–18, <https://doi.org/10.1073/pnas.1618923114>; Smitha Milli et al., "Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media," Knight First Amendment Institute, accessed February 21, 2025, <https://perma.cc/V3CU-7CQ9>.
- 24 "Does Fact-Checking Actually Work? A Critical Review," *Deutsche Welle Akademie*, October 12, 2020, <https://akademie.dw.com/en/is-fact-checking-effective-a-critical-review-of-what-works-and-what-doesnt/a-55248257>; Ethan Porter and Thomas J. Wood, "The Global Effectiveness of Fact-Checking: Evidence from Simultaneous Experiments in Argentina, Nigeria, South Africa, and the United Kingdom," *Proceedings of the National Academy of Sciences* 118, no. 37 (September 14, 2021): e2104235118, <https://doi.org/10.1073/pnas.2104235118>.

- 25 "Researchers Find That Red-Flagging Misinformation Could Slow the Spread of Fake News on Social Media," NYU Tandon School of Engineering, 2020, <https://perma.cc/76TF-WHPE>; Ernesto Calvo et al., "Chequeado en Argentina: Fact Checking y la Propagación de Noticias Falsas en Redes Sociales" (University of Maryland, 2021). <https://perma.cc/G758-PYE3>.
- 26 Sian Lee et al., "'Fact-Checking' Fact Checkers: A Data-Driven Approach," *Harvard Kennedy School Misinformation Review*, October 26, 2023, <https://perma.cc/UCQ9-HSHA>.
- 27 Cuitlahuac Castillo Camarena, "Judicial Reform in Mexico: A Comparative Between the Old and the New Process for Electing Judges | Wilson Center," December 11, 2024, <https://perma.cc/P5UD-SPNL>.
- 28 "Calendario Electoral 2024" (Instituto Nacional Electoral de Mexico, April 2024).
- 29 "Cómputos Distritales 2024," Instituto Nacional Electoral de Mexico, accessed February 21, 2025, <https://perma.cc/567L-JX3U>.
- 30 María Elena Gutiérrez-Rentería, "Mexico," in *Digital News Report 2024* (Reuters Institute for the Study of Journalism, 2024), <https://perma.cc/5KYB-GT8Y>.
- 31 "Elecciones México 2024," Observatorio de Medios Digitales, accessed February 21, 2025, <https://perma.cc/5B9H-SWGG>.
- 32 Gutiérrez-Rentería, "Mexico."
- 33 Sergio Rivera Magos and Gabriela González Pureco, "Populismo, desinformación y polarización política en la comunicación en redes sociales de los presidentes populistas latinoamericanos," *Revista Mexicana de Opinión Pública*, no. 36 (January 15, 2024): 79–107, <https://perma.cc/53GS-T3XA>; "Las Mentiras Del Cuarto Informe de Gobierno de AMLO Son Nocivas Para La Democracia," *The Washington Post*, September 2, 2022, <https://www.washingtonpost.com/es/post-opinion/2022/09/02/cuarto-informe-de-gobierno-amlo-mentiras/>; Lorena Ríos, "After Mexico's Top Streamer — the President — Left Office, YouTube Creators Hustle for Views," *Rest of World*, December 19, 2024, <https://perma.cc/KB9U-D2A4>.
- 34 Vladimir Cortés Roshdestvensky, "Mexico," in *Freedom on the Net 2024 Country Report* (Freedom House), accessed February 21, 2025, <https://perma.cc/S38M-MXY5>.
- 35 Tania Montalvo in discussion with the author, December 19, 2024.
- 36 Daniel Moreno in discussion with the author, December 19, 2024.
- 37 Roshdestvensky, "Mexico"; "How a group of Mexican YouTubers close to President López Obrador disinformed during the pandemic - Chequeado," accessed February 21, 2025, <https://perma.cc/2WV9-QLE4>; Ríos, "After Mexico's Top Streamer — the President — Left Office, YouTube Creators Hustle for Views."
- 38 Roshdestvensky, "Mexico."

