How AI Threatens Democracy

Sarah Kreps, Doug Kriner

Journal of Democracy, Volume 34, Number 4, October 2023, pp. 122-131 (Article)

Published by Johns Hopkins University Press
DOI: https://doi.org/10.1353/jod.2023.a907693

HOW AI THREATENS DEMOCRACY

Sarah Kreps and Doug Kriner

Sarah Kreps *is the John L. Wetherill Professor in the Department of Government, adjunct professor of law, and the director of the Tech Policy Institute at Cornell University.* ***Doug Kriner*** *is the Clinton Rossiter Professor in American Institutions in the Department of Government at Cornell University.*

Just a month after its introduction, ChatGPT, the generative artificial intelligence (AI) chatbot, hit 100-million monthly users, making it the fastest-growing application in history. For context, it took the video-streaming service Netflix, now a household name, three-and-a-half years to reach one-million monthly users. But unlike Netflix, the meteoric rise of ChatGPT and its potential for good or ill sparked considerable debate. Would students be able to use, or rather misuse, the tool for research or writing? Would it put journalists and coders out of business? Would it "hijack democracy," as one *New York Times* op-ed put it, by enabling mass, phony inputs to perhaps influence democratic representation?[1] And most fundamentally (and apocalyptically), could advances in artificial intelligence actually pose an existential threat to humanity?[2]

New technologies raise new questions and concerns of different magnitudes and urgency. For example, the fear that generative AI—artificial intelligence capable of producing new content—poses an existential threat is neither plausibly imminent, nor necessarily plausible. Nick Bostrom's paperclip scenario, in which a machine programmed to optimize paperclips eliminates everything standing in its way of achieving that goal, is not on the verge of becoming reality.[3] Whether children or university students use AI tools as shortcuts is a valuable pedagogical debate, but one that should resolve itself as the applications become more seamlessly integrated into search engines. The employment consequences of generative AI will ultimately be difficult to adjudicate since economies are complex, making it difficult to isolate the net effect of AI-instigated job losses versus industry gains. Yet the potential

Journal of Democracy Volume 34, Number 4 October 2023

consequences for democracy are immediate and severe. Generative AI threatens three central pillars of democratic governance: representation, accountability, and, ultimately, the most important currency in a political system—trust.

The most problematic aspect of generative AI is that it hides in plain sight, producing enormous volumes of content that can flood the media landscape, the internet, and political communication with meaningless drivel at best and misinformation at worst. For government officials, this undermines efforts to understand constituent sentiment, threatening the quality of democratic representation. For voters, it threatens efforts to monitor what elected officials do and the results of their actions, eroding democratic accountability. A reasonable cognitive prophylactic measure in such a media environment would be to believe nothing, a nihilism that is at odds with vibrant democracy and corrosive to social trust. As objective reality recedes even further from the media discourse, those voters who do not tune out altogether will likely begin to rely even more heavily on other heuristics, such as partisanship, which will only further exacerbate polarization and stress on democratic institutions.

Threats to Democratic Representation

Democracy, as Robert Dahl wrote in 1972, requires "the continued responsiveness of the government to the preferences of its citizens."[4] For elected officials to be responsive to the preferences of their constituents, however, they must first be able to discern those preferences. Public-opinion polls—which (at least for now) are mostly immune from manipulation by AI-generated content—afford elected officials one window into their constituents' preferences. But most citizens lack even basic political knowledge, and levels of policy-specific knowledge are likely lower still.[5] As such, legislators have strong incentives to be the most responsive to constituents with strongly held views on a specific policy issue and those for whom the issue is highly salient. Written correspondence has long been central to how elected officials keep their finger on the pulse of their districts, particularly to gauge the preferences of those most intensely mobilized on a given issue.[6]

In an era of generative AI, however, the signals sent by the balance of electronic communications about pressing policy issues may be severely misleading. Technological advances now allow malicious actors to generate false "constituent sentiment" at scale by effortlessly creating unique messages taking positions on any side of a myriad of issues. Even with old technology, legislators struggled to discern between human-written and machine-generated communications.

In a field experiment conducted in 2020 in the United States, we composed advocacy letters on six different issues and then used those letters to train what was then the state-of-the-art generative AI model, GPT-3,

to write hundreds of left-wing and right-wing advocacy letters. We sent randomized AI- and human-written letters to 7,200 state legislators, a total of about 35,000 emails. We then compared response rates to the human-written and AI-generated correspondence to assess the extent to which legislators were able to discern (and therefore not respond to) machine-written appeals. On three issues, the response rates to AI- and human-written messages were statistically indistinguishable. On three other issues, the response rates to AI-generated emails were lower—but only by 2 percent, on average.[7] This suggests that a malicious actor capable of easily generating thousands of unique communications could potentially skew legislators' perceptions of which issues are most important to their constituents as well as how constituents feel about any given issue.

