BROOKINGS

Report

Six Steps to Responsible AI in the Federal Government

An overview and recommendations from the U.S. experience

Darrell M. West Wednesday, March 30, 2022

Editor's Note:

This report from The Brookings Institution's Artificial Intelligence and Emerging Technology (AIET) Initiative is part of "AI Governance," a series that identifies key governance and norm issues related to AI and proposes policy remedies to address the complex challenges associated with emerging technologies.

There is widespread agreement that responsible artificial intelligence requires principles such as fairness, transparency, privacy, human safety, and explainability. Nearly all ethicists and tech policy advocates stress these factors and push for algorithms that are fair, transparent, safe, and understandable.[1]

But it is not always clear how to operationalize these broad principles or how to handle situations where there are conflicts between competing goals.[2] It is not easy to move from the abstract to the concrete in developing algorithms and sometimes a focus on one goal comes at the detriment of alternative objectives.[3]

In the criminal justice area, for example, Richard Berk and colleagues argue that there are many kinds of fairness and it is "impossible to maximize accuracy and fairness at the same time, and impossible simultaneously to satisfy all kinds of fairness."[4] While sobering, that assessment likely is on the mark and therefore must be part of our thinking on ways to resolve these tensions.

Algorithms also can be problematic because they are sensitive to small data shifts. Ke Yang and colleagues note this reality and say designers need to be careful in system development. Worrying, they point out that "small changes in the input data or in the

ranking methodology may lead to drastic changes in the output, making the result uninformative and easy to manipulate."[5]

"Algorithms also can be problematic because they are sensitive to small data shifts."

In addition, it is hard to improve transparency with digital tools that are inherently complex. Even though the European Union has sought to promote AI transparency, researchers have found limited gains in consumer understanding of algorithms or the factors that guide AI decisionmaking. Even as AI becomes ubiquitous, it remains an indecipherable black box for most individuals.[6]

In this paper, I discuss ways to operationalize responsible AI in the federal government. I argue there are six steps to responsible implementation:

- Having concrete codes of conduct
- Appropriate operational tools for promoting major ethical principles and fighting bias
- Developing clear evaluation benchmarks and metrics
- Relying upon technical standards to help with common problems
- Experimenting via pilot projects and organizational sandboxes
- Having a mix of technical nontechnical skills in the workforce

Concrete Conduct codes

There need to be codes of conduct that outline major ethical standards, values, and principles. Some principles cut across federal agencies and are common to each one. This includes ideas such as protecting fairness, transparency, privacy, and human safety.

Regardless of what a government agency does, it needs to assure that its algorithms are unbiased, transparent, safe, and capable of maintaining the confidentiality of personal records.[7]

But other parts of codes need to be tailored to particular agency missions and activities. In the domestic area, for example, agencies that work on education and health care must be especially sensitive to the confidentiality of records. There are existing laws and rights that must be upheld and algorithms cannot violate current privacy standards or analyze information in ways that generate unfair or intrusive results.[8]

In the defense area, agencies have to consider questions related to the conduct of war, how automated technologies are deployed in the field, ways to integrate intelligence analytics into mission performance, and mechanisms for keeping humans in the decisionmaking loop. With facial recognition software, remote sensors, and autonomous weapons systems, there have to be guardrails regarding acceptable versus unacceptable uses.

As an illustration of how this can happen, many countries came together in the 20th century and negotiated agreements outlawing the use of chemical and biological weapons, and the first use of nuclear weapons. There were treaties and agreements that mandated third-party inspections and transparency regarding the number and type of weapons. Even at a time when weapons of mass destruction were pointed at enemies, adversarial countries talked to one another, worked out agreements, and negotiated differences for the safety of humanity.

