CENTER FOR DEMOCRACY & TECHNOLOGY

Comments to President's Council of Advisors on Science and Technology Working Group on Generative Artificial Intelligence

31 July 2023

The Center for Democracy & Technology welcomes the opportunity to provide comments to the President's Council of Advisors on Science and Technology on the potential risks and benefits of generative artificial intelligence systems. We focus our comments in particular on the impact of generative AI on our electoral processes.

Rapid advances in generative AI are spurring creativity and innovation, but also raise significant threats for human rights. The threats to elections and democratic discourse are worth highlighting. In previous elections, operatives used robocalls to spread incorrect information about mail-in voting in an effort to suppress Black voter turnout,[1] and used (sometimes illegal) micro-targeted social media campaigns to prevent people from voting.[2] Operatives also used deceptive text messages to spread intentionally misleading voting instructions for a Kansas ballot initiative in 2022.[3] It is easy to imagine bad actors using AI to exponentially grow and personalize voter suppression or other targeting efforts, increasing their harmful impact. Today, consumers can often spot a scam email, text or robocall because it uses non-personalized language and it may have grammatical or language errors (or, in the case of robocalls, a notably automated voice). Generative AI tools will make it easier to create tailored, accurate, realistic messages that draw victims in.

Generated images can also twist public understanding of political figures and events. Recordings of public figures' voices have been manipulated to trick senior government officials into thinking they are speaking with government leaders.[4] Videos and images have been digitally altered to make public officials appear incompetent, compromised, or to misrepresent their policy positions.[5] Experts have warned how deepfakes, which are difficult to authenticate or rebut, could impact an election in the closing days of voting, when there is little time to set the record straight, or before a debate.[6] More generally, the growth of inauthentic content makes it harder for people to know what news and content

---

[1] Christine Chung, "They Used Robocalls to Suppress Black Votes. Now They Have to Register Voters.", *The New York Times*, Dec. 1, 2022, https://www.nytimes.com/2022/12/01/us/politics/wohl-burkman-voter-suppression-ohio.html.

[2] Associated Press, "Far-right influencer convicted in voter suppression scheme," *Politico, Mar. 31, 2023,* *https://www.politico.com/news/2023/03/31/far-right-influencer-convicted-in-voter-suppression-scheme-00090042*.

[3] Isaac Stanley-Becker, "Misleading Kansas abortion texts linked to Republican-aligned firm", *The Washington Post*, Aug. 2, 2022, https://www.washingtonpost.com/politics/2022/08/02/kansas-abortion-texts/.

[4] *See e.g.,* Bobby Allyn, "Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn", *NPR*, Mar. 16, 2022, https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia (the minute long deepfake video "shows a rendering of the Ukrainian president appearing to tell his soldiers to lay down their arms and surrender the fight against Russia"); Philip Oltermann, "European politicians duped into deepfake video calls with mayor of Kyiv", *The Guardian*, Jun. 25, 2022, https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko.

[5] *See e.g.,* Hannah Denham, "Another fake video of Pelosi goes viral on Facebook", *The Washington Post,* Aug. 3, 2020, https://www.washingtonpost.com/technology Ω/2020/08/03/nancy-pelosi-fake-video-facebook/ (video depicts Pelosi slurring her speech and appearing intoxicated"); Alexandra Ulmer and Anna Tong, "Deepfaking it: America's 2024 election collides with AI boom", *Reuters*, May 30, 2023, https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023- 05-30/. While running for re-election in 2019, Houston's mayor said a critical ad ran by a fellow candidate broke a Texas law that bans certain misleading political deepfakes. Ivory Hecker, "Mayor Turner calls for criminal investigation of Tony Buzbee's attack ad", *Fox 26 Houston*, Oct. 17, 2019, https://www.fox26houston.com/news/mayor-turner-calls-for-criminal-investigation-of-tony-buzbees-attack-ad.

