Centre for International
Governance Innovation

This policy brief is part of CIGI's project on freedom of thought: Legitimate Influence or Unlawful Manipulation?

Find out more at: www.cigionline.org/fot

Policy Brief No. 12 — February 2025

Generative AI, Democracy and Human Rights

David Evan Harris and Aaron Shull

Key Points

- Disinformation campaigns aimed at undermining electoral integrity are expected to play an ever larger role in elections due to the increased availability of generative artificial intelligence (AI) tools that can produce high-quality synthetic text, audio, images and videos and their potential for targeted personalization.
- As these campaigns become more sophisticated and manipulative, the foreseeable consequence is further erosion of trust in institutions and heightened disintegration of civic integrity, jeopardizing a host of human rights, including electoral rights and the right to freedom of thought.
- These developments are occurring at a time when the companies that create the fabric of digital society should be investing heavily in, but instead are dismantling, the "integrity" or "trust and safety" teams that counter these threats.
- Policy makers must hold AI companies liable for the harms caused or facilitated by their products that could have been reasonably foreseen. They should act quickly to ban using AI to impersonate real people or organizations, and require the use of watermarking or other provenance tools to allow people to differentiate between AI-generated and authentic content.

Introduction

The 2024 US presidential election gave an unexpectedly decisive win to the Republican candidate, Donald Trump. However, throughout the election season, beginning in the primaries (Nehamas 2023; Swenson and Weissert 2024; Powell 2024) and running through the final days of the campaign, US intelligence agencies (Thrush 2024) and large tech companies (AI Elections Accord 2024) warned that foreign actors were attacking the election online and weaponizing AI technologies in new ways. The election, and others around the world in 2024 (European Commission 2024) and 2023 (Meaker 2023), have been subject to disinformation campaigns and influence operations unlike any others in human history (Allyn 2024). Disinformation is not new, and neither are influence operations run by foreign states. But one thing has changed since the 2016 electoral gauntlet: the rise of generative AI tools.

Large language models, most commonly known for being the technology underlying the ChatGPT chatbot, have stormed to the forefront of public consciousness. OpenAI's ChatGPT was the first chatbot to reach large-scale public adoption. According to OpenAI, ChatGPT reached its first one million users in five days when it launched in November 2022. This same milestone took ChatGPT's contemporaries, Instagram and Netflix, 2.5 months and 3.5 years, respectively (Hu 2023).

This success has not gone unnoticed and today there is a plethora of direct competitors to ChatGPT. Google

About the Authors

David Evan Harris is a CIGI senior fellow, a Chancellor's Public Scholar at the University of California, Berkeley, and a faculty member at the Haas School of Business, where he teaches courses on artificial intelligence (AI) ethics, social movements and social media, civic technology, futures thinking and scenario planning. David is also a senior research fellow at the International Computer Science Institute, a senior policy adviser at the California Initiative for Technology and Democracy, and a senior advisor for AI and Elections to the Brennan Center for Justice at NYU. He was named to Business Insider's AI 100 list in 2023.

David previously worked as a research manager at Meta (formerly Facebook) on the responsible AI and civic integrity teams. His writings and commentary have been featured by *The Wall Street Journal, The Washington Post,* CNN, *The Guardian, BBC*, Associated Press, *Bloomberg*, *IEEE Spectrum, The Atlantic*, *Tech Policy Press* and *Adbusters*. David has advised the White House, the US Congress, the European Union, the United Nations, the North Atlantic Treaty Organization and the California State Legislature about technology policy. He has conducted research, studied or given talks in person in 37 countries and speaks fluent English, Portuguese and Spanish, as well as intermediate French.

Aaron Shull is the managing director and general counsel at CIGI. He is a senior legal executive and is recognized as a leading expert on complex issues at the intersection of public policy, emerging technology, cybersecurity, privacy and data protection.

Aaron has extensive experience building global networks of experts to enhance engagement with researchers and practitioners drawn from government, academia, industry and civil society. He recently concluded the project Reimagining a Canadian National Security Strategy, which was unprecedented in scale and scope in Canada. It engaged a multidisciplinary network of more than 250 experts to inspire updated and innovative national security and intelligence practices and offered a series of key policy recommendations to assist the Government of Canada in addressing the challenges of a new security environment.

This work was well received by senior officials in government and inspired public conversations with Canada's minister of public safety; the national security and intelligence advisor to the prime minister of Canada; the director of the Canadian Security Intelligence Service; the chief of the Communications Security Establishment; and the privacy commissioner of Canada.

now has Gemini. Meta has Llama. Anthropic has Claude. And on, and on. At their most basic level, these chatbots will take a plain language text prompt from a user and then generate new text to respond to that prompt. Given the sophistication of these tools, the text is well-written and grammatically correct, and appears substantively convincing — even if not always accurate.

There are also AI tools that can generate images or video based on text prompts. This exposes a new front in disinformation campaigns, namely, deepfakes. It is now possible to create convincing audio and video of real people saying or doing things that they did not (Bond 2024; Thebault 2024). To create a deepfake video, the creator trains a generative AI model with extensive real video footage of the target individual, enabling the system to recreate their appearance from various angles and lighting conditions. The outputs of this model are then combined with computer graphics techniques to overlay the person onto a synthetic body. Although AI accelerates the process significantly, achieving a convincing result requires time and manual adjustment of parameters to eliminate any noticeable glitches or artifacts in the final image (Adee 2020; Thompson 2024). While most video deepfake tools are cumbersome and unconvincing today, the companies developing these tools are making startling improvements nearly every month. Notably, some of these companies require the consent of the deepfaked individual in order to produce outputs,[1] while others have been distributed in an "open-source" manner that offers no such limitations (Bernaciak and Ross 2022).

