



ARTIFICIAL INTELLIGENCE

An Accountability Framework for
Federal Agencies and Other Entities

*Including Insights from
the Comptroller General's
Forum on the Oversight of
Artificial Intelligence*

Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities

What GAO Found

To help managers ensure accountability and responsible use of artificial intelligence (AI) in government programs and processes, GAO developed an AI accountability framework. This framework is organized around four complementary principles, which address governance, data, performance, and monitoring. For each principle, the framework describes key practices for federal agencies and other entities that are considering, selecting, and implementing AI systems. Each practice includes a set of questions for entities, auditors, and third-party assessors to consider, as well as procedures for auditors and third-party assessors.

Why GAO Developed This Framework

AI is a transformative technology with applications in medicine, agriculture, manufacturing, transportation, defense, and many other areas. It also holds substantial promise for improving government operations. Federal guidance has focused on ensuring AI is responsible, equitable, traceable, reliable, and governable. Third-party assessments and audits are important to achieving these goals. However, AI systems pose unique challenges to such oversight because their inputs and operations are not always visible.

GAO's objective was to identify key practices to help ensure accountability and responsible AI use by federal agencies and other entities involved in the design, development, deployment, and continuous monitoring of AI systems. To develop this framework, GAO convened a Comptroller General Forum with AI experts from across the federal government, industry, and nonprofit sectors. It also conducted an extensive literature review and obtained independent validation of key practices from program officials and subject matter experts. In addition, GAO interviewed AI subject matter experts representing industry, state audit associations, nonprofit entities, and other organizations, as well as officials from federal agencies and Offices of Inspector General.

Artificial Intelligence (AI) Accountability Framework

Data

Ensure quality, reliability, and representativeness of data sources and processing.

Data Used to Develop an AI Model

Entities should document sources and origins of data, ensure the reliability of data, and assess data attributes, variables, and augmentation/enhancement for appropriateness.

Data Used to Operate an AI System

Entities should assess the interconnectivities and dependencies of data streams that operationalize an AI system, identify potential biases, and assess data security and privacy.

Monitoring

Ensure reliability and relevance over time.

Continuous Monitoring of Performance

Entities should develop plans for continuous or routine monitoring of the AI system and document results and corrective actions taken to ensure the system produces desired results.

Assessing Sustainment and Expanded Use

Entities should assess the utility of the AI system to ensure its relevance and identify conditions under which the AI system may or may not be scaled or expanded beyond its current use.

Governance

Promote accountability by establishing processes to manage, operate, and oversee implementation.

Governance at the Organizational Level

Entities should define clear goals, roles, and responsibilities, demonstrate values and principles to foster trust, develop a competent workforce, engage stakeholders with diverse perspectives to mitigate risks, and implement an AI-specific risk management plan.

Governance at the System Level

Entities should establish technical specifications to ensure the AI system meets its intended purpose and complies with relevant laws, regulations, standards, and guidance. Entities should promote transparency by enabling external stakeholders to access information on the AI system.

Performance

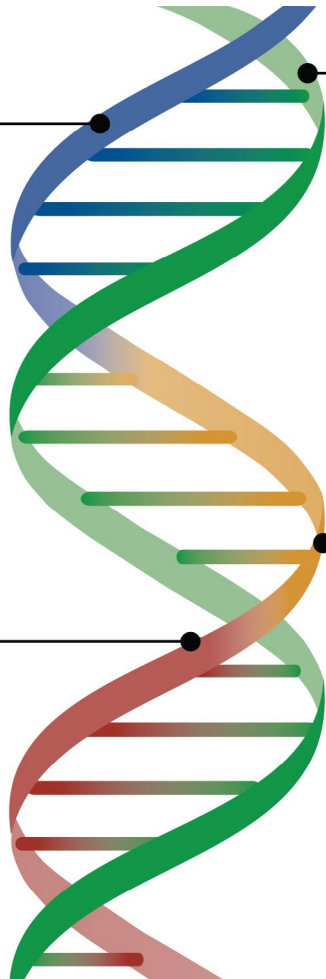
Produce results that are consistent with program objectives.

Performance at the Component Level

Entities should catalog model and non-model components that make up the AI system, define metrics, and assess performance and outputs of each component.

Performance at the System Level

Entities should define metrics and assess performance of the AI system. In addition, entities should document methods for assessment, performance metrics, and outcomes; identify potential biases; and define and develop procedures for human supervision of the AI system.



Contents

	Foreword	1
	Framework Summary	5
	Introduction	9
	Background	13
	Framework Principle 1. Governance	25
	Framework Principle 2. Data	37
	Framework Principle 3. Performance	48
	Framework Principle 4. Monitoring	59
Appendix I	Objectives, Scope, and Methodology	66
Appendix II	Insights from a Comptroller General Forum on Oversight of Artificial Intelligence	70
Appendix III	Forum Agenda	87
Appendix IV	List of Forum Participants	90
Appendix V	Information on Auditing Standards, Controls, and Procedures	92
Appendix VI	Bibliography	98
Appendix VII	GAO Contacts and Staff Acknowledgments	107
Tables		
	Table 1: Examples of AI Governance and Auditing Frameworks Developed by Foreign Governments	22
	Table 2: Examples of AI Governance Frameworks Developed by U.S. Government	22

Figures

Figure 1: The Three Waves of Artificial Intelligence	15
Figure 2: The Phases in the AI Life Cycle	17
Figure 3: Example of the Community of Stakeholders Engaged in AI Development	20
Figure A: The Five Components and 17 Principles of Internal Control	95

Abbreviation

AI	Artificial intelligence
CG	Comptroller General of the United States
DARPA	Defense Advanced Research Projects Agency
DOD	Department of Defense
ISO	International Organization for Standardization
IEEE	Institute of Electrical and Electronics Engineers
FDA	Food and Drug Administration
NIST	National Institute of Standards and Technology
ODNI	Office of the Director of National Intelligence
OECD	Organisation for Economic Co-operation and Development
OIG	Office of Inspector General
OMB	Office of Management and Budget

This is a work of the U.S. government and is not subject to copyright protection in the United States. The published product may be reproduced and distributed in its entirety without further permission from GAO. However, because this work may contain copyrighted images or other material, permission from the copyright holder may be necessary if you wish to reproduce this material separately.

The Honorable Roger Wicker
Ranking Member
Committee on Commerce, Science, and Transportation
United States Senate

Foreword

The year 2021 marks the one-hundredth year since the U.S. Government Accountability Office (GAO) was established in 1921. The work of federal agencies has changed significantly since GAO conducted its first audit, and GAO has steadfastly adapted to address changing accountability challenges over the decades. One thing has not changed—GAO still applies the same rigor toward oversight. Instead of paper ledgers and punch cards of yesterday, GAO now tackles accountability challenges through data science and emerging technologies.

In this year of GAO's centennial, it is fitting that the agency is focused on the future of audit through this accountability framework on artificial intelligence (AI). As a nation, we have yet to fully grasp the profound impacts AI is having and will have on government and the public. In a digital world that increasingly depends on algorithms to function, we are often asked—either implicitly or explicitly—to trust AI systems. But how do we know that AI is doing its job appropriately if there are no independent mechanisms to verify the system is doing so?

AI is evolving at a pace at which we cannot afford to be reactive to its complexities, risks, and societal consequences. It is necessary to lay down a framework for independent verification of AI systems even as the technology continues to advance. Auditors and the oversight community play a vital role in the trust but verify equation, and they need a toolkit to evaluate this changing technology. More importantly, organizations that build, purchase, and deploy AI need a framework to understand how AI systems will be evaluated.