- 39 Gutiérrez-Rentería, "Mexico."
- 40 Raul Duran, "Animal Político Responde a AMLO Por Acusar Campaña Contra Su Gobierno [Animal Político responds to AMLO for accusing campaign against his government]," *Debate*, February 8, 2022, <https://www.debate.com.mx/politica/Animal-Politico-responde-a-AMLO-por-acusar-campana-contra-su-gobierno-20220208-0134.html>; Jan-Albert Hoosten, "López Obrador's Anti-Press Rhetoric Leaves Mexico's Journalists Feeling Exposed," *Committee to Protect Journalists*, May 6, 2019, <https://perma.cc/S4WF-278H>; "Mexico," Reporters Without Borders, accessed February 21, 2025, <https://perma.cc/KA6K-L227>; Gutiérrez-Rentería, "Mexico."
- 41 Simon Romero, "Mexico's President Faces Inquiry for Disclosing Phone Number of Times Journalist," *The New York Times*, February 22, 2024, <https://www.nytimes.com/2024/02/22/world/americas/mexico-president-inquiry-times-journalist.html>.
- 42 "Informe Final de Violencia Electoral En México. Proceso 2023-2024" (Laboratorio Electoral, September 10, 2024), <https://perma.cc/QK6T-DSVT>.
- 43 "Mexico" in *Freedom in the World 2025*, Freedom House (2025), <https://freedomhouse.org/country/mexico/freedom-world/2025>; "Mexico," Reporters Without Borders; "Mexico Archives," *Committee to Protect Journalists*, January 31, 2025, <https://perma.cc/S9DA-88XW>.
- 44 "La RELE repudia los hechos de violencia contra periodistas y medios de comunicación en México," Comisión Interamericana de Derechos Humanos (CIDH), November 4, 2024, <https://perma.cc/LJ98-Y9CV>.
- 45 Dulce Ramos in discussion with the author, December 18, 2024.
- 46 "Signatories," IFCN Code of Principles, accessed February 21, 2025, <https://perma.cc/J3TV-TNE8>.
- 47 "How Fact-Checking Works | Transparency Center," Meta, January 7, 2025, <https://transparency.meta.com/features/how-fact-checking-works>; Jorge Ramis, "Latin American Fact-Checkers Brace for Meta's Next Moves," *Wired*, January 11, 2025, <https://perma.cc/TV4K-ZL7T>.
- 48 Paola Nalvarte, "Media Collaboration and Citizen Input Fueled Verificado 2018's Fact-Checking of Mexican Elections," *LatAm Journalism Review* by the Knight Center, July 4, 2018, <https://perma.cc/VTP7-2UM3>; Karolle Rabarison, "Announcing the 2018 Online Journalism Award Winners," *Online Journalism Awards*, September 16, 2018, <https://perma.cc/B3MV-PBDR>; "Verificado 2018," *Verificado 2018*, accessed February 21, 2025, <https://perma.cc/YS8A-VZW4>.
- 49 "How 90 Outlets Are Working Together to Fight Misinformation Ahead of Mexico's Election," The Lenfest Institute for Journalism, accessed February 21, 2025, <https://perma.cc/RP25-62KS>; #Verificado2018 [@VerificadoMX], "📞 55 1245 5032 📞📞📞 Recibimos por Whatsapp todas las cadenas y noticias falsas que te lleguen 📞📞📞📞 55 1245 5032 📞 #Elecciones2018 <https://t.co/7xEtUWVtL9>," Tweet, Twitter, June 22, 2018, <https://perma.cc/B2J5-BG4G>.

- 50 Pablo Gutiérrez, "Alba Mora Roca En El Media Party 2018: 'La Reivindicación Del Buen Periodismo Es La Cura Para Las Noticias Falsas,'" *La Nacion*, August 27, 2018, <https://perma.cc/XM78-4QYN>.
- 51 "América Latina," *Perspectivas Análisis y Comentarios Políticos* (Heinrich Böll Stiftung, November 2020), <https://perma.cc/PQZ5-4HAC>.
- 52 "Staff," *Verificado*, accessed February 21, 2025, <https://perma.cc/5LF8-5J9L>.; "La Mañanera | Verificado," *Verificado*, accessed February 21, 2025, <https://perma.cc/973L-8K4J>.
- 53 "Equipo," *Escenario Tlaxcala*, accessed February 21, 2025, <https://perma.cc/A7XD-XXJU>.
- 54 "LatamChequea: La Red de Chequeadores de Latinoamérica," *Chequeado*, accessed February 21, 2025, <https://perma.cc/8RVX-TRYJ>.
- 55 Marina Adami, "How AI-Generated Disinformation Might Impact This Year's Elections and How Journalists Should Report on It," *Reuters Institute for the Study of Journalism*, March 15, 2024, <https://perma.cc/T6DK-ZYYX>.
- 56 Georgina Zerega, "Un polémico audio con la voz de Martí Batres echa más leña al fuego en la carrera por la candidatura de Ciudad de México," *El País*, November 2, 2023, <https://perma.cc/8G3E-TLSQ>.; "Spitting Images: Tracking Deepfakes and Generative AI in Elections," accessed February 21, 2025, <https://perma.cc/9HHD-RYKQ>.
- 57 Adami, "How AI-Generated Disinformation Might Impact This Year's Elections and How Journalists Should Report on It."
- 58 Leslie Orozco and Melina Barbosa, "Contenidos Creados Con IA: El Reto Para Verificar," *Verificado*, November 8, 2023, <https://perma.cc/TS7C-NPF3>.
- 59 "Spitting Images."
- 60 Reuters Fact Check, "Verificación: Audio Falso de Sheinbaum Fue Fabricado Con Segmentos de Entrevista Antigua," *Reuters*, October 5, 2023, <https://www.reuters.com/fact-check/espanol/SB4VWGIJURMBRCOL2JUXYTEQU-2023-10-10/>.
- 61 Gabriel Nicholas and Aliya Bhatia, "Languages Left Behind: Automated Content Analysis in Non-English Languages," *Center for Democracy and Technology*, August 18, 2022, <https://perma.cc/BQP4-ZVEF>.; Mona Elswah, "Investigating Content Moderation Systems in the Global South" (*Center for Democracy & Technology*, January 30, 2024), <https://perma.cc/JBP2-4H9E>.
- 62 Gutiérrez-Rentería, "Mexico."

