CENTER FOR
DEMOCRACY
& TECHNOLOGY

Center for Democracy & Technology’s Submission to the United Nations Special Rapporteur for Freedom of Opinion and Expression

14 January 2025

The Center for Democracy & Technology (CDT) welcomes the opportunity to contribute to the upcoming thematic report on freedom of expression and elections in the digital age by the United Nations special rapporteur for freedom of opinion and expression.

Protecting freedom of expression is critical to ensuring access to the ballot. New technologies can exacerbate risks to free expression, but overbroad regulatory and corporate interventions may do the same, particularly without adequate transparency, robust testing, and independent data access to promote oversight and accountability. This submission addresses governments’ legal and regulatory efforts, companies’ election integrity measures, and transparency and data access, and offers recommendations.

I. Government legal and regulatory efforts

Several states introduced or passed national or subnational laws about the use of generative AI in elections in 2024. Such laws, particularly those prohibiting or limiting the dissemination of “deceptive” synthetic content, could easily facilitate censorship of political speech. In the United States, some proposed federal and subnational laws would have allowed legal action against individuals who disseminated “deceptive” AI content or deepfakes.[1] Attempts to legislate based on subjective measures are reminiscent of the “fake news laws” that proliferated at the end of the last decade and have since been used to punish and silence political speech around the world.[2]

Another category of regulation requires disclosure or labeling of AI-generated content, including campaign materials and content created by voters. For example, in the Philippines, a resolution requires AI disclosure in campaign materials.[3] Currently, visual watermarks or metadata can be easily removed or manipulated, reducing their effectiveness.[4] Researchers have theorized that adding labels to AI-generated content on social media may infringe upon users’ speech and decrease trust in that content while implying that unlabeled content is more trustworthy.[5] Labeling requirements on posts by regular users, instead of candidates or campaigns, have greater potential to interfere with users’ political speech. Moreover, more research is needed into the efficacy of labeling as well as its effects, including potential unintended consequences for public trust. Labeling may lead to increased public

[1] Senate Rules Committee Advances Bills to Address Harmful AI in Elections - Center for Democracy and Technology

[2] CDT’s Response to EC ‘Fake News’ Consultation: How to Tackle the Issue and Protect Free Expression? - Center for Democracy and Technology ; The Rise of Digital Authoritarianism: Fake news, data collection and the challenge to democracy | Freedom House

[3] Philippines: AI and social media guidelines for the 2025 elections issued by the COMELEC - Baker McKenzie InsightPlus

[4] Big Tech says AI watermarks could curb misinformation, but they're easy to sidestep - NBC

[5] Labeling AI-Generated Content: Promises, Perils, and Future Directions - Chloe Wittenberg, Ziv Epstein, Adam Berinsky, David Rand

confusion because unlabeled content is not necessarily accurate; deceptive content has long been created without the use of generative AI.

II. Companies' election integrity safeguards

Technology companies take varied approaches to protecting elections and individuals' rights to access authoritative information about elections such as the time, place, and manner of voting. AI companies are relatively new entrants into this field and sometimes lack foundational election integrity policies that can go far in preventing malicious uses of their services, such as instituting clear and transparent usage policies and implementing product interventions like output filters, which are automated tools designed to detect and block certain types of content the company decides its services will not produce.[6]

In contrast, social media companies have had a longer runway to develop an apparatus of automated, human, and expert-driven processes to safeguard their platforms from targeted, manipulative, and inadvertent attacks against users' rights to the ballot. What these processes look like, how well they are resourced, and how long they are in operation differ from company to company. Ahead of the "year of elections" however, several companies rolled back election safeguards and reduced critical Trust & Safety teams tasked with monitoring and enforcing election policies.[7] Rolling back these efforts prematurely in the past has led to influence operations going unmonitored during critical points of the election process, such as the certification of the vote, leading to violence.[8] As more Trust & Safety work is automated, social media companies must staff election integrity efforts and do so for a longer period to ensure adequate coverage.

