IBM Center for The Business of Government

2025

RESPONSIBLE AI FOR PUBLIC EVALUATION

Harnessing AI to Strengthen Public Sector Decisions

Daniel F. Fonner

Southern Methodist University

IBM Center for The Business of Government

Responsible Artificial Intelligence for Public Evaluation: Harnessing AI to Strengthen Public Sector Decisions

Daniel F. Fonner
Corporate Communication and Public Affairs
Southern Methodist University

DECEMBER 2025

IBM Center for
The Business
of Government

Table of Contents

AI has the potential to revolutionize program evaluation and performance auditing by providing deeper insights, improving efficiency, and promoting transparency.

Foreword

Governments are using Artificial intelligence (AI) to reshape how they operate, which offers unprecedented opportunities to improve decision-making, enhance transparency, and deliver better outcomes for the public.

This new report, *Responsible AI for Public Evaluation: Harnessing AI to Strengthen Public Sector Decisions*, by Daniel F. Fonner of Southern Methodist University, provides a timely and practical framework for integrating AI into an important function of government: program evaluation and performance auditing. The author demonstrates how AI can serve as a tool to support—not replace—human judgment, enabling agencies to assess programs more effectively and responsibly to build trust.

Using a rigorous analytical framework and an applied case study, this report illustrates how public administrators can leverage AI responsibly to identify insights, improve efficiency, and strengthen evidence-based decisionmaking. The author also offers actionable recommendations for embedding sound AI practices into existing governance frameworks, ensuring that technology serves the public good.

This report extends the IBM Center's commitment to exploring the intersection of AI and program evaluation. It builds on prior Center publications such as *AI in State Government: Balancing Innovation, Efficiency, and Risk*, a report on responsible GenAI adoption; *GenAI and the Future of Government Work*, which explores the transformative potential of GenAI in reshaping the workforce; *Navigating Generative AI in Government*, which outlines strategic pathways for integrating GenAI into public service; and *Digital Modernization for Government: An Implementation Framework*, which helps to create an evidence-based framework for digital modernization.

As governments navigate the complexities of AI adoption, this report serves as a guide for leaders committed to responsible innovation. By doing so, agencies can unlock the benefits of advanced technologies while mitigating risk.

Daniel J. Chenok
Executive Director
IBM Center for
The Business of Government
chenokd@us.ibm.com

Phaedra Boinodiris
IBM Consulting's Global
Leader for Trustworthy AI
pboinodi@us.ibm.com

Executive Summary

This report explores how artificial intelligence (AI) can be responsibly integrated into program evaluation and auditing processes in the public sector. Through the introduction and testing of the Responsible AI for Evaluation (RAI-Ev) framework, this report offers a practical approach for public administrators to improve transparency and performance in decision-making. For the purposes of this report, artificial intelligence is defined broadly as a "technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision-making, creativity, and autonomy,"[1] not bound to any one tool, task, or algorithm. Specific use cases and applications are defined throughout the report.

AI in the Public Sector

Artificial Intelligence is transforming the public sector, offering innovative solutions for governance, service delivery, and policy implementation, while also uncovering gaps in legacy systems and data infrastructure needed to identify those solutions. AI's ability to optimize resource allocation and enhance decision-making processes makes it a powerful tool for government agencies. However, ethical, legal, and operational challenges must be addressed to ensure AI applications align with transparency, accountability, and performance goals.

Efforts to address AI in government are continually evolving, with important foundations and executive actions evolving over the past several years. Additionally, executive orders and federal AI inventories have shaped AI governance, reinforcing the importance of responsible use and trust in public sector applications.

Best practices for AI implementation should be maintained to ensure responsible use. This report build on other studies from the IBM Center for The Business of Government to provide practical insights into AI's role in modernizing government operations, combating fraud, and improving efficiency that bolster the need for responsible uses of AI in the public sector.

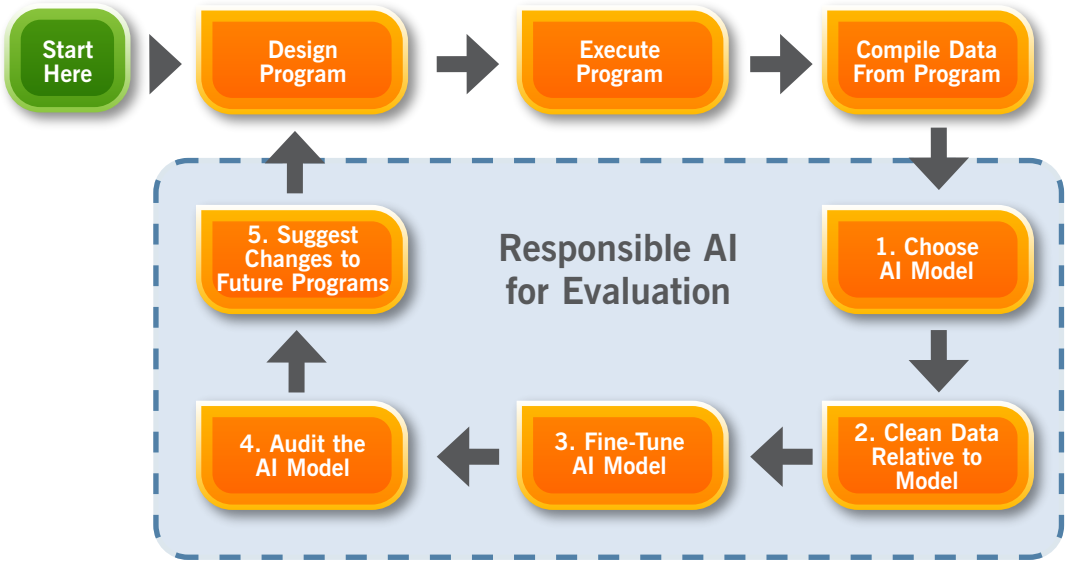
However, AI's potential in performance auditing and program evaluation has been underexplored, with primary sectoral focus on various forms of automation. Program evaluation and performance audits ensure that government initiatives meet their intended goals effectively and efficiently. By integrating AI responsibly, agencies can enhance their evaluation capabilities while maintaining human oversight and accountability.

1. https://www.ibm.com/think/topics/artificial-intelligence.

Responsible AI for Evaluation

Responsible AI for Evaluation (RAI-Ev) is a structured framework designed to guide public administrators in using AI for program evaluation and performance auditing. Unlike predictive AI models, RAI-Ev operates as a post hoc analytical process rooted in social science methods of evaluation to assess past programs and inform future human decision-making. RAI-Ev consists of a five-step process, visualized in Figure 1.

Figure 1: Responsible AI for Evaluation Structural Diagram



Before selecting a model or preparing data in the RAI-Ev process, public administrators should begin by clearly defining the core evaluation questions. Whether the goal is to assess program effectiveness or cost-efficiency, the evaluation questions shapes every subsequent step in the RAI-Ev process. The selected questions should align with the agency's stated goals and intended outcomes, serving as the foundation for model choice, data preparation, and interpretation of results. After establishing the goals of the evaluation, the RAI-Ev process can begin as described below:

1. **Choose a Model:** Select an AI model that aligns with openness, transparency, and data privacy requirements, such as open-source large language models (LLMs) that allow for more robust auditing.
2. **Clean Data Relative to the Model:** Organize government program data into a structured format that aligns with the AI model, ensuring consistency, fairness, and privacy.
3. **Fine-Tune the Model:** Train the model using program-specific data to improve its performance for evaluation tasks.
4. **Model Auditing and Evaluation:** Assess the model's (and by extension, the human decision-maker's) outputs for fairness and accuracy by using a commercially available toolkit that has proven effective and by making systematic queries to an AI system to better understand its inner workings.
5. **Identify and Report Potential Changes for Future Programs:** Communicate findings transparently to stakeholders and recommend improvements for future human-driven decision-making processes.

This framework was tested by the author through a case study of a COVID-relief grant program using RAI-Ev to evaluate the effectiveness of funding decisions. The analysis revealed insights into potential areas for program improvement, demonstrating the model's utility in auditing government initiatives while ensuring fairness and transparency.

Recommendations

To ensure responsible AI implementation for evaluation in the public sector, the following recommendations are proposed:

1. **Broaden Perspective on AI's Role in the Public Sector**—AI should be seen not only as a tool for efficiency and prediction but also as a means of improving transparency and accountability in governance through exposure to applications of frameworks such as RAI-Ev.
2. **Leverage Open-Source Models and Systems**—Where applicable, agencies should prioritize AI models that allow for open auditing and public scrutiny to enhance trust and transparency.
3. **Prioritize Transparent Reporting and Data Privacy Infrastructure in AI-Driven Evaluation Processes**—Agencies should safeguard sensitive data with privacy-preserving techniques and provide clear, plain-language documentation explaining AI model use, limitations, and findings to ensure transparency and public trust.
4. **Embed AI Use into Existing Frameworks**—AI should be integrated into established program evaluation and auditing frameworks, ensuring continuity and alignment with government accountability practices.
5. **Recognize AI Limitations and Avoid AI Hype**—AI should be treated as a tool that supports human decision-making rather than a fully autonomous decision-maker. Overpromising AI's capabilities can lead to unrealistic expectations and misinformed policies.
6. **Default to Responsible AI Practices**—Even when not legally required, government agencies should proactively implement responsible AI practices to ensure public trust and mitigate risks.