- 63 Narges Afshordi, Pearl Han Li, and Melissa Koenig, "Trusting Information from Friends: Adults Expect It but Preschoolers Do Not," *Developmental Psychology* 60, no. 6 (2024): 1161–73, <https://perma.cc/35Y2-3T5D>;; Elisa Shearer, "Friends, Family and Neighbors Are Americans' Most Common Source of Local News," *Pew Research Center*, September 26, 2024, <https://perma.cc/D235-2PXH>.
- 64 Jacob Gursky, Martin J. Riedl, and Samuel Woolley, "The Disinformation Threat to Diaspora Communities in Encrypted Chat Apps," *Brookings Institution*, March 19, 2021, <https://perma.cc/W9AC-YD7E>.
- 65 "IFCN Fact-Checking Organizations on WhatsApp," accessed February 21, 2025, <https://faq.whatsapp.com/5059120540855664>;; Alex Hern and Alex Hern UK technology editor, "WhatsApp to Impose New Limit on Forwarding to Fight Fake News," *The Guardian*, April 7, 2020, sec. Technology, <https://perma.cc/7QC4-W5K5>;; Dhanaraj Thakur et al., "Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems" (Center for Democracy & Technology, August 12, 2021), <https://perma.cc/WTT8-JWUD>.
- 66 "Meedan Final Report: Mexico 2024" (Meedan, 2024), <https://perma.cc/KKV4-S54V>.
- 67 "VerifiChat: El Olfateo de El Sabueso a Través de WhatsApp," *Animal Politico*, February 23, 2023, <https://perma.cc/6APU-6KP2>.
- 68 "Brazil Election 2022," Meedan, accessed February 21, 2025, <https://perma.cc/EB3P-H5SM>;; Katherine Pennacchio, "Desinformación En Las Últimas Elecciones En Brasil Acapara La Atención En La Conferencia Global Fact 10," *LatAm Journalism Review*, July 11, 2023, <https://perma.cc/57PT-ERJM>.
- 69 Linzer and Bhatia, "Center for Democracy & Technology's Submission to the United Nations Special Rapporteur for Freedom of Opinion and Expression."
- 70 Cristina Tardáguila, "López Obrador Launches Its Own 'Verificado' and Infuriates Fact-Checkers in Mexico," *Poynter*, July 9, 2019, <https://perma.cc/72JC-Z22Z>.
- 71 Liliana Elósegui in discussion with the author December 17, 2024.
- 72 Lee et al., "'Fact-Checking' Fact Checkers."
- 73 Daniel Funke, "In Argentina, Fact-Checkers Latest Hire Is a Bot," *Poynter*, January 11, 2018, <https://perma.cc/MF3S-SA92>.
- 74 Laura Zommer, "Welcoming AI! ... and Turning It Into an Unprecedented Ally to Tackle Disinformation in Spanish and Other Languages," *Social Innovations Journal* 23 (January 18, 2024), <https://perma.cc/2VJ2-JSGK>.
- 75 "Factchequeado, AI & Local News Challenge Cohort II," NYU Tandon School of Engineering, accessed February 21, 2025, <https://perma.cc/24CC-V9BC>.
- 76 Sebastian Lanaro, "Avances en inteligencia artificial para frenar desinformaciones:

nueva versión del Chequeabot - Chequeado," Chequeado, April 8, 2024, <https://perma.cc/9LS5-6TXL>.

- 77 Daniela Mendoza in discussion with the author.
- 78 Author's research.
- 79 Tim Harper, "CDT Joins Letter Calling for Meta to Maintain Its CrowdTangle Tool Through the Upcoming Election Cycle," *Center for Democracy & Technology*, May 8, 2024, <https://perma.cc/E9BD-AR5X>.
- 80 "NAT, la presentadora de Fórmula creada con IA, debuta oficialmente (VIDEO)," Radio Fórmula, March 23, 2023, <https://perma.cc/SH94-4D4M>.
- 81 Amy Ross Arguedas, "Public Attitudes towards the Use of AI in Journalism," Reuters Institute for the Study of Journalism, June 17, 2024, <https://perma.cc/3UR2-VNTN>.

Case Study: Foreign Interference and Decentralized Resilience in Taiwan

- 82 Workshop hosted by the author, May 2024.
- 83 Shelley Shan, "KMT's Han Kuo-Yu Elected Legislative Speaker," *Taipei Times*, February 2, 2024, <https://perma.cc/2CMK-E2HQ>.
- 84 Brian Hioe, "KMT's Han Kuo-Yu Is Taiwan's New Legislative Speaker," *The Diplomat*, February 2, 2024, <https://perma.cc/BA7Q-XTP4>.
- 85 Roshdestvensky, "Mexico."
- 86 Doublethink Lab, "2024 Taiwan Elections: Foreign Influence Observation — Preliminary Statement," Medium, February 27, 2024, <https://perma.cc/Z5JT-MYFB>.
- 87 Yang Kuang-shun and 楊光舜, "守護台灣, 就是美國的國家利益," Medium (blog), June 28, 2018, <https://perma.cc/4TWQ-LFJY>.
- 88 Doublethink Lab, "2024 Taiwan Elections: Foreign Influence Observation — Preliminary Statement."
- 89 Emanuel Maiberg, "Taiwan Claims Deepfake Audio Is Defaming a Presidential Candidate," *404 Media*, September 1, 2023, <https://perma.cc/D99E-5U3H>.; Doublethink Lab, "Artificial Multiverse: Foreign Information Manipulation and Interference in Taiwan's 2024 National Elections," *Doublethink Lab via Medium* (blog), August 13, 2024, <https://perma.cc/S2X2-4LNP>.
- 90 "Set up a national-level INGO civil association," NGO International Affairs Committee, Ministry of Foreign Affairs Address, accessed February 23, 2025, <https://perma.cc/7DQY-54MC>.
- 91 "G0v Manifesto," g0v, October 10, 2019, <https://perma.cc/BJ9D-6XAW>.

