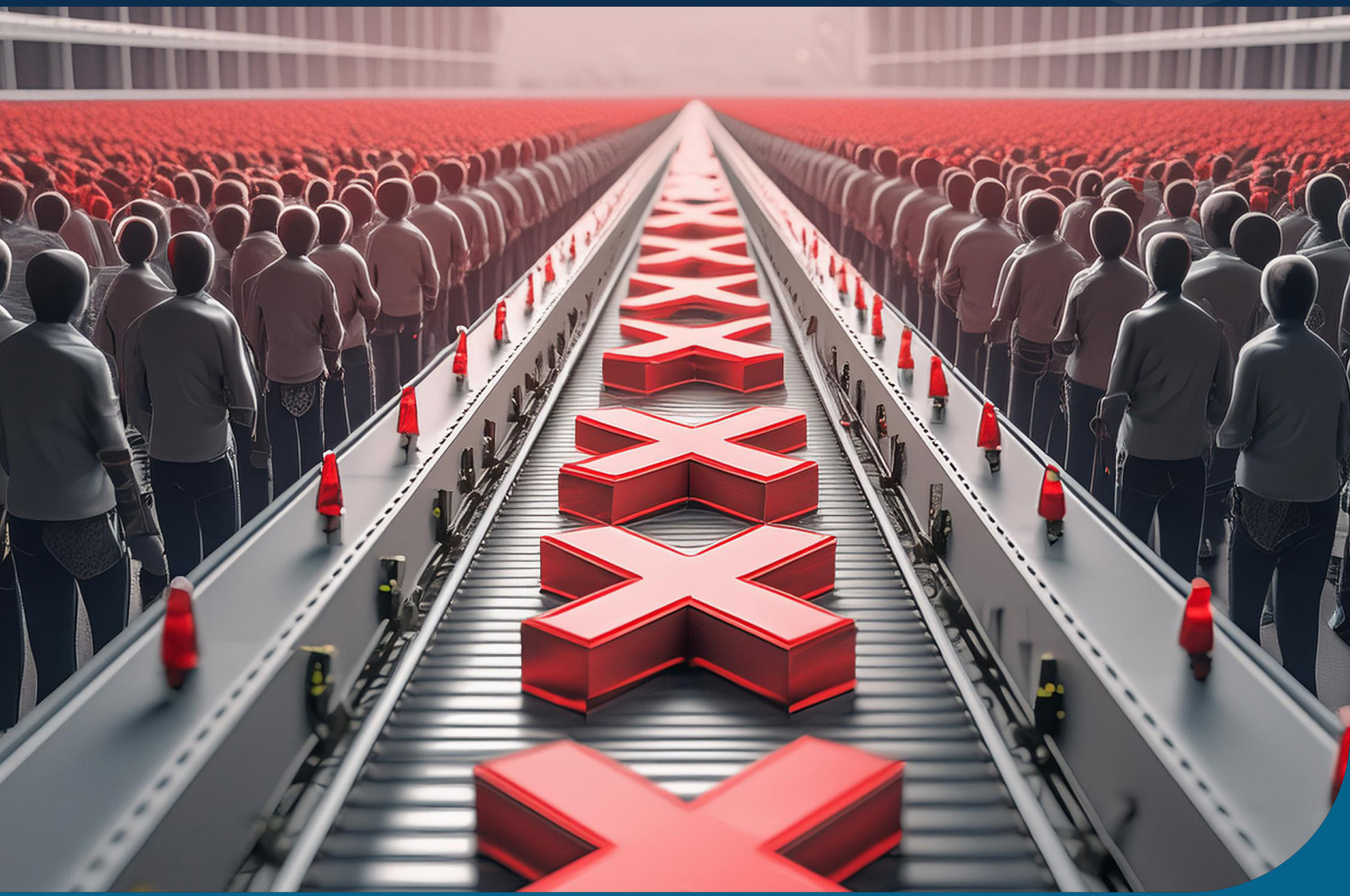
MANUFACTURING DECEIT

HOW GENERATIVE AI SUPERCHARGES INFORMATION MANIPULATION

// BEATRIZ SAAB



NED NATIONAL ENDOWMENT FOR DEMOCRACY
SUPPORTING FREEDOM AROUND THE WORLD

FORUM INTERNATIONAL FORUM FOR DEMOCRATIC STUDIES

// REPORT / JUNE 2024

MANUFACTURING DECEIT

HOW GENERATIVE AI SUPERCHARGES INFORMATION MANIPULATION

CONTENTS

Executive Summary 1

Generative AI and Information Manipulation: A Rapidly Growing Challenge for Democracies 3

How Generative AI Strengthens Information Manipulation by Authoritarians 5

Leveraging Generative AI to Support Information Space Integrity 13

Looking to the Future 18

Endnotes 19

About the Author 22

Acknowledgments 22

Photo Credits 22

Cover image generated by Adobe Firefly using prompt: "Assembly line conveyor belt that is printing 3D red X's. Large crowd of people in the distance. Factory workers packing X's in boxes."

NED NATIONAL ENDOWMENT FOR DEMOCRACY
SUPPORTING FREEDOM AROUND THE WORLD

FORUM INTERNATIONAL FORUM FOR DEMOCRATIC STUDIES

EXECUTIVE SUMMARY

Authoritarian actors have long worked to undermine democracy at a global scale by manipulating the information space, but the recent emergence of faster, more expansive, and potentially more potent "generative AI" technologies is creating new risks. With more than fifty national elections around the globe set to take place in 2024, the stakes this year are particularly high. While it may still be too early to assess whether this new technology is creating *decisive* advantages for authoritarian powers, **it is clear that they are experimenting and incorporating these tools into their strategies to undermine democracy**. That said, beyond specific manipulative information campaigns, **the deeper impact of this new technology may be felt beyond the election contexts, in citizens' loss of trust in online content or in democracy itself**.

Earlier models of artificial intelligence excel in the recognition and analysis of patterns in large-scale collections of text, audio, or visual data. Generative models of artificial intelligence surpass the capabilities of these earlier models in their ability to extrapolate from patterns to create new content. Furthermore, generative models operate in response to simple, natural-language text prompts, lowering the bar for their use and setting the stage for an even more complex and vexing information landscape.

Research indicates that authorities in countries including Russia, China, Iran, and Venezuela are experimenting with gen AI purposefully in order to manipulate the information space and undermine democracy. This new technology accelerates these efforts in at least three ways:

- **Less expensive, easier production of high-fidelity synthetic content:** Gen AI leverages patterns in data to create new content, or "synthetic media," quickly and inexpensively, including convincing-yet-false images of public figures and events.
- **Automating the production of synthetic media:** Gen AI can be used to automate key technical processes in content production, reducing the need for human engagement and oversight, and lowering barriers to scaling content creation.