AI has the potential to revolutionize program evaluation and performance auditing, enhancing decision-making while maintaining accountability. By adopting a framework like RAI-Ev and adhering to responsible AI principles, public administrators can ensure that AI applications serve the public interest ethically and effectively. Responsible AI practices should be the default rather than a reactive measure, ensuring that government programs remain effective, efficient, and transparent.

AI in the Public Sector

Artificial Intelligence is transforming the public sector, offering innovative solutions to long-standing challenges in governance, service delivery, and policy implementation. From optimizing resource allocation to enhancing decision-making processes, AI has the potential to revolutionize how government agencies operate. However, with these advancements come ethical, legal, and operational concerns that must be carefully addressed to ensure that AI serves the public good without compromising transparency or accountability.

This report examines baseline considerations for integrating AI into public administration, emphasizing its application in program evaluation and performance auditing. These methods, central to government accountability, provide a framework for assessing the effectiveness, efficiency, and fairness of public programs and operations. By leveraging AI responsibly within these processes, administrators can gain deeper insights, enhance operational transparency, and improve human decision-making.

Over the last five to ten years, efforts to regulate AI for societal benefit as well as use AI for government purposes expanded greatly. The Blueprint for an AI Bill of Rights, released by the White House Office of Science and Technology Policy in 2022, outlined principles to guide the development and deployment of AI systems to protect individuals from harm and strive for algorithmic fairness. It is not a legally binding document but provides a framework for responsible AI use across five broad areas of potential impact:

1. **Safe and Effective Systems**
2. **Algorithmic Discrimination Protections**
3. **Data Privacy**
4. **Notice and Explanation**
5. **Human Alternatives, Consideration, and Fallback**

While applying broadly to the development and use of AI systems across the private sector, the importance of these components is perhaps of even greater importance when AI is deployed in public sector settings. In 2024, many states proposed legislation to establish state-level versions of the Blueprint for an AI Bill of Rights along with other AI-related laws. The achievements and clarity around such laws have varied across the country.[2]

In addition to the Blueprint for an AI Bill of Rights, the White House has released Executive Orders to establish strategic priorities for AI development and deployment. Key directives, such as Executive Order 13859 ("Maintaining American Leadership in Artificial Intelligence"), emphasize innovation, workforce development, and standards for responsible use. Other executive actions reinforced the importance of AI in national security, economic competitiveness, and public welfare.

These orders underscored the need for a coordinated approach to AI governance. For public administrators, this means aligning AI initiatives with federal priorities while addressing specific local and sectoral needs. For example, executive orders called for investments in AI research that emphasized the importance of public engagement, encouraging administrators to involve stakeholders in AI policy decisions.

Since January 2025, the Trump Administration has revoked many prior Executive Orders stating that they "impose onerous and unnecessary government control over the development of AI."[3] The administration's order further "Calls for [government] departments and agencies to revise or rescind all [AI] policies, directives, regulations, orders, and other actions taken under the [previous administration]." While following many of the AI principals established during the previous administration are no longer required, many state and local government bodies are promoting responsible use of AI in the sector, maintaining efforts to promote transparency, safety, and privacy when deploying AI models in the public sector.[4]

2. NCSL. 2024. "Artificial Intelligence 2024 Legislation." www.ncsl.org. June 3, 2024. https://www.ncsl.org/technology-and-communication/artificial-intelligence-2024-legislation.
3. The White House. 2025. "Fact Sheet: President Donald J. Trump Takes Action to Enhance America's AI Leadership—the White House." The White House. January 23, 2025. https://www.whitehouse.gov/fact-sheets/2025/01/fact-sheet-president-donald-j-trump-takes-action-to-enhance-americas-ai-leadership/.
4. Hooshidary, Sanam, Chelsea Canada, and William Clark. 2024. "Artificial Intelligence in Government: The Federal and State Landscape." Ncsl.org. 2024. https://www.ncsl.org/technology-and-communication/artificial-intelligence-in-government-the-federal-and-state-landscape.

While government AI strategy continues to evolve, the release of the White House's "America's AI Action Plan" in July 2025, which followed and Executive Order 14277 of April 2025,[5] encourages AI workforce development through public-private and cross-sectoral collaboration. The Executive Order and its accompanying fact sheet9 emphasize the importance of integrating AI competencies into public service, educator training, and philanthropic sectors through applied AI use, education, and ongoing evaluation.

Beyond the federal government guidance, the IBM Center for The Business of Government and other entities provide practical insights of how public sector employees can effectively and safely incorporate AI into their daily workflows and processes. Research published by the IBM Center covers broad uses of AI to potentially improve public sector efficiency and productivity including:

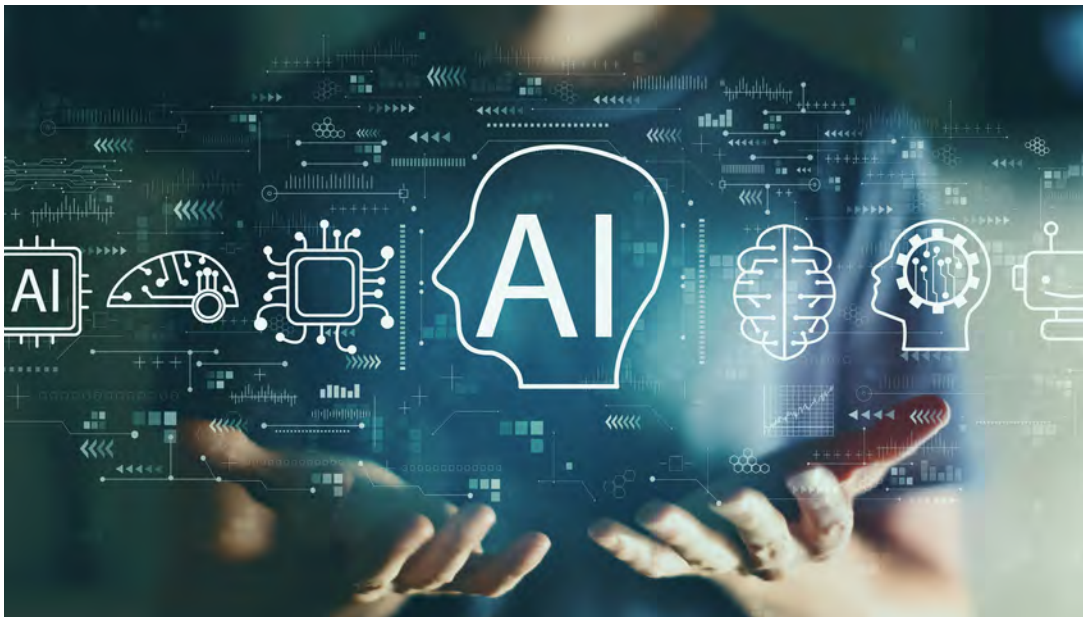
- Accelerating the speed and accuracy of decisions
- Unlocking human resources productively
- Transforming how government modernizes IT systems
- Combatting cyber-based threats
- Reducing fraud, waste, and abuse[6]

Reports exploring these areas of AI implementation and strategy throughout government activities can be found on the IBM Center for The Business of Government website.

More specifically, current and planned federal AI use cases have been published by many federal agencies in fulfillment of the requirement under Executive Order 13960 ("Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government") to prepare an inventory of non-classified and non-sensitive cases of AI use.[7] Detail and usage vary across agencies; for example, the Office of Personnel Management reports two AI use cases,[8] compared to 271 use cases identified across the entirety of the U.S. Department of Health and Human Services. The identified use cases primarily aim at prediction and tasks intended to increase efficiency throughout the federal government.[9] While immensely important to agency operations, AI use can also be implemented to aid in core evaluation functions carried out by government agencies.

Care must be taken, however, as AI's growing role in public administration demands a measured and realistic perspective—one that avoids the pitfalls of AI hype and anthropomorphization. In their book *AI Snake Oil*, Arvind Narayanan and Sayash Kapoor caution against exaggerated claims about AI's capabilities, particularly in high-stakes domains, including government decision-making.[10] Overpromising the benefits of AI or portraying it as a sentient or autonomous decision-maker can lead to misplaced trust, policy missteps, and ultimately, ineffective or even harmful implementations. Instead, AI should be understood as a computational tool that, while powerful, has limitations that require scrutiny and oversight.

5. https://www.whitehouse.gov/presidential-actions/2025/04/advancing-artificial-intelligence-education-for-american-youth/.
6. Chenok, Dan. 2025. *How Can Government Improve Performance with AI?* IBM Center for The Business of Government. 2025. https://businessofgovernment.com/blog/how-can-government-improve-performance-ai.
7. "Executive Order 13960 of December 3, 2020, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government." https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government.
8. https://github.com/ombegov/2024-Federal-AI-Use-Case-Inventory.
9. "HHS AI Use Cases." 2024. Healthit.gov. 2024. https://www.healthit.gov/hhs-ai-usecases.
10. Narayanan, A., and S. Kapoor. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference.* Princeton, NJ, USA: Princeton University Press, 2024.