- 92 "About Doublethink Lab," accessed February 23, 2025, <https://perma.cc/2DES-AU2R>.
- 93 Tseng Po-Yu 曾柏瑜 and Chen Yun-Ju 陳韻如, "An Analysis on the Impact of False Information on Taiwanese Voters," *Doublethink Lab via Medium* (blog), May 6, 2021, 2022, <https://perma.cc/2QK9-VZ5H>; Jerry Yu et al., "2022 Taiwan Election: Foreign Influence Observation Report," *Doublethink Lab via Medium* (blog), June 21, 2023, <https://perma.cc/TMQ4-R3NJ>.
- 94 Sarah Cook, "Beijing's Global Media Influence 2022: Authoritarian Expansion and the Power of Democratic Resilience," *Freedom House*, September 2022, <https://perma.cc/23GL-PQ6U>.
- 95 IORG, "Top Taiwanese Commentator Featured in PRC State Media Douyin Videos 2023Q4 – More of Candidates Hou, Ke, and Domestic Politics" (IORG, January 4, 2024), <https://perma.cc/26WP-ZG34>.
- 96 Doublethink Lab, "Artificial Multiverse."
- 97 Ibid.
- 98 Workshop hosted by the author, May 2024.
- 99 Ibid.
- 100 Doublethink Lab, "Artificial Multiverse."
- 101 Chen Yun and Jason Pan, "DPP Unveils Legislator-at-Large List, with Kuma Academy's Puma Shenese," *Taipei Times*, November 16, 2023, <https://perma.cc/XAZ4-WN3K>.
- 102 "2024 Taiwan Internet Report" (Taiwan Tetwork Information Center, October 2024), <https://perma.cc/SHR2-F5NS>; Simon Kemp, "Digital 2024: Taiwan," Digital 2024 (DataReportal, February 23, 2024), <https://perma.cc/GAK9-G8F3>.
- 103 "99 年人口及住宅普查 99 年人口及住宅普查 總報告統計結果提要分析" (Government of Taiwan, n.d.), <https://perma.cc/AEU5-AU6H>.
- 104 Ben Graham Jones, "Taiwan POWER: A Model for Resilience to Foreign Information Manipulation & Interference," *Doublethink Lab via Medium* (blog), August 9, 2024, <https://perma.cc/8XJP-KRZJ>; Wu Min-Hsuan, "Rip off the Blindfold: Let Taiwanese Civil Society Learn From Ukraine," *The Diplomat*, January 10, 2024, <https://perma.cc/H3PW-EYNW>.
- 105 Author's research.
- 106 Jessica Drun, "Taiwan's Engagement with the World: Evaluating Past Hurdles, Present Complications, and Future Prospects" (The Atlantic Council, December 20, 2022),

- <https://perma.cc/83D6-6MLS>; Michael J. Cole, "Reinvigorating Taiwan's Role as Asia's NGO Hub," Vol. 5, Issue 5., Global Taiwan Brief, March 11, 2020, <https://perma.cc/7QGS-YMGV>; Lin Liang-sheng and Sherry Hsiao, "Taiwan Can Be Hub for NGOs: Ministry," *Taipei Times*, March 19, 2018, <https://www.taipeitimes.com/News/front/archives/2018/03/19/2003689579>.
- 107 "[About Us]," MyGoPen, accessed February 23, 2025, <https://perma.cc/7M4E-RKQF>; "成立宗旨 [Mission Statement]," Taiwan FactCheck Center, accessed February 23, 2025, <https://tfc-taiwan.org.tw/mission-statement/>.
- 108 Workshop hosted by the author, May 2024.
- 109 "【錯誤】網傳文件「賴清德出席巴拉圭新任總統潘尼亞就職典禮，雙方簽署議事錄，台灣未來要幫助巴國建造社宅，[[Wrong] Documents circulated online "Lai Ching-te attended the inauguration ceremony of Paraguay's new President Pena, and both parties signed the proceedings. Taiwan will help Paraguay build social housing in the future.]" *Taiwan FactCheck Center*, August 16, 2023, <https://perma.cc/7EU9-34E5>; Li Wei-Ping, "Foreign Forgeries – an analysis of disinformation tactics leveraging Taiwan's diplomatic events," *Taiwan FactCheck Center* (blog), September 4, 2023, https://en.tfc-taiwan.org.tw/en_tfc_251/.
- 110 Pek Hua Lim, "Social Media Statistics for Taiwan [Updated 2024]," Meltwater, September 25, 2024, <https://perma.cc/6685-EU55>.
- 111 Elizabeth Lange and Doowan Lee, "Line and the Taiwanese Government Are Beating Disinformation. Can Washington Learn From Them?," *Foreign Policy*, November 23, 2020, <https://perma.cc/DHX4-N68T>.
- 112 劉致昕, 柯皓翔, and 許家瑜 [Liu Zhixin, Ke Haoxiang, and Xu Jiayu], "The Content Mill Empire Behind Online Disinformation in Taiwan," 報導者 *The Reporter*, December 25, 2019, <https://perma.cc/A5ZM-6QUL>.
- 113 "About IORG," IORG, July 17, 2024, <https://perma.cc/D5TL-RNSQ>; Julia Bergin, "Chinese Ambassadors Eager to 're-Educate' Taiwan, but Teachers Push Back on Classroom Politics," *Crikey*, August 12, 2022, <https://perma.cc/EX9C-S2Q4>.
- 114 Workshop hosted by the author, May 2024.
- 115 Chiaoning Su and Wei-Ping Li, "Three Musketeers against Mis/Disinformation: Assessing Citizen-Led Fact-Checking Practices in Taiwan," *Taiwan Insight* (blog), March 31, 2023, <https://perma.cc/DV6Y-KV3E>.
- 116 "Taiwan's fact-checkers are using AI to combat misinformation," *Taiwan FactCheck Center*, July 30, 2021, <https://perma.cc/LWK4-63VY>.
- 117 Johnson Liang and Billion Lee, "Accelerating Crowd-Sourced Fact-Checking Chatbot System," MIT Solve, June 30, 2019, <https://perma.cc/36CY-QF3M>.