As the globe moves towards greater and more sophisticated technological innovation, both domestically and in terms of military and national security, leaders must undertake talks that enshrine core principles and develop conduct codes that put those principles into concrete language. Failure to do this risks using AI in ways that are unfair, dangerous, or not very transparent.[9]

Some municipalities already have enacted procedural safeguards regarding surveillance technologies. Seattle, for example, has enacted a surveillance ordinance that establishes parameters for acceptable uses and mechanisms for the public to report abuses and offer

feedback. The law defines relevant technologies that fall under the scope of the law but also illustrates possible pitfalls. In such legislation, it is necessary to define what tools rely upon algorithms and/or machine learning and how to distinguish such technologies from conventional software that analyzes data and acts on that analysis.[10] Conduct codes won't be very helpful unless they clearly delineate the scope of their coverage.

operational tools that promote ethics and fight bias

Employees need appropriate operational tools that help them safely design and deploy algorithms. Previously, developing an AI application required detailed understanding of technical operations and advanced coding. With high-level applications, there might be more than a million lines of code to instruct processors on how to perform certain tasks. Through these elaborate software packages, it is difficult to track broad principles and how particular programming decisions might create unanticipated consequences.

"Employees need appropriate operational tools that help them safely design and deploy algorithms."

But now there are AI templates that bring sophisticated capabilities to people who aren't engineers or computer scientists. The advantage of templates is they increase the scope and breadth of applications in a variety of different areas and enable officials without strong technical backgrounds to use AI and robotic process automation in federal agencies.

At the same time, though, it is vital that templates be designed in ways where their operational deployment promotes ethics and fights bias. Ethicists, social scientists, and lawyers need to be integrated into product design so that laypeople have confidence in the

use of these tools. There cannot be questions about how these packages operate or on what basis they make decisions. Agency officials have to feel confident that algorithms will make decisions impartially and safely.

Right now, it sometimes is difficult for agency officials to figure out how to assess risk or build emerging technologies into their missions.[11] They want to innovate and understand they need to expedite the use of technology in the public sector. But they are not certain whether to develop products in-house or rely on proprietary or open-source software from the commercial market.

One way to deal with this issue is to have procurement systems that help government officials choose products and design systems that work for them. If the deployment is relatively straightforward and resembles processes common in the private sector, commercial products may be perfectly viable as a digital solution. But if there are complexities in terms of mission or design, there may need to be proprietary software designed for that particular mission. In either circumstance, government officials need a procurement process that meets their needs and helps them choose products that work for them.

We also need to keep humans in some types of AI decisionmaking loops so that human oversight can overcome possible deficiencies of automated software. Carnegie Mellon University Professor Maria De-Arteaga and her colleagues suggest that machines can reach false or dangerous conclusions and human review is essential for responsible AI.[12]

However, University of Michigan Professor Ben Green argues that it is not clear that humans are very effective at overseeing algorithms. Such an approach requires technical expertise that most people lack. Instead, he says there needs to be more research on whether humans are capable of overcoming human-based biases, inconsistencies, and imperfections.[13] Unless humans get better at overcoming their own conscious and unconscious biases, manual oversight runs the risk of making bias problems worse.

In addition, operational tools must be human-centered and fit the agency mission. Algorithms that do not align with how government officials function are likely to fail and not achieve their objectives. In the health care area, for example, clinical decisionmaking

software that does not fit well with how doctors manage their activities are generally not successful. Research by Qian Yang and her colleagues documents how "user-centered design" is important for helping physicians use data-driven tools and integrating AI into their decisionmaking.[14]

Finally, the community and organizational context matter. As argued by Michael Katell and colleagues, some of the most meaningful responsible AI safeguards are based not on technical criteria but on organizational and mission-related factors.[15] The operationalization of AI principles needs to be tailored to particular areas in ways that advance agency mission. Algorithms that are not compatible with major goals and key activities are not likely to work well.

Evaluation Benchmarks and metrics

To have responsible AI, we need clear evaluation benchmarks and metrics. Both agency and third-party organizations require a means of determining whether algorithms are serving agency missions and delivering outcomes that meet conduct codes.

One virtue of digital systems is they generate a large amount of data that can be analyzed in real-time and used to assess performance. They enable benchmarks that allow agency officials to track performance and assure algorithms are delivering on stated objectives and making decisions in fair and unbiased ways.