- **Individually-tailored content distribution:** Gen AI tools can use data from social media and other public sources of information to create individual profiles and then tailor content to those profiles. Such specially-customized content has a greater chance of impacting attitudes and beliefs.

Taken together, these examples highlight how AI technology can support authoritarians' long-term interest in undermining societal trust and the foundations of democratic governance.

At the same time, as the authoritarian threat to democracy evolves, there is growing evidence that democratic reformers can use gen AI in support of *integrity* in the information space, especially around democratic discourse in the context of elections. Some experts have suggested that the opportunities afforded by gen AI around fact checking, independent journalism, and media monitoring may be greater than those for authoritarians. Relevant examples of how civil society is leveraging gen AI include:

- **Fact checkers are employing gen AI to accelerate the verification of information.** The technology speeds the contextual research that is critical to fact checking.
- **Journalists are experimenting with gen AI to create efficiencies in their work.** Globally, journalists are using gen AI to generate interest in their work, summarize their own content, analyze large amounts of data, and more.
- **Democratic reformers are using gen AI to detect information manipulation.** Gen AI can accelerate the efforts of civil society organizations to detect information manipulation online by hastening the identification the behaviors, narratives, and tactics used by authoritarians.
- **Civil society organizations are leveraging gen AI to identify deceptive gen AI-produced content.** Gen AI is helping civil society drive public awareness and response to deceptive and harmful gen AI–produced content online.

While the balance of threats and opportunities to democracy presented by new gen AI technologies may not yet be entirely clear, experts have identified the use of gen AI tools around nearly every national election since at least mid-2023, from Bangladesh to Argentina and South Africa. In many recent elections, global authoritarian powers have deployed gen AI—or amplified gen AI content—that harms democracy. This report highlights that the threat gen AI poses to the integrity of the information space extends beyond elections, potentially undermining societal trust and with it democratic norms and standards. Future analysis and research will be critical for understanding the ways this evolving technology can support and accelerate responses from democratic reformers.

GENERATIVE AI AND INFORMATION MANIPULATION: A RAPIDLY GROWING CHALLENGE FOR DEMOCRACIES

As authoritarian regimes in Russia, China, Iran, and elsewhere actively seek to undermine trust in democracy around the world, critical changes to the information environment are aiding their efforts. The growth of generative artificial intelligence (gen AI) is among the most important of these changes, reducing the cost, time, and effort required by authoritarian actors to both mass-produce and disseminate manipulative content with the aim of smearing opponents and promoting allies, exacerbating divisions in democratic societies.

This report assesses how authoritarians are using gen AI to advance malign narratives and break down the concept of a shared truth that lies at the heart of social trust and democratic institutions. It also describes the ways in which civil society organizations have begun to use some of the same tools to push back against authoritarian distortions in the information space to empower and promote allies, enhancing responses by civil society experts, fact checkers, independent journalists, and other democratic partners.

While authoritarian actors have long worked to manipulate the information space to the detriment of democracy, the emergence of faster, more expansive, and potentially more potent gen AI technologies is already leading to profound changes in the information environment.[1] With more than fifty national

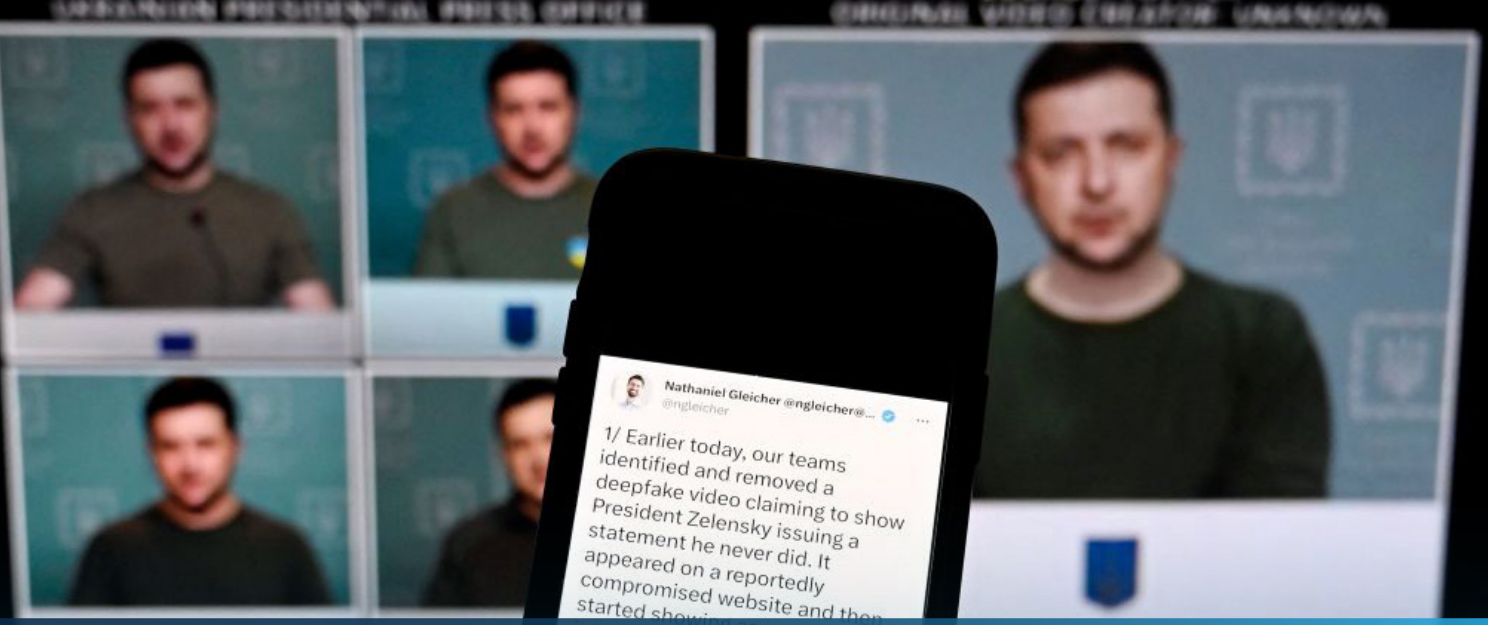
elections set to take place in 2024 around the globe, in countries comprising more than half of the world's population, the stakes are particularly high. Moreover, the threat gen AI poses to the integrity of the information space extends beyond election contexts.

"Generative models" of artificial intelligence represent a significant leap forward in the ease and speed of creation and editing of original content.[2] Traditional AI models are adept at identifying patterns in large-scale collections of text, audio, and visual data. Generative models extrapolate from those patterns to generate new content using only simple, natural-language text prompts, image, video, or audio snippets, setting the stage for an even more complex and vexing information landscape.