- 118 Elaine Chan, "From Beef Noodles to Bots: Taiwan's Factcheckers on Fighting Chinese Disinformation and 'Unstoppable' AI," *The Guardian*, June 4, 2024, <https://perma.cc/E4LU-Z2YV>.
- 119 Jon Bateman and Dean Jackson, "Countering Disinformation Effectively: An Evidence-Based Policy Guide." Carnegie Endowment, January 31, 2024. <https://perma.cc/R3DS-DNGN>.
- 120 "Auntie Meiyu," accessed February 23, 2025, <https://perma.cc/L3BZ-QDMS>; "Who is Auntie Meiyu? Taiwanese youth found a new way to challenge fake news in private [sic] chat group," *g0v via Medium* (blog), March 7, 2019, <https://perma.cc/6QY4-R3S5>.
- 121 Andrew Deck and Vittoria Elliot, "How Line Built Fact-Checking into Its Encrypted Messaging App," *Rest of World*, March 7, 2021, <https://perma.cc/CK3Y-JJVC>.
- 122 Workshop hosted by the author, May 2024.
- 123 Allie Funk, Kian Vesteinsson, and Grant Baker, "Delegitimizing the Messenger: The Assault on Fact-Checkers | Council on Foreign Relations," Council on Foreign Relations, November 4, 2024, <https://perma.cc/T3QJ-VL8C>.
- 124 Workshop hosted by the author, May 2024.
- 125 Workshop hosted by the author, May 2024.
- 126 Aaron Su, "Labours of G0v: Rethinking Work from the Perspective of Data Activists," *Taiwan Insight* (blog), October 11, 2022, <https://perma.cc/S44Q-S7JQ>.
- 127 "Public Officials Election and Recall Law-National Law Database," §51, accessed February 23, 2025, <https://web.archive.org/web/20240108015336/https://law.moj.gov.tw/LawClass/LawAll.aspx?pcode=D0020010>.
- 128 *Ibid*, §51-1.
- 129 *Ibid*, §51-3.
- 130 The Covid-era rules lapsed mid-2023.
- 131 Allie Funk, Kian Vesteinsson, and Grant Baker, "Freedom on the Net 2024: The Struggle for Trust Online" (Freedom House, 2024), <https://perma.cc/BY3Q-C7WJ>.
- 132 "Taiwan," Freedom on the Net 2024 (Freedom House, 2024), <https://perma.cc/PA7C-YR24>.
- 133 Sean Scanlan, "Taiwan President Lai Ching-Te Appoints New Ambassadors-at-Large," October 8, 2024, <https://perma.cc/6WLS-FL77>.
- 134 Nicola Smith Smith, "Schoolkids in Taiwan Will Now Be Taught How to Identify Fake News," *TIME*, April 7, 2017, <https://perma.cc/T82X-NWB8>.
- 135 Steven Butler and Iris Hsu, "Q&A: Taiwan's Digital Minister on Combatting Disinformation without Censorship," Committee to Protect Journalists, May 23, 2019,

<https://perma.cc/2JY2-7TUS>.

- 136 Kuang-shun and 楊光舜, “守護台灣, 就是美國的國家利益.”
- 137 “Taiwan” in Freedom on the Net 2024.
- 138 Russell Hsiao, “PRC Election Interference Targeting Taiwan’s 2024 Elections Tests the Anti-Infiltration Act,” Global Taiwan Institute, May 3, 2023, <https://perma.cc/U663-KDL6>.
- 139 Nick Aspinwall, “Taiwan’s War on Fake News Is Hitting the Wrong Targets,” *Foreign Policy*, January 10, 2020, <https://perma.cc/6H5Q-B3A9>.
- 140 Ibid.; “「中天關定了」國民黨控：一切都照「這套」劇本走 [“Zhongtian is closed” Kuomintang control: Everything will follow ‘this’ script],” *Yahoo News*, October 14, 2020, <https://perma.cc/MP8S-Y32Q>.
- 141 “Taiwan” in Freedom on the Net 2024.
- 142 Sam Robbins, “Can G0v Be Replicated Abroad?,” *Taiwan Insight* (blog), October 13, 2022, <https://perma.cc/9TWM-R2Y7>.