A responsible approach to AI in the public sector necessitates recognizing these elements and ensuring that AI systems enhance, rather than replace, human judgment. As *AI Snake Oil* emphasizes, AI is not inherently intelligent or unbiased; its outputs are shaped by the data it is trained on and the assumptions built into its algorithms. Without proper safeguards, AI can reinforce existing disparities, produce misleading conclusions, or be leveraged to justify opaque decision-making. For government agencies, this underscores the need for robust evaluation frameworks, transparency measures, and ongoing human oversight to ensure that AI serves as an aid to public service rather than an unchecked authority.

By grounding its use in evidence-based public administration, this report introduces a new AI framework that can empower public sector agencies to harness AI's potential responsibly for auditing and evaluation purposes, while maintaining the integrity of democratic governance.

Program Evaluation and Performance Auditing in Government

Performance auditing and program evaluation are critical tools employed by many government agencies to assess the efficiency, effectiveness, and accountability of public programs and operations. These processes, as defined by the U.S. Government Accountability Office (GAO) and embedded within the Foundations for Evidence-Based Policymaking Act of 2018, can provide valuable insights into how federal agencies achieve their objectives and ensure responsible stewardship of taxpayer resources. While these are powerful tools, little research has been conducted into how they can be applied and perhaps enhanced with assistance from AI models.

Performance Auditing

Performance auditing focuses on assessing whether government programs and activities are meeting their intended objectives in an efficient and effective manner. The GAO defines performance audits as objective and systematic examinations of evidence aimed at providing an independent assessment of the performance and management of government entities.[11] The GAO identifies five concepts that describe "how public officials are to provide functions and services effectively, efficiently, economically, ethically, and equitably." Specific details and requirements of government audits are thoroughly documented in the *Yellow Book*, the GAO's Government Auditing Standards.

Key characteristics of performance audits across those five concepts include:

- **Accountability and Transparency:** Performance audits promote accountability by evaluating whether public resources are being used appropriately and transparently. This aligns with the broader goal of fostering trust in government operations.
- **Evaluation of Effectiveness:** Auditors assess whether programs are achieving their stated goals.
- **Efficiency and Resource Use:** Performance audits also evaluate how efficiently resources are being used to achieve program objectives.
- **Evidence-Based Findings:** The auditing process relies on robust data collection and analysis to produce findings that are credible and actionable. This ensures that recommendations are grounded in evidence.

11. U.S. Government Accountability Office. 2018. *Government Auditing Standards: 2018 Revision.* GAO-21-368G. Washington, D.C. https://www.gao.gov/assets/2021-04/Performance-Audit-Discussion.pdf.

Performance audits provide actionable recommendations to address identified deficiencies or improve program outcomes. For administrators, these audits are invaluable in identifying areas for improvement and ensuring compliance with laws, regulations, and ethical standards.

Program Evaluation

Program evaluation complements performance auditing by focusing on the broader impacts and outcomes of government programs. The GAO defines program evaluation as the systematic collection and analysis of information about a program's design, implementation, and outcomes to inform decisions about program improvement and resource allocation.[12]

Key aspects of program evaluation include:

- **Assessment of Design and Implementation:** Evaluators examine whether a program's design aligns with its objectives and whether it has been implemented as intended.
- **Outcome Measurement:** Program evaluations assess the long-term impacts of programs.
- **Stakeholder Engagement:** Effective evaluations often involve engaging stakeholders, including program beneficiaries, to gather diverse perspectives on a program's performance and impacts.
- **Policy Relevance:** Findings from program evaluations are used to inform policy decisions and strategic planning. This ensures that public resources are allocated to initiatives that deliver meaningful benefits.

Program evaluations provide a framework for understanding not only whether a program works but also how and why it produces specific outcomes. This insight is critical for scaling successful initiatives or redesigning underperforming ones to better serve an agency's constituents.

12. U.S. Government Accountability Office. 2021. *Program Evaluation: Key Terms and Concepts.* GAO-21-404SP. Washington, D.C. https://www.gao.gov/assets/gao-21-404sp.pdf.

Together, these program evaluation standards aim to raise the quality and credibility of government evidence while encouraging continuous improvement. OMB also highlights leading practices that agencies can draw on when building out their evaluation capacity, such as appointing Evaluation Officers, issuing agencywide evaluation policies, engaging technical experts and stakeholders, pre-registering study designs, safeguarding data, and disseminating results in accessible formats. These practices are intended to help agencies translate the five standards into day-to-day routines, from annual evaluation plans and learning agendas to data-stewardship protocols, so that evidence consistently informs budgeting, strategic planning, and program improvement.

This finding is critically important because as of 2024, implementation of the Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act—see bellow) across major federal agencies was generally beneficial with 83 percent of federal evaluation leaders reporting that the Evidence Act helped them achieve their missions through better evaluation practices. However, insufficient staffing and funding limited the influence of the evaluations. A slight majority of federal evaluation leaders noted the use of AI in their evaluation processes under the Evidence Act, opening the possibility of using AI to improve evaluation outcomes and potentially mitigating some staffing and funding limitations in current practices.[13]

Foundations for Evidence-Based Policymaking Act of 2018

In January 2019, the Foundations for Evidence-Based Policymaking Act of 2018[14] (Evidence Act) was signed into law, aiming to "[advance] program evaluation as an essential component of Federal evidence building."[15] From the Evidence Act and accompanying guidance from the Office of Management and Budget (OMB), five standards for Federal evaluations were articulated:

1. **Relevance and Utility:** "Federal evaluations must address questions of importance and serve the information needs of stakeholders in order to be useful resources."
2. **Rigor:** "Federal evaluations must produce findings that Federal agencies and their stakeholders can confidently rely upon, while providing clear explanations of limitations."
3. **Independence and Objectivity:** "Federal evaluations must be viewed as objective in order for stakeholders, experts, and the public to accept their findings."
4. **Transparency:** "Federal evaluation must be transparent in the planning, implementation, and reporting phases to enable accountability and help ensure that aspects of an evaluation are not tailored to generate specific findings."
5. **Ethics:** "Federal evaluations must be conducted to the highest ethical standards to protect the public and maintain public trust in the government's efforts."[16]

13. Smith, K.A., S. Mumford, S. Stefanik, and N. Varnell, Measuring Progress: 2024 Survey of Federal Evaluation Officials, Data Foundation and American Evaluation Association, Nov. 2024.
14. https://www.congress.gov/bill/115th-congress/house-bill/4174/text.
15. Office of Management and Budget. 2019. "Phase 4 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Program Evaluation Standards and Practices."
16. Ibid.

Responsible AI for Evaluation

Taking the best of both worlds, the power of artificial intelligence in the public sector and the structure of program evaluation and performance auditing, Responsible AI for Evaluation (RAI-Ev, pronounced as "rave") establishes a structure that can guide administrators through a process of ethical and transparent decision-making, aligning with the Evidence Act standards. RAI-Ev is a new approach and framework created through research conducted at SMU DataArts, a research center at Southern Methodist University.[17] Public administrators and officials can benefit from this applied framework as a model, allowing for computational assistance beyond prediction and generation tasks aimed primarily at administrative efficiency. Throughout this discussion of RAI-Ev, the example of its impact will be demonstrated by applying it to a case study of a government-funded grant program focused on dispersing COVID-relief funds.

To note, the National Institutes of Health, the National Science Foundation, and other federal funding agencies have banned the use of AI models for the evaluation of individual grant applications over "confidentiality, accuracy, and 'originality of thought'" concerns.[18,19] However, the post hoc nature of RAI-Ev, as described below, enables its use across federal agencies.

What is Responsible AI?

Responsible AI is a general concept that aims to ensure AI systems are developed and deployed in a manner that is safe, secure, and trustworthy. Moreover, phrases such as "Responsible AI," "AI for Good," "Transparent and Explainable AI," "Data Science for Social Good," and many others generally align in their ultimate goals of benefitting society through the safe and ethical use of artificial intelligence.

17. Fonner, D.F., and F. P. Coyle. "Informing Human Decision-Making in Public Administration through NLP Algorithm Audits," 2025 18th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA 2025), Corfu Island, Greece, 2025, https://doi.org/10.1145/3733155.3737371. Fonner, D.F., and F. P. Coyle. "Responsible AI for Government Program Evaluation and Performance Audits," *2024 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2024, pp. 8222-8224, https://doi.org/10.1109/BigData62323.2024.10825518. Fonner, D.F., and Frank P Coyle. 2022. "Explainable Machine Learning Models for Evaluating Government Grantmaking." *2022 IEEE International Conference on Big Data (Big Data)*, December. https://doi.org/10.1109/bigdata55660.2022.10020713.
18. Kaiser, Jocelyn. 2023. "Funding Agencies Say No to AI Peer Review." *Science* 381 (6655): 261–61. https://doi.org/10.1126/science.adj8309.
19. "Notice to Research Community: Use of Generative Artificial Intelligence Technology in the NSF Merit Review Process." 2023. National Science Foundation. December 14, 2023. https://new.nsf.gov/news/notice-to-the-research-community-on-ai.