Furthermore, despite the increasing accessibility of deepfake technology, the general public still struggles to recognize manipulated content (American Bankers Association 2025). This poses a significant threat to democracies, as malicious actors exploit this uncertainty to disseminate synthetic or falsified information, eroding trust. Such exploitation can have psychological, reputational and economic harm implications (Government of Canada 2023). This is troubling, because when looking at how disinformation campaigns are constructed, it is clear these technologies will supercharge disinformation campaigns and influence operations. Democracies should expect that disinformation campaigns seeking to undermine electoral integrity will become much more sophisticated and widespread as a consequence of the increasing availability of progressively higher-quality generative AI tools.

While many people have pointed out that some predictions of the impact of AI-powered election interference in 2024 did not come true (Kapoor and Narayanan 2024), it is important to recognize that the technology is still in its early stages, and traces of its impact can already be seen around the world (Elliott 2024). Futurist Roy Amara described this phenomenon in "Amara's Law," which states that "we tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run" (quoted in Lin 2024).

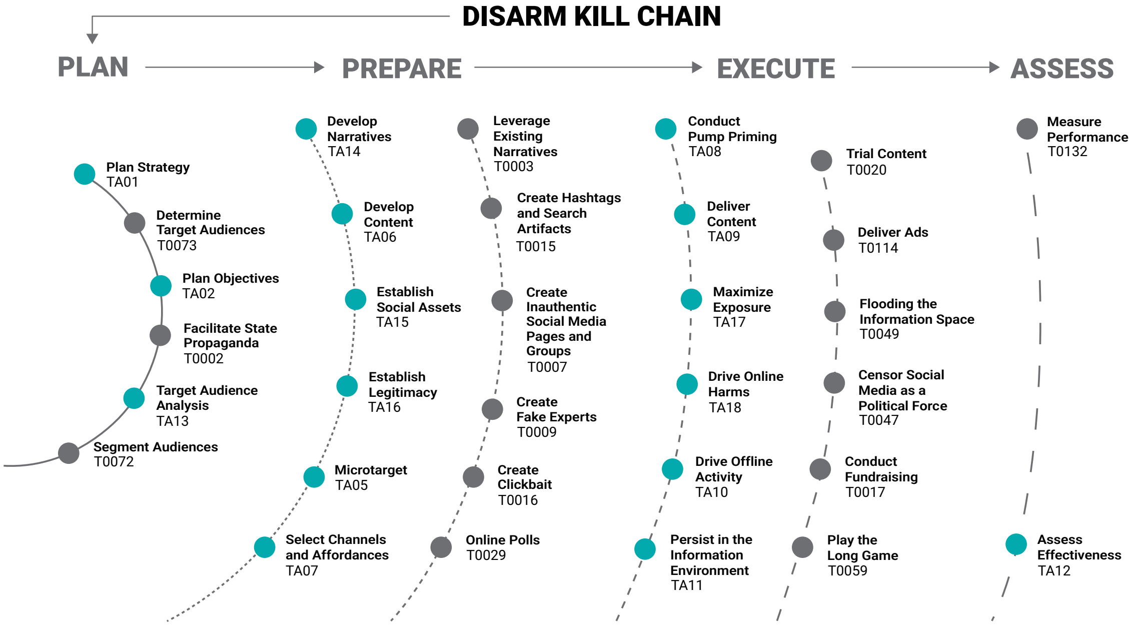
Additionally, encrypted messaging platforms, one of the best mechanisms for distributing AI-generated election interference and manipulation efforts, are at present very difficult to study due to the lack of transparency about user complaints of abuse from the companies who run these services. A recent study notes that "messaging platforms, which market themselves as spaces for private conversation, increasingly serve as arenas for intense political activity, including electoral campaigns. The architectures and features of applications like WhatsApp, Telegram, and Viber make them particularly useful as vectors for political propaganda, or information calculated to manipulate public opinion" (Olaizola Rosenblat, Trauthig and Woolley 2024, 1). It will be important for researchers and policy makers to continue to put pressure on these companies to preserve encryption capabilities, while at the same time closely monitoring for signals of abuse and taking rapid and transparent action when accounts generate repeat violations of platform policies.

Disinformation Is Set to Get Much Worse

How disinformation campaigns actually work has been described elsewhere in detail. Individual campaigns will vary, but inevitably involve some measure of narrative creation and propagation, social media manipulation, use of state-controlled media or proxies, the exploitation of authentic grievances, covert influence and the instigation of conflict between groups. The more sophisticated of these can be buttressed by cyber operations, typically "hacks" and "leaks," whereby an email

[1] See, for example, HeyGen's "Acceptable Use and Moderation Policy": www.heygen.com/moderation-policy.