GAO recognizes this need. The AI accountability framework is the latest example of GAO's foresight in providing forward-looking, prospective work for Congress and the American people on the most important national issues. This is the first of many steps on the journey of AI accountability. GAO looks forward to seeing this framework in use by federal agencies, and to working with the oversight community, researchers, industry, and the Congress to bring verifiable AI oversight to the cross-cutting work that GAO will undertake during its second century.

All of us at GAO who have worked on this report are grateful to the forum participants (see app. IV) and others who contributed to our work. We thank them for sharing their insights, experience, and time.

A handwritten signature in black ink, appearing to read "Taka". The signature is stylized with a large, sweeping loop at the beginning.

Taka Ariga, Chief Data Scientist and Director of Innovation Lab
Science, Technology Assessment, and Analytics
U.S. Government Accountability Office

A handwritten signature in black ink, appearing to read "T.M. Persons". The signature is written in a cursive, slightly slanted style.

Timothy M. Persons, PhD, Chief Scientist and Managing Director
Science, Technology Assessment, and Analytics
U.S. Government Accountability Office

A handwritten signature in black ink, appearing to read "Stephen J. Sanford". The signature is written in a cursive, slightly slanted style.

Stephen Sanford, Managing Director
Strategic Planning and External Liaison
U.S. Government Accountability Office

How to Use This Report

This report describes an accountability framework for artificial intelligence (AI). The framework is organized around four complementary principles and describes key practices for federal agencies and other entities that are considering and implementing AI systems. Each practice includes a set of questions for entities, auditors, and third-party assessors to consider, along with audit procedures and types of evidence for auditors and third-party assessors to collect. The report is organized into the following sections:

Summary

For each principle, a one-page summary of the key practices is provided.

Introduction

The introduction provides an overview of AI, the Comptroller General Forum on Oversight of Artificial Intelligence, and the approach we used in developing the framework.

Background

Background information on AI is presented, including defining AI and its life cycle, discussing AI technical and societal implications, and providing characteristics and examples of existing governance frameworks.

AI Accountability Framework

For each principle of the framework, we present descriptions of key practices and associated audit questions and procedures.

Appendixes

Appendix I describes the objectives, scope, and methodology used to carry out this work. Appendix II summarizes the findings from the Comptroller General Forum on Oversight of Artificial Intelligence. The remaining appendixes provide additional information about the forum and its participants, excerpts from the *Government Auditing Standards* and the *Standards for Internal Control in the Federal Government*, and a bibliography.



FRAMEWORK **SUMMARY**



1. Governance

To help entities promote accountability and responsible use of AI systems, GAO identified key practices for establishing governance structures and processes to manage, operate, and oversee the implementation of these systems.

Key Practices

Governance at the Organizational Level

- 1.1 Clear goals:** Define clear goals and objectives for the AI system to ensure intended outcomes are achieved.
- 1.2 Roles and responsibilities:** Define clear roles, responsibilities, and delegation of authority for the AI system to ensure effective operations, timely corrections, and sustained oversight.
- 1.3 Values:** Demonstrate a commitment to values and principles established by the entity to foster public trust in responsible use of the AI system.
- 1.4 Workforce:** Recruit, develop, and retain personnel with multidisciplinary skills and experiences in design, development, deployment, assessment, and monitoring of AI systems.
- 1.5 Stakeholder involvement:** Include diverse perspectives from a community of stakeholders throughout the AI life cycle to mitigate risks.
- 1.6 Risk management:** Implement an AI-specific risk management plan to systematically identify, analyze, and mitigate risks.

Governance at the Systems Level

- 1.7 Specifications:** Establish and document technical specifications to ensure the AI system meets its intended purpose.
- 1.8 Compliance:** Ensure the AI system complies with relevant laws, regulations, standards, and guidance.
- 1.9 Transparency:** Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system.

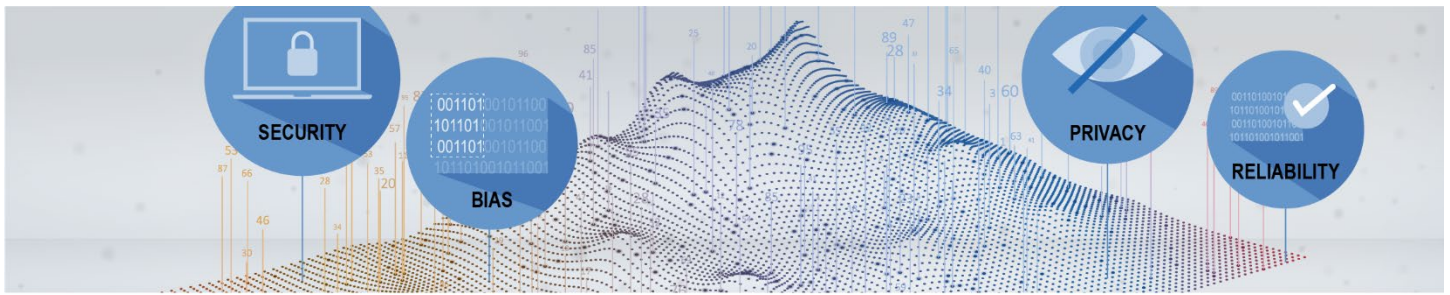


Example of an AI Governance Structure

In 2020, the Department of Defense (DOD) established an AI Executive Steering Group, which was created as the senior governance body to provide coordination and oversight of DOD's AI policies and activities. The Executive Steering Committee oversees nine subcommittees, one of which is on ethics. That subcommittee is responsible for providing practical guidance on how to apply the ethical principles for AI adopted by DOD to the different phases of the AI life cycle.

Selected Discussion from the Comptroller General Forum

- Entities should implement governance structures for AI systems that incorporate organizational values, consider risks, assign clear roles and responsibilities, and involve multidisciplinary stakeholders.
- Entities should define a governance structure that includes clear goals and objectives, which translates into systems requirements and performance metrics.
- Entities should include diverse perspectives from technical and non-technical communities throughout the AI life cycle to anticipate and mitigate unintended consequences including potential bias and discrimination.



2. Data

To help entities use data that are appropriate for the intended use of each AI system, GAO identified key practices to ensure data are of high quality, reliable, and representative.

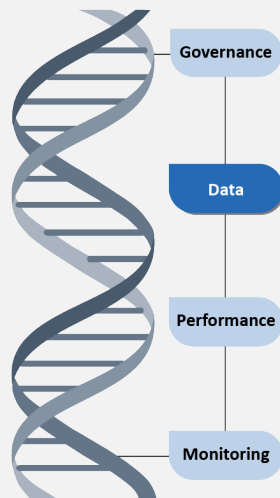
Key Practices

Data used for Model Development

- 2.1 Sources:** Document sources and origins of data used to develop the models underpinning the AI system.
- 2.2 Reliability:** Assess reliability of data used to develop the models.
- 2.3 Categorization:** Assess attributes used to categorize data.
- 2.4 Variable selection:** Assess data variables used in the AI component models.
- 2.5 Enhancement:** Assess the use of synthetic, imputed, and/or augmented data.

Data Used for System Operation

- 2.6 Dependency:** Assess interconnectivities and dependencies of data streams that operationalize the AI system.
- 2.7 Bias:** Assess reliability, quality, and representativeness of all the data used in the system's operation, including any potential biases, inequities, and other societal concerns associated with the AI system's data.
- 2.8 Security and privacy:** Assess data security and privacy for the AI system.