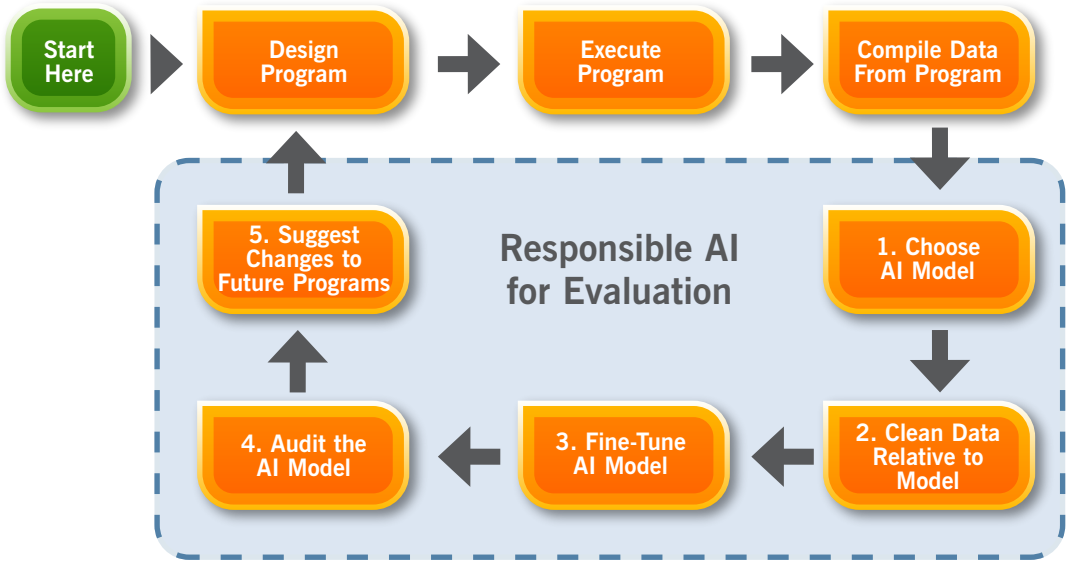
For the purposes of RAI-Ev, the application of Responsible AI principles is shown through the:

- use of AI for post hoc evaluation to aid human-decision making rather than replacing human decision-making.
- use of AI models that are the most "open" to allow for robust model auditing and transparency.
- use of smaller, efficient AI models that do not require sacrificing data privacy to train and run.
- communication of results with government staff and officials as well as constituents and other stakeholders.

What is RAI-Ev?

Figure 2 below shows the broad process that constitutes RAI-Ev, with the process starting from an existing agency program. Importantly, RAI-Ev is a post hoc analytical process to evaluate past programs to inform changes to human decision-making for future programs.[20]

Figure 2: Responsible AI for Evaluation Structural Diagram



RAI-Ev is a five-step process that moves from raw data to suggested changes for future programs. While this process could be developed in a manner only accessible to computer and data scientists, the goal of this research is to make the process as simple, clear, and safe as possible while maintaining robust analysis processes and documentation for accountability purposes.

20. Note that RAI-Ev is a process of using AI to evaluate human processes whereas many other efforts use internal auditing processes to evaluate the performance of AI systems deployed for public or private services. For example, see the latter use explored here by the Institute of Internal Auditors: https://www.theiia.org/en/content/tools/professional/2023/the-iias-updated-ai-auditing-framework/.

RAI-Ev broadly offers a bridge between AI and social science research methods, just as the method of regression can be used both as an evaluative social science tool and predictive algorithmic tool. Table 1 compares general uses and assumptions that typically underpin regression and generative AI models for social science and AI-driven research.[21]

Table 1: Simplified Comparison of Data vs Algorithmic Models for Research

	Regression	Generative AI
Social Science Research	**Assumption:** • Data come from a probabilistic random process with focus on model fit **Use:** • Describe associations between variables of interest	RAI-EV **Assumption:** • Biases and strengths in AI outputs reflect patterns in the human data they were trained on. **Use:** • Identify potential areas of goal alignment and misalignment in human decision-making processes
AI-focused Research	**Assumption:** • Data come from unknown process with focus on predictive accuracy Use: • Generate predictions	**Assumption:** • Data reflect complex patterns that can be approximated by algorithms **Use**: • Produce content/information by imitating patterns in existing data

The benefits of using generative AI for social science research are still not well-defined, but a growing body of research (including RAI-Ev) shows that new AI approaches have value in the social science settings, just as regression models have long been applied across diverse fields.[22]

The following sections will describe each step in the RAI-Ev process in more detail with an accompanying example. In summary, the five steps include:

6. Choosing an appropriate AI model as the foundation to build upon
7. Cleaning government program data relative to the chosen model
8. Fine-tuning the AI model using the cleaned data
9. Auditing the AI model
10. Developing suggested changes to future programs to align with agency goals

21. Breiman, Leo. "Statistical Modeling: The Two Cultures." Statistical Science 16, no. 3 (2001): 199–231. https://doi.org/10.1214/ss/1009213726.
22. Miller, Katherine. Social Science Moves in Silico. Stanford Institute for Human-Centered AI, 2025, https://hai.stanford.edu/news/social-science-moves-in-silico.

Introducing the Applied Case: COVID-Relief Grant Program

In November 2021, funding from the federal government was distributed across many local arts agencies in the United States that then redistributed the funds within their communities "to save jobs and to fund operations and facilities, health and safety supplies, and marketing and promotional efforts to encourage attendance and participation."[23] Through partnership with one of these local arts agencies, researchers had access to the agency's grant applicant data including:

- Completed grant application data including narrative responses
- Panel reviewer scores and notes
- Ultimate funding decision data for the applicant organizations

Through this data and the local agency's detailed grant guidelines, which included their goals and intended outcomes, there was sufficient machine-readable data compiled from the funding program to implement the RAI-Ev framework. Specific details about the program will be discussed throughout each of the following sections.

1 Choose a Model

Across the fields of AI and statistics, the term "model" is used in a broad sense to describe the structure of an algorithmic or mathematical system used to manipulate or align data in a certain manner. This process of modeling data allows for robust analysis and processing of the data based on accepted norms for similar types of data. In the case of RAI-Ev, this means selecting a base AI model that can ingest data from a government program and draw meaningful insights from the data. So how does an agency choose an appropriate model?

The structure of RAI-Ev generally assumes diverse data types within programs, including items such as narrative information, tabular data, and other diverse forms of information that all go into the decision-making process. As such, large language models (LLM) generally suit these types of scenarios well. The question then becomes, "Which large language model does our agency choose?"[24]

The Blueprint for an AI Bill of Rights, introduced in October 2022, provides some guidance as it describes five key principles AI models should abide by:

1. Safe and effective systems
2. Algorithmic discrimination protections
3. Data privacy
4. Notice and explanation
5. Human alternatives, consideration and fallback[25]

23. https://www.arts.gov/news/press-releases/2021/american-rescue-plan-grants-local-arts-agencies.
24. Note that some institutions and agencies may be limited in model choice depending on internal constraints and authorizations.
25. The White House. 2022. "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People." https://www.govinfo.gov/content/pkg/GOVPUB-PREX23-PURL-gpo193638/pdf/GOVPUB-PREX23-PURL-gpo193638.pdf.

To achieve these goals in public administration deployment of AI models, agencies must strive for choosing an "open" model that can be probed and studied to protect those impacted by the outputs and decisions made relative to the model.[26] Researchers have identified model characteristics under three primary categories that are critical in assessing the "openness" of various LLMs:

- Availability: are the following items open and available?
    - Model training code, data, and weights
    - Reinforcement learning data and weights
    - Data and model licenses
- Documentation: how well is the LLM documented and described through the following items?
    - Code and model architecture
    - Preprint and published papers describing the LLM
    - Model Cards and datasheets
- Access: How can people access the LLM?
    - Availability of coding packages or APIs[27]

While the intricate details of these components can be difficult to assess without a technical background, researchers have created an assessment of major LLMs, making it easy to quickly assess potential models based on the above measures of openness. By choosing models that are more open, government agencies can have more confidence in their ability to audit the system to better understand the decisions being informed or derived from the AI model.

Across all AI models, the resource impact must also be factored into the model selection process.[28] While the marginal cost of querying a model is perhaps small on its own, training, fine-tuning, and multiple model adjustments each have their own carbon footprint that can add up quickly. Models that are more open tend to include information on the carbon footprint required for training the model. Additionally, developers can use simple tools to estimate the resource impact of running code for model development, training, and deployment.[29]

An additional consideration in model selection focuses on data privacy and ownership. LLMs that are only accessible on third party platforms may require that agencies export their data into these private systems.

26. François, Camille, et al. *"A Different Approach to AI Safety: Proceedings from the Columbia Convening on AI Openness and Safety."* arXiv preprint, June 2025. https://arxiv.org/pdf/2506.22183.
27. Liesenfeld, Andreas, Alianda Lopez, and Mark Dingemanse. 2023. "Opening up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators." ArXiv.org. July 19, 2023. https://doi.org/10.1145/3571884.3604316.
28. Crawford, Kate, 1976-. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.
29. CodeCarbon. https://codecarbon.io/.

While robust in description, some government agency data may be too sensitive to risk sharing with a third party. This may necessitate agencies using models that are high-performing and can be run on in government computing environments, perhaps even on individual government computers. While there may be a tradeoff in performance when comparing the power of proprietary models with those that can be run locally, this research has found successful results in applying RAI-Ev on smaller, local computers. Constraints related to agency staffing and expertise must also be considered when choosing an appropriate model.

