Deloitte.

Developing and deploying trustworthy AI in Government

MAKING AN IMPACT THAT MATTERS since 1845

Introduction

Artificial intelligence (AI) and machine learning are powerful tools with the potential to revolutionize how the government delivers services to citizens. AI can help cut traffic by 25%, answer citizen questions to free up human workers, and even detect fraud in government programs, just to name a few examples.[1] But as with any powerful tool, AI comes with risks. AI and autonomous decision systems without sufficient controls can adversely affect the lives of real people. For example, consider the case of teachers being fired yet no one being able to explain why the AI-enabled evaluation system made that recommendation;[2] or a man detained for 30 hours due to a faulty facial recognition match;[3] or a woman's long-term health care support cut from 56 to 32 hours a week due to a coding error in an algorithm. The common theme in all three incidents is a result of unintentional bias, causing implicit generalizations and unintended outcomes.

To help address and mitigate those impacts, organizations need to develop controls and mechanisms to manage AI technology—enabling it to advance society while functioning in a trustworthy, equitable, and ethical manner.

AI algorithms do not have ethics in themselves; people have ethics. If government organizations want people to trust that the outcomes of their AI tools are ethical, they need to build principles into the development of the tools themselves.

AI systems can pose diverse ethical risks

Despite significant benefits, the rapid use of AI in government has raised some moral, social, and ethical concerns. Key risks associated with designing, developing, and deploying AI include:

Issues with bias, fairness, transparency, equity, and explainability: AI is increasingly used to make important hiring, economic, medical, and education decisions. However, an algorithm is only as good as the dataset used to train it. If the underlying data is biased or unrepresentative, the algorithm will reflect and propagate that bias. Furthermore, in many scenarios, government organizations need to be able to explain the rationale behind the recommendations provided by AI systems, which can often be challenging.

Use of AI for disinformation and digital manipulation: AI can be used to spread disinformation. For example, it can create "deepfakes," in which politicians, public figures, journalists, or others can be made to appear as if they have said things they did not.

Enables collection of private data that can be misused: Vast datasets about citizen profiles can be used to provide seamless and proactive services to citizens, but the same datasets can be used to surveil and alter the behavior of individuals found unacceptable to governments and corporations.

Lack of governance and accountability: Government agencies are large organizations tackling complex problems. The size and complexity can complicate the governance and accountability as it relates to AI. If roles and responsibilities are not clearly articulated across organizational elements as well as with agency leadership, IT staff, vendors and auditors, critical issues that impact AI's performance may be overlooked.

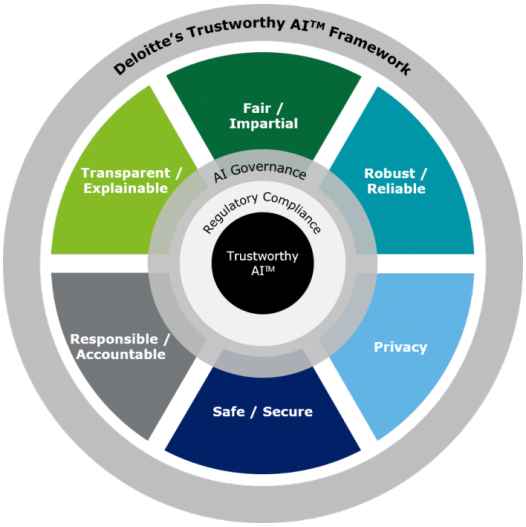
Governing artificial intelligence

Recognizing the gravity of AI and its risks, governments are taking action. Using their roles as both regulators and buyers and users of AI, many governments are implementing various solutions to address the associated risks. For instance, in its role as regulator, the European Union's General Data Protection Regulation (GDPR) enacted strict controls over cross-border data transmissions, giving citizens the right to be "forgotten" and mandating organizations, including government agencies, to provide "data protection by design" and "data protection by default." Similarly, the US government has used its role as buyer to shape AI development. In December 2020, the US government issued the Executive Order on Trustworthy AI that provides nine guiding principles to develop, use, and deploy AI in federal agencies. As a buyer, the US Department of Defense (DoD) started to embed AI ethics principles in RFPs and RFIs.[4]

These efforts often remain at a high level or exist only as one-off solutions. We live in a digital world where these technologies are already in use, and a high-profile failure of one of those uses can impact trust in AI across the board. So how can AI be further safeguarded so its benefits can be realized? The starting point is to shift the discussion from why people mistrust AI when it fails, to understanding why they can trust AI when it works. In short, AI needs to earn people's trust.