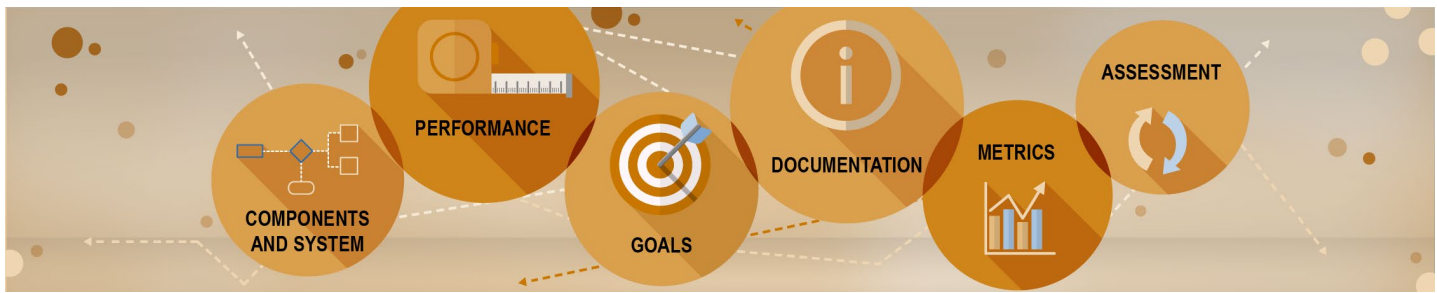
Example of Data Reliability

In 2019, the European Union Agency for Fundamental Rights released the report *Data Quality and Artificial Intelligence – Mitigating Bias and Error to Protect Fundamental Rights*. The report emphasizes the need for high-quality data and algorithms in machine learning systems and AI, and how transparency about data used in AI systems may help to prevent rights violations. The report also explains how AI systems use data, provides examples of how biases could be introduced, and provides examples of how low-quality data might affect accuracy and outcomes. Criteria for assessing data quality listed in the report include completeness, accuracy, consistency, timeliness, duplication, validity, availability, and whether the data are fit for the purpose.

Selected Discussion from the Comptroller General Forum

- Entities should provide documentation describing how training and testing data have been acquired or collected, prepared, and updated to demonstrate data quality and reliability.
- Entities should test data used in AI systems for biases. Biases may be introduced unintentionally during data collection and labeling.
- Entities should monitor data after deploying AI systems to identify potential data drift, which can lead to unintended consequences.

Source: GAO, majcot/stock.adobe.com (header); GAO (illustration). | GAO-21-519SP



3. Performance

To help entities ensure AI systems produce results that are consistent with program objectives, GAO identified key practices for ensuring that systems meet their intended purposes.

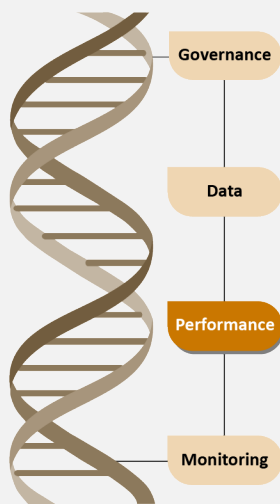
Key Practices

Performance at the Component Level

- 3.1 Documentation:** Catalog model and non-model components, along with operating specifications and parameters.
- 3.2 Metrics:** Define performance metrics that are precise, consistent, and reproducible.
- 3.3 Assessment:** Assess the performance of each component against defined metrics to ensure it functions as intended and is consistent with program goals and objectives.
- 3.4 Outputs:** Assess whether outputs of each component are appropriate for the operational context of the AI system.

Performance at the System-Level

- 3.5 Documentation:** Document the methods for assessment, performance metrics, and outcomes of the AI system to provide transparency over its performance.
- 3.6 Metrics:** Define performance metrics that are precise, consistent, and reproducible.
- 3.7 Assessment:** Assess performance against defined metrics to ensure the AI system functions as intended and is sufficiently robust.
- 3.8 Bias:** Identify potential biases, inequities, and other societal concerns resulting from the AI system.
- 3.9 Human supervision:** Define and develop procedures for human supervision of the AI system to ensure accountability.



Example of Performance Documentation

Industry and nonprofit entities provided several examples of how entities can document performance by recording several aspects of AI systems, including intended use, specifications, testing methodology and test results, ethical considerations, and evaluation. Each of those examples includes questions or factors for consideration to guide entities in designing, developing, and deploying AI systems.

Selected Discussion from the Comptroller General Forum

- Entities should document requirements—including performance metrics—for the AI system throughout the life cycle.
- Entities should document methods to assess performance—which can include input-output tests, stress tests, and evaluations of model drift—to ensure AI systems meet their intended goals.
- Entities should provide access to performance test results, change logs, and other documentation describing updates and key design choices, and provide a copy of the model or algorithm code to third-party assessors of AI systems.

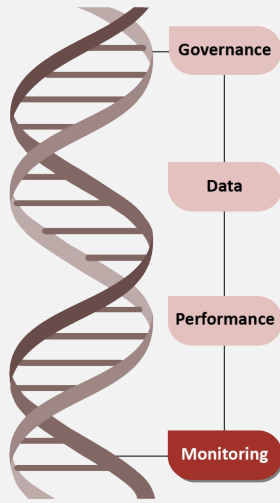
Sources: GAO, treenabeena/stock.adobe.com (header); GAO (illustration) | GAO-21-519SP.



4. Monitoring

To help entities ensure reliability and relevance of AI systems over time, GAO identified key practices for monitoring performance and assessing sustainment and expanded use.

Key Practices



Continuous Monitoring of Performance

- 4.1 Planning:** Develop plans for continuous or routine monitoring of the AI system to ensure it performs as intended.
- 4.2 Drift:** Establish the range of data and model drift that is acceptable to ensure the AI system produces desired results.
- 4.3 Traceability:** Document results of monitoring activities and any corrective actions taken to promote traceability and transparency.

Assessing Sustainment and Expanded Use

- 4.4 Ongoing assessment:** Assess the utility of the AI system to ensure its relevance to the current context.
- 4.5 Scaling:** Identify conditions, if any, under which the AI system may be scaled or expanded beyond its current use.

Example of Monitoring

In 2020, the World Economic Forum released the *Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations*, which includes guidance on data monitoring and a discussion of ongoing monitoring, review, and tuning of AI algorithms and models. The guidance suggests updating AI systems based on changes in the operational environment, as well as documenting when and how the update took place, and the impact it had on the model outputs.

Selected Discussion from the Comptroller General Forum

- Entities should continuously monitor and evaluate the AI system to ensure it addresses program objectives.
- Entities should monitor changes in the data and models to ensure relevance and appropriateness.
- Entities should continuously monitor the AI system to ensure the system is appropriate in its current operating context.

Source: GAO. | GAO-21-519SP

Introduction

AI is a transformative technology with applications ranging from medical diagnostics and precision agriculture, to advanced manufacturing and autonomous transportation, to national security and defense.¹ It also holds substantial promise for improving the operations of government agencies. An Executive Order issued in December 2020, *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*, states that federal agencies have begun to use AI to “accelerate regulatory reform”, “combat fraud, waste, and abuse committed against taxpayers” and “identify information security threats,” among other purposes.² More recently, guidance from the National Security Council has also recognized the importance of AI in shaping the economic and military balance among the world’s leading powers.³

However, AI systems pose unique challenges for independent assessments and audits to promote accountability because their inputs and operations are not visible to the user. Such a system can be an opaque “black box,” either because the inner workings of the software are inherently very difficult to understand, or because vendors do not reveal them for proprietary reasons. This lack of transparency limits the ability of auditors and others to detect error or misuse and ensure equitable treatment of people affected by AI systems.⁴

Bias is not specific to AI, but the use of AI has the potential to amplify existing biases and concerns related to civil liberties, ethics, and social disparities. Biases arise from the fact that AI systems are created using data that may reflect preexisting biases or social inequities. The U.S. government, industry leaders, professional associations, and others have begun to develop principles and frameworks to address these concerns,

¹Office of Science and Technology Policy (OSTP), Exec. Office of the President, *American Artificial Intelligence Initiative: Year One Annual Report*, (Feb. 2020).