[6] James Bickerton, "Deepfakes Could Destroy the 2024 Election", *Newsweek*, Mar. 24, 2023, https://www.newsweek.com/deepfakes-could-destroy-2024-election-1790037.

they can trust, such that even authentic content is undermined. Journalists, whistleblowers, and human rights defenders are experiencing these effects already, facing higher hurdles than ever before to establish and defend their credibility.[7]

While the rise of affordable AI-generated content poses new threats to public discourse, policy interventions must be approached with care. This is because there are many legitimate reasons why people use software to generate and alter content: from laypeople and artists using AI to make creative works, to people engaging in parody, actors being de-aged in a movie, voices being sampled for a music track, or researchers altering images of North American and European cities to show what they would look like if they faced the same bombardment as the cities attacked in the Syrian war.[8] Barring or heavily restricting such activities would harm free expression, creativity, and innovation, and would quickly run afoul of the First Amendment.

Efforts to restrict or condition the distribution of generative images may also suppress protected expressive activities. To give one example, in recent years a number of companies and stakeholders have come together in the Content Authenticity Initiative, an impressive undertaking that allows photographers and other content creators to attach immutable provenance signals showing the authenticity of their work (such as details of the image's creator, date/time/location, tracked edits and more).[9] This is a creative solution to help newspapers, human rights watchdogs and others reassure the public about the authenticity and provenance of images they create and display. But *mandating* the use of such an authenticity standard (or prohibiting the distribution of materials without such standards) would be deeply problematic, because it would suppress the posting and sharing of lawful images whose creators lacked the resources or awareness to use a provenance tool, who face safety risks if their work can be traced back to them, or who simply do not want to do so.

The challenges of regulating deepfakes does not mean policymakers must sit idle. To the contrary, the PCAST Working Group on Generative AI can recommend concrete steps to increase transparency and accountability in the design, development and use of generative AI tools. The Working Group should consider how federal agencies and executive actions can advocate for and employ best practices and novel innovations to address potential harms.

Mandating transparency & disclosures of AI risks. Transparency is critical to helping users understand when they are interacting with AI generated content that may be false or misleading. Because some generative AI tools are designed to mimic human interaction, it may not always be clear to individuals when they are, for example, interacting with an actual campaign representative or just seeing AI-generated outputs. Generative AI tools are—by design—remarkably good at producing content that appears to be generated by a human. As discussed above, that reality increases the risks

[7] Sam Gregory, "Tracing trust: Why we must build authenticity infrastructure that works for all", *Witness*, May 2020, https://blog.witness.org/2020/05/authenticity-infrastructure/.
[8] Tiffany Hsu, "As Deepfakes Flourish, Countries Struggle With Response", *The New York Times*, Jan. 22, 2023, https://www.nytimes.com/2023/01/22/business/media/deepfake-regulation-difficulty.html.
[9] *See* Content Authenticity Initiative, https://contentauthenticity.org/.

associated with deep fakes, fraud, and widespread digital influence operations.[10]

Partially in response to these risks, companies and people have begun building systems to detect whether content is created by a generative AI system, though these are currently largely ineffective.[11] But developers of generative AI systems can facilitate the detection of AI-generated content by enabling their software to embed "watermarks."[12] If developers of generative AI systems were to commit to watermarking their outputs, it would be easier for users to know when they are seeing synthetic content. Some initial progress has been made on this front: In July 2023, seven leading AI companies made commitments to the White House that they would develop and deploy detection technology like watermarks for audio and visual content.[13]

Watermarking is not a perfect solution, however; even if all major AI companies were to watermark their outputs, users hoping to evade detection (perhaps in order to deceive a target audience) might turn to open-source generative AI systems configured to not watermark outputs. Users could also de-watermark text by, for example, passing watermarked outputs through another piece of software that paraphrases the text.[14] The tug-of-war between watermarking systems and users who might try to evade the watermarks is an area of active research. But watermarking may nonetheless provide a benefit, allowing detection of a significant amount of AI-generated content that a user might encounter online.

As a starting point, the PCAST Working Group could recommend that the developers of AI systems used in high-risk settings disclose how their tools are developed and designed, to subject them to independent testing using frameworks based on principles such as those set out in the Blueprint for an AI Bill of Rights and the NIST AI Risk Management Framework, and to share the analysis of those tests, including with the public and independent researchers, balancing concerns about the potential privacy and safety aspects of such disclosures. Again the recent White House commitments move usefully in this direction by promising to publicly report model limitations and consequent societal risks. Such steps would increase transparency and support meaningful public dialogue about how tools are developed and governed.