Figure 1: DISARM Kill Chain


DISARM KILL CHAIN
PLAN
PREPARE
EXECUTE
ASSESS
Plan Strategy TA01
Determine Target Audiences T0073
Plan Objectives TA02
Facilitate State Propaganda T0002
Target Audience Analysis TA13
Segment Audiences T0072
Develop Narratives TA14
Develop Content TA06
Establish Social Assets TA15
Establish Legitimacy TA16
Microtarget TA05
Select Channels and Affordances TA07
Leverage Existing Narratives T0003
Create Hashtags and Search Artifacts T0015
Create Inauthentic Social Media Pages and Groups T0007
Create Fake Experts T0009
Create Clickbait T0016
Online Polls T0029
Conduct Pump Priming TA08
Deliver Content TA09
Maximize Exposure TA17
Drive Online Harms TA18
Drive Offline Activity TA10
Persist in the Information Environment TA11
Trial Content T0020
Deliver Ads T0114
Flooding the Information Space T0049
Censor Social Media as a Political Force T0047
Conduct Fundraising T0017
Play the Long Game T0059
Measure Performance T0132
Assess Effectiveness TA12

Source: Strategic Communications, Task Forces and Information Analysis Data Team (2023, 29). Visualization of the DISARM framework's threat actor "kill chain" (conceptual model of a cyberattack) for a "red team" (simulated adversary). Green dots represent overarching tactics (TA) at a given stage; black dots represent techniques (T) used under a given tactic.

compromise or some other exploit results in the release of private information that is either embarrassing or damaging. To date, one of the most prominent examples of the latter occurred when Russian intelligence agencies hacked the computer network of the Democratic National Committee in 2016 and leaked salacious internal documents and emails to various public facing outfits, including Wikileaks (Office of Public Affairs 2018).

Figure 1 depicts the "Kill Chain" of the Disinformation Analysis and Risk Management (DISARM) open-source framework. The DISARM Framework "provides a common language to combat disinformation, for defenders to coordinate, share data, analysis, and act in synchrony,"[2] and the Kill Chain, originally a military concept and adapted for cybersecurity, is a model that details the stages of a disinformation campaign (Strategic Communications, Task Forces and Information Analysis Data Team 2023, 4). The stages include collective measures taken to plan, prepare, execute and then assess the efficacy of these campaigns. Given how disinformation campaigns are constructed, there is almost no stage that will not be rendered more effective by the use of generative AI. Developing convincing narratives no longer requires native language skills or deep understanding of social cleavages. Creating realistic, believable, tailored content no longer requires graphic design teams or commensurate marketing expertise; it is just a few text prompts away. Given the unsatisfactory nature of the existing tools to address this budding reality, it is completely foreseeable that disinformation — especially during elections — is set to get much, much worse (Elliott 2024).

Although 2024's elections did not have any single blockbuster AI use case that appears to have had such a significant impact that it might have changed an election outcome, there is definitely evidence of the technology being put to use (ibid.). During the Democratic primary, President Joe Biden recorded a message that was then delivered through a computerized auto-dialer (a "robocall") to prospective voters in New Hampshire's primary, discouraging them from voting in the primary, by saying that they only got one vote to cast and needed to save it for the November general election against Donald Trump. In essence, they were told by the President of the United States not to waste their one and only vote during the primary. The problem, of course, is

[2] See www.disarm.foundation/.

Box 1: The Human, Civil and Political Rights at Stake

Universal Declaration of Human Rights

Article 21

1. Everyone has the right to take part in the government of his country, directly or through freely chosen representatives.

2. Everyone has the right of equal access to public service in his country.

3. The will of the people shall be the basis of the authority of government; this will shall be expressed in periodic and genuine elections which shall be by universal and equal suffrage and shall be held by secret vote or by equivalent free voting procedures.

International Covenant on Civil and Political Rights

Article 25

Every citizen shall have the right and the opportunity, without any of the distinctions mentioned in article 2 and without unreasonable restrictions:

a. To take part in the conduct of public affairs, directly or through freely chosen representatives;

b. To vote and to be elected at genuine periodic elections which shall be by universal and equal suffrage and shall be held by secret ballot, guaranteeing the free expression of the will of the electors;

c. To have access, on general terms of equality, to public service in his country.

European Convention on Human Rights

Protocol 1, article 3

The High Contracting Parties undertake to hold free elections at reasonable intervals by secret ballot, under conditions which will ensure the free expression of the opinion of the people in the choice of the legislature.

Charter of Fundamental Rights of the European Union

Article 39

1. Every citizen of the Union has a right to vote and to stand as a candidate at elections to the European Parliament in the Member State in which he or she resides, under the same conditions as nationals of that State.

2. Members of the European Parliament shall be elected by direct universal suffrage in a free and secret ballot.

Article 40

Every citizen of the Union has the right to vote and to stand as a candidate at municipal elections in the Member State in which he or she resides under the same conditions as nationals of that State.

that the information was false, and that Biden did not record that message. But it sounded like him, mimicked his intonation, and even borrowed some of his signature language — for example, in describing the early vote as “a bunch of malarky” (CNN 2024).

In addition to such deceptive uses, AI also has an impact through what scholars have dubbed the “liar’s dividend” (Chesney and Keats Citron 2018), a phenomenon by which people can dismiss authentic claims as false. One of the clearest examples of this was Trump’s assertion that images of large crowds at Kamala Harris’s rallies were fake, although, in fact, they were real (Joffe-Block 2024). Policy makers must expect that over the coming elections, the use of deepfake technology will be more ubiquitous, more effective and more damaging to the democratic process, if policies are not quickly put in place to mitigate its harms.