²Exec. Order No. 13,960, *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government* (Dec. 3, 2020), 85 Fed. Reg. 78,939, (Dec. 8, 2020).

³National Security Council, Exec. Office of the President, *Interim National Security Strategic Guidance*, (Mar. 2021).

⁴Select Committee on Artificial Intelligence of the National Science and Technology Council, *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update* (Washington, D.C.: Executive Office of the President, June 2019).

but there is limited information on how these will be implemented to allow for third-party assessments and audits of AI systems.⁵

The December 2020 Executive Order directed federal agencies to be guided by nine common principles for the design, development, acquisition, and use of AI, emphasizing that AI use must be lawful and respectful of our nation's values; purposeful and performance-driven; accurate, reliable, and effective; safe, secure, and resilient; understandable; responsible and traceable; regularly monitored; transparent; and accountable.⁶ The Order also established a government-wide process for implementing these principles by directing the Office of Management and Budget to develop common policy guidance. In February 2020, the Department of Defense (DOD) adopted its own set of five AI ethical principles, with the goal of ensuring DOD's use of AI is responsible, equitable, traceable, reliable, and governable.⁷ The United States is also a member of the Organisation for Economic Co-operation and Development, which issued its *Recommendation of the Council on Artificial Intelligence* in May 2019. It includes five principles for responsible stewardship of trustworthy AI. Those principles have now been adopted by 46 countries.⁸

The *Government Auditing Standards* (commonly referred to as the Yellow Book) also provides guidance and notes that "those charged with

⁵A. Jobin, M. Ienca, and E. Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence*, vol. 1 (2019): pp. 389–399.

⁶85 Fed. Reg. 78,939, 78,940 (Dec. 8, 2020).

⁷Department of Defense (DOD), *Ethical Principles for Artificial Intelligence* (Washington, D.C.: Feb. 24, 2020), accessed June 2, 2021, https://www.ai.mil/docs/Ethical_Principles_for_Artificial_Intelligence.pdf. In addition to adopting ethical principles, DOD's Office of Inspector General issued an audit report in June 2020, which recommended that DOD establish an AI governance framework that includes a standard definition of AI, a central repository for AI projects, and a security classification guide. Report available at <https://media.defense.gov/2020/Jul/01/2002347967/-1/-1/1/DODIG-2020-098.PDF>. In July 2020, the Office of the Director of National Intelligence (ODNI) released the *Principles of Artificial Intelligence Ethics for the Intelligence Community* to guide personnel on whether and how to develop and use AI. See <https://www.dni.gov/index.php/features/2763-principles-of-artificial-intelligence-ethics-for-the-intelligence-community>.

⁸Organisation for Economic Co-operation and Development (OECD), *Recommendation of the Council on Artificial Intelligence OECD/LEGAL/0449* (Paris, France: adopted May 22, 2019).

governance, and the public need to know whether (1) management and officials manage government resources and use their authority properly and in compliance with laws and regulations; (2) government programs are achieving their objectives and desired outcomes; and (3) government services are provided effectively, efficiently, economically, ethically, and equitably.”⁹ These aspects are essential for ensuring accountability and transparency over government programs and processes. In addition, the *Standards for Internal Control in the Federal Government* (commonly referred to as the Green Book, and hereafter referred to as Federal Internal Control Standards) note that management and other personnel should design and implement an effective control system to provide reasonable assurance that an entity’s objectives will be achieved.¹⁰

To understand how to enable third-party assessments and audits of AI systems, the Comptroller General of the United States (CG) convened a Forum on the Oversight of Artificial Intelligence on September 9 and 10, 2020. The purpose of the forum was to bring together experts from across the federal government, industry, academia, and the nonprofit sectors. These experts discussed how recent principles and frameworks on the use of AI can be operationalized into practices for managers and supervisors of these systems, as well as third-party assessors. The forum included topics such as governance factors to consider in auditing AI systems, criteria auditors can use in assessing AI systems, issues and challenges in auditing AI systems in the public sector, and testing AI systems for bias and equity. The emphasis of the CG Forum was on discussing substantive approaches third-party assessors and auditors should take to develop credible assurance assessments of AI systems.

In this report, we present our AI Accountability Framework, including the results of the CG Forum. The audience for this framework is broad and includes federal agencies, state and local governments, industry, civil liberties advocacy groups, academia, and research institutions.

We developed the framework based on the following sources: (1) literature on accountability, governance frameworks, and principles on the use of AI; (2) presentations by and comments made by forum experts during the CG Forum; (3) interviews with subject matter experts including

⁹GAO, *Government Auditing Standards 2018 Revision Technical Update April 2021*, [GAO-21-368G](#) (Washington, D.C.: Apr. 14, 2021).

¹⁰GAO, *Standards for Internal Control in the Federal Government*, [GAO-14-704G](#) (Washington, D.C.: Sept. 10, 2014).

federal auditors and program managers, a state auditor, civil liberties advocates, industry representatives and legal counsel, developers, privacy experts, and data scientists; (4) GAO auditing standards and federal internal controls; (5) technical review of the framework and an outline of the forum findings by forum participants, including officials from three federal agencies and two Offices of Inspector General; and (6) internal review by GAO subject matter experts. Based on these sources, we developed four principles, which address Governance, Data, Performance, and Monitoring. For each principle, the framework includes the following: key practices, which we developed by synthesizing information and identifying at least two sources that noted the importance of a certain practice in implementing AI systems; key questions, which we developed from information provided during the CG forum, interviews with experts, and documents; and audit procedures, which we developed by reviewing the types of evidence noted in the Government Auditing Standards. We defined a practice as a key practice if at least two independent sources described it as important for implementing an AI system. This report focuses on the current AI technology, which relies heavily on machine learning methods. The text box below shows the organization of the framework.

A Guide for Reading the Framework

This framework is organized around four complementary principles:

Principles

Governance

Promote accountability by establishing processes to manage, operate, and oversee implementation.

Data

Ensure quality, reliability, and representativeness of data sources, and processing.

Performance

Produce results that are consistent with program objectives.

Monitoring

Ensure reliability and relevance over time.

For each principle, the Framework provides the following:

Key Practices for entities using AI systems.

Key Questions for entities, auditors, and third-party assessors.

Audit Procedures with the types of evidence for auditors and third-party assessors to collect.

The key practices address an AI system's technical performance, societal concerns as it relates to its use, or both. The questions are for entities to consider as they develop, select and implement AI systems. They also reflect the types of issues third-party assessors and auditors may inquire about when assessing the implementation of a practice. The audit procedures include the types of evidence for auditors and third-party assessors to collect and assess.

Source: GAO. | GAO-21-519SP

We conducted our work from February 2020 to June 2021 in accordance with all sections of GAO’s Quality Assurance Framework that are relevant to our objective. That framework requires that we plan and perform the engagement to obtain sufficient and appropriate evidence to meet our stated objective and to discuss any limitations in our work. We believe that the information and data obtained, and the analysis conducted, provide a reasonable basis for any conclusions in this product.

Background

Applications of AI are found in everyday technologies, such as video games, web searching, spam filtering, and voice recognition. More broadly, AI has applications across a variety of sectors, including transportation, health care, education, finance, defense, and cybersecurity.¹¹

Defining AI

The term AI has a range of meanings in the scientific literature. Recently, the U.S. government included a definition of AI in the “National Artificial Intelligence Initiative Act of 2020,” but this definition has not yet been widely adopted within the data science community.

Section 5002 of the National Defense Authorization Act for Fiscal Year 2021, defines AI as: a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to—(A) perceive real and virtual environments; (B) abstract such perceptions into models through

¹¹Congressional Research Service, *Overview of Artificial Intelligence*, IF-10608, ver. 3 (Oct. 24, 2017).

analysis in an automated manner; and (C) use model inference to formulate options for information or action.¹²

For the purpose of the Framework, we rely on a set of generalized characteristics of AI that are broader than the recently enacted definition.