In essence, traditional AI is like an automatic labeling machine that—with a bit of human training—can, for instance, distinguish and label photos of various species of birds; generative AI models conduct the same analysis and produce hundreds or thousands of new images or even videos of the same types of birds, and can do so within seconds on the basis of a simple text prompt. Of course, if the use of gen AI was limited to replicating pictures of birds, there would be little concern about the technology's potential to help undermine democracy. One generative model, Microsoft's VASA-1, creates realistic talking-head-style videos with synchronized facial and lip movements in mere seconds, using just two inputs: a photo of an individual's face and an audio clip for that individual to "voice."[3]

Therefore, it is essential that civil society groups and other democratic actors—including governments and the private sector—understand this technology and its potential application for authoritarian ends. They should also consider how to best fight back by employing such tools as part of discourse-focused efforts to secure the integrity of the information space—both during election cycles and between them, when democracies are typically less attentive to information manipulation by authoritarians.

Traditional AI is like an automatic labeling machine that can distinguish and label photos of various species of birds; generative AI models conduct the same analysis but then can manufacture entirely new, artificial images of birds.



HOW GENERATIVE AI STRENGTHENS INFORMATION MANIPULATION BY AUTHORITARIANS

Researchers and advocates have raised significant concerns regarding authoritarian regimes' experimentation with gen AI technologies around elections.[4] These concerns are particularly prevalent given the increasing speed and scale of public discourse that often strains the resources available to fact checkers, journalists, and others working to counter information manipulation and amplify high-quality information.

Globally influential authoritarian powers such as Russia and China—among others—exploit an information space that is evolving faster than the relevant governing or regulatory mechanisms, expanding the reach of such manipulative efforts across much of the world.

Information Manipulation and Authoritarian Powers

Information manipulation can be described as a **threat to democratic values, institutions, or political processes** that seeks to exert **hidden, malign influence on public attitudes and behaviors** through the information space, and is **conducted in an intentional and coordinated manner by authoritarian regimes or their proxies**. In many cases, the content in question is authentic but has been removed from its original context or edited in order to confuse the public.

Globally influential authoritarian powers such as Russia and China—among others—exploit an information space that is evolving faster than the relevant governing or regulatory mechanisms, expanding the reach of such manipulative efforts across much of the world. These regimes employ sophisticated information manipulation strategies to undermine democracy by controlling which narratives and perspectives dominate public discourse, while suppressing reasoned debate on key issues, bolstering their autocratic allies, and undercutting prodemocratic actors.

Building on existing antidemocratic tactics in the information space, gen AI tools may be used for a variety of purposes, including:

- To smear participants in the political process, as was the case in Slovakia when an audio "deepfake"—AI-generated content that simulates the actions or speech of a real person—targeting a candidate was released online during the official media quiet period just before the country's September 2023 elections, preventing fact checkers and journalists from debunking the hoax in real time;[5]
- To polish and "launder" reputations, such as that of Indonesia's Prabowo Subianto, whose campaign used gen AI to portray the former special forces commander and alleged human rights abuser as a "cuddly grandfather;"[6]
- To promote and prioritize narratives that exacerbate societal divides, such as during India's election campaign, which saw gen AI-enhanced information campaigns used to inflame tensions with the country's Muslim minority;[7] and
- To analyze massive quantities of data about internet users, which can enable powerful, microtargeted information campaigns at the individual level. The addition of gen AI to the authoritarian toolbox may make similar efforts more convincing and potent.[8]

Information manipulation by authoritarians—regardless of whether gen AI tools are employed—has multiple aims. In the short to medium term, it is meant to advance specific narratives that bolster public perception of antidemocratic actors and harm perceptions of prodemocracy candidates, civil society activists, journalists, and others who have the power to hold authoritarians and their allies accountable. In the long term, beyond any campaign period, information manipulation is intended to undermine the concept of knowable truth and a shared reality, damaging societal trust and attacking the foundations of democratic systems.[9] The addition of gen AI tools to the authoritarian toolbox is making these efforts easier, less costly, and possibly more effective.[10]

In the long term, beyond any campaign period, information manipulation is intended to undermine the concept of knowable truth and a shared reality, damaging societal trust and attacking the foundations of democratic systems.

One component of the long-term attack on truth is a progressive divergence in how a society evaluates the validity of information online: the debut of generative AI raises the bar for what is considered real and legitimate information, and lowers the bar precipitously for what is considered likely to be false. This effect is called the "liar's dividend"[11] and suggests that as the public's awareness of gen AI information manipulation increases and people become more accustomed to AI-generated media, malicious actors can dismiss embarrassing truths as entirely false more easily.[12] Some experts have warned that the dissolution of societal trust may be gen AI's most powerful, long-term impact on democracy.[13]

These concerns take on even greater significance in the context of authoritarian actors' ongoing efforts to undermine democracy at a global level.[14] Current detection and attribution efforts indicate that authorities in Iran,[15] Venezuela,[16] Russia, and China are using gen AI to advance information manipulation campaigns in backsliding democracies worldwide. Research by the NATO Strategic Communications Centre of Excellence identified a fully automated group of 130 Telegram channels and thirteen websites, all in Russian, leveraging generative AI to "generate arbitrary noise" and repost political news. The low-cost network, which was involved in influence operations in Ukraine, reached approximately fifty-thousand individuals monthly.[17] In addition, a recent report by OpenAI, which runs the AI chatbot ChatGPT, demonstrated how actors linked to Russia, Iran, and China engage in "covert influence operations that sought to use AI in support of their activity," using ChatGPT to gather and analyze data about potential targets for information manipulation to code manipulative websites and produce manipulative images.[18]

Gen AI tools offer authoritarian actors and their allies an opportunity to reduce the required cost and technical capability to create and distribute high-fidelity, false content—which is difficult to distinguish from genuine content—quickly and at scale. In fact, such tools may eliminate or drastically reduce the tradeoffs that were previously inherent in the production of high-fidelity manipulative content, which has traditionally taken comparatively more time and effort to produce but which may be more convincing to individuals. One example is in translation between Romance and Germanic languages and tonal languages such as Mandarin Chinese, which Gen AI does much better than earlier AI models.

Experts have identified the use of gen AI tools around nearly every national election since mid-2023, from Bangladesh to Argentina and South Africa,[19] and are attempting to assess whether the new technology is creating a *differential* impact for authoritarians both in an election context and for democracy more broadly. What is clear, however, is that authoritarians are experimenting with these tools and incorporating them into their strategies to undermine democracy. These efforts simultaneously weaken public trust in online content, society, and democracy itself.

The following sections describe three specific ways in which authoritarians are using gen AI to enhance their efforts to manipulate the information space, and outline some of the potential pitfalls for the democratic response.