In the same way, generative AI could strike a double blow against the quality of democratic representation by rendering obsolete the public-comment process through which citizens can seek to influence the actions of the regulatory state. Legislators necessarily write statutes in broad brushstrokes, granting administrative agencies considerable discretion not only to resolve technical questions requiring substantive expertise (e.g., specifying permissible levels of pollutants in the air and water), but also to make broader judgements about values (e.g., the acceptable tradeoffs between protecting public health and not unduly restricting economic growth).[8] Moreover, in an era of intense partisan polarization and frequent legislative gridlock on pressing policy priorities, U.S. presidents have increasingly sought to advance their policy agendas through administrative rulemaking.

Moving the locus of policymaking authority from elected representatives to unelected bureaucrats raises concerns of a democratic deficit. The U.S. Supreme Court raised such concerns in *West Virginia v. EPA* (2022), articulating and codifying the major questions doctrine, which holds that agencies do not have authority to effect major changes in policy absent clear statutory authorization from Congress. The Court may go even further in the pending *Loper Bright Enterprises v. Raimondo* case and overturn the *Chevron* doctrine, which has given agencies broad latitude to interpret ambiguous congressional statutes for nearly three decades, thus further tightening the constraints on policy change through the regulatory process.

Not everyone agrees that the regulatory process is undemocratic, however. Some scholars argue that the guaranteed opportunities for public participation and transparency during the public-notice and comment period are "refreshingly democratic,"[9] and extol the process as "democratically accountable, especially in the sense that decision-making is kept above board and equal access is provided to all."[10] Moreover, the advent of the U.S. government's electronic-rulemaking (e-rulemaking) program in 2002 promised to "enhance public participation . . . so as

to foster better regulatory decisions" by lowering the barrier to citizen input.[11] Of course, public comments have always skewed, often heavily, toward interests with the most at stake in the outcome of a proposed rule, and despite lowering the barriers to engagement, e-rulemaking did not alter this fundamental reality.[12]

Despite its flaws, the direct and open engagement of the public in the rulemaking process helped to bolster the democratic legitimacy of policy change through bureaucratic action. But the ability of malicious actors to use generative AI to flood e-rulemaking platforms with limitless unique comments advancing a particular agenda could make it all but impossible for agencies to learn about genuine public preferences. An early (and unsuccessful) test case arose in 2017, when bots flooded the Federal Communications Commission with more than eight-million comments advocating repeal of net neutrality during the open comment period on proposed changes to the rules.[13] This "astroturfing" was detected, however, because more than 90 percent of those comments were not unique, indicating a coordinated effort to mislead rather than genuine grassroots support for repeal. Contemporary advances in AI technology can easily overcome this limitation, rendering it exceedingly difficult for agencies to detect which comments genuinely represent the preferences of interested stakeholders.

Threats to Democratic Accountability

A healthy democracy also requires that citizens be able to hold government officials accountable for their actions—most notably, through free and fair elections. For ballot-box accountability to be effective, however, voters must have access to information about the actions taken in their name by their representatives.[14] Concerns that partisan bias in the mass media, upon which voters have long relied for political information, could affect election outcomes are longstanding, but generative AI poses a far greater threat to electoral integrity.

As is widely known, foreign actors exploited a range of new technologies in a coordinated effort to influence the 2016 U.S. presidential election. A 2018 Senate Intelligence Committee report stated:

> Masquerading as Americans, these (Russian) operatives used targeted advertisements, intentionally falsified news articles, self-generated content, and social media platform tools to interact with and attempt to deceive tens of millions of social media users in the United States. This campaign sought to polarize Americans on the basis of societal, ideological, and racial differences, provoked real world events, and was part of a foreign government's covert support of Russia's favored candidate in the U.S. presidential election.[15]