Characteristics of AI

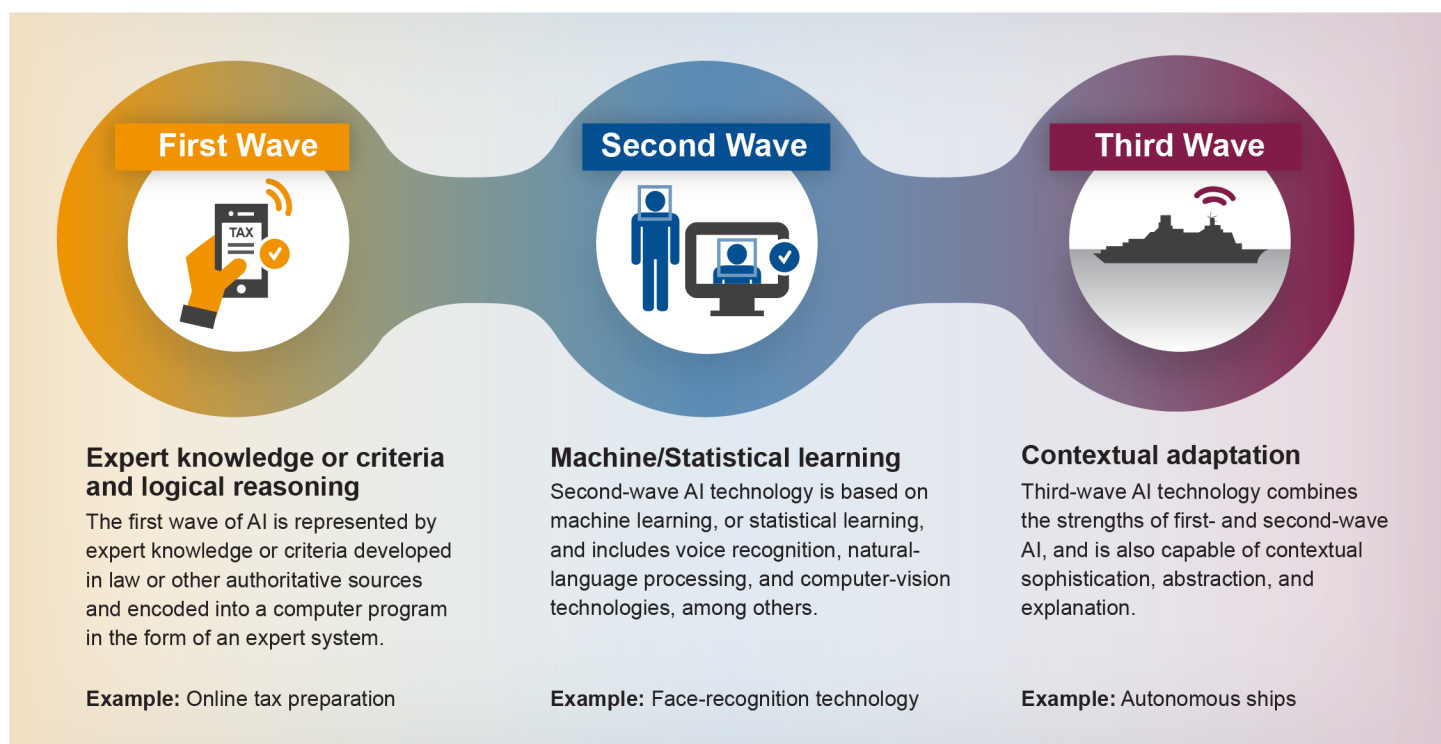
In *Artificial Intelligence: Emerging Opportunities, Challenges, and Implications*, we described AI as having three distinct waves of development, as shown in figure 1.¹³ Early AI technologies, the first wave, were often expert or rules-based systems, whereby a computer is programmed based on expert knowledge or criteria and produces outputs consistent with its programming. Current, second-wave AI systems, based on machine learning begin with data and infer rules or decision procedures that predict specified outcomes. Third-wave AI systems combine the strengths of first- and second-wave AI systems, while also being capable of contextual sophistication, abstraction, and explanation. Additionally, third-wave AI systems are not only capable of adapting to new situations, but also are able to explain to users the reasoning behind these decisions.¹⁴

¹²William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021 (NDAA FY21), Pub. L. No. 116-283, § 5002, 134 Stat. 3388 (2021). The National Artificial Intelligence Initiative Act of 2020 was enacted as Division E of the NDAA FY21. The act also creates the National AI Initiative. Its purpose is to ensure continued U.S. leadership in AI research and development; lead the world in the development and use of trustworthy AI systems in the public and private sectors; prepare the present and future U.S. workforce for the integration of AI systems across all sectors of the economy and society; and coordinate ongoing AI research, development, and demonstration activities among the civilian agencies, the Department of Defense, and the Intelligence Community to ensure that each informs the work of the others. Among other things, the AI Initiative will engage in interagency planning and coordination of federal AI research, development, demonstration, and standards engagement among the civilian agencies, DOD, and the Intelligence Community. The act also established the National AI Initiative Office within the White House Office of Science and Technology Policy, which, among other things, is to serve as the point of contact and conduct public outreach to diverse stakeholders on federal AI activities. Further, the act establishes an Interagency Committee, creates a National AI Advisory Committee, requires an AI impact study on the workforce, establishes a National AI Research Resource Task Force, and it calls for collaboration and stakeholder involvement of the private sector throughout the federal government, among other activities. NDAA FY21, Pub. L. No. 116-283, § 5101-06, 134 Stat. 3388 (2021).

¹³GAO, *Artificial Intelligence: Emerging Opportunities, Challenges, and Implications*, [GAO-18-142SP](#) (Washington, D.C.: Mar. 28, 2018).

¹⁴In 2018, the Department of Defense announced that it had begun to work on a multi-year strategy to develop third-wave technologies of AI.

Figure 1: The Three Waves of Artificial Intelligence



Source: Defense Advanced Research Projects Agency (DARPA) information; Art Explosion (art). | GAO-21-519SP

As noted in a recent GAO report on AI in health care, recent machine learning systems begin with data—generally in large amounts—and infer rules or decision procedures that aim to predict specified outcomes. This inference happens when the system is able to train itself using data to increase the accuracy of its predictions. Increased availability of large data sets and computing power has enabled recent advances in machine learning such as voice recognition by personal assistants on smart phones (an example of natural language processing) and image recognition (an example of computer vision).¹⁵

AI Life Cycle

According to a review of literature, the life cycle of an AI system can involve several phases: design, development, deployment, and

¹⁵GAO, *Artificial Intelligence in Health Care: Benefits and Challenges of Technologies to Augment Patient Care*, [GAO-21-7SP](#) (Washington, D.C.: Nov. 30, 2020).

continuous monitoring.¹⁶ As shown in figure 2, the various phases involve articulating the system’s concepts, collecting and processing data, building one or more models, validating the system, continuously assessing its impact and, if necessary, retiring an AI system from production.¹⁷

¹⁶See OECD, *Artificial Intelligence in Society* (OECD Publishing: Paris, France, revised Aug. 2019), accessed Apr. 4, 2021, <https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm>; and see Select Committee on Artificial Intelligence of the National Science and Technology Council, *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update* (Washington, D.C.: June 2019); [GAO-21-7SP](#).

¹⁷OECD, *Artificial Intelligence in Society*.

