DAY ONE PROJECT | INSTITUTE FOR PROGRESS

Creating an AI Testbed for Government

Tina Huang

January 2022

The Day One Project Progress Studies Policy Accelerator is a joint initiative between the Federation of American Scientists & the Institute for Progress

Summary

The United States should establish a testbed for government-procured artificial intelligence (AI) models used to provide services to U.S. citizens. At present, the United States lacks a uniform method or infrastructure to ensure that AI systems are secure and robust. Creating a standardized testing and evaluation scheme for every type of model and all its use cases is an extremely challenging goal. Consequently, unanticipated ill effects of AI models deployed in real-world applications have proliferated, from radicalization on social media platforms to discrimination in the criminal justice system.[1,2] Increased interest in integrating emerging technologies into U.S. government processes raises additional concerns about the robustness and security of AI systems. Establishing a designated federal AI testbed is an important part of alleviating these concerns. Such a testbed will help AI researchers and developers better understand how to construct testing methods and ultimately build safer, more reliable AI models. Without this capacity, U.S. agencies risk perpetuating existing structural inequities as well as creating new government systems based on insecure AI systems — both outcomes that could harm millions of Americans while undermining the missions that federal agencies are entrusted to pursue.

Challenge and Opportunity

In 2018, news broke that Amazon was testing a hiring algorithm, tasked with distinguishing top talent from large pools of applications, that turned out to be sexist. The algorithm penalized applications containing the word "women" (such as "women's soccer team") and demoted applicants who graduated from all-female universities. But the algorithm was not originally designed to reject or downgrade women. It learned to do this from a decade of Amazon hiring data, based on the historic dominance of men in the tech sector. The AI system concluded from these data that male applicants must be preferred over female applicants, and therefore pushed men to the top of the pile — an outcome that would have further perpetuated the tech-sector gender gap.[3]

Amazon's sexist hiring algorithm is one of the many incidents demonstrating AI system "brittleness", i.e., the failure of such systems to perform as intended. Over 100 types of similarly problematic AI incidents have been recorded to date. Some of these incidents caused severe and even fatal harm to humans.[4] By contrast, AI models that do perform as intended, particularly in new conditions, are considered robust.[5] The more robust an AI system is, the more reliable and trustworthy it is for human use.

[1] Tufecki, Z. (2018). YouTube, the Great Radicalizer. *The New York Times*, March 10.
[2] Wiggers, K. (2020). Study finds crime-predicting judicial tool exhibits gender bias. *VentureBeat,* December 10.
[3] Dastin, J. (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, October 10.
[4] Partnership on AI. (n.d.). AI Incident Database.
[5] Rudner, T.G.J.; Toner, H. (2021). *Key Concepts in AI Safety: Robustness and Adversarial Examples*. Center for Security and Emerging Technology. March.

What makes an AI system brittle or robust? AI comprises three core ingredients: data, computing power, and algorithms. The role of data is especially important in machine learning (ML), a type of AI in which algorithms are taught to learn from data.[6] Due to recent surges of available data and expansion of computing power needed to train increasingly elaborate algorithms, ML is an especially fast-growing and much-hyped subset of AI.[7,8] But ML models that prove to be successful with one dataset may perform poorly with another. For example, researchers found that FDA-approved medical devices using AI to detect pneumothorax (a collapsed lung) was highly accurate at one clinical site, but experienced degraded performance when evaluated at a second clinical site. This was due to the differences in patient demographics across sites, revealing that ML models sometimes self-learn patterns that work only in specific conditions.[9]

Ensuring the robustness of AI systems, especially within the U.S. government, is a matter of urgent importance. A key challenge in evaluating robustness is a lack of transparency and understanding around how and why a model comes to a decision. If an AI system fails, system designers often cannot explain what happened. And if an AI system succeeds, system designers often cannot ensure that it was for the right reasons.[10] This "black box" problem makes quality-checking and improving AI models difficult.

AI models are also vulnerable to a variety of adversarial attacks, wherein bad actors deploy malicious tactics to cause a model to perform unwanted behaviors. Adversarial tactics include "poisoning" (i.e., tampering with) AI training data, changing the training environment so a model learns the wrong lesson, or extracting a model to expose any classified or sensitive underlying training data.[11] Recently, the Defense Advanced Research Projects Agency (DARPA) released a public toolkit, GARD, to help AI developers learn defensive techniques against adversarial attacks on AI models.[12] While it is too early to know how effective GARD will be, the new tools are a step in the right direction considering very little research has gone into investigating how to defend AI models from adversarial attacks.[13]

Even as the challenges facing AI systems have come to light, the U.S. government has been actively deploying such systems in efforts to better serve Americans. The Social Security Administration is using AI tools to identify citizens likely to qualify for

[6] Buchanan, B.; Miller, T. (2017). *Machine Learning for Policymakers: What It Is and Why It Matters*. Belfer Center for Science and International Affairs.