Consistent and well-resourced staffing is of special importance given that many of the automated interventions companies introduce are not tested robustly. For example, efforts to protect users from gender-based violence routinely miss targeted attacks against women of color politicians. More than 1 in 5 tweets sent in reply to or referring to Black and Asian-American women candidates for federal office in the U.S. in 2024 included offensive language directed at the candidate.[9] Commonly used automated tools to detect targeted and toxic speech are often counterintuitively prone to flagging women of color's speech as toxic more often than their white and male counterparts, resulting in over-moderating and silencing their speech and under-moderating targeted attacks against them, leading to many women of color politicians choosing not to run for office.[10]

Automated systems tasked with applying election integrity policies are also routinely undertested in non-English languages and often have critical vulnerabilities. There are well-documented risks of

[6] CDT Brief - Election Integrity Recommendations for Generative AI Developers ; Judge blocks new California law cracking down on election deepfakes | AP News

[7] Big Tech companies reveal trust and safety cuts in disclosures to Senate Judiciary Committee; X Fires Its Election Team Before a Huge Election Year | WIRED

[8] CDT Weighs In on Meta Oversight Board's case on Takedown of Speech Calling for Attack on Brazil's National Congress

[9] Report – Hated More: Online Violence Targeting Women of Color Candidates in the 2024 US Election - Center for Democracy and Technology

[10] How Automated Tools Discriminate Against Black Language

influence operations targeting voters who speak languages other than English leading to their disenfranchisement.[11] Yet, online services do not train or test their automated systems using natively created datasets, instead relying on machine-translated prompts, making them easy to circumvent and particularly error-prone.[12] Take for example a finding from OpenAI's red teaming process where external experts found that the company's output filters installed to prevent malicious use of OpenAI services failed when prompts were input in Farsi.[13] In a similar vein, Global Witness found that automated advertising review tools accepted over 50% of posts violating its election policies in Irish as compared to 20% in English ahead of a close election in 2024.[14] By not investing in high-quality training and testing datasets in non-English languages, users who don't speak English and those who live outside of the U.S. and EU encounter online information environments of a worse quality despite being large markets for these companies. This results in increased barriers for voters to easily access authoritative and credible information in their language, which is particularly important when voters operate in "data voids" or environments where limited information is available and networked fora such as social media platforms are essential for voters to access information.[15]

Finally, even within the English-speaking paradigm, AI companies rolling out user-facing applications have not invested in developing and testing their services for voters with disabilities. A study conducted by CDT found that popular chatbots fall short of providing accurate information to voters who have disabilities.[16] Of the five chatbots tested, every model hallucinated at least once and produced incorrect information such as describing a law or voting process that did not exist. Sixty-one percent of responses were insufficient to some degree with over one-third of responses including blatantly incorrect information.[17]

III. Limited transparency and data access

Limited transparency and independent access to data limit public understanding of how technology poses risks to elections, the efficacy of proposed solutions, and the risk that interventions may pose to human rights. In 2018, a multistakeholder effort led by civil society resulted in the Santa Clara Principles. The Principles helped establish baseline transparency and procedural norms among major social media companies.[18] The emergent culture of transparency had its limits, but the absence of a similar approach among AI developers and deployers has highlighted its value. Little is known about AI companies' electoral integrity policies and practices, when they are invoked, and how robustly they are

[11] Election Disinformation in Different Languages is a Big Problem in the U.S. - Center for Democracy and Technology; Spanish-speaking voters are in the crosshairs of disinformation agents | CyberScoop; Deepfakes are being used to dub adverts into different languages | New Scientist.

[12] Lost in Translation: Large Language Models in Non-English Content Analysis - Center for Democracy and Technology

[13] OpenAI's red team: the experts hired to 'break' ChatGPT

[14] TikTok still approving election disinformation ads in Ireland | Global Witness

[15] How online misinformation exploits 'information voids' — and what to do about it

[16] Brief - Generating Confusion: Stress-Testing AI Chatbot Responses on Voting with a Disability - Center for Democracy and Technology

[17] Brief - Generating Confusion: Stress-Testing AI Chatbot Responses on Voting with a Disability - Center for Democracy and Technology

[18] In 2021, the Principles were revised through a global multistakeholder process and now reflect the demands and contexts of civil society around the globe. https://santaclaraprinciples.org/

tested for efficacy and for ensuring respect for free expression and other rights. AI companies are reluctant to share critical usage data,[19] making it difficult to identify real-world harms and develop informed policy.

At the same time, the access that does exist is increasingly being curtailed, or offered in formats that limit its usefulness. In 2023, Twitter (now X) announced that it would no longer offer free access to its API and made access to data that independent researchers previously relied on prohibitively expensive.[20] In 2024, Meta shut down CrowdTangle against the urging of CDT and 50 other civil society organizations.[21] CDT's assessment of social media companies' political advertising repositories found a mix of formats (web-based interfaces v. downloadable CSVs) and frequent exclusion of ad creative (i.e. the content of the ad), both of which make the repositories difficult to search and to compare.[22] CDT has heard from other civil society actors that the lack of machine-readable political advertising repositories limited their ability to research the use of generative AI in elections.

