

The Paradox of Unwanted Democratic AI

Seth D. Baum
Global Catastrophic Risk Institute
<http://gcri.org> * <https://sethbaum.com>

Forthcoming in *AI & Society*, DOI 10.1007/s00146-026-03386-y.
This version 21 September 2026.

Suppose that (1) there is an AI system designed, in democratic fashion, to follow some measure of the aggregate values of humanity, and (2) according to this measure of the aggregate values of humanity, it is preferred that this AI system does not exist. In this situation, what should the AI system do? Arguably, the system should self-destruct, because that is what its values dictate. It follows that there cannot be an unwanted democratic AI system. This is the paradox of unwanted democratic AI.

The AI system in (1) is a common idea in AI ethics, related to concepts such as value alignment, human compatibility, and social choice ethics (Baum 2020; 2025). This approach to AI ethics may be regarded as democratic because the use of aggregate values to guide AI system behavior is analogous to the use of voting to guide the behavior of democracies. In contrast, the top-down imposition of values on the AI system may be regarded as authoritarian or dictatorial because the system's values are dictated by some narrower group or even just a single individual.

Now consider a project for the research, development, and deployment of AI systems. The project expresses that it wishes to advance democracy, including in the global contest between democracy and authoritarianism; it further expresses that it views its AI as a contribution to democracy. However, the public does not support this, and instead prefers that the project did not exist. In this situation, what should the AI project do? Arguably, it should disband itself on grounds that there can be no unwanted democratic AI project. This is another form of the paradox of unwanted democratic AI, for AI projects instead of AI systems.

This is not a mere thought experiment. At present, it is common for AI projects, or at least their leadership, to express democratic values, including in the context of a competition against authoritarian dictatorship (e.g., Altman 2024; Amodei 2026). The projects referenced here—OpenAI and Anthropic—are both based in the United States. US public opinion on AI is mixed and increasingly negative (e.g., Kennedy et al. 2025; Smith 2026), and there is growing civic advocacy against some aspects of AI projects, especially data centers (Shapero 2026). This is not an unambiguous case of unwanted democratic AI, but there are clear signs pointing in that direction.

Some future scenarios could go further. In a remarkable essay titled “Eat your AI slop or China wins”, Bellafiore (2025) argues that the US should embrace aggressive AI development to outcompete China, even if that causes major harms such as “mass unemployment, retreat into delusional virtuality, learned helplessness among all who can no longer function without ChatGPT, and so forth”. Such conditions would place an extreme burden on the public. Given that conditions are not yet nearly this bad, and still public opinion is already generally negative, it seems likely that if conditions were to worsen in these ways, public opposition to AI would substantially intensify.

Other future AI scenarios would likely garner more public support. AI is a very broad category; many AI applications are uncontroversial and have robust public support. For example, Americans today show strong support for the use of AI in weather forecasting, medicine development, and identifying cases of fraud and financial crime (Kennedy et al. 2025). It may therefore follow that the

public would especially support future scenarios involving narrow AI designed for some specific applications instead of artificial general intelligence (AGI).

AGI is also heavily associated with extreme catastrophic risks. Suffice to say, scenarios in which AI causes extreme catastrophe are unlikely to garner public support. Given uncertainty about which AI technologies might cause catastrophe, democratic AI faces important questions about how to weigh catastrophic risks alongside other factors. Already, there has been strong bipartisan public support for legislation addressing catastrophic AI risks (e.g., Shelter 2025), suggesting a low public appetite for catastrophic AI risk and potentially also for AGI instead of narrow AI.

Meanwhile, AI also has other implications for democracy, at least for certain conceptions of democracy. In political science research, democracy as the aggregation of citizen preferences via elections is a minimalist or “thin” conception; more expansive or “thicker” conceptions also include other aspects such as citizen participation in public deliberation and issue advocacy (Han 2014; Bächtiger et al. 2018; Coeckelbergh 2023).

AI is relevant to these other aspects of democracy, often with harmful effects. Specifically, AI may: facilitate the creation and propagation of misinformation at scale, such as via synthetic video deepfakes, which reduces citizens’ ability to form accurate political preferences and maintain the accountability of their elected governments (Coeckelbergh 2023); function as an information platform, thereby reducing the advertising revenues of news media companies, which further reduces citizens’ ability to maintain accountability of elected governments (Jungherr 2023); facilitate the creation and popularization of ballot initiatives and other citizen policy initiatives, thereby crowding out traditional human-run issue advocacy groups (Altman 2026); and automate significant portions of the economy, concentrating wealth and political power within a narrow elite that owns and operates the automation technology (Jungherr 2023).