The Harms of “Truth Decay”

The rising use of this technology indicates further “truth decay” and a continued rise in “bespoke realities.” The RAND Corporation defines truth decay as the erosion of the role of facts and data in public discourse (Kavanagh and Rich 2018, chapter 2). Four trends characterize this concept: growing disagreement about facts and their interpretation, a blurred distinction between opinion and fact, the rising prominence of opinion and personal experience over factual evidence, and a decline in trust in once-reliable sources of information (ibid.).

Individuals impacted by “truth decay” inhabit a unique realm of perception termed a “bespoke reality,” a concept coined in 2019 by Renée DiResta, an associate research professor at the McCourt School of Public Policy at Georgetown University. She describes it as a consequence of a “Cambrian explosion of bubble realities” (DiResta 2019), where communities develop distinct norms, trusted sources and factual frameworks. These bespoke realities are personalized overlays of the world shaped by individual desires and preferences. Advancements in technology, including the internet, social media, and augmented and virtual reality, have magnified and diversified this phenomenon (French 2023). Those “realities” are set to become much more convincing with the increasing access to generative AI tools of higher and higher quality.

As these campaigns become more sophisticated and manipulative, the foreseeable consequence will be a further erosion of trust in institutions and a heightened disintegration of civic integrity, which in turn will jeopardize a host of human rights, including electoral rights and the right to freedom of thought (see Box 1).

Electoral rights are the very fabric of democracy. While the content or implementation of these rights might vary marginally from country to country, the essence is that democratic polity is made of a bundle of rights, which include the right to vote and to have fair elections, non-discrimination, freedom of association, freedom of expression, access to information, and privacy when voting, as enshrined in, for example, the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights, the European Convention for the Protection of Human Rights and Fundamental Freedoms, and the Charter of Fundamental Rights of the European Union[3] (see Box 1).

Fundamental Rights at Risk

It is not hard to conceive of a not-too-distant state of affairs where the information ecosystem is so poisoned by AI-generated content that the right to fair elections and the right to access information is rendered meaningless.

This technological evolution could also undermine other fundamental human rights. The right to freedom of thought is a core human right that is enumerated under international law. It has three essential elements:

- the right to keep our thoughts private so that we may not be coerced into revealing them;
- freedom from manipulation; and
- a prohibition on penalizing a person for their thoughts or opinions alone (Shull 2024).

It is the ability to manipulate that makes these technologies so potentially offensive to the

[3] *Universal Declaration of Human Rights,* GA Res 217A (III), UNGAOR, 3d Sess, Supp No 13, UN Doc A/810 (1948) [*UDHR*], online: <https://documents.un.org/doc/resolution/gen/nr0/043/88/pdf/nr004388.pdf>; *International Covenant on Civil and Political Rights*, 16 December 1966, 999 UNTS 171 (entered into force 23 March 1976), online: <www.ohchr.org/en/professionalinterest/pages/ccpr.aspx>; *European Convention for the Protection of Human Rights and Fundamental Freedoms*, 4 November 1950, 213 UNTS 221, ETS No 5 (entered into force 3 September 1953) [*European Convention on Human Rights*], online: <www.echr.coe.int/documents/d/echr/convention_eng#page=19>; EC, *Charter of Fundamental Rights of the European Union*, [2012] OJ, C 326/391 [*Charter*], online: <https://eur-lex.europa.eu/eli/treaty/char_2012/oj/eng>.

right to freedom of thought. If an individual is manipulated through targeted fake content, turbocharged by generative AI, it is likely to result in instances of the breach of this fundamental right. Indeed, it is almost impossible to conceive of an instance where fake information or content is passed off as genuine and not being used to manipulate an individual in some way.

The very essence of what makes us human is our freedom to think what we wish, and this freedom to think and to choose our elected representatives is what makes democracy work. It cannot be subject to further erosion because of technological evolution. However, at the same time that these technologies are becoming more powerful and more persuasive, the companies that form the global distribution channels for this content are hollowing out the safety mechanisms that could check the most harmful impacts of these tools' misuse.

According to the *Big Tech Backslide* report from Free Press — the most comprehensive effort so far to document the divestment from efforts to protect social media and messaging platforms from abuse — "the largest social-media companies have deprioritized content moderation and other user trust and safety protections, including rolling back platform policies that had reduced the presence of hate, harassment and lies on their networks. These companies have also laid off critical staff and teams tasked with maintaining platform integrity" (Benavidez 2023, 3).

One significant factor enabling this transition was the 2022 acquisition of Twitter by Elon Musk (Toh and Liu 2023), which led to an astounding layoff of approximately 80 percent of the company's staff, including a team responsible for AI ethics (Knight 2022). Mark Zuckerberg appeared emboldened by these efforts, expressing admiration for Musk's having made Twitter a "leaner" company (Gendron 2023). Zuckerberg himself called 2023 Meta's "year of efficiency," conducting multiple rounds of layoffs and disbanding his own company's "Responsible AI" team (Picciotto 2023). In early 2025, Zuckerberg announced that Meta would be ceasing all fact-checking operations in favour of a community-driven moderation system reliant on community-voted content labelling (Isaac and Schleifer 2025).

Meta, Amazon, Alphabet and Twitter "ravaged" their "trust and safety" and "integrity" teams, making major cuts to teams fighting online misinformation and hate speech across their companies. In a noteworthy cut, Meta eliminated a fact-checking tool that teams had been building for more than half a year (Field and Vanian 2023).