To build trust in AI, make Trustworthy AI™

Trust is fragile—earned over time and lost in an instant. Because of this, a consistent framework to ethical issues can make sure an organization doesn't overlook any key factors that underpin trust in AI. Deloitte's Trustworthy AI framework lays out six key dimensions to build trust in AI. (See "The Dimensions of Trustworthy AI" callout box for more details.) The framework is designed to help agencies identify and mitigate potential risks related to AI ethics at every stage of the AI development lifecycle.



The Dimensions of Trustworthy AI

- **Fair and impartial:** AI must be designed and trained to follow a fair and consistent process that takes the bias out of the decisions. It must include internal and external checks to reduce discriminatory bias. Further, to reduce historical biases in data, it's important to use training datasets that are diverse in terms of race, gender, ethnicity, and nationality.
- **Transparent and explainable**: Agencies should emphasize creation of algorithms that are transparent and can be explained to people who are being impacted by those algorithms. The Defense Advanced Research Projects Agency (DARPA) has launched a program on Explainable AI that aims to produce more explainable models, maintain a high level of accuracy, and enable human users to understand models to effectively build trust in AI.
- **Responsible and accountable:** AI systems need policies about who is responsible and accountable for their output or decision making. This will increasingly become more important as AI is used in critical applications such as disease diagnosis, autonomous vehicles, and deciding whether a citizen is eligible for a certain service.
- **Safe and secure:** For AI to be trustworthy, it must be protected from cybersecurity risks that could manipulate the models and result in digital or physical harm. As agencies increase dependency and use of AI for critical services, it would create further incentives for attackers to target the AI systems. Through the Joint Common Foundation (JCF) cloud, the Department of Defense (DoD) is adopting the DevSecOps approach that embeds security and privacy controls from the start of the software development cycle. "What the JCF is trying to do is democratize the whole process of DevSecOps for Artificial Intelligence/Machine Learning and make it easier to secure and rapidly authorize AI/ML capabilities," says Tom Morton, Deputy Director for the JCF.[6]
- **Privacy:** Privacy is critical for AI since the sophisticated insights generated by AI systems often stem from data that is more detailed and personal. Trustworthy AI must comply with data regulations and only use data for the stated and agreed-upon purposes. For instance, Australian federal agencies must conduct privacy impact assessment (PIA) for "high privacy risk projects."[7] A PIA is a systematic assessment that identifies the impact on the privacy of individuals and provides recommendations for mitigating, managing, or eliminating privacy risks through a privacy-by-design approach.[8]
- **Robust and reliable:** AI needs to be at least as robust and reliable as the traditional systems, processes, or people it is augmenting or replacing. It must generate consistent and reliable outputs, especially as it is scaled.

Putting principles into practice: Operationalize and institutionalize ethical AI

Formulating principles or guidelines is helpful, but not sufficient. To bring real value to citizens, AI needs to operate at scale without reliance on manual reviews or interested individuals to stay trustworthy. The following are strategies that can be used for operationalizing and institutionalizing ethical AI.

Institutionalizing governance: Agencies need governance structures responsible for implementing ethical AI practices. The diverse nature of AI means governance may stretch across portfolios of many C-suite executives and agency leaders. This can be positive because it means existing governance structures related to privacy, cybersecurity, compliance, and other data-related risks can be brought together into an AI ethics task force or data governance board. If such structures do not exist, an organization can establish one by tapping employees with skills related to risk, compliance, privacy, and analytics.