Examining how existing criminal and civil laws map onto harms created by new tools, and filling gaps. In some instances, the appropriate framework to address harms created by generative AI (and other AI systems) may be enforcement of existing laws. For example, people who use AI to perpetrate scams could be prosecuted for fraud, extortion, or harassment; face investigation by the Federal Trade

[10] *See* Josh A. Goldstein, et al., *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations* (Jan. 2023), https://arxiv.org/pdf/2301.04246.pdf

[11] Armin Alimardani & Emma Jane, We Pitted ChatGPT Against Tools for Detecting AI-Written Text, and the Results are Troubling, The Conversation (Feb. 19, 2023), https://theconversation.com/we-pitted-chatgpt-against-tools-fordetecting-ai-written-text-and-the-results-are-troubling-199774.

[12] *See* John Kirchenbauer, et al., *A Watermark for Large Language Models* (Jun. 6, 2023), https://arxiv.org/pdf/2301.10226.pdf.

[13] AI companies including OpenAI, Google, Meta, Microsoft and Amazon made commitments to the White House on July 21, 2023 to implement measures such as watermarking AI-generated content, though they did not release specifics about how the technology will work or when the measures would take effect. *See* Diane Bartz and Krystal Hu, *OpenAI, Google, others pledge to watermark AI content for safety - White House,* Thomson Reuters (July 21, 2023), https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/.

[14] *See* Kalpesh Krishna, et al., *Paraphrasing Evades Detectors of AI-Generated Text, But Retrieval is an Effective Defense* (Mar. 23, 2023) https://arxiv.org/pdf/2303.13408.pdf.

Commission for unfair and deceptive trade practices or the Federal Elections Commission for violating campaign laws; or face civil litigation for claims such as fraud, intentional infliction of emotional distress, harassment, defamation and intellectual property violations. The PCAST Working Group should evaluate where these existing legal frameworks adequately address emerging harms.[15] For example, election laws could be amended to require disclosure of the use of AI to generate campaign materials.

Courts and policymakers will also have to consider how to allocate responsibility among the multiple actors that may be involved in the creation of content using generative AI. For example, the developer of a model, the deployer of a chatbot using that model, and the user who provided a prompt may all plausibly bear at least partial responsibility for the creation of a particular deepfake or voter-suppressive content. A regulatory regime that allows them all to point the finger at each other may not provide the needed remedy or other recourse.

As regulators and courts begin to grapple with these and other complex issues, PCAST can shine a light and drive public discourse—including identifying gaps and recommendations for how to fill them. This could include recommending that the President commission reports by the GAO or federal agencies, or create an expert Commission to advance such work.

Advancing best practices for responsible design and governance of generative AI systems. There is an urgent need for companies developing generative AI systems to develop robust safety processes and other governance measures, as many of their CEOs have publicly acknowledged.[16] This can include steps ranging from well-developed content policies and technical safeguards that limit the creation of certain high-risk content or uses of the technology;[17] robust pre- and post-release testing to identify and address bias and potential harms; improved interfaces, labeling, and product descriptions to better educate users about the systems' limitations and risks of inaccurate results;[18] and safeguarding systems against security threats.

Governments in different countries are pressing companies on what these steps should look like.[19] Whether or not these steps are ripe for legislation, the PCAST Working Group can play a role in driving

[15] Four federal agencies recently announced their efforts to enforce existing laws to protect the American public from AI-related harms. *See* Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems, Apr. 25, 2023, https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf.

[16] *See e.g.,* Sam Altman, Oversight of A.I.: Rules for Artificial Intelligence Hearing before the U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, & the Law, 118th Cong. (2023), https://www.judiciary.senate.gov/committee- activity/hearings/oversight-of-ai-rules-for-artificial-intelligence; Sundar Pichai, "Why Google thinks we need to regulate AI", *Financial Times*, Jan. 20, 2020, https://www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04; Brad Smith, "Meeting the AI moment: advancing the future through responsible AI", Microsoft, Feb. 2, 2023, https://blogs.microsoft.com/on-the-issues/2023/02/02/responsible-ai-chatgpt-artificial-intelligence/.