Figure 2: The Phases in the AI Life Cycle



Source: GAO. | GAO-21-519SP

Machine learning systems learn from data, known as the training set, in order to perform a task.¹⁸ Further, to confirm the AI system is functioning as intended, developers must iteratively evaluate and validate the predictions made by the AI.¹⁹

According to OECD, the phases of the AI life cycle are often iterative and are not necessarily sequential. For example, the decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.²⁰ GAO has also highlighted the iterative and incremental nature of software development in its *Best Practices for Agile Adoption and Implementation*.²¹ The guide notes that in developing software, organizations should establish appropriate life-cycle activities which integrate planning, design, development, and testing to effectively measure progress continually, reduce technical and programmatic risk, and respond to feedback from stakeholders.

Technical and Societal Implications

Implementing AI systems involves assessing technical performance, as well as identifying and mitigating any societal concerns. For example, to manage technical performance, AI technical stakeholders—data scientists, data engineers, developers, cybersecurity specialists, program managers, and others—will have to ensure that the AI system solves the problem initially identified; uses data sets appropriate for the problem; selects the most suitable algorithms; and evaluates and validates the system to ensure it is functioning as intended. Without such assurances,

¹⁸Researchers use several methods to train machine learning algorithms, including: supervised machine learning—the data scientist presents an algorithm with labeled data or input; the algorithm identifies logical patterns in the data and uses those patterns to predict a specified answer to a problem. For example, an algorithm trained on many labeled images of cats and dogs could then classify new, unlabeled images as containing either a cat or a dog. In unsupervised machine learning—the data scientist presents an algorithm with unlabeled data and allows the algorithm to identify structure in the inputs, for example by clustering similar data, without a preconceived idea of what to expect. In this technique, for example, an algorithm could cluster images into groups based on similar features, such as a group of cat images and a group of dog images, without being told that the images in the training set are those of cats or dogs. For additional information, see: GAO, *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development*, [GAO-20-215SP](#) (Washington, D.C.: Dec. 20, 2019, reissued Jan. 31, 2020).

¹⁹[GAO-21-7SP](#).

²⁰OECD, *Artificial Intelligence in Society*.

²¹See GAO, *Agile Assessment Guide: Best Practices for Agile Adoption and Implementation*, [GAO-20-590G](#) (Washington, D.C.: Sept. 28, 2020).

AI systems may result in unintended consequences. The text box below shows an example of unintended consequences resulting from an AI model used to predict health care needs.

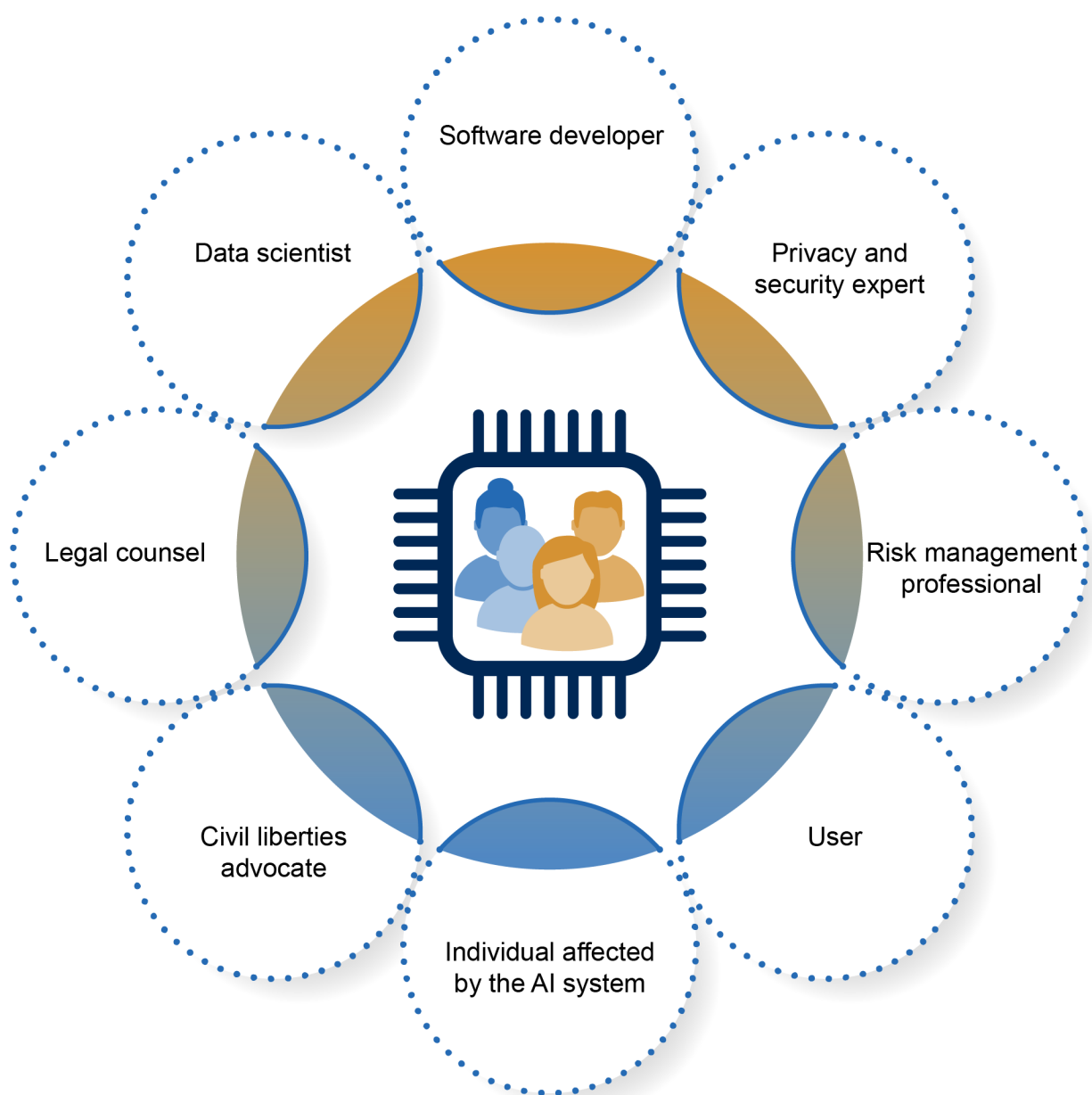
Unintended Consequences Resulting from an AI Predictive Model Used in Health Care Management

AI has been used in the U.S. health care system as a tool to identify patients with complex needs. In a recent study, researchers identified biased outputs in an AI model used to identify patients at risk for negative outcomes. A predictive model was used to generate a risk score to identify patients who could benefit from “high-risk care management” programs. Although the model excluded the patients’ races, researchers observed racial bias in the model predictions. Researchers compared Black and White patients who received the same risk scores and found that Black patients were at a higher risk of negative health outcomes than White patients. The model produced biased risk scores because the developers used health care expenses as a proxy for health care needs. However, health care expenses do not represent health care needs across racial groups, because Black patients tend to spend less on health care than White patients for the same level of needs, according to the study. For this reason, the model assigned a lower risk score to Black patients, resulting in that group being under-identified as potentially benefitting from additional help, despite having similar health care needs.

Source: GAO summary of Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464 (2019), pp. 447-453 | GAO-21-519SP.

As shown in figure 3, in addition to the AI technical stakeholders noted above, a broader community of stakeholders—policy and legal experts, subject matter experts, and individuals using the AI system or impacted by its use, among others—is engaged in the development of AI.

Figure 3: Example of the Community of Stakeholders Engaged in AI Development



Source: GAO. | GAO-21-519SP

Each stakeholder plays a role in ensuring that any ethical, legal, economic, or social concerns raised by the AI system are identified, assessed, and appropriately mitigated. For example, AI systems could

perpetuate historical biases, such as underrepresentation of groups based on race, socioeconomic status, or gender. The community of stakeholders should provide input on potential societal concerns during design and development to ensure the AI system is appropriate for the use case and representative of the intended population. Stakeholder input can help to address unintended consequences. The text box below shows a recent example of unintended consequences that can be introduced by a predictive policing software trained on crime data that may not be representative of all crimes that occur, which can reproduce patterns of systemic bias.