The above examples all involve AI harming democracy. It is also possible for AI to benefit democracy, such as by making online conversations less divisive (Argyle et al. 2023). However, the current literature suggests that the overall effect of AI on democracy may be strongly negative. This places additional pressure on democratic AI projects. For them to support democracy, it may be necessary to do more than just maintain favorable public opinion: they may need to also address other potential harms to democracy from AI. Otherwise, their AI projects may substantially weaken democracy, and democracy may likewise be better off without them.

The term “democratic” AI, with scare quotes, may be used for AI projects that express democratic values while weakening democracy.

Despite the threats of “democratic” AI, it is important to account for the geopolitical dynamics. Bellafiore (2025) spells out what is arguably the greatest challenge for democratic AI: the prospect of competition from authoritarian regimes. The concern is that AI may become such a powerful technology that having an advantage in AI could confer global domination. This complicates the picture. “Democratic” AI projects may, on their own, substantially weaken democracy, and yet democracies may still have reason to want them, because without them, democracy may get overrun by competing authoritarian regimes.

This is a very bleak scenario for democracy. It implies that democracy is doomed to suffer an extreme decline and the only question is how extreme the decline will be. In this scenario, the “democratic” AI is the lesser of two evils; it is the only hope for maintaining any semblance of democracy, however limited it may be. However, this scenario depends heavily on the character of future AI technology. Current AI technology does not confer global domination. The prospects for future AI technology to be that powerful is deeply uncertain and controversial (Bengio et al. 2026).

If future AI technology will not confer global domination, then democracy does not face a bleak choice between two evils. In that case, competition with authoritarian regimes on AI resembles

competition in other economic sectors, such as renewable energy and electric vehicles, as well as in other military technologies, such as drones and stealth aircraft. These domains of competition certainly matter, but they can be more readily balanced against the health of democracy. If the choice is between maintaining a stronger democracy or gaining market share in one sector of the economy, it may be entirely reasonable to choose democracy by heavily regulating or disbanding “democratic” AI projects.

Even if AI technology could confer global domination, there is still another hopeful path that democracy can pursue. Altman (2024) and Bellafiore (2025) claim that “there is no third option” between democratic and authoritarian AI, but this is not true. The third option is diplomacy. Specifically, it is diplomacy to avoid, on a multilateral or even global basis, AGI and other extreme AI technologies. Diplomacy may be difficult, but it is unambiguously an option. Diplomacy does not guarantee a positive outcome, but neither does the aggressive pursuit of democratic AI. Furthermore, given the potential for AI to be unwanted and to substantially weaken democracy, diplomacy may be the only option for achieving strongly democratic outcomes.

Finally, one must account for the major conflicts of interest. AI projects like OpenAI and Anthropic are not neutral actors within democracies. They are private corporations that act, to a significant extent, in pursuit of market share and financial profit. Their institutional success can conflict with the democratic values they sometimes express. This occurs, for example, when they take financial investments from authoritarian governments (Robison 2025).

The AI projects’ conflict of interest calls into question their commitment to democracy. Do they actually support democracy, or are they just saying it for their own benefit? Maintaining a pro-democracy appearance could help the projects dodge regulations and attract support from democratic governments. If and when AI becomes so powerful that it confers global domination, then the AI projects could drop their pro-democracy ruse, because at that point democracies will have lost the ability to keep it in check.

AI projects’ conflict of interest gives democracies further reason to heavily regulate or even ban some of them, especially projects for the pursuit of AGI and other extreme forms of AI technology. Alternatively, democratic governments could consider nationalizing these AI projects, bringing them under full democratic control. Such measures may not be in the financial interest of the AI projects or their investors, but if the measures are better for democracy, they may nonetheless be worth pursuing.

To revisit the paradox: Should democratic AI exist? The answer is clearly yes for some limited forms of AI, such as narrow AI for weather forecasting or financial crime detection. However, for other forms of AI, especially for more extreme potential future forms such as AGI, the answer may well be no; ditto for AI projects pursuing such extreme AI.