But this cross-cutting approach can pose a challenge because an executive's main efforts are likely to be elsewhere. Agencies need a leader responsible for keeping a focus squarely on ethical AI practices and fostering the coordination to make it happen. States such as California are turning to their Chief Data Officers (CDOs) to help them think through AI, the data, as well as the ethics and trustworthiness of the AI being developed. Similarly, DoD's Joint Artificial Intelligence Center (JAIC) created the position of Head of AI Ethics policy in 2020.[8] This person is responsible for operationalizing the DoD's ethical AI principles. The department also established a DoD-wide interdisciplinary Responsible AI Subcommittee that collaborates on prorogating Responsible AI across DoD.[9] In comparison, other agencies, such as Health and Human Services, created a new position, Chief Artificial Intelligence Officer (CAIO). Whether an organization taps an existing C-suite leader or creates a new position, having a role with singular focus implementing AI principles is one way to achieve expected results.

Embedding ethical AI in organizational culture: Ethical AI needs to be ingrained in organizational culture just like the mission and vision of agencies. Many private and public sector organizations conduct workshops, open houses, and events to make sure every employee understands the mission and vision—shared values that bind the organization. Agencies do not have to look further than cybersecurity awareness campaigns they would launch internally. A decade back, organizations hardly paid attention to cybersecurity, and now every employee is expected to have a grasp of cyber risks that exist.

For AI, the JAIC has started the journey of embedding ethical AI practices across the DoD. The center launched a Responsible AI Champions pilot program that took 15 cross-functional individuals through an experiential learning journey on AI ethics. The pilot identified tactics to operationalize AI and create a network of ambassadors for Responsible AI. AI ethics training has been incorporated into the department's broader workforce education strategy and was delivered to a group of acquisition and product capability managers in October 2020.[10]

Monitoring AI using tools and resources: Government agencies can learn from financial institutions that put algorithm risk guidance into practice at scale. After the global financial crisis exposed the risks of inaccurate algorithm-driven models, the Federal Reserve and Office of the Comptroller of the Currency (OCC) issued the Supervisory Guidance on Model Risk Management (SR 11-7). It required the identification and estimation of adverse consequences of inaccurate or misused models. SR 11-7 obligates users understand the limitations of models and avoid using models for uses other than originally intended. It mentions models should be validated regularly to ensure they are performing as expected.[11] The guidance further emphasizes developing and maintaining governance, policies, and controls to manage and mitigate risks of models.[12] This operationalizing of guidance led many banks to actively publish their model risk management practices in their annual reports.[13]

Just like financial institutions did in response to SR11-7 regulation, government organizations can deploy solutions to continuously monitor AI models, detect biases, vulnerabilities, and validate that AI solutions are operating as intended. Monitoring at the scale of government operations requires a systemic quality assurance process that includes not just the typical elements of software quality, but also the six dimensions of our Trustworthy AI framework. The scale and volume of data necessitate automation as the key to the quality process. At the same time, the six dimensions of Trustworthy AI represent real, human values and the monitoring process cannot be left only to technological tools. Further, leaders and data scientists should champion the testing and monitoring of AI models to help ensure conformance to the values an organization wishes to uphold.[14]

Operationalizing AI: Where to start

Government leaders should consider starting now to prepare their organizations to build and encourage trustworthy AI. Agencies can take a three-pronged approach by setting-up a governance structure, cultivating an organizational culture that nurtures ethical AI, and continuing to use and develop tools and resources to monitor AI systems.

Institutionalizing governance is not just about setting up a team or committee to lead the implementation of ethical AI. Ensuring data integrity, providing tactical guidance to AI developers, and supporting and encouraging an independent governance structure can make governing AI more effective and feasible:

- **Focus on data management and governance:** An organization's ability to skillfully work with data is critical to AI quality and explainability. Effective data governance systems can help increase confidence in AI and leaders' preparedness to manage its ethical implications.
- **Optimize tools and guidance for AI developers:** While a framework provides high-level direction, AI developers need essential guidance at a granular level. For example, while all six dimensions of Trustworthy AI are critical, agencies may need to manage tradeoffs. Often there is a trade-off between accuracy (robustness) and explainability. To manage those tradeoffs, an organization can use a tool by which developers evaluate the importance of explainability for a given solution and compare it against the required accuracy.
- **Independent governance structure:** Like the risk management function of financial institutions, agencies should strive to create an independent governance structure. This allows agencies to ethically govern AI by avoiding conflict and building effective control mechanisms to help prevent negative impacts of unethical AI.
- **Embedding ethical AI in organizational culture** requires building awareness of available guidance, encouraging diversity of thought, and cultivating trust between humans and machines.
- **Encourage diversity of thought:** Data capabilities and governance processes create checkpoints to evaluate adherence to a variety of standards, but without a diversity of thoughts, important insights may not surface. Ensuring and creating a culture supportive of diversity of thought increases the likelihood problems are flagged during the design and review phases to reduce unintentional biases that can crop up—and not after the deployment of AI solutions.
- **Build trust in employees** – For a responsible and ethical AI, agencies must work on building trust in employees for AI technologies. AI systems that are human-centered, designed with intuitive user interfaces, and extend and augment the capabilities of employees can gain the trust of employees more rapidly.

Monitoring AI calls for periodically validating and auditing AI to make sure AI is performing as expected. However, critical AI systems should allow the option of manual overrides, or even human confirmation before taking action to help further reduce risk.

Building trustworthy AI is a complex yet worthwhile task. It requires not only focusing on all six dimensions, but also establishing a governance structure that implements the framework with buy-in from leadership, employees, vendors, and end-users. Agencies that invest in building trust in AI can seize the benefits of AI to help achieve mission outcomes, improve human experience, and deliver efficient services.

Contact us to learn more:

Ed Van Buren
Executive Director, Deloitte AI Institute for Government
Principal
Deloitte Consulting LLP
emvanburen@deloitte.com
+1 571 882 5170

Tasha Austin
Director, Deloitte AI Institute for Government
Principal
Deloitte & Touche LLP
laustin@deloitte.com
+1 202 270 8379

Deloitte AI Institute for Government

The Deloitte AI Institute for Government is a hub of innovative perspectives, groundbreaking research, and immersive experiences focused on artificial intelligence (AI) and its related technologies for the government audience. Through publications, events, and workshops, our goal is to help government use AI ethically to deliver better services, improve operations, and facilitate economic growth.

We aren't solely conducting research—we're solving problems, keeping explainable and ethical AI at the forefront and the human experience at the core of our mission. We live in the Age of With—humans with machines, data with actions, decisions with confidence. The impact of AI on government and its workforce has only just begun.

1 https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/public-sector/lu-government-trends-2020.pdf
2 https://ainowinstitute.org/AI_Now_2018_Report.pdf
3 http://www3.weforum.org/docs/WEF_Global_Technology_Governance_2020.pdf
4 https://www.meritalk.com/articles/dod-begins-embedding-ai-ethics-principles-in-rfps/
5 https://www.ai.mil/blog_07_16_20-jaic_pushes_the_envelope_with_devsecops_jcf.html
6 https://www.themandarin.com.au/139670-new-resource-helps-agencies-know-when-to-conduct-privacy-impact-assessments/
7 https://www.oaic.gov.au/privacy/guidance-and-advice/guide-to-undertaking-privacy-impact-assessments/
8 https://www.meritalk.com/articles/jaic-hires-alka-patel-to-lead-ai-ethics/
9 https://www.ai.mil/blog_02_26_21-ai_ethics_principles-highlighting_the_progress_and_future_of_responsible_ai.html
10 https://www.ai.mil/blog_01_05_21-the_ai_ethics_journey_will_hit_new_heights_in_2021.html
11 https://bigdatapath.wordpress.com/2020/05/22/streamlining-the-ai-model-risk-management-process/
12 https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/risk/lu-model-risk-management.pdf
13 https://info1.exlservice.com/hubfs/EXL-WP-SR-11-07-Compliance-in-Model-Risk-%20Management.pdf
14 https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/chatterbox-placemat.pdf

Deloitte.