Unintended Consequences of Predictive Policing Technology

Local law enforcement agencies are using predictive policing software to identify likely targets for police intervention. The intended benefits are to prevent crime in specific areas and improve resource allocation of law enforcement. In one study, researchers demonstrated that the tool disproportionately identifies low-income or minority communities as targets for police intervention regardless of the true crime rates. Applying a predictive policing algorithm to a police database, the researchers found that the algorithm behaves as intended. However, if the machine learning algorithm was trained on crime data that are not representative of all crimes that occur, it learns and reproduces patterns of systemic biases. According to the study, the systemic biases can be perpetuated and amplified as police departments use biased predictions to make tactical policing decisions.

Sources: GAO summary based on [GAO-18-142SP](#) and Kristian Lum and William Isaac, "To Predict and Serve," Significance, vol. 13 (2016): pp. 14-19 | GAO-21-519SP.

Foreign and Domestic
Governance Frameworks

In recent years, both foreign and domestic stakeholders have developed governance and auditing frameworks, in part, to address the technical and societal issues in using AI in the public sector. Table 1 below provides examples of governments that have developed frameworks for implementing AI.

Table 1: Examples of AI Governance and Auditing Frameworks Developed by Foreign Governments

Government	Framework title	Date	Description
Canada	Directive on Automated Decision-Making	Apr. 2019	Requirements for Canadian government departments to use automated decision systems in a manner that reduces risks and leads to more efficient, accurate, consistent, and interpretable decisions.
Singapore	Model AI Governance Framework (2nd Edition)	Jan. 2020	Guidance on responsible use of AI in four key areas: internal governance structures and measures, determining AI decision-making model, operations management, and stakeholder communications.
United Kingdom	Information Commissioner's Office, Guidance on AI and Data Protection	July 2020	Guidance on data protection compliance for AI, including methodology to audit AI applications and ensure they process personal data fairly.

Source: GAO analysis of foreign government frameworks. | GAO-21-519SP

Note: See also <https://oecd.ai/countries-and-initiatives> for an interactive database of AI policies and initiatives provided by OECD.AI, powered by EC/OECD (2021), STIP Compass database, accessed June 9, 2021.

These international frameworks provide key information on assessing and mitigating the risks associated with deploying automated decision systems; ensuring data privacy and fairness; and establishing governance processes including monitoring human supervision of such systems.

Table 2 provides a list of U.S. agencies and departments that have released principles and guidance on implementing AI in the federal government.

Table 2: Examples of AI Governance Frameworks Developed by U.S. Government

U.S. government entities	Framework title	Date	Description
National Institute of Standards and Technology	U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools ^a	Aug. 2019	The plan provides guidance for federal agencies on how to bolster AI standards-related knowledge, leadership, and coordination among agencies that develop or use AI; promote focused research on the trustworthiness of AI systems; support and expand public-private partnerships; and engage with international parties.
Department of Defense	Artificial Intelligence Governance Plan	May 2020	The plan outlines the DOD AI governance structures to establish and advance policies. This includes initiatives to identify and integrate AI technologies, tools, and systems across DOD in support of the National Defense Strategy and the DOD CIO's Digital Modernization Strategy. ^b

U.S. government entities	Framework title	Date	Description
Food and Drug Administration	Artificial Intelligence/Machine Learning-Based Software as a Medical Device Action Plan	Jan. 2021	The plan outlines the FDA's next steps toward furthering oversight for artificial intelligence/machine learning based software as a medical device.
Office of the Director of National Intelligence	Artificial Intelligence Ethics Framework for the Intelligence Community	June 2020	The framework provides guidance on how to procure, design, build, use, protect, consume, and manage AI and related data in alignment with ethical principles adopted by the Intelligence Community.
Office of Management and Budget	Memorandum M-21-06, Guidance for Regulation of Artificial Intelligence Applications	Nov. 2020	The Memorandum sets out policy considerations that should guide, to the extent permitted by law, regulatory and non-regulatory approaches to AI applications developed and deployed outside of the federal government. In addition, it contains 10 principles for the stewardship of AI applications.
White House	Executive Order No. 13,960, Promoting the Use of Trustworthy AI in the Federal Government	Dec. 2020	The Executive Order includes principles that direct federal agencies to ensure that the design, development, acquisition, and use of AI is done in a manner that protects privacy, civil rights, civil liberties, and American values (Office of Management and Budget is directed to develop common policy guidance across agencies for implementing the principles).

Source: GAO analysis of U.S. government documents. | GAO-21-519SP

^aAccording to NIST officials, this document is a guidance or a reference document. Aspects of this guidance—such as the nine focus areas for AI standards—may help inform the development of a governance framework and processes.

^bIn addition, DOD issued its Ethical Principles for AI in Feb 2020, with the goal of assisting the U.S. military in ensuring its AI development and use is responsible, equitable, traceable, reliable, and governable.

These AI Governance frameworks identify principles for implementing AI in the U.S. government. They note that AI systems should be responsible,

equitable, traceable, reliable, and governable.²² These frameworks also highlight the need for developing standards, metrics and guidelines for accuracy, explainability,²³ and safety of such systems. One recent framework provides more detailed information on technical questions to consider and procedures to undertaken when assessing AI systems.²⁴

²²One agency—DOD—provides definitions for these terms. For example, it defines *responsible* as human beings exercising appropriate levels of judgment and responsibility for the development, deployment, use, and outcomes of DOD’s AI systems; *equitable* as taking deliberate steps to avoid unintended bias in the development and deployment of combat or non-combat AI systems that would inadvertently cause harm to persons; *traceable* as having an engineering discipline that is sufficiently advanced such that technical experts possess an appropriate understanding of the technology, development processes, and operational methods of its AI systems, including transparent and auditable methodologies, data sources, and design procedure and documentation; *reliable* as having an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured across their entire life cycle within that domain of use; and *governable* as designed and engineered to fulfill their intended function while possessing the ability to detect and avoid unintended harm or disruption, and for human or automated disengagement or deactivation of deployed systems that demonstrate unintended escalatory or other behavior. See: Department of Defense, Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* (Washington, D.C.: Oct., 2019).

²³Explainability refers “the details and reasons a model gives to make its functioning clear or easy to understand.” See A. Barredo Arrieta, N. Diaz Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado González, S. Garcia, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI,” *Information Fusion*, vol. 58 (2020), pp. 82-115.

²⁴Office of the Director of National Intelligence (ODNI), *Artificial Intelligence Ethics Framework for the Intelligence Community version 1.0* (Washington, D.C.: June 2020), accessed June 22, 2021, <https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community>.



FRAMEWORK
PRINCIPLE

1. ■

GOVERNANCE

Key Governance Practices

Governance Principle:

Promote accountability by establishing processes to manage, operate, and oversee implementation.

Source: GAO. | GAO-21-519SP

Management, and those charged with oversight of AI, can use governance structures and processes to manage risk, demonstrate the importance of integrity and ethical values, and ensure compliance with relevant laws, regulations, and guidance. GAO developed nine key practices for this principle, grouped into two categories:

- **Organizational level:** Governance at the organizational level helps entities ensure oversight and accountability and manage risks of AI systems.
- **System level:** Governance at the system level helps entities ensure AI systems meet performance requirements.