While unprecedented in scope and scale, several flaws in the influence campaign may have limited its impact.[16] The Russian operatives' social-

media posts had subtle but noticeable grammatical errors that a native speaker would not make, such as a misplaced or missing article—telltale signs that the posts were fake. ChatGPT, however, makes every user the equivalent of a native speaker. This technology is already being used to create entire spam sites and to flood sites with fake reviews. The tech website *The Verge* flagged a job seeking an "AI editor" who could generate "200 to 250 articles per week," clearly implying that the work would be done via generative AI tools that can churn out mass quantities of content in fluent English at the click of the editor's "regenerate" button.[17] The potential political applications are myriad. Recent research shows that AI-generated propaganda is just as believable as propaganda written by humans.[18] This, combined with new capacities for microtargeting, could revolutionize disinformation campaigns, rendering them far more effective than the efforts to influence the 2016 election.[19] A steady stream of targeted misinformation could skew how voters perceive the actions and performance of elected officials to such a degree that elections cease to provide a genuine mechanism of accountability since the premise of what people are voting on is itself factually dubious.[20]

Threats to Democratic Trust

Advances in generative AI could allow malicious actors to produce misinformation, including content microtargeted to appeal to specific demographics and even individuals, at scale. The proliferation of social-media platforms allows the effortless dissemination of misinformation, including its efficient channeling to specific constituencies. Research suggests that although readers across the political spectrum cannot distinguish between a range of human-made and AI-generated content (finding it all plausible), misinformation will not necessarily change readers' minds.[21] Political persuasion is difficult, especially in a polarized political landscape.[22] Individual views tend to be fairly entrenched, and there is little that can change people's prior sentiments.

The risk is that as inauthentic content—text, images, and video—proliferates online, people simply might not know what to believe and will therefore distrust the entire information ecosystem. Trust in media is already low, and the proliferation of tools that can generate inauthentic content will erode that trust even more. This, in turn, could further undermine perilously low levels of trust in government. Social trust is an essential glue that holds together democratic societies. It fuels civic engagement and political participation, bolsters confidence in political institutions, and promotes respect for democratic values, an important bulwark against democratic backsliding and authoritarianism.[23]

Trust operates in multiple directions. For political elites, responsiveness requires a trust that the messages they receive legitimately

represent constituent preferences and not a coordinated campaign to misrepresent public sentiment for the sake of advancing a particular viewpoint. Cases of "astroturfing" are nothing new in politics, with examples in the United States dating back at least to the 1950s.[24] However, advances in AI threaten to make such efforts ubiquitous and more difficult to detect.

For citizens, trust can motivate political participation and engagement, and encourage resistance against threats to democratic institutions and practices. The dramatic decline in Americans' trust in government over the past half century is among the most documented developments in U.S. politics.[25] While many factors have contributed to this erosion, trust in the media and trust in government are intimately linked.[26] Bombarding citizens with AI-generated content of dubious veracity could seriously threaten confidence in the media, with severe consequences for trust in the government.

Mitigating the Threats

Although understanding the motives and technology is an important first step in framing the problem, the obvious next step is to formulate prophylactic measures. One such measure is to train and deploy the same machine-learning models that generate AI to detect AI-generated content. The neural networks used in artificial intelligence to create text also "know" the types of language, words, and sentence structures that produce that content and can therefore be used to discern patterns and hallmarks of AI-generated versus human-written text. AI detection tools are proliferating quickly and will need to adapt as the technology adapts, but a "Turnitin"-style model—like those that teachers use to detect plagiarism in the classroom—may provide a partial solution. These tools essentially use algorithms to identify patterns within the text that are hallmarks of AI-generated text, although the tools will still vary in their accuracy and reliability.

Even more fundamentally, the platforms responsible for generating these language models are increasingly aware of what it took many years for social-media platforms to realize—that they have a responsibility in terms of what content they produce, how that content is framed, and even what type of content is proscribed. If you query ChatGPT about how generative AI could be misused against nuclear command and control, the model responds with "I'm sorry, I cannot assist with that." OpenAI, the creator of ChatGPT, is also working with external researchers to democratize the values encoded in their algorithms, including which topics should be off limits for search outputs and how to frame the political positions of elected officials. Indeed, as generative AI becomes more ubiquitous, these platforms have a responsibility not just to create the technology but to do so with a set of values that is

ethically and politically informed. The question of who gets to decide what is ethical, especially in polarized, heavily partisan societies, is not new. Social-media platforms have been at the center of these debates for years, and now the generative AI platforms are in an analogous situation. At the least, elected public officials should continue to work closely with these private firms to generate accountable, transparent algorithms. The decision by seven major generative AI firms to commit to voluntary AI safeguards, in coordination with the Biden Administration, is a step in the right direction.

The platforms responsible for generating language models are increasingly aware of what it took many years for social-media platforms to realize—that they have a responsibility in terms of what content they produce, how that content is framed, and even what type of content is proscribed.