[17] For example, OpenAI claims that its image generator DALL-E cannot create images of public figures, and that it restricts any "scaled" usage of its products for political purposes, such as the use of its AI to send out mass personalized emails to constituents. Reporters testing these claims have found significant exceptions and workarounds. Alexandra Ulmer and Anna Tong, "Deepfaking it: America's 2024 election collides with AI boom", *Reuters*, May 30, 2023, https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30/.

[18] Michal Luria, "Your ChatGPT Relationship Status Shouldn't Be Complicated", *WIRED*, Apr. 11, 2023, https://www.wired.com/story/chatgpt-social-roles-psychology/.

[19] Ryan Browne, "Europe takes aim at ChatGPT with what might soon be the West's first A.I. law. Here's what it means", *CNBC*, May 15, 2023, https://www.cnbc.com/2023/05/15/eu-ai-act-europe-takes-aim-at-chatgpt-with-landmark-regulation.html. The White House issued the AI Bill of Rights in October 2022 and the National Institute of Standards and Technology (NIST) followed in January 2023 with an AI Risk Management Framework, and officials have spoken about ways in which these map onto the risks posed by generative AI. *See* Blueprint for an AI Bill of Rights, https://www.whitehouse.gov/ostp/ai-bill-of-rights/; National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

forward these efforts—and, most critically, helping to ensure they are not taking place behind closed doors with only companies in attendance, but instead with meaningful participation from civil society and independent sources of expertise.

Scaling agencies' capacity to address deepfakes and boost authentic sources of information. It has long been said that the best remedy to combat undesirable speech is counterspeech[20]—but in our cacophonous information ecosystem, it takes work for counterspeech to be effective. There are steps policymakers can take to mature the systems that can help individuals better understand content authenticity and identify reliable sources of information. As one step, the government could increase funding and other efforts to support the development of technologies that assist in deepfake detection.[21] Policymakers could also support and foster awareness of voluntary efforts to authenticate content, funding research projects through the National Science Foundation and other programs, or raising awareness of private sector efforts to encourage the quick development of such work.[22]

PCAST should recommend that the Administration significantly ramp up efforts to equip key institutions so they can identify and debunk manipulated content that threatens national security, financial markets, election administration, public health and similar priority areas. The bipartisan Deepfake Task Force Act proposed last Congress provides a good bipartisan example of a path forward. That bill proposed the creation of a task force comprised of government and non-government experts to "investigate the feasibility of, and obstacles to, developing and deploying standards and technologies for determining digital content provenance", and created "a formal mechanism for interagency coordination and information sharing to facilitate the creation and implementation of a national strategy to address the growing threats posed by digital content forgeries."[23]

Capacity-building efforts could also include funding training, providing resources, and using oversight pressure to ensure public institutions take steps to best earn public trust when they speak out. To give one simple example, CDT research revealed that only 1 in 4 official election websites uses the trusted ".gov" domain managed by DHS, while other election officials use ".com" domains that can be easily spoofed.[24] The result is to undermine the role of such websites as a source for people to access trusted information about the administration of elections. Funding, education and oversight could help election officials address this simple vulnerability.

---

[20] *Whitney v. Cal.*, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring) ("If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence.").

[21] *See, e.g.,* IOGAN Act, Pub. L. No. 116-258 (2020), directing the National Science Foundation and the National Institute of Standards and Technology (NIST) to support research on generative adversarial networks.

[22] Shirin Ghaffary, "What will stop AI from flooding the internet with fake images?, *Vox*, Jun. 3, 2023, https://www.vox.com/technology/23746060/ai-generative-fake-images-photoshop-google-microsoft-adobe.

[23] Section 5709 of the National Defense Authorization Act of 2020 also took steps to improve government agency awareness and competency to address deepfakes. It directed the Director of National Intelligence to produce a report on the technological capabilities of foreign actors with respect to "machine-manipulated media, machine generated text, generative adversarial networks, and related machine-learning technologies", and analysis of the counter-technologies that have been or could be developed and deployed to address such uses, among other factors. National Defense Authorization Act of 2020, Pub. L. No. 116-92 (2019).

[24] William T. Adler et al, "Only 1 in 4 Election Websites Uses the .gov Domain. That's a Problem — and an Opportunity," *Center for Democracy & Technology*, Oct. 19, 2022, https://cdt.org/insights/only-1-in-4-election-websites-uses-the-gov-domain-thats-a-problem-and-an-opportunity/.