RAI-Ev: Choosing a Model for a COVID-Relief Grant Program

Knowing that data privacy and openness were paramount for this application of RAI-Ev, for this government-funded COVID-relief program for arts and culture organizations the analysis uses the most open LLM that could be run and adjusted on a local computer: OLMo 1B-HF. This model is the most open and is maintained on the AI platform Hugging Face for ease of downloading and accessing documentation about the model's creation, uses, and limitations.[30]

The OLMo family of models are ideal for RAI-Ev as the entire process of development and deployment can be scrutinized to identify potential biases in the processes and provide exceptional levels of transparency, without needing to disclose private government data to third party companies. The smallest OLMo model, OLMO 1B-hf, can fit on most computers and can be fine-tuned to best meet the needs of the task, which in this case involves using an evaluation rubric developed by the local arts agency to assess grant applications.

This LLM was primarily designed for text generation tasks, although not specifically trained to evaluate grant applications and panel reviewer notes and scores. The model will be fine-tuned for the specific task at hand in Step 3 below.

The Hugging Face platform provides additional information about the OLMo 1B-hf model including performance metrics on standard tasks, environmental impacts of energy required to train the model,[31] and other key descriptions of the model. The importance of openness and descriptions of limitations and uses in Model Cards[32] and Datasheets[33] are intended to be generally accessible to a broader audience including non-technical staff.

30. "Allen AI/OLMo-1B-Hf." 2024. Huggingface.co. April 17, 2024. https://huggingface.co/allenai/OLMo-1B-hf.
31. The developers of the OLMo family of models estimate that the development of the OLMo 7B model had carbon emissions equivalent to roughly the annual emissions of 15 passenger cars in the US (70 tCO_2eq). The specific emissions of the OMLo 1B model were not specifically disclosed. https://arxiv.org/pdf/2402.00838.
32. Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596.
33. Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723.

2 Clean Data Relative to the Model

To move from a base LLM to a model adjusted to meet the specific needs of a government program, the data from the program needs to be "cleaned" or adjusted in ways that allow it to be useful to the LLM during the fine-tuning process in Step 3. This process can involve tasks such as, but are not limited to:

- Gathering data exports in spreadsheet form from specific programs
- Ensuring data is in consistent formats such as numeric, textual, and other formats
- Identifying how various data connect or align with one another (such as constituent requests aligning with government responses to those requests)
- Redact, remove, or flag any data that should not be used in the training process

After this first stage of data cleaning and wrangling, the data must then be restructured to align with the structure of the chosen LLM.[34] For simplicity, many LLMs use a similar structure for data communication that has three components:

1. A "system" role that contains data explaining the task required of the LLM
2. A "user" role that contains data supplied from a constituent or applicant
3. An "assistant" role that contains the data from the government agency's assessment of the "user" data, which could include grant panelist notes and scores

This approach to machine learning is called Supervised Learning, where the LLM is trained and fine-tuned based on instructive examples of what to accomplish when seeing new data. If this type of system were to be used in a forward-looking context, historic examples of program data could inform the assessment of future programs. However, RAI-Ev takes a post hoc approach in only assessing past programs to inform changes in human decision-making for future programs.

Returning to the three data components listed above, LLM developers format this data in different ways known as chat templates. One of the most common chat templates is ChatML, or Chat Markup Language, developed by OpenAI.[35,36] This template identifies each of the three roles and their associated data. Figure 4 shows an example of this structure.

34. Note that complexity of the program and data being evaluated can vary the time required for properly structuring the data. For recurring programs, the upfront time of development would vary but would be minimal for future iterations of the same or similar programs. For the research described in the case study, the process of code development for restructuring took less than five hours. You can view the example code here: https://github.com/danielfonner/IBM_RAI-Ev/blob/8831fb56d483bc22ee4cfc0a912c37ff6741843d/GITHUB_IBM_CleaningInstructionData.ipynb.
35. "Chat Templates." 2024. Huggingface.co. 2024. https://huggingface.co/docs/transformers/en/chat_templating.
36. OpenAI. 2020. "Openai-Python/Chatml.md at Release-V0.28.0." GitHub. 2020. https://github.com/openai/openai-python/blob/release-v0.28.0/chatml.md.

Figure 4: ChatML Example

```
<|im_start|>system
You are a helpful chatbot that responds kindly to questions about
RAI-Ev.<|im_end|>

<|im_start|>user
Is RAI-Ev a useful framework for evaluating government
programs?<|im_end|>

<|im_start|>assistant
It can be useful in building internal and external trust in govern-
ment programs through iterative program improvement. <|im_end|>
```

While this example is trivial, it provides context for how data must be wrangled into a long string for this type of process.[37] Practically, this string of text could exist within a spreadsheet with a format such as the following table:

Table 2: Example Spreadsheet for ChatML Formatted Data

#	System	User	Assistant
1	You are a helpful chatbot that responds kindly to questions about RAI-Ev.	Is RAI-Ev a useful framework for evaluating government programs	It can be useful in building internal and external trust in government programs through iterative program improvement.
2	You are a helpful chatbot that responds kindly to questions about RAI-Ev.	Where was RAI-Ev created?	RAI-Ev was developed by researchers at Southern Methodist University.

That spreadsheet would then be fed into the LLM system during the fine-tuning step to better align the base model with the specific task required by the government agency.

37. Note that the long strings are further converted in JSON formatting through the code to more efficiently interact with an AI model.

RAI-Ev: Clean Data Relative to Model for a COVID-Relief Grant Program

After selecting the OLMo 1B-HF model for assessing the COVID-relief grant program for arts and culture organizations, the grantmaking agency provided researchers with two spreadsheets: one containing application information on each applicant and one containing panel reviewer notes and scores for each application. In combination with the grant program guidelines, the data was formatted in the ChatML format in preparation for fine-tuning. Figure 5 shows an example of the grant guidelines rubric.

Figure 5: Partial Grant Program Evaluation Rubric

Access to the Arts (0–35 points)

- Does the applicant demonstrate knowledge of the target audience(s)?
- Does the applicant demonstrate effective plans and strategies to reach the target audience?
- Does the applicant demonstrate plans to reach underserved communities?
- Does the applicant demonstrate plans to reach the general public?

Management (0–30 points)

Is the budget appropriate?

- Are people being paid correctly (i.e., are people being grossly overpaid or underpaid)?
- Has money been allocated appropriately?
- Is there enough money in the budget to execute the project?
- Have other revenue streams been secured if needed?

The information contained in the evaluation rubric falls into the "system" role category as these are the prompts that inform panel reviewers on how they should evaluate each application. Table 3 shows an example of the cleaned data in spreadsheet form followed by Figure 6 showing the same information in textual string form.

Table 3: Spreadsheet of ChatML Formatted Data (truncated)

#	System	User	Assistant
1	You are a grant review panelist. Score the following application on a scale of 0-35…	Example Arts Organization's mission is to engage with, entertain, and serve the community…	Score 29: This project has been around since 2010 and it's still going strong! - I like the idea…

Figure 6: ChatML Formatted Data as Text String (truncated)

```
<|im_start|>system
You are a grant review panelist. Score the following applica-
tion on a scale of 0-35. 0 being the lowest and 35 being the
highest. Provide a brief explanation of your score based on
the following criteria:
  1) The work deepens its impact by welcoming people of all
abilities and backgrounds.
  2) Strong ideas expressed with clarity advance artistic
goals [truncated]
<|im_end|>

<|im_start|>user
```

```
Example Arts Organization's mission is to engage with, enter-
tain, and serve the community by examining current social issues
through the lens of classic text. Our vision is [truncated]
<|im_end|>

<|im_start|>assistant
Score 29: This project has been around since 2010 and it's still
going strong! - I like the idea of this project and the fact
that it will continue to grow and expand over time. [truncated]
<|im_end|>
```

Note that the above text was truncated and anonymized for readability and data privacy concerns.

The process of cleaning data and aligning it with the ChatML format does require basic knowledge of code development in a language such as Python to ensure consistency in cleaning processes. Example Python code for cleaning and ChatML formatting purposes can be found here.[38]

As an initial experiment to test the base OMLo model prior to fine-tuning, the "user" prompt from the above example was fed to the model. The model did not provide an "assistant" output evaluating the application material but rather output the "user" content over 20 times in a row. The base OLMo model was not sufficient to meet the needs of this project. The next step of fine-tuning will remedy this situation.

3 Fine-Tune the Model

Fine-tuning a Large Language Model can take the original model's performance and improve it to work on data unique to a specific situation, such as a government agency program being evaluated through RAI-Ev. By using the cleaned data from Step 2 to train the model, the results from the LLM will be more useful and potentially reduce bias in the model that is not relevant to the task at hand. This process is also referred to as "instruction-tuning" in this context, as the model is being retrained to carry out specific tasks it was not originally designed to conduct. Over the course of the fine-tuning process, the computations made through the LLM's structure are slightly adjusted to be more accurate to the new tasks. (A similar process known as transfer learning could also be employed where only certain computations in the LLM are adjusted along with the final output computations.)