Finally, digital-literacy campaigns have a role to play in guarding against the adverse effects of generative AI by creating a more informed consumer. Just as neural networks "learn" how generative AI talks and writes, so too can individual readers themselves. After we debriefed the state legislators in our study about its aims and design, some said that they could identify AI-generated emails because they know how their constituents write; they are familiar with the standard vernacular of a constituent from West Virginia or New Hampshire. The same type of discernment is possible for Americans reading content online. Large language models such as ChatGPT have a certain formulaic way of writing—perhaps having learned a little too well the art of the five-paragraph essay.

When we asked the question, "Where does the United States have missile silos?" ChatGPT replied with typical blandness: "The United States has missile silos located in several states, primarily in the central and northern parts of the country. The missile silos house intercontinental ballistic missiles (ICBMs) as part of the U.S. nuclear deterrence strategy. The specific locations and number of missile silos may vary over time due to operational changes and modernization efforts."

There is nothing wrong with this response, but it is also very predictable to anyone who has used ChatGPT somewhat regularly. This example is illustrative of the type of language that AI models often generate. Studying their content output, regardless of the subject, can help people to recognize clues indicating inauthentic content.

More generally, some of the digital-literacy techniques that have already gained currency will likely apply in a world of proliferating AI-generated texts, videos, and images. It should be standard practice for everyone to verify the authenticity or factual accuracy of digital content

across different media outlets and to cross-check anything that seems dubious, such as the viral (albeit fake) image of the pope in a Balenciaga puffy coat, to determine whether it is a deep fake or real. Such practices should also help in discerning AI-generated material in a political context, for example, on Facebook during an election cycle.

Unfortunately, the internet remains one big confirmation-bias machine. Information that seems plausible because it comports with a person's political views may be less likely to drive that person to check the veracity of the story. In a world of easily generated fake content, many people may have to walk a fine line between political nihilism—that is, not believing anything or anyone other than their fellow partisans—and healthy skepticism. Giving up on objective fact, or at least the ability to discern it from the news, would shred the trust on which democratic society must rest. But we are no longer living in a world where "seeing is believing." Individuals should adopt a "trust but verify" approach to media consumption, reading and watching but exercising discipline in terms of establishing the material's credibility.

New technologies such as generative AI are poised to provide enormous benefits to society—economically, medically, and possibly even politically. Indeed, legislators could use AI tools to help identify inauthentic content and also to classify the nature of their constituents' concerns, both of which would help lawmakers to reflect the will of the people in their policies. But artificial intelligence also poses political perils. With proper awareness of the potential risks and the guardrails to mitigate against their adverse effects, however, we can preserve and perhaps even strengthen democratic societies.

NOTES

1. Nathan E. Sanders and Bruce Schneier, "How ChatGPT Hijacks Democracy," *New York Times*, 15 January 2023, *www.nytimes.com/2023/01/15/opinion/ai-chatgpt-lobbying-democracy.html*.

2. Kevin Roose, "A.I. Poses 'Risk of Extinction,' Industry Leaders Warn," *New York Times*, 30 May 2023, *www.nytimes.com/2023/05/30/technology/ai-threat-warning.html*.

3. Alexey Turchin and David Denkenberger, "Classification of Global Catastrophic Risks Connected with Artificial Intelligence," *AI & Society* 35 (March 2020): 147–63.

4. Robert Dahl, *Polyarchy: Participation and Opposition* (New Haven: Yale University Press, 1972), 1.

5. Michael X. Delli Carpini and Scott Keeter, *What Americans Know about Politics and Why it Matters* (New Haven: Yale University Press, 1996); James Kuklinski et al., "'Just the Facts Ma'am': Political Facts and Public Opinion," *Annals of the American Academy of Political and Social Science* 560 (November 1998): 143–54; Martin Gilens, "Political Ignorance and Collective Policy Preferences," *American Political Science Review* 95 (June 2001): 379–96.

6. Andrea Louise Campbell, *How Policies Make Citizens: Senior Political Activism and the American Welfare State* (Princeton: Princeton University Press, 2003); Paul Martin and Michele Claibourn, "Citizen Participation and Congressional Responsiveness: New Evidence that Participation Matters," *Legislative Studies Quarterly* 38 (February 2013): 59–81.

7. Sarah Kreps and Doug L. Kriner, "The Potential Impact of Emerging Technologies on Democratic Representation: Evidence from a Field Experiment," *New Media and Society* (2023), *https://doi.org/10.1177/14614448231160526*.

