DAY ONE PROJECT

A National Framework for AI Procurement

Eva Zhang
Grant Gordon
Katie Jonsson

June 2021

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 and do not reflect the views and opinions of the Day One Project or its S&T Leadership Council.

Summary

As artificial intelligence (AI) applications for public use have proliferated, there has been a large uptick in challenges associated with AI safety and fairness. These challenges are due in part to poor transparency in and standardization of AI procurement protocols, particularly for public-use applications. In this memo, we propose a federal framework—orchestrated through the Office of Federal Procurement Policy (OFPP) situated in the Office of Management and Budget (OMB)—to standardize and guide AI procurement in a safer, fairer manner. While this framework is designed for federal implementation, it is important to recognize that many decisions on AI usage are made by municipalities. The principles guiding the federal framework outlined herein are intended to also help guide development and implementation of similar frameworks for AI procurement at the local level.

Challenge and Opportunity

The increasing prevalence of AI applications in private and public life has raised concerns about their fairness—and indeed, AI applications across sectors have been found lacking. AI used in facial-recognition technology, for instance, has the highest error rates among some minority groups.[1] Reliance on such technology for law-enforcement surveillance, airport passenger screening, and employment and housing decisions results in anti-minority bias. The problem exists despite creation of internal "AI fairness" teams at many tech companies, perhaps because of "biases baked into algorithmic models that adopt the norms, values, and assumptions of their developers."[2] Others attribute persistent bias in AI algorithms to the fact that the racial and gender composition of the teams built to regulate AI development often matches the racial and gender composition of the developers.[3] Regardless of the cause, there is a clear and pressing need for external—i.e., government—action to address the bias problem by providing standards for private development of AI algorithms and by enforcing those standards in public applications of AI.

The federal government is taking some steps towards improving regulation of AI—but these must go further. Agencies including the National Institute of Standards and Technology and the U.S. Food and Drug Administration are developing general safety standards for AI. However, there is not yet a framework to support systematic inclusion of fairness review as a criterion for public-use contracts around AI technology. The Joint Artificial Intelligence Center (JAIC) exists to help the Department of Defense (DOD) take advantage of emerging AI capabilities. This effort

[1] A. Najibi, "Racial Discrimination in Face Recognition Technology," Harvard University, October 24, 2020.
[2] The Brookings Institution, "Race, artificial intelligence, and systemic inequalities," June 19, 2020, https://www.brookings.edu/events/webinar-race-artificial-intelligence-and-systemic-inequalities/.
[3] K. Johnson, "OpenAI and Stanford researchers call for urgent action to address harms of large language models like GPT-3," *VentureBeat,* February 9, 2020.

is lacking because the DOD, in addition to other agencies with procured AI technologies, lacks baseline metrics and standards to gauge fairness.[4] No agency has to yet to create working guidelines for AI procurement and continuous deployment by public agencies. Absent such guidelines, AI technologies will continue to introduce harms and bias in public domains ranging from housing loan accreditation to national security. DOD officials have publicly acknowledged the essential importance of AI fairness in driving adoption of AI capabilities throughout government.[5]

To make fairness review a standard component of AI procurement and deployment, a regular process is needed to vet the model, funding sources, and development teams behind AI applications proposed for government use. The process should consider factors such as the data used to train the AI model, the diversity of the development team, and whether the application exhibits any apparent biases in practice. The review process and evaluation criteria should be made public so that application developers and procurement personnel can prepare accordingly and so that external observers can be sure that due diligence is being made to ensure responsible use of AI for public benefit.

Because AI applications are inherently dynamic—learning as they go in response to new training data and user feedback—the *performance* of these applications may change even as the underlying algorithms remain constant. It is also important to establish protocols for transparent, ongoing monitoring of AI applications approved for public use after procurement and initial deployment. Review should track not only application performance at procurement time, but also monitor deployment: especially for applications where unnoticed performance issues and/or distribution shifts may result in major trust and safety issues.

Either the Office of Federal Procurement Policy (OFPP) or the Government Accountability Office (GAO) would be well positioned to implement a new vetting system, given their experience evaluating contracts for a wide range of public-use applications. If either agency is assigned the role of vetting AI applications on behalf of the entire federal government, it will be important for that agency to stay informed on key changes in governmental priorities for AI technology applications. That agency should also leverage relevant efforts and expertise from the National Institute of Standards and Technology, the General Services Administration (GSA), the National Artificial Intelligence Initiative, among others.

An alternative to having one agency implement a standard vetting framework on behalf of the entire federal government would be for individual assessment teams at different federal agencies to apply that framework in-house, at the point of purchase. Either way, such a framework will ensure that federal oversight of AI applications for public use is strong and consistent, and will enable the U.S. government to ethically take advantage of rapidly advancing AI capabilities for public benefit.

[4] D.C. Tarraf et al., *The Department of Defense Posture for Artificial Intelligence: Assessment and Recommendations*. Rand Corporation, 2019.

[5] P. Tucker, "China Is 'Danger Close' to US in AI Race, DOD AI Chief Says" *Defense One*, March 23, 2021.

