

Enabling Civil Society's Role in Benchmarking AI

By Benjamin Jensen and Ian Reynolds

KEY TAKEAWAYS

- Benchmarks, which are studies that evaluate AI performance on domain-specific tasks, are a key tool for evaluating how foundation models and AI agents will impact society.
- Civil society organizations can play a fundamental role in ensuring that AI model benchmarking and evaluation processes are robust, accountable, and work in the public, democratic interest.
- Transparency, accountability, and domain expertise are critical in designing AI benchmarks. Benchmark design must consider a range of stakeholders, including civil society, public institutions, and the private sector.
- Congress can play a key role in creating a legislative environment that enables civil society associations to meaningfully contribute to AI development and push the technology's governance in a more democratic direction.
- AI governance efforts should be bottom up and be attuned to public needs in contexts ranging from local government services to national security and foreign policymaking.

BACKGROUND AND CONTEXT

AI agents are set to shape how decisions across broader society are made. As AI is further integrated into critical decision environments, it is essential that it be transparent, accountable, and robust. To assure that AI agents work in the best interest of the American people, and adhere to shared democratic values, research from the CSIS Futures Lab argues that the United States' civil society associations can play a key role in governing AI. Civil society organizations, if properly empowered, can inject important inputs into the process of benchmarking and evaluating AI, helping to support the technology's adoption across U.S. society in a fashion that works for the practical needs of Americans.

CSIS's Futures Lab suggests that building a bottom-up process of benchmarking AI models, rooted in the domain-specific needs of civil society organizations and the associative potential of American citizens, can assist in improving AI literacy across society while also improving the performance of the technology. Thus far, the Trump administration has focused on promoting an innovation-forward AI policy; however, the administration must also ensure that, in the context of AI, issues relevant to the public good are not swept aside by more powerful corporate actors and political interests.

LEGISLATIVE AND POLICY IMPLICATIONS

Executive guidance under both the Biden and Trump administrations has sought to promote AI adoption across U.S. society. Moreover, congressional action has attempted to democratize access to AI through initiatives such as the CREATE AI Act. As the federal government further pursues robust national AI policy, to truly ensure that AI works for the broader American public, legislators must incentivize civil society organizations to play a role in AI governance in general and model benchmarking and evaluation in particular. Public policy should endeavor to enable diverse sets of civil society associations, from local community organizations to mainstream universities and think tanks, to assist in providing domain-specific expertise to processes of benchmarking and evaluating AI models. The result can be a form of new associative action in which diverse actors across American society come together to ensure that advances in AI support democratic values and the broader social good.

RECOMMENDATIONS

- Congress should consider providing legislative support to civil society associations through funding organizations such as think tanks, community groups, and universities oriented toward benchmarking efforts and increasing AI literacy across civil society.
- Congress should consider creating tax incentives for collaborative benchmarking, require transparent evaluation reporting standards, and encourage voluntary commitments from AI firms to include civil society groups in technology development efforts.
- Congress should consider holding routine hearings on model evaluation results and processes to ensure accountability and transparency.
- Legislative efforts should encourage foundation-led independent funding efforts to support cross-disciplinary, bottom-up, benchmarking efforts.

Contact Information

For more information, contact: **Chloe Himmel** at 202.775.3186 or chimmel@csis.org.

CHALLENGES AND RISKS

First, a clear risk following from integrating AI across U.S. society is not having the proper tools to evaluate its performance in critical decision domains. Second, further dependence on digital technologies like AI runs the risk of contributing to social isolation among the public, facilitating the rise of anti-democratic outcomes and polarization. A third risk is the potential co-option of tech development and deployment by corporate organizations that are driven by financial interests and not motivated to ensure AI agents are accountable and transparent. In combination, these three factors could contribute to detrimental impacts on AI's performance, along with an undesirable degradation of democratic governance. Ensuring that AI evaluation and benchmarking is driven from the bottom up, and is transparent and accountable, can help to address these risks. It is in the United States' long-term interest to ensure that AI is not a driver of social isolation, but instead a focal point of associative action.

Additional Resources

Ben Jensen and Ian Reynolds, *Why Tocqueville Would Embrace AI Benchmarking: Charting a Path for the Future of Democracy in the Age of Artificial Intelligence* (Washington, DC: CSIS, July 2025), <https://www.csis.org/analysis/why-tocqueville-would-embrace-ai-benchmarking-charting-path-future-democracy-age>.

CSIS Futures Lab, "Critical Foreign Policy Decisions Benchmark," Digital Series, <https://www.csis.org/programs/futures-lab/projects/critical-foreign-policy-decisions-benchmark>.