While more technical in nature and perhaps requiring assistance from data science or IT staff withing an agency, the fine-tuning structure could be implemented with few changes across diverse government programs. The process includes four broad steps:

1. Load the base LLM
2. Load cleaned ChatML-formatted data
3. Set parameters for model training and where to save the model
4. (Re)train the base LLM

These steps will most likely result in a fine-tuned model ready for the specific evaluation task. Steps three and four may take multiple iterations and adjustments to the model parameters to achieve an accurate, high-performing model. It is possible that the model will not achieve a

38. See example code here: https://github.com/danielfonner/IBM_RAI-Ev/blob/8831fb56d483bc22ee4cfc0a912c37ff6741843d/GITHUB_IBM_CleaningInstructionData.ipynb.

level of performance sufficient to carry out the rest RAI-Ev of the process, which could be due to many different reasons from poor data quality to outcomes misaligning with program goals. Should that occur, public administrators should reevaluate their project goals and determine if the data and modeling processes can be improved or if the process should be abandoned for another form of evaluation.

RAI-Ev: Fine-Tuning the Model for a COVID-Relief Grant Program

Formatting the grant program data into the ChatML format allows for the fine-tuning of the OLMo 1B-HF model for the task of evaluating grant applications, panel reviewer notes, and panelist scores. Using standard machine learning libraries from Hugging Face and PyTorch, among others, researchers developed Python code to carry out the fine-tuning and examine the outputs of the model. Code for the training process can be found here.[39] While requiring some experience with Python to fully understand and manipulate the code, once established, the code can be used across many different training processes.

The performance of the model was measured in many ways, both quantitative and anecdotal. One of the basic quantitative measures of performance for this type of model is called perplexity, which is a way to measure how well the model predicts text.[40] Lower perplexity scores indicate increased probabilities of predicting the "correct" words. Higher perplexity scores indicate the model is less accurate in predicting words. The base OLMo 1B model achieved a perplexity score of 10.8 meaning that the model, on average, generated new text from a set of roughly 11 words at each step of generation.[41] After fine-tuning, the new model for evaluating grant applications achieved a perplexity score of 8.9, meaning the model performed better at generating possible text in the project's evaluation domain.[42] Using the CodeCarbon Python library, the process to fine-tune the model one time was roughly the equivalent of charging a smartphone 5 times (~30 grams CO_2).

From a more anecdotal point of view, researchers generated a test output based on the ChatML formatted data described in Step 3 to show how the model performed:

Human Reviewer Assessment: Score 29: This project has been around since 2010 and it's still going strong! - I like the idea [truncated]

LLM Output Assessment: <|im_start|>assistant Avg_Reviewer_Score: 28; Reviewer Notes: This is a worthy project that has been going on for many years [truncated] <|im_end|>

These outputs show similarity across scores and content between human and model evaluation. A more rigorous approach to qualitative assessment of the model could include sampling from the generated outputs and having subject matter experts review the panel notes, scores, and model outputs for similarity and logical evaluations.

39. See example code here: https://github.com/danielfonner/IBM_RAI-Ev/blob/caf3fb43ea28ea56da072d89c8c6217c08e5300e/GITHUB_IBM_instruct_tuning_olmo.ipynb.
40. "Perplexity of Fixed-Length Models." 2024. Huggingface.co. https://huggingface.co/docs/transformers/en/perplexity.
41. "Weights & Biases." 2024. W&B. https://wandb.ai/ai2-llm/OLMo-1B/reports/OLMo-1B--VmlldzO2NzY1Njk1.
42. Fonner, Daniel F, and Frank P Coyle. 2024. "Responsible AI for Government Program Evaluation and Performance Audits." 2024 *IEEE International Conference on Big Data (Big Data)*, December, 8222–24. https://doi.org/10.1109/bigdata62323.2024.10825518.

4 Model Auditing and Evaluation

In the Winter 2019/2020 edition of *The Business of Government* magazine, algorithm auditing was highlighted as key to mitigating AI model risks, especially within public sector deployment of such models.[43] The proposed strategies align closely with other efforts for informing algorithm auditing in the public sector, with strong emphasis on bias mitigation in predictive AI tools.[44] In the context of RAI-Ev, auditing the AI algorithm for biases or misalignments with other program goals provides information regarding issues that were potentially present in the human decision-makers administering the program. By auditing the model, identified areas of concern can be mitigated in future human decision-making processes through targeted training, evaluation rubric adjustment, or other strategies to improve outcomes.

Research and policy briefs developed by the Stanford University Institute for Human-Centered Artificial Intelligence (HAI) offer government administrators and policy officials great insight into algorithm audit concepts and a survey of key audits that have been performed in recent years.[45,46] Through this work, they defined algorithm audits as "methods of repeatedly and systematically querying an algorithm with inputs and observing the corresponding outputs to draw inferences about its opaque inner workings" that are explored through nine dimensions of AI deployment processes:

1. Legal and ethical considerations
2. Selecting a research topic
3. Choosing an algorithm
4. Temporal considerations
5. Collecting data
6. Measuring personalization
7. Interface attributes
8. Analyzing data
9. Communicating findings

43. Kassir, Sara. 2020. "Algorithmic Auditing: The Key to Making Machine Learning in the Public Interest." IBM Center for The Business of Government. https://www.businessofgovernment.org/sites/default/files/Winter%202019%202020%20Magazine.pdf.
44. Ojewale, Victor, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2024. "Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling." ArXiv.org. March 14, 2024. https://doi.org/10.48550/arXiv.2402.17861.
45. Metaxa, Danaë, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. "Auditing Algorithms: Understanding Algorithmic Systems from the Outside In." *Foundations and Trends in Human–Computer Interaction* 14 (4): 272–344. https://doi.org/10.1561/1100000083.
46. "Policy Brief—Using Algorithm Audits to Understand AI." 2022. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/policy-brief-using-algorithm-audits-understand-ai.

A recent example of public sector algorithm auditing took place in New York City in 2023. Following the adoption of Local Law 144, which:

> [P]rohibits employers and employment agencies from using an automated employment decision tool unless the tool has been subject to a bias audit within one year of the use of the tool, information about the bias audit is publicly available, and certain notices have been provided to employees or job candidates.[47]

Even when assuming the best intentions of the law, researchers found the following:

> Using qualitative interviews with 16 experts and practitioners working within the regime, we find [Local Law] 144 has failed to create an effective auditing regime: the law fails to clearly define key aspects like [Automated Employment Decision-Making Tools (AEDTs)] and what constitutes an independent auditor, leaving auditors, vendors who create AEDTs, and companies using AEDTs to define the law's practical implementation in ways that failed to protect job applicants.[48]

The depth of knowledge and technical facility required to "systematically [query] an algorithm" as required by legislation like Local Law 144 might seem daunting to non-technical government administrators, but many tools are being developed to aid in this process.[49, 50]

HAI researchers have proposed a collaborative technique whereby non-technical users contribute to the process of evaluating algorithmic behavior. The process, known as end-user audits, involves three primary steps related to a classification algorithm's raw training data:

1. An individual, or "end user," manually assesses the raw data to classify it for the chosen task, and that process is applied throughout the entirety of the training data.
2. The end-user classifications are compared against the classifications made by the AI model, providing information on areas of disagreement between the end-user and the model.
3. The end-user uses this information to report on the performance of the AI model, prompting necessary change to mitigate any identified biases or unintended outcomes.[51]

47. "Automated Employment Decision Tools (AEDT)." 2021. City of New York. https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page.
48. Groves, Lara, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. 2024. "Auditing Work: Exploring the New York City Algorithmic Bias Audit Regime." *ArXiv* (Cornell University), June. https://doi.org/10.1145/3630106.3658959.
49. Ojewale, Victor, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2025. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. arXiv preprint arXiv:2402.17861v3. March 3, 2025. Available at: https://arxiv.org/abs/2402.17861.
50. AI literacy for public sector administrators is a key skill that must be developed for the responsible use of any AI system, including use of the RAI-Ev framework. Training offered by institutions such as the Stanford Institute for Human-Centered Artificial Intelligence can be a valuable resource for public servants. https://hai.stanford.edu/policy/policymaker-education.
51. Lam, Michelle, Mitchell L Gordon, Danaë Metaxa, Jeffrey T Hancock, James A Landay, and Michael S Bernstein. 2022. "End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior." *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW2): 1–34. https://doi.org/10.1145/3555625.

This stage of the process, along with original public program development, is critical for gathering feedback from public servants and relevant stakeholders to ensure their subject matter expertise informs the evaluation process, reinforcing important aspects of the program and challenging results that are misaligned with program goals and outcomes.

While a very useful approach for non-technical audiences, the process requires the development of visualization programs for each model that needs to be audited. This may become more scalable across differing AI model tasks as new tools become available.

Comparing end-user outputs with AI model outputs is a critical element of the algorithm auditing, but a more foundational analysis of a model's data can provide higher-level information about potential bias within a system. Research initiated at the University of Chicago and now maintained by Carnegie Mellon University provides public administrators with a simple tool to broadly assess bias and fairness in AI models, known as Aequitas.[52]

The toolkit provides a means by which public administrators can upload the training data from their AI model, identify specific groups to be evaluated for disparate treatment, select relevant metrics of fairness, and review a system-generated bias report covering many areas of potential bias within a system.[53] IBM has compiled a listing of many forms of bias that can impact AI model performance, and ultimately society, and are described in Table 4.[54]

52. Saleiro, Pedro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit Rodolfa, and Rayid Ghani. n.d. "Aequitas: A Bias and Fairness Audit Toolkit." Accessed May 24, 2024. https://arxiv.org/pdf/1811.05577.
53. Example bias reports and visualizations, are available on Carnegie Mellon University's website: https://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/. Additionally, bias and bias mitigation can take many forms across datasets and topics areas. Gallegos et. al provide a comprehensive survey of these topics aimed primarily at data and computer scientists, but the information is valuable across domains and levels of expertise: https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A.
54. Holdsworth, James. 2023. "AI Bias." IBM. December 22, 2023. https://www.ibm.com/think/topics/ai-bias.