Case Study: AI Incident Tracking by Global Civic Space Actors

- 143 Daniel Schiff, Kaylyn Jackson Schiff, and Christina Walker (PDID) in discussion with the author, November 26, 2024.
- 144 Sarah Grevy Gotfredsen and Kaitlyn Dowling, “Meta Is Getting Rid of CrowdTangle—and Its Replacement Isn’t as Transparent or Accessible,” *Columbia Journalism Review*, July 9, 2024, <https://perma.cc/DB4A-B45Q>; Harper, “CDT Joins Letter Calling for Meta to Maintain Its CrowdTangle Tool Through the Upcoming Election Cycle.”
- 145 John Naughton, “Closing the Stanford Internet Observatory Will Edge the US towards the End of Democracy,” *The Guardian*, June 29, 2024, <https://perma.cc/B35E-KCUV>; Adler and Jain, “Seismic Shifts.”
- 146 Zeve Sanderson and Cristina de la Puerta (AI Political Archive) in discussion with the author, December 10, 2024.
- 147 Neil Chilson, “The Matter of Disclosure and Transparency of Artificial Intelligence-Generated Content in Political Advertisements,” Abundance Institute, September 2024, <https://perma.cc/T3PU-L73R>.
- 148 Jon Roozenbeek, Sander Van der Linden, and Thomas Nygren, “Prebunking Interventions Based on the Psychological Theory of ‘Inoculation’ Can Reduce Susceptibility to Misinformation across Cultures. | HKS Misinformation Review,” *Misinformation Review*, February 3, 2020, <https://perma.cc/UHX5-LJUG>.
- 149 Dina Sadek, Meredith Furbish, and Max Rizzuto, “DFRLab Launches the 2024 Foreign

- Interference Attribution Tracker - DFRLab," DFRLab, October 23, 2024, <https://perma.cc/764T-RLYD>.
- 150 Christina P. Walker, Daniel S. Schiff, and Kaylyn Jackson Schiff, "Merging AI Incidents Research with Political Misinformation Research: Introducing the Political Deepfakes Incidents Database," *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 21 (2024): 23053–58, <https://perma.cc/ZQ5X-F6MD>.
- 151 Daniel Schiff, Kaylyn Jackson Schiff, and Christina Walker (PDID) in discussion with the author, November 26, 2024.
- 152 Sean McGregor, "When AI Systems Fail: Introducing the AI Incident Database," Partnership on AI, November 18, 2020, <https://perma.cc/59RJ-4DBH>.
- 153 Sean McGregor and Kevin Paeth (AIID) in discussion with the author, December 13, 2024.
- 154 Walker, Schiff, and Schiff, "Merging AI Incidents Research with Political Misinformation Research."
- 155 Victoria Turk, "How We're Tracking AI Incidents around Global Elections," *Rest of World*, April 16, 2024, <https://perma.cc/U5ZL-P3M3>.
- 156 Sean McGregor and Kevin Paeth (AIID) in discussion with the author, December 13, 2024.
- 157 "List of Taxonomies," AI Incident Database, accessed February 23, 2025, <https://perma.cc/U32J-YNLJ>.
- 158 Camille François, "Actors, Behaviors, Content: A Disinformation ABC" (Transatlantic Working Group, September 20, 2019), <https://perma.cc/M7ZW-PCBW>.
- 159 "Editor's Guide," AI Incident Database, accessed February 23, 2025, <https://perma.cc/QX5N-M2M2>; Sean McGregor and Kevin Paeth (AIID) in discussion with the author, December 13, 2024.
- 160 Sean McGregor and Kevin Paeth (AIID) in discussion with the author, December 13, 2024.
- 161 "Spitting Images: Tracking Deepfakes and Generative AI in Elections Methodology," German Marshall Fund of the United States, October 10, 2024, <https://perma.cc/Q59N-WWVX>.
- 162 "Launching the AI Political Archive," NYU's Center for Social Media and Politics, July 8, 2024, <https://perma.cc/DTA7-RMWG>.
- 163 Walker, Schiff, and Schiff, "Merging AI Incidents Research with Political Misinformation Research."
- 164 Adler and Jain, "Seismic Shifts."