[7] OpenAI. (2018). AI and Compute. May 16.

[8] Holst, A. (2021). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. *Statista*. Jun 7.

[9] Wu, E., Wu, K., Daneshjou, R. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 27, 582–584 (2021). https://doi.org/10.1038/s41591-021-01312-x

[10] Buchanan, B.; Miller, T. (2017). *Machine Learning for Policymakers.*

[11] Strout, N. (2019). The 3 major security threats to AI. C4ISRNET, September 10.

[12] Miller, A. (2022). DARPA's New Public Tools Teach AI Developers to Defend Against Attacks. Air Force Magazine. January 7.

[13] National Security Commission on Artificial Intelligence. (2019). *Interim Report*. November.

benefits.[14] The Department of Veterans Affairs is applying AI to decrease veterans' wait times for benefits.[15] The Department of Education recently launched an automated chatbot to assist students seeking information about their federal student loans.[16] The National Park Service (NPS) is planning to pilot an NPS ranger, which will be powered by AI to guide visitors exploring U.S. national parks.[17]

To help facilitate the adoption of AI across government, the General Services Administration (GSA) established an AI Center of Excellence (COE).[18] Part of accelerating the adoption of AI across the government, according to the AI COE, is encouraging agencies to reframe how they view technology in their daily processes.[19] A significant challenge facing many agencies is the administrative burden of paperwork: a burden that exceeds nine billion hours annually across all federal agencies.[20] In December 2021, President Biden signed an Executive Order highlighting the need for the government to use technologies such as AI systems to "implement services that are simple to use, accessible, equitable, protective, transparent, and responsive" for all Americans.[21]

Increasing federal usage of AI systems, coupled with the vulnerabilities of AI models to internal malfunction and external attack, means that it is of utmost importance to create a testbed for government uses of AI. A testbed, broadly speaking, is typically a facility staffed with subject-matter experts who conduct rigorous evaluations of a certain type of technology using a variety of tools and methodologies. While there are not yet standardized metrics or benchmarks to evaluate what makes an AI system robust or secure, establishing a federal AI testbed will help push the field in the right direction.

An AI testbed will also serve as a tool to build trust between the U.S. government and its citizens. A University of Oxford survey found that 57% of Americans have little or no confidence in the U.S. government to build AI, and that 84% of Americans believe AI should be carefully managed.[22] A testbed is one opportunity for the U.S. government to demonstrate its commitment to carefully managing AI systems, particularly those used to serve the American people.

[14] Freeman Engstrom, D.; Ho, D.E.; Sharkey, C.; Cuéllar, M.-F. (2020). *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. Administrative Conference of the United States, February.
[15] U.S. Department of Veterans Affairs. (2019). VA launches National Artificial Intelligence Institute. December 5.
[16] Koenig, R. (2019). Meet Aidan, the U.S. Education Department's Financial Aid Chatbot. EdSurge, December 4.
[17] National Park Service. (n.d.). Personal Assistant Pilot Project: NPS Ranger.
[18] General Services Administration. (2020). Accelerate Adoption of Artificial Intelligence to Discover Insights at Machine Speed.
[19] Neupane, A.; Lane, B.; Ewing, E.; Kinnard, K. (2020). Moving Beyond Modernization: Adapt to Better Serve the Public. General Services Administration, February 20.
[20] The White House. (2021). Executive Order on Transforming Customer Experience and Delivery to Rebuild Trust in Government. December 13.
[21] The White House. (2021). Executive Order on Transforming Customer Experience and Delivery to Rebuild Trust in Government. December 13.
[22] Zhang, B.; Dafoe, A. (2019). *Artificial Intelligence: American Attitudes and Trends*. Center for the Governance of AI, January.

Plan of Action

This proposal focuses on creating a testbed specifically for government-procured AI systems — from simple rules-based models to models learning from data — that are used to provide services to Americans. Such services include those previously mentioned, from identifying eligible citizens for government benefits to exploring a U.S. national park. The testbed would evaluate the robustness and security of these AI systems, ensuring such systems maximize benefits and reduce potential harm against Americans.

A natural home for such a testbed is the National Institute for Standards and Technology (NIST), given that agency's mission to "promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life" aligns with the purpose of a federal AI testbed.[23] NIST already has experience in managing testbeds for other technologies and sectors. As a model, NIST can look to the Argonne Leadership Computing Facility's AI testbed set to launch in early 2022.[24]

A NIST-managed AI testbed should collaborate closely with the GSA's AI COE. An AI testbed could directly support the AI COE's mission to accelerate adoption of AI in government. Specifically, an AI testbed would help the AI COE work with agencies to identify and implement AI solutions to various technical challenges.[25] The AI COE, in turn, could provide subject-matter experts to staff the NIST testbed. The result would be a strong collaborative network between the testbed workforce at NIST and the GSA staff who directly support government agencies in adopting AI systems.