These cuts to trust and safety teams have gone so deep that a new industry has taken shape to sell back these functions as a service to big tech companies (Elliott 2023). Start-ups such as Cove,[4] Cinder and Safety Kit[5] were all founded by alumni of the same big tech companies that made these cuts. In an interesting twist, each of the companies advertises on its website how it is using AI to support content-moderation efforts.

While there is reason to think that this new industry's approach could be promising, it is by no means a mature or complete substitute for the human beings whose jobs were eliminated by these platforms. There is also reason to be concerned that the pendulum could eventually swing too far in the opposite direction, and that these types of AI tools could eventually become "too reliable" for content moderation, "providing a mechanism for suppression masquerading as moderation" (Barrett and Hendrix 2024).

To be fair, it's not pure greed on the part of CEOs themselves driving these cuts — activist investors such as Altimeter Capital's Brad Gerstner play a significant role in driving these decisions. Gerstner wrote in October 2022 that "it is a poorly kept secret in Silicon Valley that companies ranging from Google to Meta to Twitter to Uber could achieve similar levels of revenue with far fewer people" (Gerstner 2022).

Protecting the Information Ecosystem

When facing pressures like these, all roles that generate costs and not profits are strong candidates for the chopping block. Without strong laws making the protection of the information environment a priority for these companies, the companies making big investments in areas such as AI ethics and content moderation get punished.

This means that policy makers must act.

[4] See https://getcove.com/.

[5] See www.safetykit.com.

Europe has jumped ahead of the rest of the world's democracies by passing not only their path-breaking AI Act in 2024, but also their 2022 Digital Services Act (DSA).[6] Although the DSA does not mention the words "artificial intelligence," it has extensive provisions that refer and apply to the algorithmic systems (often referred to also as "AI") that underpin online platforms, including social media and search engines.

An important aspect of the DSA, in relation to the right to freedom of thought, is that it is designed, as is all EU law, to operate in accordance with the Charter of Fundamental Rights of the European Union,[7] which includes both the right to freedom of thought and the right to mental integrity. That charter has significant overlap with, and a lineage that can be traced in part back to, the Universal Declaration of Human Rights (Anderson and Murphy 2011).

In article 34, the DSA lays out that very large online platforms (VLOPs) and very large online search engines (VLOSEs) must conduct risk assessments to evaluate "any actual or foreseeable negative effects for the exercise of fundamental rights," as well as "any actual or foreseeable negative effects on civic discourse and electoral processes" and, further, "the protection of public health and minors and serious negative consequences to the person's physical and mental well-being" (paras 1[b], 1[c], 1[d], respectively). In assessing these risks, article 34, paragraph 2, outlines that the platforms must consider, among other factors, "(a) the design of their recommender systems and any other relevant algorithmic system; (b) their content moderation systems; and (c) the applicable terms and conditions and their enforcement." In March 2024, the European Commission released new draft guidelines under the DSA "for the mitigation of systemic risks online for elections," giving much more detail about their expectations of DSA enforcement in the context of imminent European elections and developments in generative AI technology since the DSA came into force in 2022 (European Commission 2024a). These guidelines — drafted after much of the substance of the EU AI Act was agreed on — are critically important, because their guidance is effective immediately, whereas the substantive provisions of the EU AI Act will take up to three years to come into force fully.

The guidelines have several critical provisions supporting freedom of thought in the context of AI. Most notable is a requirement in article 39 that "providers of VLOPs and VLOSEs whose services can be used for the creation of deceptive, biased, false or misleading generative AI content....Ensure that generative AI content, and other types of synthetic and manipulated media, is detectable — notably by using sufficiently reliable, interoperable, effective and robust techniques and methods, such as **watermarks**" (emphasis in original; ibid., 26).

While not all generative AI providers are VLOPs or VLOSEs, many of them are (Meta, Google, Microsoft and X, for example), and the guidelines mean that they need to take steps immediately to begin marking their AI-generated content in robust ways. This requirement is significant, because while most of the major AI companies made commitments to mark AI-generated content in the White House Voluntary AI Commitments of July 2023 (The White House 2023), and again in February 2024 at the Munich Security Conference under a new "AI Elections Accord,"[8] the companies have failed to live up to their voluntary commitments (Harris and Norden 2024; Kroet 2024; Ahmed et al. 2025).

Policy makers elsewhere in the world should quickly follow suit and make broader laws requiring all generative AI developers to include difficult-to-remove watermarks in their AI-generated content. This type of watermark is difficult to remove and can be ingested by social media and messaging platforms and turned into a label that lets people know if content was generated by AI. Equally important is placing requirements on camera and phone manufacturers to include authenticity watermarks in authentic images and audio and video recordings, and to give people the option to place digital signatures on content that they create. These types of policies will push people toward online experiences where they are able to understand much more about the provenance of most of the content they see and hear, which will give users a better idea of what is authentic and what is synthetic. California's new AI Transparency Act is a good start down this road (Kemp 2024), as will likely be the code of practices currently under development that will provide details

[6] See EC, *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC*, [2022] OJ, L 277/1, online: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065#d1e3513-1-1>.

[7] *Charter, supra* note 3.

[8] See www.aielectionsaccord.com/.

on implementation of the required provenance and watermarking provisions of the EU AI Act.