Gen AI tools offer authoritarian actors and their allies an opportunity to reduce the required cost and technical capability to create and distribute high-fidelity, false content quickly and at scale.

LESS EXPENSIVE AND EASIER PRODUCTION OF HIGH-FIDELITY SYNTHETIC CONTENT

With a few strokes of the keyboard, gen AI tools can produce convincing "synthetic" content. In contrast to traditional methods of information manipulation commonly used by authoritarian actors, gen AI leverages patterns in data drawn from millions of social media posts, online articles, images, and videos to create entirely new content, or "synthetic media," based on as little as a simple text prompt. Both the manipulation of existing content such as so-called "deepfakes" and the wholesale creation of content through gen AI fall within the concept of synthetic media.[20]

The emergence of this easily produced synthetic content may exacerbate the vulnerabilities of open, democratic information environments. Malign actors increasingly leverage this technology to produce high-fidelity, photo- or audio-realistic content that is challenging to identify as false with the human eye, further stoking the public's uncertainty about the veracity of online media. This new capacity is being deployed across authoritarians' existing and ever-expanding networks of online trolls and automated "bot" accounts, which are already promoting AI-generated content to advance narratives that serve the interests of authoritarian powers and their allies.[21]

In a recent example of likely AI-generated synthetic content that was intended to influence public opinion, actors associated with the People's Republic of China (PRC) attempted to manipulate voters during Taiwan's January 2024 elections by massively amplifying false narratives on popular Taiwanese social media and messaging platforms. These operations aimed to weaken support for Taiwanese independence and, more narrowly, to undermine support for candidates from the Democratic Progressive Party (DPP), which takes a harder line against Beijing's influence in Taiwan.

Researchers found at least two PRC-affiliated threat actors attempting to interfere in Taiwan's elections. One was connected to the PRC's extensive network of fake social media profiles,[22] referred to as Spamouflage or Dragonbridge.[23] During the election campaign, this network focused on promoting a book—suspected to have been written using gen AI—that amplified false claims of adultery and immoral behavior among DPP leaders. The book was used as a script for gen AI–produced social media videos advancing its core assertions and narrative, which were promoted at scale.[24] While the DPP retained the presidency, it lost its outright majority in Taiwan's legislature, and several malign narratives that were strengthened amid the campaign may persist.

Although it is difficult to identify a causal link between the election outcome and the use of gen AI to influence Taiwanese voters, this case shows how one global authoritarian power is already using gen AI–produced content to augment their interference in democratic processes. Elsewhere, in Pakistan, the political party of former prime minister Imran Khan, Pakistan Tehreek-e-Insaf (PTI)—who along with his party had been banned from running for parliament—used generative AI to "deepfake" Khan delivering a speech, rallying his supporters from his prison cell.[25] In Pakistan's recent general election, the PTI won a plurality of votes.

Whether this development proves differential or decisive has yet to be determined, partly because the sample size of elections that have featured such manipulation remains small. That said, the growing public discourse about synthetic media campaigns as well as the rising skepticism toward online media—the liar's dividend—fueled by the increasingly difficult challenge of distinguishing between real and synthetic material, seems likely to further erode public trust in elections and democratic institutions.

Generative AI "deepfake" depiciton of Imran Khan delivering a televised speech.

AUTOMATING THE PRODUCTION OF SYNTHETIC MEDIA

In addition to enhancing of the quality of synthetic content, gen AI technologies may facilitate an increase in quantity, reducing the cost and effort required for automated production and enabling information manipulation campaigns to be produced at unprecedented speed and scale.[26] As a result, democracies may face deluges of rapidly-deployed and exponentially-expanding deceptive content that drowns out genuine, high-quality information, as exemplified by the PRC's information campaign around Taiwan's elections.

Russian information campaigns have worked to amplify the Kremlin's perspective and within minutes spin high-profile political events to their advantage,[27] but such major campaigns are not commonplace, likely due to the complexity of coordinating hundreds or thousands of accounts, channels, and platforms, which is necessary to elevate a topic significantly in public discourse. Yet, gen AI tools remove some of the biggest barriers to scaling content generation and distribution, as they allow the automation of key technical processes of content production with far less need for human engagement or oversight than earlier approaches to manipulative efforts.

For example, digital investigators from the German Federal Foreign Office uncovered a Russian-backed information manipulation network specializing in "doppelgänger media outlets"—which trick users into believing false information by spoofing the look of real, trusted news sites—and social media accounts on X.[28] Between December 20, 2023, and January 20, 2024, they identified over fifty-thousand fake user accounts that appeared to have coordinated propaganda in German, including more than a million German-language posts. On some days,

Russian-Backed Information Manipulation Networks

These networks trick users into believing false information by spoofing the look of real, trusted news sites and social media accounts on X. Digital investigators from the German Federal Foreign Office identified:

50K+
Fake user accounts

1M+
German-language posts

200K
Per day

2
Per second

(Between December 20, 2023, and January 20, 2024)

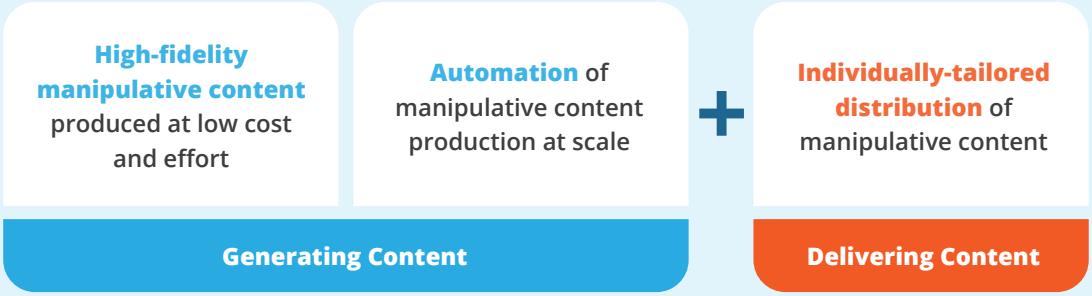
the experts at the Foreign Office registered 200,000 of these short messages, which equates to about two messages per second—a digital barrage designed to manipulate the public en masse. Given the speed and volume of the posts and accounts involved, German authorities suspected that gen AI tools were used for account creation, content production, and dissemination. While the ultimate impact of the effort is difficult to ascertain, it would probably not have been possible without gen AI.

Gen AI's capacity to automate content production and dissemination at scale may change the character of authoritarian regimes' social media manipulation and online influence efforts in noticeable ways. For instance, experts are already observing new behaviors from bot networks,[29] including the ability to create distinct "personalities" that develop over time.