- 165 Developers [@XDevelopers], "Starting February 9, We Will No Longer Support Free Access to the Twitter API, Both v2 and v1.1. A Paid Basic Tier Will Be Available Instead 🍷," Tweet, Twitter, February 2, 2023, <https://perma.cc/R SX8-DKLD>.
- 166 Gotfredsen and Dowling, "Meta Is Getting Rid of CrowdTangle—and Its Replacement Isn't as Transparent or Accessible"; Harper, "CDT Joins Letter Calling for Meta to Maintain Its CrowdTangle Tool Through the Upcoming Election Cycle."
- 167 Gabriel Nicholas, "Grounding AI Policy: Towards Researcher Access to AI Usage Data" (Center for Democracy & Technology, August 13, 2024), <https://perma.cc/YQG7-AQ2U>.
- 168 Emerson Brooking, Meredith Furbish, Max Rizzuto, and Dina Sadek in discussion with the author, December 5, 2024.
- 169 Sadek, Furbish, and Rizzuto, "DFRLab Launches the 2024 Foreign Interference Attribution Tracker - DFRLab."
- 170 Sean McGregor and Kevin Paeth (AIID) in discussion with the author, December 13, 2024.
- 171 Zeve Sanderson and Cristina de la Puerta (AI Political Archive) in discussion with the author, December 10, 2024.
- 172 Ibid.
- 173 Tim Harper et al., "Rules of the Road: Political Advertising on Social Media in the 2024 U.S. Election Report — Rules of the Road: Political Advertising on Social Media in the 2024 U.S. Election" (Center for Democracy & Technology, September 19, 2024), <https://perma.cc/68S3-LSBH>.
- 174 Claire Pershan and Amaury Lesplingart, "Full Disclosure: Stress Testing Tech Platforms' Ad Repositories" (Mozilla Foundation, April 16, 2024), <https://foundation.mozilla.org/en/research/library/full-disclosure-stress-testing-tech-platforms-ad-repositories/>.
- 175 Kevin Schaul, Pranshu Verma, and Cat Zakrzewski, "See Why AI Detection Tools Can Fail to Catch Election Deepfakes," *The Washington Post*, August 15, 2024, <https://www.washingtonpost.com/technology/interactive/2024/ai-detection-tools-accuracy-deepfakes-election-2024/>.
- 176 Loreben Tuquero, "AI Detection Tools for Audio Deepfakes Fall Short. How 4 Tools Fare and What We Can Do Instead," *Poytner*, March 21, 2024, <https://perma.cc/23EC-LDYT>.
- 177 Daniel Schiff, Kaylyn Jackson Schiff, and Christina Walker (PDID) in discussion with the author, November 26, 2024.
- 178 "Spitting Images: Methodology."
- 179 Daniel Schiff, Kaylyn Jackson Schiff, and Christina Walker (PDID) in discussion with the author, November 26, 2024.
- 180 Ibid.

- 181 Adrienne Goldstein (GMF's Spitting Images) in discussion with the author, December 11, 2024.; "Spitting Images: Methodology."
- 182 Turk, "How We're Tracking AI Incidents around Global Elections."
- 183 Zeve Sanderson and Cristina de la Puerta (AI Political Archive) in discussion with the author, December 10, 2024.
- 184 "Abundance Institute AI and Elections Tracker," Abundance Institute, accessed February 23, 2025, <https://perma.cc/B5UP-658G>.
- 185 Adrienne Goldstein (GMF's Spitting Images) in discussion with the author, December 11, 2024.
- 186 Neil Chilson, Taylor Barkley, and Logan Whitehair, "The Abundance Institute AI and Elections Update," Substack newsletter, *Now + Next* (blog), May 9, 2024, <https://nowandnext.substack.com/p/the-abundance-institute-ai-and-elections>.
- 187 Taylor Barkley and Logan Whitehair (Abundance Institute) in discussion with the author, December 10, 2024.
- 188 Taylor Barkley, Neil Chilson, and Logan Whitehair, "Part 2: The Abundance Institute AI and Elections Update," Substack newsletter, *Now + Next* (blog), August 7, 2024, <https://nowandnext.substack.com/p/part-ii-the-abundance-institute-ai>.
- 189 Taylor Barkley and Logan Whitehair (Abundance Institute) in discussion with the author, December 10, 2024.
- 190 Emerson Brooking, Meredith Furbish, Max Rizzuto, and Dina Sadek (DFRLab) in discussion with the author, December 5, 2024.
- 191 Brooking, Furbish, Rizzuto, and Sadek (DFRLab) in discussion with the author.
- 192 Taylor Barkley and Logan Whitehair (Abundance Institute) in discussion with the author, December 10, 2024.
- 193 Adrienne LaFrance, "When the Telephone Was Dangerous - The Atlantic," *The Atlantic*, September 6, 2015, <https://perma.cc/EL5U-SRBB>.
- 194 Schiff, Jackson Schiff, and Walker (PDID) in discussion with the author.
- 195 Shannon Bond, "A Political Consultant Faces Charges and Fines for Biden Deepfake Robocalls," *NPR*, May 23, 2024, <https://perma.cc/DT5M-DYRC>.; Kat Tenbarge, "The AI Taylor Swift Endorsement Trump Shared Was Originally a Pro-Biden Facebook Meme," *NBC News*, September 13, 2024, <https://www.nbcnews.com/tech/tech-news/ai-taylor-swift-endorsement-trump-shared-was-originally-biden-meme-rcna170945>.
- 196 Zeve Sanderson, Solomon Messing, and Joshua A. Tucker, "Misunderstood Mechanics: How AI, TikTok, and the Liar's Dividend Might Affect the 2024 Elections," *Brookings*, January 22, 2024, <https://perma.cc/2NVJ-BRCG>.