Watermarking is not, however, a silver bullet. It will not work in all cases, and some types of synthetic content that are particularly dangerous should be banned altogether. One is the use of AI for impersonation. The US Federal Trade Commission (FTC) recently banned government and business impersonation, and is now seeking public comments on a potential ban on impersonation of individuals (FTC 2024). Deepfakes largely overlap with this category, and have already been used to interfere in elections in noteworthy ways in the United States, as with the earlier cited example of a voice mimicking Biden's voice being used in a robocall discouraging voting (Ramer 2024); AI-generated images, even if marked as inauthentic, can be immensely disruptive as well, as happened with fake photos of Trump being arrested that were shared millions of times on social media in March 2023 (Stanley-Becker and Nix 2023).

There are many other solutions that should be required of social media and messaging platforms to protect freedom of thought. One solution, already in use in many parts of the world, is independent fact-checking. The European Union's final election guidelines require that VLOPs and VLOSEs take "measures to provide users with more contextual information on the content and accounts they engage with," including the placement of "fact-checking labels on identified disinformation and FIMI [foreign information manipulation and interference] content provided by independent fact-checkers and fact-checking teams of independent media organisations."[9]

Fact-checking is already implemented by some social media companies, but implementations are often partial or do not meet the scale of the problem. Facebook and Instagram perversely exempt politicians altogether from fact-checking, which created a bizarre situation where, upon his November 2022 announcement that he was running for president in the 2024 elections, Donald Trump secured nearly two years of immunity from fact-checking (Duffy 2022).

Freedom of thought in the online world, however, is not only implicated in decisions from large platforms about how to handle content made by AI. Freedom of thought can also become compromised within the design structure of platforms themselves, and specifically, the way that AI algorithms are used to rank and recommend certain content to certain users. Many researchers have highlighted the prevalence of echo chambers and the propensity for social algorithms to group biased individuals together and rapidly spread information between them (Cinelli et al. 2021). This practice is deployed by social platforms to maximize engagement and advertising revenue, but it can just as easily cut an individual off from valuable perspectives beyond their current biases. A new wave of researchers has begun attempts to address this fundamental issue with design solutions. An author of the Neely Center Design Code for Social Media says that "our design code is 'content neutral,' meaning that it does not require a platform to make decisions about which content to allow or amplify. Instead, it anchors on signals from users as to what they prefer and/or consider to be higher quality content" (quoted in Skacan 2024). Elements of this design code are already making their way into policy circles, including via Minnesota's proposed Prohibiting Social Media Manipulation Act.[10] Design changes could modify features that feed or reward "psychological factors such as social comparison, the need for social validation, and the fear of missing out" that researchers identify as driving social media addiction (Perez-Lozano and Espinosa 2024), reduce exposure to types of content unwanted by the user and give people better ways to control their privacy settings.

At the most basic level, the shift needed is a balanced move toward the democratization of social media platforms — toward a world in which democratically elected governments can act to make AI and communications tools serve, above all, the needs of everyday people, while protecting their human rights.

Acknowledgements

The authors wish to thank Susie Alegre, Owen Doyle and Lynn Schellenberg for their valuable work in reviewing this manuscript.

[9] EC, *Annex to the Communication to the Commission. Approval of the content of a draft Communication from the Commission on Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes pursuant to the Digital Services Act,* C(2024) 2121 final at 8.

[10] See US, HF 4400, *Prohibiting Social Media Manipulation Act,* 93rd Sess, Minn, 2024, online: <www.revisor.mn.gov/bills/bill.php?b=house&f=HF4400&ssn=0&y=2024>.

Works Cited

Adee, Sally. 2020. "What Are Deepfakes and How Are They Created? Deepfake technologies: What they are, what they do, and how they're made." *IEEE Spectrum*, April 29. Updated March 8, 2024. https://spectrum.ieee.org/what-is-deepfake.

Ahmed, Abdiaziz, Owen Doyle, David Evan Harris and Lawrence Norden. 2025. "Tech Companies Pledged to Protect Elections from AI — Here's How They Did." Brennan Center for Justice, February 13. www.brennancenter.org/our-work/research-reports/tech-companies-pledged-protect-elections-ai-heres-how-they-did.

AI Elections Accord. 2024. "Technology Industry to Combat Deceptive Use of AI in 2024 Elections." Press release, February 16. www.aielectionsaccord.com/uploads/2024/02/Press-Release-AI-Elections-Accord-16-Feb-2024.pdf.

Allyn, Bobby. 2024. "An Elon Musk-backed political group is posting fake Kamala Harris ads on Facebook." NPR, October 30. www.npr.org/2024/10/30/g-s1-31042/elon-musk-kamala-harris-facebook.

American Bankers Association. 2025. "Study finds most people can't spot deepfakes." *ABA Banking Journal*, February 12. https://bankingjournal.aba.com/2025/02/study-finds-most-people-cant-spot-deepfakes/.

Anderson, David and Cian C. Murphy. 2011. "The Charter of Fundamental Rights: History and Prospects in Post-Lisbon Europe." European University Institute Working Paper LAW 2011/08. https://hdl.handle.net/1814/17597.

Barrett, Paul M. and Justin Hendrix. 2024. "Is Generative AI the Answer for the Failures of Content Moderation?" *Wired*, April 3. www.techpolicy.press/is-generative-ai-the-answer-for-the-failures-of-content-moderation/.