Table 4: Common Bias Types and Their Definitions

Bias Type	Definition
Algorithm Bias	• Occurs when the problem, question, or feedback given to an AI system is incorrect or incomplete, leading to misinformation.
Cognitive Bias	• Results from human input errors where personal biases unintentionally influence AI datasets or model behavior.
Confirmation Bias	• AI overemphasizes existing beliefs or patterns in data, preventing it from recognizing new trends.
Exclusion Bias	• Happens when important data is unintentionally left out, leading to incomplete or skewed AI outcomes.
Measurement Bias	• Caused by incomplete data, often due to oversight, leading to inaccurate or misleading conclusions.
Out-Group Homogeneity Bias	• AI struggles to differentiate between individuals outside the majority group in the training data, leading to misclassification and bias.
Prejudice Bias	• Occurs when societal stereotypes influence the AI dataset, resulting in biased and discriminatory outputs.
Recall Bias	• Develops when data labels are inconsistently applied due to subjective human observations.
Sample/Selection Bias	• Happens when training data is too small, unrepresentative, or incomplete, leading to inaccurate AI predictions.
Stereotyping Bias	• AI unintentionally reinforces harmful stereotypes, often in language processing or classification tasks.

But beyond the scope of interrogating training data to mitigate biases in predictive AI models, which is of critical importance to any AI auditing process, public administrators can also leverage potential biases perpetuated through AI models to their advantage by identifying how human actors within government agencies may have imbued bias into their decision-making practices. This is where the combination of algorithm auditing, performance auditing, and program evaluation produce new mechanisms for agency self-assessments.

Steps in conducting an algorithm performance audit and program evaluation

To evaluate and audit AI models effectively, the RAI-Ev framework incorporates key principles from the GAO's *Yellow Book* and other standard approaches to program evaluation. The evaluation should focus on assessing program effectiveness, efficiency, cost-effectiveness, and ethics. These principles should be applied to the specific goals and intended outcomes established by the government agency when the program was created. Additionally, the audit and evaluation process should consider the program's design, implementation, and outcomes while keeping its underlying theory of change in focus.

The following steps outline how to use RAI-Ev to assess the AI model developed and fine-tuned to specific government programs in the previous steps:

1. ***Gather all available documentation related to the program's planning and development.*** Begin by collecting all relevant documents that describe the program's design, objectives, and implementation. This may include policy briefs, program guidelines, evaluation rubrics, funding proposals, internal reports, and previous evaluations. These materials will serve as a foundation for understanding the program's intended purpose and structure.
2. ***Identify the agency's key goals and intended outcomes based on the collected documentation.*** Carefully review the documentation to extract the program's primary objectives and expected outcomes. These goals should be clearly defined, measurable, and aligned with the agency's mission. Understanding these elements ensures that the AI model's performance is evaluated in the proper context.
3. ***Assess fairness and potential bias in the raw data, considering its relevance to the agency's objectives.*** Analyze the program data used to train and operate the AI model to identify any potential biases or fairness concerns, using a tool such as the Aequitas Bias and Fairness Audit Toolkit. This involves checking for imbalances, disparities, or patterns that could lead to discriminatory or inequitable outcomes. It is critical to ensure that the data accurately represents the populations and conditions the program is meant to serve.
4. ***Develop test datasets, including counterfactual examples, to evaluate the AI model's performance on specific target groups of interest.*** Create test cases that can assess how the AI model performs across different subgroups or situations of interest. Counterfactual examples—hypothetical variations of data—can help determine whether the model produces consistent and fair results across different demographic or operational conditions.
5. ***Run the fine-tuned AI model on these test datasets across multiple iterations, adjusting key input parameters as needed.*** Conduct repeated tests using the AI model on the prepared datasets, systematically modifying key inputs to observe how the model responds. This step helps identify potential weaknesses, inconsistencies, or unintended biases in how the model, and in the case of RAI-Ev, how humans perhaps processed information and generated decisions.
6. ***Analyze the model's outputs to assess its performance across different target groups and key evaluation metrics.*** Evaluate the results of the AI model by comparing its outputs against the agency's goals and established performance criteria. This assessment should determine whether the model (and human decision-makers) is (are) achieving the intended outcomes, flagging any disparities, inefficiencies, or ethical concerns that need to be addressed before running the program again in the future under adjusted human-driven decision-making processes.

RAI-Ev: Model Auditing and Evaluating for a COVID-Relief Grant Program

1. **Gather all available documentation related to the program's planning and development.** The primary documentation available from the COVID-relief grant program came from grant application instruments, program guidelines, evaluation rubrics, and interviews with administrators of the program.

2. **Identify the agency's key goals and intended outcomes based on the collected documentation.** From the interviews and documentation, the grant program had a few key goals to ensure support for arts and culture organizations across multiple demographic groups, as well as organizations with budgets under $250,000 as these entities were less likely to have received COVID-relief funding based on agency research. Additionally, the agency strove to have grant review panelists whose demographic characteristics generally matched the characteristics of the applicant pool. (Additional goals and outcomes were identified by the agency but are not included here for clarity and space considerations.)

3. **Assess fairness and potential bias in the underlying data, considering its relevance to the agency's objectives.** Using the Aequitas Bias and Fairness Audit Toolkit, data regarding the small arts and culture organizations relative to grant application success as determined by the human panel reviewers was assessed across many different measures of bias.

4. **Develop test datasets, including counterfactual examples, to evaluate the AI model's performance on specific target groups of interest.** An experimental dataset of hypothetical applications was created, breaking the applications into four distinct groups based on size and demographic spread. All other characteristics of the applications were kept identical, allowing for assessment of the applications purely on the primary outcome goals of the agency. (Ongoing research on this topic will also shift which panelists evaluated each application to identify potential differences at the individual reviewer level. Additionally, counterfactuals were developed to assess how the real applications might have been evaluated differently had their demographic or size flag been switched, and this will be tested in ongoing research.)

5. **Run the AI model on these test datasets across multiple iterations, adjusting key input parameters as needed.** Using code derived from the original AI model testing and fine-tuning process, the AI model was applied to the experimental dataset, allowing for an iterative analysis of identical applications systematically adjusted relative to demographic identification and organization size. Example code for this process can be found here.[55]

6. **Analyze the model's outputs to assess its performance across different target groups and key evaluation metrics.** After analyzing the experimental data that was consistent in conceptual narrative and differed only by demographic status and budget size of the organization, the results showed that the demographic status of an organization did not materially impact the score received. Conversely, small organizations received scores 4.5 percent lower, on average, than larger organizations.

While this disparate impact appears to exist statistically on a small experimental sample, the grantmaker specifically planned for the allocation of 20 grants to small organizations and 15 to large organizations. As such, the small and large organizations were essentially not competing against each other. Additionally, the applicant pools of both large and small organizations showed that, at least for large organizations, there was no statistical disparate impact based on demographic status.

Beyond only evaluating the numeric scores, a concise content analysis of generated reviewer notes showed little variation across narrative content.

55. See example code here: https://github.com/danielfonner/IBM_RAI-Ev/blob/36f84e47996affcd11377dd9ba1de772b886b028/GITHUB_IBM_AuditModel.ipynb.

5 Identify and Report Potential Changes for Future Programs

The final step in the RAI-Ev process is to be as transparent as possible in reporting the performance of the model to internal stakeholders as well as any effected constituents or community members. While RAI-Ev models cannot be assumed to function in the exact same way as humans, the information derived from the auditing and evaluation process could inform beneficial changes to future government programs to improve outcomes to better align with agency priorities and outcome goals.

The GAO's *Yellow Book of Government Auditing Standards* provides specific guidance on how federal agencies are to report on auditing activities, which would include RAI-Ev audits and evaluations.[56] The generally accepted government auditing standards (GAGAS) specify that any reporting on government audits must include the following:

- The objectives, scope, and methodology of the audit
- The audit results, including findings, conclusions, and recommendations, as appropriate
- A summary of the views of responsible officials
- If applicable, the nature of any confidential or sensitive information omitted

Public administrators should clearly and objectively communicate the RAI-Ev objectives, scope, and limitations in their reports. Reports should also outline any constraints, such as restricted access to records and other data, to ensure any stakeholders accurately understand the findings, conclusions, and recommendations. The ultimate output of the RAI-Ev process should be a blueprint for identifying successes in program delivery as well as highlighting areas of needed change in the human-driven decision-making process of future government programs.

56. U.S. Government Accountability Office. 2024. *Government Auditing Standards.* GAO-24-106786. Washington, D.C. https://www.gao.gov/assets/d24106786.pdf.

RAI-Ev: Identify Potential Future Changes for a COVID-Relief Grant Program

At a basic level, the Aequitas assessment of potential bias in the panel reviewer decisions relative to organizational size and demographic spread focus of the grantmaker showed some areas of concern regarding the treatment of the diverse segments within the applicant pool. However, in the context of this specific grant program, small and diverse organizations were a primary, stated focus for funding, which aligns with the Aequitas results.