8. Elena Kagan, "Presidential Administration," *Harvard Law Review* 114 (June 2001): 2245–2353.

9. Michael Asimow, "On Pressing McNollgast to the Limits: The Problem of Regulatory Costs," *Law and Contemporary Problems* 57 (Winter 1994): 127, 129.

10. Kenneth F. Warren, *Administrative Law in the Political System* (New York: Routledge, 2018).

11. Committee on the Status and Future of Federal E-Rulemaking, American Bar Association, "Achieving the Potential: The Future of Federal E-Rulemaking," 2008, *https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=2505&context=facpub*.

12. Jason Webb Yackee and Susan Webb Yackee, "A Bias toward Business? Assessing Interest Group Influence on the U.S. Bureaucracy," *Journal of Politics* 68 (February 2006): 128–39; Cynthia Farina, Mary Newhart, and Josiah Heidt, "Rulemaking vs. Democracy: Judging and Nudging Public Participation That Counts," *Michigan Journal of Environmental and Administrative Law* 2, issue 1 (2013): 123–72.

13. Edward Walker. "Millions of Fake Commenters Asked the FCC to End Net Neutrality: 'Astroturfing' Is a Business Model," *Washington Post* Monkey Cage blog, 14 May 2021, *www.washingtonpost.com/politics/2021/05/14/millions-fake-commenters-asked-fcc-end-net-neutrality-astroturfing-is-business-model/*.

14. Adam Przeworksi, Susan C. Stokes, and Bernard Manin, eds., *Democracy, Accountability, and Representation* (New York: Cambridge University Press, 1999).

15. Report of the Select Committee on Intelligence United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Senate Report 116–290, *www.intelligence.senate.gov/publications/report-select-committee-intelligence-united-states-senate-russian-active-measures*.

16. On the potentially limited effects of 2016 election misinformation more generally, see Andrew M. Guess, Brendan Nyhan, and Jason Reifler, "Exposure to Untrustworthy Websites in the 2016 US Election," *Nature Human Behavior* 4 (2020): 472–80.

17. James Vincent, "AI Is Killing the Old Web, and the New Web Struggles to be Born," *The Verge*, 26 June 2023, *www.theverge.com/2023/6/26/23773914/ai-large-language-models-data-scraping-generation-remaking-web*.

18. Josh Goldstein et al., "Can AI Write Persuasive Propaganda?" working paper, 8 April 2023, *https://osf.io/preprints/socarxiv/fp87b*.

19. Sarah Kreps, "The Role of Technology in Online Misinformation," Brookings Institution, June 2020, *www.brookings.edu/articles/the-role-of-technology-in-online-misinformation*.

20. In this way, AI-generated misinformation could greatly heighten "desensitization"—the relationship between incumbent performance and voter beliefs—undermining democratic accountability. See Andrew T. Little, Keith E. Schnakenberg, and Ian R. Turn-

er, "Motivated Reasoning and Democratic Accountability," *American Political Science Review* 116 (May 2022): 751–67.

21. Sarah Kreps, R. Miles McCain, and Miles Brundage, "All the News that's Fit to Fabricate," *Journal of Experimental Political Science* 9 (Spring 2022): 104–17.

22. Kathleen Donovan et al., "Motivated Reasoning, Public Opinion, and Presidential Approval" *Political Behavior* 42 (December 2020): 1201–21.

23. Mark Warren, ed., *Democracy and Trust* (New York: Cambridge University Press, 1999); Robert Putnam, *Bowling Alone: The Collapse and Revival of American Community* (New York: Simon and Schuster, 2000); Marc Hetherington, *Why Trust Matters: Declining Political Trust and the Demise of American Liberalism* (Princeton: Princeton University Press, 2005); Pippa Norris, ed., *Critical Citizens: Global Support for Democratic Governance* (New York: Oxford University Press, 1999); Steven Levitsky and Daniel Ziblatt, *How Democracies Die* (New York: Crown, 2019).

24. Lewis Anthony Dexter, "What Do Congressmen Hear: The Mail," *Public Opinion Quarterly* 20 (Spring 1956): 16–27.

25. See, among others, Pew Research Center, "Public Trust in Government: 1958–2022," 6 June 2022, *www.pewresearch.org/politics/2022/06/06/public-trust-in-government-1958-2022/#:~:text=Only%20two%2Din%2Dten%20Americans,least%20most%20of%20the%20time*.

26. Thomas Patterson, *Out of Order* (New York: Knopf, 1993); Joseph N. Cappella and Kathleen Hall Jamieson, "News Frames, Political Cynicism, and Media Cynicism," *Annals of the American Academy of Political and Social Science* 546 (July 1996): 71–84.