Benavidez, Nora. 2023. *Big Tech Backslide: How Social-Media Rollbacks Endanger Democracy Ahead of the 2024 Elections*. Florence, MA: Free Press. www.freepress.net/big-tech-backslide-report.

Bernaciak, Catherine and Dominic A. Ross. 2022. "How Easy Is It to Make and Detect a Deepfake?" *SEI Blog*, March 14. https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/.

Bond, Shannon. 2024. "A political consultant faces charges and fines for Biden deepfake robocalls." NPR, May 23. www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative.

Chesney, Robert and Danielle Keats Citron. 2018. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." Preprint, SSRN, July 24. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954.

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi and Michele Starnini. 2021. "The echo chamber effect on social media." *Proceedings of the National Academy of Sciences* 118 (9): e2023301118. https://pubmed.ncbi.nlm.nih.gov/33622786/.

CNN. 2024. "Hear fake Biden robocall urging voters not to vote in New Hampshire." January 23. www.cnn.com/videos/politics/2024/01/23/robocall-fake-biden-new-hampshire-primary-nn-vpx.cnn.

DiResta, Renée. 2019. "Mediating Consent." *Ribbonfarm* (blog), December 17. www.ribbonfarm.com/2019/12/17/mediating-consent/.

Duffy, Kate. 2022. "Facebook won't be fact-checking Donald Trump now he's announced he's running for president in 2024." Business Insider, November 16. www.businessinsider.com/facebook-trump-fact-checking-halted-presidential-run-2024-announced-meta-2022-11.

Elliott, Vittoria. 2023. "Big Tech Ditched Trust and Safety. Now Startups Are Selling It Back As a Service." *Wired*, November 6. www.wired.com/story/trust-and-safety-startups-big-tech/.

———. 2024. "The Year of the AI Election Wasn't Quite What Everyone Expected." *Wired*, December 26. www.wired.com/story/the-year-of-the-ai-election-wasnt-quite-what-everyone-expected/.

European Commission. 2024. "Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act." Press release, April 29. https://ec.europa.eu/commission/presscorner/detail/fen/ip_24_2373.

Field, Hayden and Jonathan Vanian. 2023. "Tech layoffs ravage the teams that fight online misinformation and hate speech." CNBC, May 26. www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html.

French, David. 2023. "Welcome to Our New 'Bespoke Realities.'" *The New York Times*, November 30. www.nytimes.com/2023/11/30/opinion/political-reality-algorithms.html.

FTC. 2024. "FTC Proposes New Protections to Combat AI Impersonation of Individuals." Press release, February 15. www.ftc.gov/news-events/news/press-releases/2024/02/ftc-proposes-new-protections-combat-ai-impersonation-individuals.

Gendron, Will. 2023. "Mark Zuckerberg praised Elon Musk's push 'to make Twitter a lot leaner.'" Business Insider, June 10. www.businessinsider.com/meta-mark-zuckerberg-praises-elon-musk-changes-twitter-layoffs-2023-6.

Gerstner, Brad. 2022. "Time to Get Fit — an Open Letter from Altimeter to Mark Zuckerberg (and the Meta Board of Directors)." Medium, October 24. https://medium.com/@alt.cap/time-to-get-fit-an-open-letter-from-altimeter-to-mark-zuckerberg-and-the-meta-board-of-392d94e80a18.

Government of Canada. 2023. *The Evolution of Disinformation: A Deepfake Future*. World Watch: Expert Notes series publication No. 2023-10-01. October. www.canada.ca/en/security-intelligence-service/corporate/publications/the-evolution-of-disinformation-a-deepfake-future.html.

Harris, David Evan and Lawrence Norden. 2024. "Meta's AI Watermarking Plan Is Flimsy, at Best: Watermarks are too easy to remove to offer any protection against disinformation." *IEEE Spectrum*, March 4. https://spectrum.ieee.org/meta-ai-watermarks.

Hu, Krystal. 2023. "ChatGPT sets record for fastest-growing user base — analyst note." Reuters, February 2. www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

Isaac, Mike and Theodore Schleifer. 2025. "Meta Says It Will End Its Fact-Checking Program on Social Media Posts." *The New York Times*, January 7. www.nytimes.com/live/2025/01/07/business/meta-fact-checking.

Joffe-Block, Jude. 2024. "Why false claims that a picture of a Kamala Harris rally was AI-generated matter." NPR, August 14. www.npr.org/2024/08/14/nx-s1-5072687/trump-harris-walz-election-rally-ai-fakes.

Kapoor, Sayash and Arvind Narayanan. 2024. "We Looked at 78 Election Deepfakes. Political Misinformation Is Not an AI Problem." *Toward a Better Internet* (blog), December 13. https://knightcolumbia.org/blog/we-looked-at-78-election-deepfakes-political-misinformation-is-not-an-ai-problem.

Kavanagh, Jennifer and Michael D. Rich. 2018. *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*. Santa Monica, CA: RAND. www.rand.org/pubs/research_reports/RR2314.html.

Knight, Will. 2022. "Elon Musk Has Fired Twitter's 'Ethical AI' Team." *Wired*, November 4. www.wired.com/story/twitter-ethical-ai-team/.

Kroet, Cynthia. 2024. "How has the Digital Services Act been enforced, one year on?" Euronews, July 19. www.euronews.com/next/2024/07/19/how-has-the-digital-services-act-been-enforced-one-year-on.