Plan of Action

The federal government should establish a standard framework for vetting AI applications proposed for public use. The primary goal is to increase safety and fairness of AI use across federal agencies, making AI adoption process safer, more effective, and more transparent for application developers and end users alike. A secondary goal is for this framework to inspire similar frameworks at the local level, where municipalities and cities face similar safety and fairness challenges in leveraging AI.

The Appendix of this memo contains a table summarizing key attributes of AI models that should be considered in a vetting process. The table is intended as a starting point for development of a standard federal evaluation framework. The table is based on recommendations from the sizable body of research literature related to AI fairness and safety that now exists. A particularly influential work is "Model Cards for Model Reporting",[6] which recommends details (e.g., intended use, training data, etc.) that should be published alongside any AI model intended for public use. Other factors that a federal evaluation framework could consider include robustness features (i.e., provisions put into place to ensure that a model performs as well on testing data as it does on testing data), developer profiles, and plans for ongoing monitoring.

Once the framework is developed, the OFPP should require its use by any agency seeking to enter into a contract for a product or service that relies in whole or in part on an AI model. While we recognize that more rigorous vetting alone will not solve all problems associated with AI model biases and fragility, we believe that implementing a standard evaluation framework across the federal government will help advance the dialogue around bias in AI, and will provide more transparency in the often obfuscated procurement process.

We further recommend that federal agencies create new or additional positions for staff tasked with overseeing procurement, deployment, and ongoing monitoring of AI applications. Oversight staff should have technical understanding of AI, as well as experience in diversity, equity, and inclusion. AI subject matter expertise in different domains, such as natural language processing and computer visions, can be particularly valuable. AI oversight positions could be modeled on oversight positions in other sectors, such as institutional review board (IRB) positions for procurement in personalized health monitoring.

[6] M. Mitchell et al., "Model Cards for Model Reporting," *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*: 220–229, 2019.

Precedent for AI oversight positions comes from a $9.9 million contract awarded by the GSA to the companies Deloitte and Esper in 2020. The contract called on the companies to review machine-learning support for agency regulatory reviews, through the Commercial Solutions Opening (CSO).[7] Under a CSO, federal agencies may open a call for proposals on a certain use, have subject-matter experts (e.g., from Deloitte and Esper) review those proposals, and then competitively select proposals based expert feedback. We propose that permanent AI oversight positions be created—either at OFPP to serve multiple agencies or distributed across federal agencies—to support processes such as CSO review. Persons in these positions could also weigh in as needed on tailored standards for AI in different subject areas, and could provide long-term monitoring and feedback on deployment of approved AI applications in practice.

Conclusion

The growing ubiquity of AI applications is accompanied by growing challenges in AI transparency and safety. The federal government can help address these challenges by (1) requiring that agencies adopt a transparent, standardized procurement framework for evaluating all AI applications proposed for public use, and (2) creating designated positions for subject-matter experts who can assist with framework implementation and related needs.

[7] D. Nyczepir, "GSA adds machine learning support for agency regulatory review," *FedScoop*, September 21, 2020.

Frequently Asked Questions

Why focus on procurement?

The European Commission recently delineated four levels of AI risks in its Regulatory Framework Proposal on Artificial Intelligence: unacceptable, high, limited, and minimal.[8] Unacceptable risk is defined as systems considered to pose a "clear threat to the safety, livelihoods and rights of people". High risk includes applications intended for public use, such as in critical infrastructure, educational training, employment decisions, law enforcement, and criminal justice. Limited and minimal risk include applications such as chatbots.

Most high-risk AI applications widely used in the public domain are subject to procurement-review processes. Strengthening oversight of AI applications during procurement is therefore an effective way to mitigate AI risks and harms. But policymakers must take care to ensure that stronger oversight does not unduly limit agencies' ability to take advantage of AI capabilities that could yield significant public benefits. Developing a standard, transparent evaluation framework will help advance the national dialogue on safer AI while also sending clear guidance to application developers on what they must do to meet federal expectations for product quality, fairness, and safety.

Why is it important that a standardized evaluation framework consider factors besides performance metrics?

While standardizing reporting and evaluation of performance metrics is crucial to realizing safer and more transparent AI, static measures of fairness are insufficient for evaluating inherently dynamic AI applications. For instance, information on the training data used to develop an application may be important should problems be found with the training data down the line. A dataset of 80 million images introduced by the Massachusetts Institute of Technology (MIT) in 2006 was found in 2020 to contain multiple racist and derogatory labels.[9] Knowing if an application was trained on that dataset would enable oversight teams to flag that application for deeper review.

Would the proposed evaluation framework use metrics to set "fairness" thresholds?

No. The goal of the proposed framework is to help application developers and end users alike better understand and evaluate potential biases and safety issues in AI models proposed for public use. Objective, specific, and quantitative metrics for evaluating the performance of AI models are valuable. A number of such metrics already exist, and federal agencies such as the National Institute of Standards and Technology and the Food and Drug Administration are continuing to propose and refine additional metrics. However, we do not believe that model

[8] European Commission, *Regulatory framework proposal on Artificial Intelligence*, 2021.