INDIVIDUALLY-TAILORED CONTENT DISTRIBUTION

Gen AI has the potential to simplify the production and delivery of false content that is highly tailored for specific audiences. By leveraging this technology, authoritarian actors may generate content aimed not at bucketed profiles, which remain broad and thus less convincing, but at individual users, based on thousands of data points scraped from social media and other public sources of information. Such content is more likely to play effectively to personal emotions, biases, experiences, and relationships that make people susceptible to manipulative information campaigns,[30] though early evidence is split on whether gen AI–produced content is more convincing than similarly targeted authentic content in this regard.[31] Authoritarian regimes may exploit this capability to advance their agendas, using gen AI to craft and spread messages that aim to bolster their legitimacy, discredit opponents, or manipulate public opinion about key issues related to the practice of democracy.

Generative AI Capabilities Leveraged by Authoritarian Actors



According to a recent RAND report, researchers in China are working on a gen AI–based system for "precision cognitive attacks," which would use highly tailored information operations to target individuals or small groups. Such an approach would likely employ "ChatGPT's powerful data processing capabilities and high autonomy to enable it to conduct preference analysis and subsequent related information production and information delivery, supporting 'precision cognitive attacks' based on 'personalized user portraits.'"[32] While the deployment of these systems is currently speculative, it takes little imagination to conceive how they might work—particularly since the revelations about Cambridge Analytica's efforts to gather data on Facebook users for targeted political messaging. This approach could be particularly useful for interfering in subnational elections, which are singled out for manipulation less frequently than national contests due to their greater number and the diversity of relevant campaign issues at the subnational level.

BUILDING AWARENESS TO THE EVOLVING CHALLENGE

The emergence of gen AI as a tool for information manipulation by authoritarians may not fundamentally change the nature of democratic responses to such efforts. For example, it is still critical that civil society organizations take the lead in identifying and combating malign information campaigns, whether or not they enjoy the support of their governments.

Still, building awareness about the capabilities of gen AI tools may be helpful for the public at-large, as such an approach can essentially serve to "pre-bunk" the new capabilities that these tools offer authoritarians and prepare citizens for what they might see around elections or at other critical moments of public discourse outside of an election period. At the same time, there is a risk that overemphasizing the power and ubiquity of gen AI would strengthen the liar's dividend. In other words, the more people are warned about the threat of highly deceptive content, the more they may come to distrust even authentic media.[33]

This danger makes it especially important for civil society organizations, technology companies, and governments to advance their efforts to determine whether content is authentic and to inform the public about the motives and interests of the authoritarian actors engaged in AI-backed manipulation campaigns. If citizens understand who is pushing them and in what direction, they will be better equipped to resist this pressure and potentially to push back against it themselves.

Building awareness about the capabilities of gen AI tools may be helpful for the public at-large, as such an approach can essentially serve to "pre-bunk" the new capabilities that these tools offer authoritarians.

LEVERAGING GENERATIVE AI TO SUPPORT INFORMATION SPACE INTEGRITY

As authoritarians use gen AI to amplify their efforts to manipulate the public, democracy activists may have an opportunity to employ the same technology to support the integrity of the information space, especially to promote democratic discourse during election campaigns. Fact checkers, independent journalists, narrative researchers, nongovernmental organizations, and media monitors—often operating in coalitions during elections to compete with well-funded authoritarian information campaigns—are now exploring how gen AI systems might help them move faster and more effectively to counter information manipulation. Some experts have suggested that the potential advantages afforded by such technologies to democracy advocates may be even greater than those of authoritarians.[34] It is critical to note that in the absence of broadly adopted ethical standards in the use of gen AI systems, prodemocracy actors must be transparent about their use, and should include human review throughout any project to ensure accuracy.

GENERATIVE AI FOR FACT CHECKING

Fact checkers are employing gen AI to expedite the verification of information. The technology also enhances their ability to identify patterns in large volumes of text, images, and videos, and to detect trends in behavior over time, which is

useful in predictive analysis of authoritarian efforts to manipulate information. In addition, a growing number of fact checkers are building fact-checking chatbots that operate in closed settings such as messaging apps, allowing users to forward messages and links from their own chats and quickly verify the veracity of such content in response. While many of these systems already used traditional AI capabilities to understand and interpret fact-checking requests, the addition of gen AI is accelerating this trend—which requires significant human effort and review—by hastening the collection of basic information surrounding the questionable content or underlying narrative.

For example, Gwara Media in Kharkiv, Ukraine, has developed a chatbot called Perevirka,[35] which specializes in debunking Kremlin messaging online. Users send in a text, photo, video, or link that they want Gwara to check, and they receive a response immediately if the item has already been investigated. If not, Gwara's team of human fact checkers reviews the request, uses a gen AI–powered chatbot to quickly conduct initial research and gather contextual information about the issue in question, and produces a new fact-check in a timely manner. Cofacts, a Taiwanese civic tech community that works to counter information manipulation and which played an important role during the 2024 Taiwanese elections, is similarly experimenting with the use of gen AI for their chatbot-based fact-checking system.[36] The system operates primarily on closed-door messaging apps such as Facebook Messenger and Line, and uses gen AI to provide more substantive responses to user submissions, such as suggestions about how to verify the specific content in review, as well as basic arguments against identified manipulative narratives.

While these country-level efforts are promising, important work is also being done at the international level. The global nonprofit Meedan has developed chatbot software that uses traditional AI to group similar content (images, videos, and text) together and match them to fact-checks, and is experimenting with using Gen AI to generate predictable variants of fact-checked claims, group content by narrative, and improve communication with users.[37] This tool is "white labeled" and can be adopted by other organizations under their own branding and banner.

Fact checkers, independent journalists, narrative researchers, nongovernmental organizations, and media monitors are now exploring how gen AI systems might help them move faster and more effectively to counter information manipulation.