There are many components and complexities to designing, implementing, and maintaining a testbed. To start, an AI testbed should staff an interdisciplinary workforce consisting not only of AI researchers, computer scientists, data scientists, and engineers but also ethicists, economists, lawyers, anthropologists, and other liberal arts professionals. AI will influence every sector, with second, third, and fourth order effects on society. Staff with diverse expertise and perspectives must be at the table for a fully comprehensive understanding of the implications of a model.

Once the testbed is established, NIST should require agencies to provide an information packet for each model submitted for testing. This packet should include an exhaustive description of the model such as, but not limited to, the intended purpose of the model; who designed the model (whether an organization or an individual); the data used to train, test, and validate the model; any information on how, when, and where the training data were obtained; the proxies (if any) used by the model to indicate success; the demographic(s) the model impacts; the model type; and other critical information. Specifications for the information packet could

[23] National Institute of Standards and Technology. (n.d.).
[24] Argonne Leadership Computing Facility. (n.d.). ALCF AI Testbed.
[25] General Services Administration. (2020). Accelerate Adoption of Artificial Intelligence.

also be based on specifications that have already been proposed by AI researchers for “model cards for models” or “datasheets for datasets”.[26,27]

Additionally, the testbed should not create a uniform testing or metrics structure. A one-size-fits-all approach will not work since AI models can have very different use cases and disparate effects in a given sector. Instead, an interdisciplinary team of researchers should determine on a case-by-case basis how each model should be evaluated. Whatever testing regime is determined most appropriate by the research team should have the goal of obtaining critical information for a final evaluation “report card”. The report card should detail the testing methodology and the methodology rationale; provide test results; describe implications of the model on relevant populations; identify ways the model could malfunction; identify model limitations; recommend how the model should be monitored once deployed; and summarize any other concerns the research team had about the model regarding its ability to perform in an equitable, transparent, and fair manner.

Conclusion

We live in a reality where development and adoption of AI across all sectors will continue apace. AI system brittleness and vulnerability to attack have engendered significant concern about the detrimental impacts AI systems may impose on society. Americans, as a result, are deeply concerned about the management of AI systems. The U.S. government has an imperative and opportunity to improve development of AI systems while simultaneously demonstrating its commitment to serving Americans and building public trust and confidence.

Introducing an AI testbed at NIST is a step in the right direction. The testbed will generate many opportunities for agencies to collaborate on how to safely adopt AI technologies. In particular, such a testbed would complement the work of GSA’s AI COE in identifying technical challenges and advising on solutions for agencies seeking to adopt AI systems.

In developing and deploying an AI testbed, perfect cannot be the enemy of the good. While creating a uniform testing system that works for every model, use case, and sector would be ideal, it may not be feasible in the near future. The ubiquitous, multi-use nature of AI makes it a uniquely challenging technology to monitor and control. But the U.S. government can leverage its unique resources and power to create a testbed that can help guide and shape how these technologies are integrated within its own ecosystem.

[26] Mitchell, M.; et al. (2019). Model Cards for Model Reporting. Association for Computing Machinery.
[27] Gebru, T.; et al. (2021). Datasheets for Datasets. *Communications of the ACM*.

About the Author

Tina Huang is the policy program manager at the Stanford Institute for Human-Centered Artificial Intelligence (HAI). At HAI, she oversees numerous programs designed to equip policymakers in the U.S. and abroad with the knowledge and skills necessary to make informed decisions on various emerging technologies. Previously, she was an AI Policy Fellow for Rep. Jerry McNerney and a research analyst at Georgetown's Center for Security and Emerging Technology where she focused on AI talent and military use of AI. Tina is also an advisor to Girl Security, a non-profit preparing girls and gender minorities for national security careers. Tina earned her B.A. in international studies from Emory University and M.A. in security studies from the Georgetown Walsh School of Foreign Service.

The views and opinions expressed in this memo reflect only those of the author and not Stanford University or the Stanford Institute for Human-Centered Artificial Intelligence.

About the Day One Project

The Federation of American Scientists' Day One Project is dedicated to democratizing the policymaking process by working with new and expert voices across the science and technology community, helping to develop actionable policies that can improve the lives of all Americans. For more about the Day One Project, visit **dayoneproject.org.**

The Day One Project offers a platform for ideas that represent a broad range of perspectives across S&T disciplines. The views and opinions expressed in this proposal are those of the author(s) and do not reflect the views and opinions of the Day One Project or its S&T Leadership Council.