[9] K. Johnson, "MIT takes down 80 Million Tiny Images data set due to racist and offensive content," *VentureBeat*, July 1, 2020.

metrics, captured at time of evaluation, are by themselves sufficient to ensure the long-term safety and fairness of AI applications. The proposed framework goes beyond metrics by introducing additional information about the environment in which an application was developed and provisions for continuous monitoring. Domain-specific evaluation criteria may also be appropriate for some applications.

Are there any existing frameworks for AI procurement?

We are unaware of any existing frameworks for AI procurement in the United States. Overseas, the United Kingdom and the World Economic Forum have jointly proposed Guidelines for AI procurement.[10] The European Commission has issued a White Paper on Data Ethics in Public Procurement of AI-based Services and Solutions.[11]

Acknowledgement

We would like to thank the Day One team advisors for their help and insight during this process. We would also like to thank Chris Meserole, Director of Research and Policy at Brookings Institute for his valuable feedback, and everyone else who provided feedback in the formation of this memo. Finally, we would like to acknowledge the AI ethics research community for the work that has been essential for development of the framework outlined in this memo, particularly that of Model Cards for Model Reporting.

[10] Office for Artificial Intelligence; Department for Digital, Culture, Media & Sport; Departmennt for Business, Energy & Industrial Strategy. "Guidelines for AI procurement," June 8, 2020.

[11] G. Hasselbalch, B.K. Olsen, and P. Tranberg, *White Paper on Data Ethics in Public Procurement of AI-based Services and Solutions*, DataEthics.eu, 2020.

Appendix

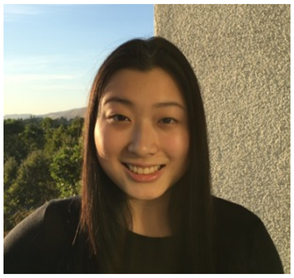
Key attributes of AI models that should be considered in evaluation

Attribute	Details
Model Specification	● Brief description and image of model type or architecture. For example: "Long-short memory network with forget gates, trained with binary cross entropy loss." ● Model task and possible uses. For example: "Time sequence (binary classification)." ● Read-only access or blackbox-sampling function to the API for independent agency audits, when appropriate.
Datasets Used	● Information on: o Datasets used for both training and deployment. o Dataset size, features, known limitations, and any third-party annotations or factors of these datasets. o Whether the datasets used were procured with individual consent. o Summary of and sample data from datasets. o Any factors or attributes of sensitive groups for the particular task for which the AI application was developed. For instance, in tasks such as computer vision[12], factors might include gender, age, race, and Fitzpatrick skin type, hardware factors (e.g., camera type and lens type), and environmental factors (e.g., lighting and humidity). In tasks such as speech processing, factors might include vernaculars (e.g., African American Vernacular English (AAVE)), age, and race.
Metrics and Evaluation	● Raw performance metrics. For instance, accuracy, precision, and recall. ● Metrics to assess performance across factors, especially those recommended by NIST for fairness review. For instance, metrics could assess accuracy rates across

[12] Mitchell, M.; et al. (2019). Model Cards for Model Reporting.

	different groups (accuracy parity), and true positive rate / false positive rate (TPR/FPR) across groups.
Quality Monitoring	● Plans for continued assessment, including frequency of formal reevaluation. Generally appropriate monitoring intervals could be three months, six months, or annually, depending on the model use case. ● Planned quality-monitoring tools and metrics. ● When applicable, disclosure of a black-box sampling API that agencies can use to independently audit proprietary model performance. See third bullet point under "Model Specification", above.
Team Profile	● Profile of application-development team. This should include reporting of diversity metrics across groups relevant to the application-targeted task, such as age, gender, and race. While the profile of a development team should not preclude adoption of an application, a large body of psychology research demonstrates that teams with diverse developers are more likely to identify and improve on model biases. The development-team profile will complement other components of the evaluation framework by flagging possible blind spots that may warrant deeper review.

About the Authors



Eva Zhang is a senior and AI researcher at Stanford pursuing a B.S in Mathematics and M.S in Computer Science. She previously designed, led, and co-taught a computer science class on interpretability and fairness at Stanford. She has spent time working on projects in tech. policy, research, and engineering at the Stanford Machine Learning Group, World Bank Group, Google, NASA, and the Compton Pledge.

Grant Gordon is a Stanford student currently on leave at Sequoia, working to support Sequoia's founders on narrative, product and strategy. He also works to support student founders through BASES and Cardinal Ventures, a nonprofit start-up accelerator. Formerly, he worked for the Research Director of Rebuild by Design, and one of the competitions winners, ONE Urbanism + Architecture, designing Community Engagement strategies for climate infrastructure projects.

Katie Jonsson is an undergraduate at Stanford University majoring in International Relations. She is interested in the intersection of innovation, technology and government. She has spent the past year as Chief of Staff at an early-stage defense start-up and working at the Stanford Internet Observatory.

About the Day One Project

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