INDEPENDENT JOURNALISM IN THE AGE OF GENERATIVE AI

Journalists are experimenting with gen AI in a number of work areas, including to create efficiencies in the journalistic process, customize currently-existing online news products more effectively, differentiate their work in a more pronounced manner, hasten the editorial process, help citizens better understand complex and voluminous topics, analyze large quantities of data, translate languages, and for other, more routine tasks.[38] For example, in Zimbabwe, the Center for Innovation and Technology (CITE) is using gen AI to increase its capacity, for instance by prompting the systems to suggest

CITE's AI news presenter, Alice, reading the headlines.

headlines and write article summaries, while also debuting an AI newsreader which has garnered significant attention in Zimbabwe.[39]

Private companies are also playing a prominent role in supporting journalists' experimentation with gen AI. For instance, Microsoft has recently launched collaborations with news organizations to integrate gen AI into journalism, seeking to innovate and create financially sustainable newsrooms.[40] Also, Google has instituted a program for a handful of independent publishers, providing them with beta access to an unreleased gen AI platform that facilitates the creation of news content by summarizing key background information.[41] These initiatives aim to inform, lead, and scale AI solutions in journalism, supporting viable business models and audience growth while attempting to maintain ethical standards. It is important to note that there are many unresolved questions about which forms of gen AI usage should be considered acceptable in the journalistic sphere.[42]

Investigative journalists who rely on open-source intelligence methods are conducting their own experiments with gen AI tools and techniques. Bellingcat, an investigative journalism collective, has shown how gen AI can automate and speed up the collection and analysis of vast datasets, and how gen AI chatbots can be used for geolocation,[43] enabling journalists to uncover patterns in authoritarian messaging that might be missed by manual analysis. Gen AI can also help in verifying facts and cross-referencing information quickly, which is crucial in investigative journalism.[44]

GENERATIVE AI AND THE DETECTION OF INFORMATION MANIPULATION

Civil society organizations and their allies are using gen AI to detect information manipulation campaigns, as these systems enhance their ability to identify false or inauthentic content efficiently and hasten the response times of fact checkers and journalists working to push back on malign narratives and circulate high-quality information. At the same time, a new class of detection tools is under development to detect gen AI-produced content, whether audio, visual, or text-based, with differing levels of success for each modality.

Civil society organizations and their allies are using gen AI to detect information manipulation campaigns.

Detecting information manipulation

Gen AI can accelerate the efforts of research organizations to spot information manipulation online by helping to identify the behaviors, narratives, and tactics authoritarian state actors use. For example, GLOBSEC, a global think tank based in Slovakia, has used the gen AI-powered monitoring tool Gerulata Juno to analyze Russian influence in Slovakia's information space.[45]

Several private firms that monitor the information space for authoritarian information manipulation are also using gen AI to strengthen their efforts. LetsData is based in Ukraine and operates globally, processing data from thousands of websites and social media platforms to look for early evidence of Russian state-backed information operations. Their system automatically alerts relevant partners and government actors to enable rapid responses. LetsData uses GenAI to enhance its cross-platform and cross-language tracking capabilities. Previously, retraining detection algorithms in multiple languages for various platforms took weeks. With GenAI, LetsData can adapt to track new malign narratives across various countries and platforms without retraining the system each time.

Detecting gen AI–produced content

Some organizations are using gen AI technology to help the public identify and respond to deceptive gen AI–produced content. For example, Democracy Reporting International uses gen AI to train models to detect information manipulation online through a tool called Disinfo Radar, which helps other civil society organizations prepare for, identify, and respond to manipulative campaigns.[46]

Numerous private-sector organizations are working together in an initiative called the Coalition for Content Provenance and Authenticity (C2PA). The coalition, which includes established companies like Adobe, Google, and Microsoft as well as newer players such as Truepic, uses technical means to identify gen AI–produced content that various platforms can then label. Labeling of malign content has been shown in some contexts to reduce people's belief in and sharing of that content.[47]

Civil society organizations, however, face a few significant barriers when attempting to use gen AI for detecting authoritarian information operations. First, many organizations lack the technical expertise needed to implement and manage advanced AI detection models, and it can be prohibitively costly to hire skilled data scientists, for whom the private sector is also competing. Second, most detection models are developed by academics or private companies and are not readily accessible to technologically under-resourced nonprofit organizations. Finally, detection of gen AI–produced content is a developing area, and no tool is 100 percent reliable, leaving civil society without a simple, trusted, one-stop shop or service to address a clear and urgent need.

Many organizations lack the technical expertise needed to implement and manage advanced AI detection models, and it can be prohibitively costly to hire skilled data scientists, for whom the private sector is also competing.

LOOKING TO THE FUTURE

While the ultimate impact of gen AI technologies is not yet clear, authoritarian actors are already using them in ways that present a serious challenge to democracy during election cycles and beyond. By facilitating the creation at scale of high-fidelity synthetic media—which can diminish audiences' ability to discern what is inauthentic from that which is real—gen AI tools are directly serving authoritarians' long-term interest in undermining societal trust and the foundations of democratic governance. Future analysis and research might focus on the ways this evolving technology affects both elections and public discourse outside of election periods.

At the same time, democracy advocates have a critical opportunity to leverage these technologies for positive applications. Many civil society organizations, fact checkers, and journalists have begun to harness the power of gen AI to resist authoritarian manipulation. As efforts of authoritarian powers to manipulate the information space continue to evolve, it will be critical for civil society to experiment with gen AI tools, which may prove an important accelerator and amplifier for securing democracy through the information space in the long run.

ENDNOTES

1 Rhiannon Williams, "Humans May Be More Likely to Believe Disinformation Generated by AI," *MIT Technology Review*, 28 June 2023, www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation-generated-by-ai/.

2 Josh A. Goldstein, Girish Sastry, Micah Musser, Renée DiResta, Matthew Gentzel, and Katerina Sedova, "Forecasting Potential Misuses of Language Models for Disinformation Campaigns—and How to Reduce Risk," Center for Security and Emerging Technology, January 2023, https://cset.georgetown.edu/article/forecasting-potential-misuses-of-language-models-for-disinformation-campaigns-and-how-to-reduce-risk/.

3 Benj Edwards, "Microsoft's VASA-1 Can Deepfake a Person with One Photo and One Audio Track," *Ars Technica,* 19 April 2024, https://arstechnica.com/information-technology/2024/04/microsofts-vasa-1-can-deepfake-a-person-with-one-photo-and-one-audio-track/.

4 Tate Ryan-Mosley, "How Generative AI Is Boosting the Spread of Disinformation and Propaganda," *MIT Technology Review,* 4 October 2023, www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/.

5 Morgan Meaker, "Slovakia's Election Deepfakes Show AI Is a Danger to Democracy," *Wired,* 3 October 2023, www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/.

6 Kate Lamb, Fanny Potkin, and Ananda Teresia, "Generative AI May Change Elections This Year. Indonesia Shows How," Reuters, 8 February 2024, www.reuters.com/technology/generative-ai-faces-major-test-indonesia-holds-largest-election-since-boom-2024-02-08/.

7 "As India Election Underway, Meta Approves Series of Violent, Inflammatory, Islamophobic AI-Generated Ads Targeting Voters," Ekō and India Civil Watch International (ICWI), May 2024, https://aks3.eko.org/pdf/Meta_AI_ads_investigation.pdf.

8 Almog Simchon, Matthew Edwards, and Stephan Lewandowsky, "The Persuasive Effects of Political Microtargeting in the Age of Generative AI," PsyArXiv Preprints, 28 January 2024, https://osf.io/preprints/psyarxiv/62kxq.