To be effective, performance benchmarks should distinguish between substantive and procedural fairness. The former refers to equity in outcomes, while the latter involves the fairness of the process, and many researchers argue that both are essential to fairness. Work by Nina Grgic-Hlaca and colleagues, for example, suggests that procedural fairness needs to "consider the input features used in the decision process, and evaluate the moral judgments of humans regarding the use of these features." They use a survey to validate their conclusions and find that "procedural fairness may be achieved with little cost to outcome fairness".[16]

Joshua New and Daniel Castro of the Center for Data Innovation suggest that "error analysis" can lead to better AI outcomes. They call for three kinds of analysis (manual review, variance analysis, and bias analysis). Comparing "actual and planned behavior" is important as is identifying cases where "systematic errors occur".[17] Building those types of assessments into agency benchmarking would help guarantee safe and fair AI.

A way to assure useful benchmarking is through open architecture that enables data sharing and open application programming interfaces (API). Open source software helps others keep track of how AI is performing and data sharing enables third-party organizations to assess performance. APIs are crucial to data exchange because they help with data sharing and integrating information from a variety of different sources. AI often has impact in many areas so it is vital to compile and analyze data from several domains so that its full impact can be evaluated.

Technical standards

Technical standards represent a way for skilled professionals to agree on common specifications that guide product development. Rather than having each organization develop its own technology safeguards, which could lead to idiosyncratic or inconsistent designs, there can be common solutions to well-known problems of safety and privacy protection. Once academic and industry experts agree on technical standards, it becomes easy to design products around those standards and safeguard common values.

An area that would benefit from having technical standards is fairness and equity. One of the complications of many AI algorithms is the difficulty of measuring fairness. As an illustration, fair housing laws prohibit financial officials from making loan decisions based on race, gender, and marital status in their assessments.

"One of the complications of many AI algorithms is the difficulty of measuring fairness."

Yet AI designers either inadvertently or intentionally can find proxies that approximate these characteristics and therefore allow the incorporation of information about protected categories without the explicit use of demographic background.[18]

AI experts need technical standards that guard against unfair outcomes and proxy factors that allow back-door consideration of protected characteristics. It does not help to have AI applications that indirectly enable discrimination by identifying qualities associated with race or gender and incorporating them in algorithmic decisions. Making sure this does not happen should be a high priority for system designers.

pilot projects and organizational sandboxes

Pilot projects and organizational sandboxes represent ways for agency personnel to experiment with AI deployments without great risk or subjecting large numbers of people to possible harm. Small scale projects that can be scaled up when preliminary tests go well protect AI designers from catastrophic failures while still offering opportunities to deploy the latest algorithms.

Federal agencies typically go through several review stages before launching pilot projects. According to Dillon Reisman and colleagues at AI Now, there are pre-acquisition reviews, initial agency disclosures, comment periods, and due process challenges periods. Throughout these reviews, there should be regular public notices so vendors know the status of the project. In addition, there should be careful attention to due process and disparate analysis impact.

As part of experimentation, there needs to be rigorous assessment. Reisman recommends opportunities for "researchers and auditors to review systems once they are deployed."[19] By building assessment into design and deployment, it maximizes the chance to mitigate harms before they reach a wide scale.

Workforce capacity

The key to successful AI operationalization is a well-trained workforce where people have a mix of technical and nontechnical skills. AI impact can range so broadly that agencies require lawyers, social scientists, policy experts, ethicists, and system designers in order to assess all its ramifications. No single type of expertise will be sufficient for the operationalization of responsible AI.

For that reason, agency executives need to provide funded options for professional development so that employees gain the skills required for emerging technologies.[20] As noted in my previous work, there are professional development opportunities through four-year colleges and universities, community colleges, private sector training, certificate programs, and online courses, and each plays a valuable role in workforce development.[21]

Federal agencies should take these responsibilities seriously because it will be hard for them to innovate and advance unless they have a workforce whose training is commensurate with technology innovation and agency mission. Employees have to stay abreast of important developments and learn how to implement technological applications in their particular divisions.

Technology is an area where breadth of expertise is as important as depth. We are used to allowing technical people to make most of the major decisions in regard to computer software. Yet with AI, it is important to have access to a diverse set of skills, including those of a non-technical nature. A Data and Society article recommended that it is crucial to invite "a broad and diverse range of participants into a consensus-based process for arranging its constitutive components."[22] Without access to individuals with societal and ethical expertise, it will be impossible to implement responsible AI.