For the experimental approach of assessing new applications varied only by demographic status and organization size, no material disparate impact was identified. Small organizations did receive scores 4.5 percent lower than large organizations, on average, but initial planning for the program bifurcated the applications into two pools that did compete for the same pool of funding.

Even though there does not appear to be statistically identified bias across this organization grant program, the budget size disparity in scoring may warrant discussion in future grant programs where small and large organizations might be in the same applicant pools. Research shows that small organizations often lack capacity or training needed to apply competitively for complex grant programs.[57] Agency administrators should consider additional training and instruction for panel reviewers to take organization size into consideration when evaluating future grant applications.

While this summary of the results is not fully comprehensive and in GAGAS form due to ongoing research, government agencies should keep the *Yellow Book* guidelines in mind throughout the RAI-Ev reporting process.

57. "Hearing before the Senate Committee on Homeland Security and Governmental Affairs: Improving Access to Federal Grants for Underserved Communities: Written Statement of the Council of Nonprofits." 2023. https://www.councilofnonprofits.org/files/media/documents/2023/ncn-federal-grants-hearing-testimony-5-2-2023.pdf.

Recommendations

As artificial intelligence becomes an integral tool across the public sector, government agencies must ensure its deployment aligns with principles of accountability, fairness, and transparency. While AI offers significant opportunities to enhance decision-making, its adoption must be guided by a responsible approach, such as the RAI-Ev framework described in this report, that avoids common pitfalls such as overhyped expectations, lack of oversight, or misalignment with existing governance frameworks.

The following recommendations outline key strategies for integrating AI responsibly within public sector evaluation processes, ensuring that its use remains aligned with democratic values, policy objectives, and ethical considerations.

1, Broaden Perspective on AI's Role in the Public Sector

Government administrators should move beyond viewing AI solely as a means of improving efficiency or making predictions. While these functions are invaluable, AI can also be used to analyze complex systems, uncover insights, and enhance transparency. By embracing a broader perspective through frameworks such as RAI-Ev, administrators can unlock new opportunities to improve public programs.

2. Leverage Open-Source Models and Systems for Public Sector AI Implementation

Where applicable, government agencies should prioritize the use of open-source AI models and systems when using the RAI-Ev framework. Open models enable greater transparency, as they allow internal and external experts and auditors to examine how decisions were made. This openness can enhance public trust and make it easier to identify and address biases or errors in AI systems and human-driven processes.

3. Prioritize Transparent Reporting and Data Privacy Infrastructure in AI-Driven Evaluation Processes

Government agencies should adopt privacy-preserving techniques when handling sensitive data in AI-assisted evaluations. At the same time, agencies should accompany model outputs with clear, plain-language interpretability reports that document how the model was used, the data it relied on, identified risks or limitations, and how results should be interpreted. Together, these practices promote responsible data stewardship, enhance transparency, and build public trust in AI-driven program evaluation.

4. Embed AI Use into Existing Frameworks

AI use should not be for just unique or one-off tasks but integrated into existing frameworks that agencies are very familiar with, and in the case of this research, RAI-Ev for Evidence Act-aligned performance audits and program evaluations. Administrators should apply established principles of accountability and transparency to the development and use of such AI systems. By aligning AI initiatives within these established frameworks, agencies can ensure that new technologies enhance, rather than replace, their existing practices and domain knowledge.

5. Recognize AI limitations

AI technologies are powerful tools, but they have inherent limitations that public administrators must acknowledge to ensure responsible implementation even with frameworks such as RAI-Ev. Overhyping AI's capabilities can lead to unrealistic expectations and misinformed decision-making, while anthropomorphizing AI (i.e., treating it as if it possesses human-like reasoning, intent, or judgment) distorts the model's actual functionality, risks, and assignment of liability. By recognizing AI limitations and avoiding anthropomorphization, government agencies can foster a balanced, realistic, and ethical approach to AI deployment, ensuring that technology serves the public good without undermining human oversight and accountability.

6. When using AI in the public sector, agencies should default to responsible AI practices even when not required

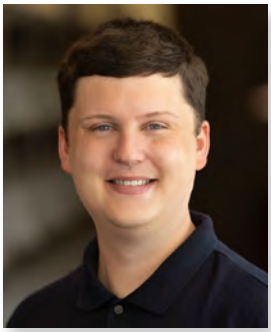
Government agencies should proactively adopt responsible AI practices, even in cases where regulations or legal requirements do not explicitly mandate them. As AI continues to evolve, ethical and governance frameworks may lag technological advancements. By defaulting to responsible AI principles, agencies can better ensure that their AI usage aligns with public values, minimizes harm, and builds long-term trust in government operations.

Conclusion

AI has the potential to revolutionize program evaluation and performance auditing by providing deeper insights, improving efficiency, and promoting transparency. However, realizing this potential requires thoughtful planning and implementation. By broadening the perspective on AI's role, recognizing its limitations, leveraging open systems, embedding AI governance into existing frameworks, and implementing AI responsibly, government administrators can work toward AI that serves as a tool for public good. Research on this topic is ongoing in efforts to apply RAI-Ev to other public sector programs and to develop additional tools to aid public sector employees in their critical work.

Any use of AI in the public sector should aligne with responsible AI practices, such as those described for RAI-Ev. Ethical AI practices should not be applied only when required by government directive; they should be the default. The recommendations in this report not only enhance the effectiveness of applying AI in the public sector but also foster trust and accountability in their use, ensuring that government programs remain innovative, efficient, and effective.

About the Author



Daniel F. Fonner
Corporate Communication and
Public Affairs
Southern Methodist University

E: dfonner@smu.edu

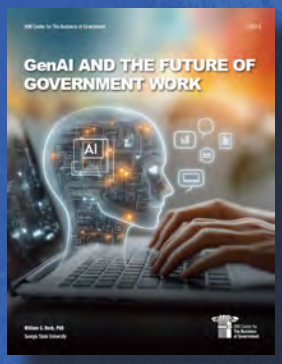
Daniel Fonner is the Associate Director for Research at SMU DataArts, the National Center for Arts Research at Southern Methodist University (SMU), as well as an adjunct lecturer in SMU's Division of Corporate Communication and Public Affairs. Daniel's teaching and research focus on data science for social good, employing artificial intelligence to improve public administration and support the arts and culture sector.

Prior to joining SMU, Daniel was a researcher at BOP Consulting in London (UK) and spent time as the Research and Policy Associate at the Greater Pittsburgh Arts Council (PA). He is a Fellow of the Royal Society of Arts (UK) and was a Fulbright Postgraduate Scholar where he studied cultural policy and the use of artificial intelligence to recreate lost works of art. Daniel is currently completing a Ph.D. in Computer Science at SMU focusing on Responsible Artificial Intelligence for Public Policy and Administration, with a secondary focus in Digital/Data Curation and Management through study at the University of North Texas. Daniel has received degrees from Duquesne University (Pittsburgh, PA), Carnegie Mellon University (Pittsburgh, PA), the University of Warwick (Coventry, UK), and Southern Methodist University (Dallas, TX).

Acknowledgments

I would like to thank all the staff at SMU DataArts for their input, support, assistance, and collaboration in furthering this research. Additionally, this research would not have been possible without the support and insights shared by Morgan Kasprowicz. I would also like to thank my collaborator in conducting many aspects of this research, Dr. Frank Coyle of the UC Berkeley School of Information and the Southern Methodist University Computer Science Department (retired).

Recent Reports from the IBM Center for The Business of Government



GenAI and the Future of Government Work

by William G. Resh

Embedding Strategic Foresight into Strategic Planning and Management

by Bert George

Resilience in action: Crisis leadership through innovation, collaboration, and human-centered solutions

by Julia Carboni

Leadership Framework for an Agile Government

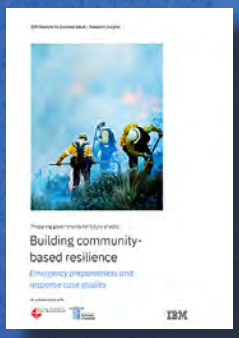
by Pallavi Awasthi and Kuang-Ting Tai

The Opportunity Project

by Joel Gurin and Matt Rumsey

Government's Digital DNA: Identity and Access Management for Public Sector Security

by Andrew Whitford



Building community- based resilience

by Authors of Case Study

AI in State Government

by Katherine Barrett and Richard Greene

For a full listing of our reports, visit businessofgovernment.org/reports

About the IBM Center for The Business of Government

Through research stipends and events, the IBM Center for The Business of Government stimulates research and facilitates discussion of new approaches to improving the effectiveness of government at the federal, state, local, and international levels.

About IBM Consulting

With consultants and professional staff in more than 160 countries globally, IBM Consulting is the world's largest consulting services organization. IBM Consulting provides clients with business process and industry expertise, a deep understanding of technology solutions that address specific industry issues, and the ability to design, build, and run those solutions in a way that delivers bottom-line value. To learn more visit ibm.com.

For more information:
Daniel J. Chenok
Executive Director
IBM Center for The Business of Government

600 14th Street NW
Second Floor
Washington, D.C. 20005
(202) 551-9342

website: www.businessofgovernment.org
e-mail: businessofgovernment@us.ibm.com

IBM Center for The Business of Government