Future research is further jeopardized by the increasing politicization of work relating to elections and free expression in the digital age. Threats to researchers and fact checkers are chilling critical research and our broader understanding of how technology affects democracy and democratic processes such as elections. The rise in strategic lawsuits against public participation (SLAPPs) poses a threat to journalists and researchers around the world.[23] Political elites and governments have smeared and investigated fact-checking organizations worldwide, from Egypt to South Korea, while trying to capture greater authority to control online content in the guise of fact-checking themselves. In India, the Bombay High Court found an attempt by the government to gain greater authority to conduct fact checking unconstitutional and a risk to free expression and press freedom.[24] In the U.S., certain government officials and non-state actors directed legal harassment and unfounded allegations of bias and censorship towards the Stanford Internet Observatory and the Election Integrity Partnership, leading to the discontinuation of their work during the 2024 elections.[25] In a further reflection of the politicization of fact-checking, in January 2025 Meta announced that it would end its third-party fact-checking model in the U.S., a program that was originally implemented after mis- and disinformation spread online during the 2016 U.S. election.[26] Fact-checking can be a crucial way to improve access to quality information while respecting freedom of expression, and it is not yet clear that the Community Notes-style system Meta announced it will adopt will be a sufficient replacement.

[19] Grounding AI Policy: Towards Researcher Access to AI Usage Data - Center for Democracy and Technology

[20] https://x.com/XDevelopers/status/1621026986784337922 ; Twitter's $42,000-per-Month API Prices Out Nearly Everyone | WIRED

[21] CDT Joins Letter Calling for Meta to Maintain its CrowdTangle Tool Through the Upcoming Election Cycle - Center for Democracy and Technology

[22] CDT Brief - Rules of the Road: Political Advertising on Social Media in the 2024 U.S. Election

[23] New global report: How are courts responding to SLAPPs?

[24] Delegitimizing the Messenger: the Assault on Fact-Checkers | Council on Foreign Relations ; Bombay High Court officially strikes down Centre's Fact Check Unit, calls amended IT Rules 'unconstitutional' - The Hindu

[25] Seismic Shifts: How Economic, Technological, and Political Trends are Challenging Independent Counter-Election-Disinformation In ; Election Integrity Partnership targeted election misinformation, not conservatives - Poynter

[26] More Speech and Fewer Mistakes ; Increasing Our Efforts to Fight False News - Meta

IV. Recommendations

CDT makes the following recommendations:

- Companies should pursue meaningful transparency to communicate clearly how they approach election integrity efforts.
    - Online services in general should disclose when, for what reason, and for how long election integrity efforts are ongoing and what these efforts entail.
    - AI companies should adopt the Santa Clara Principles and disclose what their election integrity policies are, how their policies are enforced and tested, and when they are in place.
    - Online services should test their election integrity policies in languages other than English including by investing in high-quality benchmarks, red-teaming, and other forms of external engagement and disclosing to users which languages their systems are trained and tested in.

- Companies, funders, and research institutions should support independent research by creating privacy-preserving mechanisms to increase data access and provide other resources for researchers.
    - In line with forthcoming requirements under the Digital Services Act's Article 40, companies should develop a privacy-preserving and anonymized mechanism to provide data access to independent researchers. Part of this can include equipping researchers with machine-readable data related to political advertisements and when and how users use services to seek voting-related information on their platform, including per language and region. Companies should also make available data on rates of use of output filters and other interventions (e.g., how often an output filter or some sort of automated tool blocked access to certain types of content or a certain prompt) to facilitate greater inquiry into the efficacy of election integrity interventions.
    - Funders, research institutions, and nonprofits should create shared resources and practices for researchers under attack. These might include pools for legal defense, cybersecurity assistance, and proactively developed communications plans for responding to coordinated attacks. Companies should provide material support and access for these programs.

- Companies should invest in multistakeholder engagement with election experts, civil society, and other affected stakeholders.
    - Companies should invest in fact-checking partnerships and provide funding to independent fact-checking organizations.
    - Companies should promote and direct users to external authoritative sources of election-related information. They should consult with independent civil society organizations to determine what approach and sources are appropriate in a given country's political context.