Lin, Pohan. 2024. "Amara's Law and Its Place in the Future of Tech." IEEE Computer Society, September 6. www.computer.org/publications/tech-news/trends/amaras-law-and-tech-future.

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

Nehamas, Nicholas. 2023. "DeSantis Campaign Uses Apparently Fake Images to Attack Trump on Twitter." *The New York Times*, June 8. www.nytimes.com/2023/06/08/us/politics/desantis-deepfakes-trump-fauci.html.

Office of Public Affairs. 2018. "Grand Jury Indicts 12 Russian Intelligence Officers for Hacking Offenses Related to the 2016 Election." Press release, July 13. Washington, DC: Department of Justice. www.justice.gov/opa/pr/grand-jury-indicts-12-russian-intelligence-officers-hacking-offenses-related-2016-election.

Olaizola Rosenblat, Mariana, Inga K. Trauthig and Samuel C. Woolley. 2024. *Covert Campaigns: Safeguarding Encrypted Messaging Platforms from Voter Manipulation*. October. New York, NY: NYU Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/wp-content/uploads/2024/10/NYU-CBHR-Covert-Campaigns_FINAL-FINAL-Sep29.pdf.

Perez-Lozano, Delia and Francisco Saucedo Espinosa. 2024. "Social Media Addiction: Challenges and Strategies to Promote Media Literacy." In *Social Media and Modern Society — How Social Media Are Changing the Way We Interact with the World Around*, edited by Ján Višňovský and Jana Majerová, chapter 12. London, UK: IntechOpen. www.intechopen.com/chapters/1191216.

Picciotto, Rebecca. 2023. "Facebook-parent Meta breaks up its Responsible AI team." CNBC, November 18. www.cnbc.com/2023/11/18/facebook-parent-meta-breaks-up-its-responsible-ai-team.html.

Powell, Catherine. 2024. "Deepfake of Kamala Harris Reups Questions on Tech's Self-Regulation." Council on Foreign Relations, August 1. www.cfr.org/blog/deepfake-kamala-harris-reups-questions-techs-self-regulation.

Ramer, Holly. 2024. "Political consultant behind fake Biden robocalls says he was trying to highlight a need for AI rules." AP News, February 26. https://apnews.com/article/ai-robocall-biden-new-hampshire-primary-2024-f94aa2d7f835ccc3cc254a90cd481a99.

Shull, Aaron. 2024. "Safeguarding Privacy and Preserving Freedom of Thought in Canada: A Call to Action." Opinion, Centre for International Governance Innovation, March 28. www.cigionline.org/articles/safeguarding-privacy-and-preserving-freedom-of-thought-in-canada-a-call-to-action/.

Skacan, Sabrina. 2024. "Neely Center Efforts Turn Design Codes into Policy." University of Southern California Marshall School of Business, April 1. www.marshall.usc.edu/posts/neely-center-design-codes-attracts-policymakers.

Stanley-Becker, Isaac and Naomi Nix. 2023. "Fake images of Trump arrest show 'giant steps' for AI's disruptive power." *The Washington Post,* March 22. www.washingtonpost.com/politics/2023/03/22/trump-arrest-deepfakes/.

Strategic Communications, Task Forces and Information Analysis Data Team. 2023. *1st EEAS Report on Foreign Information Manipulation and Interference Threats: Towards a framework for networked defence*. February. Brussels, Belgium: European External Action Service. https://euvsdisinfo.eu/uploads/2023/02/EEAS-ThreatReport-February2023-02.pdf.

Swenson, Ali and Will Weissert. 2024. "New Hampshire investigating fake Biden robocall meant to discourage voters ahead of primary." Associated Press, January 22. https://apnews.com/article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5.

The White House. 2023. "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI." Press release, September 12. https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

Thebault, Reis. 2024. "Deepfake Kari Lake video shows coming chaos of AI in elections." *The Washington Post*, March 24. www.washingtonpost.com/politics/2024/03/24/kari-lake-deepfake/.

Thompson, Stuart A. 2024. "A.I. Can Now Create Lifelike Videos. Can You Tell What's Real?" *The New York Times*, September 9. www.nytimes.com/interactive/2024/09/09/technology/ai-video-deepfake-runway-kling-quiz.html.

Thrush, Glenn. 2024. "3 U.S. Intelligence Agencies Warn of Election Interference Efforts by Russia and Iran." *The New York Times*, November 4. www.nytimes.com/2024/11/04/us/politics/russia-iran-election-interference.html.

Toh, Michelle and Juliana Liu. 2023. "Elon Musk says he's cut about 80% of Twitter's staff." CNN, April 12. www.cnn.com/2023/04/12/tech/elon-musk-bbc-interview-twitter-intl-hnk/index.html.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

Credits

Managing Director and General Counsel Aaron Shull
CIGI Senior Fellow and Project Co-Leader Susie Alegre
Director, Program Management Dianna English
Program Manager Jenny Thiel
Publications Editor Lynn Schellenberg
Graphic Designer Sepideh Shomali

KONRAD ADENAUER STIFTUNG

This policy brief was made possible thanks to the financial support of the Konrad-Adenauer-Stiftung (KAS) Canada.

The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Centre for International Governance Innovation or its Board of Directors.

For publications enquiries, please contact publications@cigionline.org.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

CIGI