9 Nobel Prize, "Nobel Prize Lecture: Maria Ressa, Nobel Peace Prize 2021," YouTube, 17 December 2021, www.youtube.com/watch?v=NsWVb2AUl5Y&t=1s.

10 Simchon, Edwards, and Lewandowsky, "The Persuasive Effects of Political Microtargeting in the Age of Generative AI."

11 Lena-Maria Böswald, Beatriz Almeida Saab, and Jan Nicola Beyer, *What a Pixel Can Tell: Text-to-Image Generation and Its Disinformation Potential*, Democracy Reporting International, September 2022, https://democracyreporting.s3.eu-central-1.amazonaws.com/images/6331fc834bcd1.pdf.

12 Josh A. Goldstein and Andrew Lohn, "Deepfakes, Elections, and Shrinking the Liar's Dividend," Brennan Center for Justice, 23 January 2024, www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend.

13 John Ternovski, Joshua Kalla, and P. M. Aronow, "The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments," *Journal of Online Trust & Safety* 1, no. 2 (28 February 2022), https://tsjournal.org/index.php/jots/article/view/28.

14 Josh A. Goldstein, Girish Sastry, Micah Musser, Renée DiResta, Matthew Gentzel, and Katerina Sedova, "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations," arXiv, 10 January 2023, https://arxiv.org/pdf/2301.04246.pdf.

15 Microsoft Threat Intelligence, "Staying ahead of threat actors in the age of AI," 14 February 2024, www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

16 Iria Puyosa and Mariví Marín Vázquez, *Deepening the Response to Authoritarian Information Operations in Latin America,* National Endowment for Democracy, November 2023, www.ned.org/wp-content/uploads/2023/12/NED_FORUM-Deepening-Response-Latin-America.pdf.pdf.

17 "Hijacking Reality: The Increased Role of Generative AI in Russian Propaganda," NATO Strategic Communications Centre of Excellence, *Virtual Manipulation* 1 (June 2024), https://stratcomcoe.org/pdfjs/?file=/publications/download/VM_210x2975_FINAL_DIGITAL_PDF.pdf.

18 Ben Nimmo, "AI and Covert Influence Operations: Latest Trends," OpenAI, May 2024, https://downloads.ctfassets.net/kftzwdyauwt9/5lMxzTmUclSOAcWUXbkVrK/3cfab518e6b10789ab8843bcca18b633/Threat_Intel_Report.pdf.

19 For more information, please consult: "2024 AI Elections Tracker," Rest of World, https://restofworld.org/2024/elections-ai-tracker/.

20 "Building a Glossary for Synthetic Media Transparency Methods, Part 1: Indirect Disclosure," Partnership on AI, 19 December 2023, https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure/#Glossary.

21 Allie Funk, Adrian Shahbaz, and Kian Vesteinsson, "Freedom on the Net 2023: The Repressive Power of Artificial Intelligence," Freedom House, 3 October 2023, https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence#the-repressive-power-of-artificial-intelligence.

22 Donie O'Sullivan, Curt Devine, and Allison Gordon, "China Is Using the World's Largest Known Online Disinformation Operation to Harass Americans, a CNN Review Finds," CNN, 13 November 2023, https://edition.cnn.com/2023/11/13/us/china-online-disinformation-invs/index.html.

23 Albert Zhang, "As Taiwan Voted, Beijing Spammed AI Avatars, Faked Paternity Tests, and 'Leaked' Documents," *the Strategist*, Australian Strategic Policy Institute, 18 January 2024, www.aspistrategist.org.au/as-taiwan-voted-beijing-spammed-ai-avatars-faked-paternity-tests-and-leaked-fake-documents/.

24 Tsai Yung-yao and Jonathan Chin, "China Is Posting Fake Videos of President: Sources," *Taipei Times,* 11 January 2024, www.taipeitimes.com/News/front/archives/2024/01/11/2003811930.

25 Siladitya Ray, "Imran Khan—Pakistan's Jailed Ex-Leader—Uses AI Deepfake To Address Online Election Rally," *Forbes,* 18 December 2023, www.forbes.com/sites/siladityaray/2023/12/18/imran-khan-pakistans-jailed-ex-leader-uses-ai-deepfake-to-address-online-election-rally/?sh=47d5fbbb5903.

26 Jan Nicola Beyer and Lena-Maria Böswald, *On the Radar: Mapping the Tools, Tactics, and Narratives of Tomorrow's Disinformation Environment,* Democracy Reporting International, June 2022, https://democracyreporting.s3.eu-central-1.amazonaws.com/images/62c8333ec3aea.pdf.

27 Givi Gigitashvili, "Russian-Language Telegram Channels Foment Tensions in Georgia," DFRLab, 1 May 2024, https://dfrlab.org/2024/05/01/russian-language-telegram-channels-foment-tensions-in-georgia/.

28 Marcel Rosenbach and Christoph Schult, "Baerbocks Digitaldetektive decken russische Lügenkampagne auf" [Baerbock's digital detectives uncover Russian lying campaign], *Spiegel,* 26 January 2024, https:/www.spiegel.de/politik/deutschland/desinformation-aus-russland-auswaertiges-amt-deckt-pro-russische-kampagne-auf-a-765bb30e-8f76-4606-b7ab-8fb9287a6948?giftToken=956f67a3-3158-41e2-ba69-8b67d74da941.

29 William Marcellino, et al., *The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0: Next-Generation Chinese Astroturfing and Coping with Ubiquitous AI,* RAND Corporation, September 2023, www.rand.org/content/dam/rand/pubs/perspectives/PEA2600/PEA2679-1/RAND_PEA2679-1.pdf.

30 Thor Benson, "This Disinformation Is Just for You," *Wired,* 1 August 2023, www.wired.com/story/generative-ai-custom-disinformation/.

31 Amirsiavosh Bashardoust, Stefan Feurriegel, and Yash Raj Shrestha, "Comparing the Willingness to Share for Human-Generated vs. AI-Generated Fake News," arXiv, 12 February 2024, https://arxiv.org/pdf/2402.07395; and Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani, "AI Model GPT-3 (Dis)informs Us Better than Humans," *Science Advances* 9, no. 26, 28 June 2023, www.science.org/doi/10.1126/sciadv.adh1850.

32 Nathan Beauchamp-Mustafaga, "Exploring the Implications of Generative AI for Chinese Military Cyber-Enabled Influence Operations," testimony before the US-China Economic and Security Review Commission, RAND Corporation, 1 February 2024, www.rand.org/content/dam/rand/pubs/testimonies/CTA3100/CTA3191-1/RAND_CTA3191-1.pdf.