- 197 Sanderson and de la Puerta (AI Political Archive) in discussion with the author.
- 198 Aliya Bhatia and William T. Adler, "CDT Weighs In on Meta Oversight Board's Case on Takedown of Speech Calling for Attack on Brazil's National Congress," *Center for Democracy and Technology*, March 23, 2023, <https://perma.cc/9VYU-UCH4>.
- 199 Joel Kaplan, "More Speech and Fewer Mistakes," Meta, January 7, 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>; Tessa Lyons, "Increasing Our Efforts to Fight False News," Meta, June 21, 2018, <https://perma.cc/42UA-37Y6>.
- 200 Casey Newton, "Meta Just Flipped the Switch That Prevents Misinformation from Spreading in the United States," *Platformer*, January 14, 2025, <https://perma.cc/3SXA-UQNL>.
- 201 Michael Zelenko (Rest of World) in discussion with the author, December 17, 2024.
- 202 Forthcoming CDT research on campaigns' uses of AI.
- 203 Taylor Barkley, Neil Chilson, and Logan Whitehair, "Part V: The Abundance Institute AI and Post-Election Update," Substack newsletter, *Now + Next* (blog), November 12, 2024, <https://nowandnext.substack.com/p/part-v-the-abundance-institute-ai>.
- 204 Goldstein (GMF's Spitting Images) in discussion with the author.
- 205 Zelenko (Rest of World) in discussion with the author.
- 206 Brooking, Furbish, Rizzuto, and Sadek (DFRLab) in discussion with the author.
- 207 Kevin Paeth et al., "Lessons for Editors of AI Incidents from the AI Incident Database" (arXiv, September 24, 2024), <https://perma.cc/88T3-SZYK>.
- 208 Schiff, Jackson Schiff, and Walker (PDID) in discussion with the author.
- 209 Zelenko (Rest of World) in discussion with the author.
- 210 Sanderson and de la Puerta (AI Political Archive) in discussion with the author.
- 211 "Disinformation and Human Rights Explained," *Global Partners Digital*, June 1, 2023, <https://perma.cc/KM5N-DLWW>; "International Standards: Right to Information," *ARTICLE 19*, April 5, 2012, 19, <https://perma.cc/88SJ-33W8>.
- 212 Odanga Madung and Brian Obilo, "Inside the Shadowy World of Disinformation-for-Hire in Kenya," Mozilla Foundation, September 2, 2021, <https://foundation.mozilla.org/en/blog/fellow-research-inside-the-shadowy-world-of-disinformation-for-hire-in-kenya/>.
- 213 Reilly Anne Dempsey Willis, "Habermasian Utopia or Sunstein's Echo Chamber? The 'Dark Side' of Hashtag Hijacking and Feminist Activism," *Legal Studies* 40, no. 3 (September 2020): 507–26, <https://perma.cc/X3RX-RQDW>.
- 214 Valentin Châtelet, "French Prime Minister Faces Onslaught of Online Attacks," *DFRLab*, February 20, 2024, <https://perma.cc/9JF2-LU9X>.

- 215 Bobby Chesney and Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review* 107, no. December 2019 (2019), <https://perma.cc/259J-KQ6T>.
- 216 Philip Bump, "Mark Robinson Offers up the 2024 Version of the I-Was-Hacked Defense," *Washington Post*, September 20, 2024, <https://www.washingtonpost.com/politics/2024/09/20/mark-robinson-comments-cnn-story-north-carolina/>.
- 217 Nilesh Christopher, "An Indian Politician Says Scandalous Audio Clips Are AI Deepfakes. We Had Them Tested," *Rest of World*, July 5, 2023, <https://perma.cc/Y4P3-3CZP>.
- 218 Schiff, Jackson Schiff, and Walker (PDID) in discussion with the author.
- 219 Kaylyn Jackson Schiff, Daniel S. Schiff, and Natália S. Bueno, "The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?," *American Political Science Review* 119, no. 1 (February 2025): 71–90, <https://perma.cc/T2G6-PXRA>.
- 220 Sander van der Linden, "Misinformation: Susceptibility, Spread, and Interventions to Immunize the Public," *Nature Medicine* 28 (2022): 460–67, <https://perma.cc/6DZE-KPKB>.
- 221 Michael Engle, "LibGuides: Misinformation, Disinformation, and Propaganda: Personal Bias," Cornell University Library, January 24, 2025, <https://perma.cc/Z47J-HJ6U>.; Michal Piksa et al., "The Impact of Confirmation Bias Awareness on Mitigating Susceptibility to Misinformation," *Frontiers in Public Health* 12 (2024), <https://perma.cc/PJ7W-LWSU>.
- 222 Ione Wells and Jessica Cruz, "Why 'Comrade Kamala' Memes Are Taking off among Latino Exiles," *BBC*, September 23, 2024, <https://perma.cc/7KPP-3J92>.
- 223 Schiff, Jackson Schiff, and Walker (PDID) in discussion with the author.
- 224 Sara Fischer, "Scoop: Google Won't Add Fact-Checks despite New EU Law," *Axios*, January 16, 2025, <https://www.axios.com/2025/01/16/google-fact-check-eu>.
- 225 Zelenko (Rest of World) in discussion with the author.
- 226 "A Tech Accord to Combat Deceptive Use of AI in 2024 Elections."

Recommendations

- 227 Tom Stafford, "Do Community Notes Work?," LSE Impact Blog (blog), January 14, 2025, <https://perma.cc/LJD3-8NWA>.
- 228 Miranda Bogen and Amy Winecoff, "Applying Sociotechnical Approaches to AI Governance in Practice" (Center for Democracy & Technology, May 15, 2024), <https://perma.cc/UFT8-3ER5>.




 cdt.org

 cdt.org/contact

 **Center for Democracy & Technology**

1401 K Street NW, Suite 200

Washington, D.C. 20005

 202-637-9800

 @CenDemTech