Thanks to James Seddon for his outstanding research assistance on this project.

---

public. The conclusions and recommendations of any Brookings publication are solely those of its author(s), and do not reflect the views of the Institution, its management, or its other scholars.

Microsoft provides support to The Brookings Institution's Artificial Intelligence and Emerging Technology (AIET) Initiative. The findings, interpretations, and conclusions in this report are not influenced by any donation. Brookings recognizes that the value it provides is in its absolute commitment to quality, independence, and impact. Activities supported by its donors reflect this commitment.

Report Produced by **Center for Technology Innovation**

Footnotes

1. 1 For more information, see Darrell M. West and John R. Allen, Turning Point: Policymaking in the Era of Artificial Intelligence, Brookings Institution Press, 2020.
2. 2 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," Arkiv, November 17, 2016.
3. 3 Zana Bucinca, Maja Barbara Malaya, and Krzysztof Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making," *ACM Human-Computing Interaction*, April, 2021.
4. 4 Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth, "Fairness in Criminal Justice Risk Assessments," University of Pennsylvania, May 30, 2017.
5. 5 Ke Yang, "A Nutritional Label for Rankings," *Proceedings of 2018 International Conference on Management of Data,* 2018.
6. 6 Alexander Wulf and Ognyan Seizov, "Artificial Intelligence and Transparency: A Blueprint for Improving the Regulation of AI Applications in the EU," *European Business Law Review*, 2020, pp. 611-640.
7. 7 Nripsuta Ani Saxena, et al., "How Do Fairness Definitions Fare? Testing Public Attitudes Towards Three Algorithmic Definitions of Fairness in Loan Allocations," *Artificial Intelligence*, February 20, 2020.
8. 8 Alessandro Mantelero, "Ai and Big Data: A Blueprint fore a Human Rights, Social and Ethical Impact Assessment," *Computer Law and Security Review*, August, 2018, pp. 754-772.
9. 9 John R. Allen and Darrell M. West, "It is Time to Negotiate Global Treaties on Artificial Intelligence," Brookings TechTank, March 24, 2021.
10. 10 Meg Young, Michael Katell, and P. M. Krafft, "Municipal Surveillance Regulation and Algorithmic Accountability, *Big Data and Society*, July-December, 2019, pp. 1-14.
11. 11 Booz Allen Hamilton, "Artificial Intelligence Risk Management Framework," 2021.
12. 12 Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova, "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores," *Computing-Human Interactions*, April 25-30, 2020.
13. 13 Ben Green, "The Flaws of Policies Requiring Human Oversight of Government Algorithms," unpublished paper, University of Michigan, undated.
14. 14 Qian Yang, Aaron Steinfeld, and John Zimmerman, "Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes," *CHI Conference on Human Factors in Computing Systems Proceedings*, May 4-9, 2019, Glasgow, Scotland, United Kingdom.
15. 15 Michael Katell, et al., "Toward Situated Interventions for Algorithmic Equity: Lessons from the Field," *Conference on Fairness, Accountability, and Transparency*, January 27-30, 2020, Barcelona, Spain.

16. 16 Nina Grgic-Hlaca, Muhammed Bilal Zafar, Krishna Gummadi, and Adrian Weller," Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning," *Association for the Advancement of Artificial Intelligence*, 2018.
17. 17 Joshua New and Daniel Castro, "How Policymakers Can Foster Algorithmic Accountability," Center for Data Innovation, May 21, 2018.
18. 18 Sam Corbett-Davies and Sharad Goel, "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," *Computers and Society*, August 14, 2018.
19. 19 Darrell M. West, *The Future of Work: Robots, AI, and Automation*, Brookings Institution Press, 2018.
20. 20 Darrell M. West, "Using Artificial Intelligence and Machine Learning to Reduce Government Fraud," Brookings Institution, September 10, 2021.
21. 21 Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf, "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest, *Data and Society*, undated.
22. 22 Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability, AI Now, April, 2018.