33 Josh A. Goldstein and Andrew Lohn, "Deepfakes, Elections, and Shrinking the Liar's Dividend."

34 "Lessons from Ukraine: How AI Is Accelerating the Response to Authoritarian Information Manipulation," *Power 3.0,* 21 February 2024, www.power3point0.org/2024/02/21/lessons-from-ukraine-how-ai-is-accelerating-the-response-to-authoritarian-information-manipulation/.

35 For more information, please see "'PEREVIRKA' Is a Bot That Detects Fake News," Gwara Media, https://gwaramedia.com/en/perevirka-do-it-together/.

36 Albert Zhang, "As Taiwan Voted, Beijing Spammed AI Avatars, Faked Paternity Tests, and 'Leaked' Documents."

37 For more information, please visit Meedan's website: https://meedan.com/.

38 David Caswell, "AI and Journalism: What's Next?" Reuters Institute for the Study of Journalism, 19 September 2023, https://reutersinstitute.politics.ox.ac.uk/news/ai-and-journalism-whats-next.

39 Alice, the AI Presenter, "US Flags Rampant Rights Abuses in Zimbabwe," CITE, 23 April 2024, https://cite.org.zw/us-flags-rampant-rights-abuses-in-zimbabwe/.

40 Noreen Gillespie, "Here's How We're Working with Journalists to Create the Newsrooms of the Future with AI," *Microsoft on the Issues* (blog), Microsoft, 5 February 2024, https://blogs.microsoft.com/on-the-issues/2024/02/05/journalism-news-generative-ai-democracy-forward/.

41 Mark Stenberg, "Google Is Paying Publishers to Test an Unreleased Gen AI Platform," *Adweek,* 27 February 2024, www.adweek.com/media/google-paying-publishers-unreleased-gen-ai/.

42 Shirin Anlen and Raquel Vazquez Llorente, "Using Generative AI for Human Rights Advocacy," Witness, 28 June 2023, https://blog.witness.org/2023/06/using-generative-ai-for-human-rights-advocacy/.

43 Dennis Kovtun, "Can AI Chatbots Be Used for Geolocation?," Bellingcat, 14 July 2023, www.bellingcat.com/resources/2023/07/14/can-ai-chatbots-be-used-for-geolocation/.

44 Craig Silverman, "Key Tools and Approaches for Using AI in OSINT and Investigations," *Digital Investigations* (blog), 6 March 2024, https://digitalinvestigations.substack.com/p/key-tools-and-approaches-for-using.

45 Anushka Kaushik, "The War on Ukraine: A Look at (Underemphasized) Russian Cyber Operations," GLOBSEC, 10 February 2023, www.globsec.org/sites/default/files/2023-07/Cyber%20Brief%20Russian%20Cyber%20Operations_0.pdf.

46 For more information, please see: "Disinfo Radar," https://disinforadar.com/tools/.

47 Chloe Wittenberg, Ziv Epstein, Adam J. Berinsky, and David G. Rand, "Labeling AI-Generated Content: Promises, Perils, and Future Directions," Massachusetts Institute of Technology, 27 March 2024, https://mit-genai.pubpub.org/pub/hu71se89/release/1.

ABOUT THE AUTHOR

Beatriz Saab is a Digital Democracy Research Officer at Democracy Reporting International, a nonprofit organization based in Berlin, Germany. She is involved in projects that monitor digital platforms, identify information manipulation and harmful speech, and examine the impact of artificial intelligence on democratic discourse online. In addition, she participates in advocacy groups that ensure the effective regulation of digital platforms in the European Union. She earned a Master's in Public Policy from the Hertie School in Berlin. Follow her on X: @bea_saab.

ACKNOWLEDGMENTS

The author appreciates the contributions of the International Forum's staff and leadership, including Christopher Walker, John K. Glenn, Kevin Sheives, John Engelken, Amaris Rancy, and Maya Recanati, all of whom played important roles in the editing and publication of this report. The author also wishes to thank Péter Krekó, Chris Beall, Sam Gregory, Theo Dolan, and Beth Kerley for lending their expertise and knowledge to further sharpen and refined the analysis. The author is grateful for the support offered by representatives from Maldita.es, IREX, Zinc Network's Open Information Partnership, and IRI's Beacon Project for their invaluable insight. Particular acknowledgment goes to Adam Fivenson whose support and vision for this project were vital to its completion. Special thanks are also due to Tyler Roylance for his careful and expert copyediting of the text. Finally, the Forum wishes to acknowledge Factor3 Digital for their efforts and invaluable support in designing this report for publication.

PHOTO CREDITS

Cover image: Please see credit on page 3.

Page 3: Photo by Hsu Tsun-Hsu/Getty Images (This image shows protesters holding placards with messages that read "reject red media" and "safeguard the nation's democracy" during a 2019 rally against pro-China media in Taiwan).

Page 5: Photo by Olivier Douliery/Contributor/Getty Images

Page 9: Photo by YouTube video screenshot from Imran Khan's official YouTube channel.

Page 13: Photo by Bay Ismoyo/AFP/Getty Images

Page 15: Photo by YouTube video screenshot from CITE's official YouTube channel.

Page 18: Photo by picture alliance/Contributor/Getty Images

FORUM | INTERNATIONAL FORUM FOR DEMOCRATIC STUDIES

The International Forum for Democratic Studies at the National Endowment for Democracy (NED) is a leading center for analysis and discussion of the theory and practice of democracy around the world. The Forum complements NED's core mission—assisting civil society groups abroad in their efforts to foster and strengthen democracy—by linking the academic community with activists from across the globe. Through its multifaceted activities, the Forum responds to challenges facing countries around the world by analyzing opportunities for democratic transition, reform, and consolidation. The Forum pursues its goals through several interrelated initiatives: publishing the *Journal of Democracy*, the world's leading publication on the theory and practice of democracy; hosting fellowship programs for international democracy activists, journalists, and scholars; coordinating a global network of think tanks; and undertaking a diverse range of analytical initiatives to explore critical themes relating to democratic development.

NED | NATIONAL ENDOWMENT FOR DEMOCRACY

SUPPORTING FREEDOM AROUND THE WORLD

The National Endowment for Democracy (NED) is a private, nonprofit foundation dedicated to the growth and strengthening of democratic institutions around the world. Each year, NED makes more than 1,700 grants to support the projects of nongovernmental groups abroad who are working for democratic goals in more than 90 countries. Since its founding in 1983, the Endowment has remained on the leading edge of democratic struggles everywhere, while evolving into a multifaceted institution that is a hub of activity, resources, and intellectual exchange for activists, practitioners, and scholars of democracy the world over.

1201 Pennsylvania Avenue, NW
Suite 1100
Washington, DC 20004
(202) 378-9700
ned.org

@thinkdemocracy

ThinkDemocracy

International Forum for Democratic Studies