AI projects that pursue extreme AI while professing to be democratic have a high standard to meet. From the citizen’s vantage point, the conflict of interest is alarming. These projects should do what they can to support democracy, such as by: restraining the spread of tools that enable the creation and spread of misinformation and that can marginalize citizen policy initiatives; funding independent journalism that holds powerful actors accountable; supporting policy that reduces concentrations of wealth and power; supporting diplomacy to reduce geopolitical competition for AI; and being prepared to shut down the AI project or at least restrict its activities if it is unwanted and unhelpful for democracy.

For their part, citizens can help by organizing themselves into advocacy coalitions (Han 2014). These coalitions could pursue a suite of complementary objectives: advancing sound AI policies to ensure that AI company behavior follows the public will; putting pressure on AI companies to have corporate governance in the public interest (Cihon et al. 2021); advancing broader policies for reducing the corruption of democracies by large corporations and wealthy individuals (Page and Gilens 2020); and electing politicians who will also support all of the above. Both the organizing itself and the

potential outcomes of the organizing would help to strengthen democracy while making it easier to achieve sound AI policies.

References

- Altman D (2026) The AI democracy dilemma. *Journal of Democracy* 37(1):32-44
- Altman S (2024) Who will control the future of AI? *Washington Post*, 25 July, <https://www.washingtonpost.com/opinions/2024/07/25/sam-altman-ai-democracy-authoritarianism-future>
- Amodei D (2026) The adolescence of technology: Confronting and overcoming the risks of powerful AI. <https://www.darioamodei.com/essay/the-adolescence-of-technology>
- Argyle LP, Bail CA, Busby EC, Gubler JR, Howe T, Rytting C, Sorensen T, Wingate D (2023) Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences* 120(41):e2311627120, DOI 10.1073/pnas.2311627120
- Bächtiger A, Dryzek JS, Mansbridge J, Warren ME (Editors) (2018). *The Oxford Handbook of Deliberative Democracy*. Oxford University Press, Oxford
- Baum SD (2020). Social choice ethics in artificial intelligence. *AI & Society* 35(1): 165-176, DOI 10.1007/s00146-017-0760-1
- Baum SD (2025) Manipulating aggregate societal values to bias AI social choice ethics. *AI and Ethics* 5(2): 1615-1627, DOI 10.1007/s43681-024-00495-6
- Bellafiore R (2025) Eat your AI slop or China wins. *The New Atlantis* 81:62-69. <https://www.thenewatlantis.com/publications/eat-your-ai-slop-or-china-wins>
- Bengio Y, Clare S, Prunkl C, Andriushchenko M, Bucknall B, Murray M, et al. (2026). *International AI Safety Report 2026*. UK Department of Science Innovation and Technology, London, <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>
- Coeckelbergh M (2023) Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics* 3:1341-1350.
- Cihon P, Schuett J, Baum SD (2021). Corporate governance of artificial intelligence in the public interest. *Information* 12:275, DOI 10.3390/info12070275.
- Han H (2014) *How Organizations Develop Activists: Civic Associations and Leadership in the 21st Century*. Oxford University Press, Oxford
- Jungherr A (2023) Artificial intelligence and democracy: A conceptual framework. *Social Media + Society* 9(3):1-14, DOI 10.1177/20563051231186353
- Kennedy B, Yam E, Kikuchi E, Pula I, Fuentes J (2025) How Americans view AI and its impact on people and society. *Pew Research Center*, 17 September. <https://www.pewresearch.org/science/2025/09/17/how-americans-view-ai-and-its-impact-on-people-and-society>
- Page BI, Gilens M (2020) *Democracy in America? What Has Gone Wrong and What We Can Do About It*. University of Chicago Press, Chicago
- Robison K (2025) Leaked memo: Anthropic CEO says the company will pursue Gulf State investments after all. *Wired*, 21 July, <https://www.wired.com/story/anthropic-dario-amodei-gulf-state-leaked-memo>
- Shapero J (2026) Anti-data center measures gain traction at state, local level. *The Hill*, 23 April. <https://thehill.com/policy/technology/5843665-data-center-backlash-grows>
- Shelter M (2025) New poll: 84% of New Yorkers support RAISE Act; Hochul's approval surges. *Empire Report*, 9 June. <https://empirereportnewyork.com/new-poll-84-of-new-yorkers-support-raise-act-hochuls-approval-surges>

Smith A (2026) Poll: Majority of voters say risks of AI outweigh the benefits. NBC News, 10 March.
<https://www.nbcnews.com/politics/politics-news/poll-majority-voters-say-risks-ai-outweigh-benefits-rcna262196>