Governance at the Organizational Level

Managers should establish and maintain an environment throughout the entity that sets a positive attitude toward internal controls. The *Federal Internal Control Standards* note that “the oversight body and management set the tone at the top and throughout the organization by their example.”²⁵ Similarly, forum participants highlighted the need for entities to establish governance structures for AI systems that incorporate organizational values, consider risks, assign clear roles and responsibilities, and involve multidisciplinary stakeholders. To help entities establish governance structures and processes, as well as mitigate risks of implementing AI systems, GAO identified six key practices. The following provides detail on these practices based primarily on forum discussions, literature review, and expert interviews.

1.1 Clear goals: Define clear goals and objectives for the AI system to ensure intended outcomes are achieved.

1.2 Roles and responsibilities: Define clear roles, responsibilities, and delegation of authority for the AI system to ensure effective operations, timely corrections, and sustained oversight.

According to literature we reviewed and discussions with forum participants, a key first step in the life cycle of an AI system is to clearly define its goals and objectives. The goals and objectives should be

²⁵[GAO-14-704G](#).

specific and measurable to enable management to identify, analyze, and respond to risks related to achieving those objectives.²⁶ Objectives may relate to the effectiveness and efficiency of the AI system, the reliability of using the information generated by the system, and compliance with applicable laws and regulations. Management should also define roles and responsibilities and delegate authority for various stages of the AI system's life cycle, including design, development, deployment, assessment, and monitoring. The roles and responsibilities of personnel should be appropriate and clearly understood, according to Executive Order 13,960 and forum participants.²⁷

1.3 Values: Demonstrate a commitment to values and principles established by the entity to foster public trust in responsible use of the AI system.

1.4 Workforce: Recruit, develop, and retain personnel with multidisciplinary skills and experiences in design, development, deployment, assessment, and monitoring of AI systems.

Deficit of AI Talent in the Federal Government

As noted by the final report of the National Security Commission on Artificial Intelligence, "The human talent deficit is the government's most conspicuous AI deficit and the single greatest inhibitor to buying, building, and fielding AI-enabled technologies for national security purposes. This is not a time to add a few new positions in national security departments and agencies for Silicon Valley technologists and call it a day. We need to build entirely new talent pipelines from scratch."

Source: National Security Commission on Artificial Intelligence, *Final Report* (Washington, D.C. Mar. 1, 2021). | GAO-21-519SP

As management establishes the governance processes, it should also consider that "accountability is driven by the tone at the top and supported by a commitment to integrity, ethical values, organizational structure, and expectations of competence," according to the *Federal Internal Control Standards*.²⁸ Management should demonstrate its commitment to values and principles through directives, attitudes, and behaviors. One way to do so is by ensuring the entity recruits and retains competent personnel who have skills and experiences with design, development, assessment, and monitoring of the AI system. (See sidebar for a discussion on current talent deficits in the federal workforce.) The literature we reviewed and forum participants stressed the importance of having a multidisciplinary team that can address the technical aspects and the societal impacts of the AI system, whether the system is developed in house or procured.²⁹

²⁶ODNI, *Artificial Intelligence Ethics Framework*; Info-comm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC), *Model Artificial Intelligence Governance Framework, Second Edition* (Singapore: Jan. 21 2020).

²⁷Exec. Order No. 13,960, *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government* (Dec. 3, 2020), 85 Fed. Reg. 78,939, 78,940 (Dec. 8, 2020).

²⁸GAO-14-704G.

²⁹ODNI, *Artificial Intelligence Ethics Framework*; World Economic Forum, *Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations*, (Cologne/Geneva, Switzerland: Jan. 2020).

1.5 Stakeholder involvement: Include diverse perspectives from a community of stakeholders throughout the AI life cycle to mitigate risks.

Forum participants and the literature we reviewed also emphasize the importance of including diverse perspectives throughout the AI life cycle. Strategies to incorporate diverse perspectives include establishing collaborative processes and multidisciplinary teams that involve subject matter experts in data science, software development, civil liberties, privacy and security, legal counsel, and risk management. These processes and teams should also engage with individuals who may be using or operating the system, or who may be affected by the AI system. The engagement of such a community of stakeholders can increase the likelihood of effectively identifying known and unknown risks, problematic assumptions, and limitations associated with the AI system, according to forum participants and the ODNI framework.³⁰

1.6 Risk management: Implement an AI-specific risk management plan to systematically identify, analyze, and mitigate risks.

Example of a Risk Management Tool

The Government of Canada's Directive on Automated Decision-Making requires government agencies to use the Algorithmic Impact Assessment tool to assess risks and impacts prior to the production of any automated decision-making system. The assessment provides a questionnaire to help agencies assess and mitigate the risks associated with deploying such a system, and consider the impact level of an AI system by reviewing its underlying capabilities and algorithms.

Source: Government of Canada's Directive on Automated Decision-Making. | GAO-21-519SP

Implementing a risk management plan for the AI system helps managers systematically identify, analyze, and mitigate risks associated with achieving the defined goals and objectives. Entities should consider implementing the risk management plan throughout the AI system life cycle to ensure risk mitigation is continuous and timely. For example, prior to developing the AI system, managers should first consider whether a particular use case is appropriate and identify the level of risk and potential societal harm.³¹ (See sidebar for an example of a risk management tool.) Furthermore, risk management should distinguish between risks inherent to the business or subject matter and those associated with the AI system. For example, financial institutions may face risks associated with financial crimes, investments, and compliance. Such risks should be factored in; however, to ensure appropriate risk

³⁰ODNI, *Artificial Intelligence Ethics Framework*.

³¹According to the 2021 *Final Report* of the National Security Commission on Artificial Intelligence, "agencies should institute specific oversight and enforcement practices, including...a mechanism that would allow thorough review of the most sensitive/high-risk AI systems to ensure auditability and compliance with responsible use and fielding requirements..." National Security Commission on Artificial Intelligence, *Final Report*, (Washington, D.C.: Mar. 1 2021). In addition, according to one forum participant, entities should consider mitigating risks by limiting the scope of the AI system when there is not sufficient confidence that the stated goals and objectives can be achieved.

management and mitigation, they should not be comingled with those introduced by the AI system. Moreover, in the case of entities procuring AI systems from vendors, contracts should include provision for appropriate access to data, models, and parameters to enable sufficient oversight and auditing.

Governance at the System Level

Establishing governance structures and processes at a system level helps managers ensure the AI system achieves its intended outcomes and complies with relevant laws and regulations. GAO identified three key practices for system-level governance. The following provides detail on these practices from forum discussions, literature, and expert interviews.

1.7 Specifications: Establish and document technical specifications to ensure the AI system meets its intended purpose.

The *Federal Internal Control Standards* state that documentation is a necessary part of an internal control system and is required for the effective design, implementation, and operation.³² The Singapore *Model Artificial Intelligence Governance Framework (2nd Edition)* suggests incorporating descriptions of the AI system's design and expected behavior into the documentation of technical specifications as a way to demonstrate accountability to individuals and regulators.³³ Management should use judgment in determining the extent of documentation that is necessary to provide sufficient assurance that AI objectives will be met. The level and nature of documentation vary based on the complexity of the operational processes the entity performs, and the risk level of the AI system. For example, entities may identify certain AI systems as high risk,

³²[GAO-14-704G](#).

³³IMDA and PDPC, *Model Artificial Intelligence Governance Framework*.

thus requiring a higher level of documentation, according to a forum participant.³⁴

1.8 Compliance: Ensure the AI system complies with relevant laws, regulations, standards, and guidance.

Entities can also take a proactive approach to ensuring compliance by considering applicable laws and regulations, industry standards, and guidance from federal agencies and other entities. Federal laws and regulations specific to AI are largely still under development, but data privacy and non-discrimination laws are likely to be relevant for AI systems that process personally identifiable information or sensitive data. Numerous private and public organizations have developed guidance on incorporating ethical principles such as fairness, accountability, transparency, and safety in AI use.³⁵ Considering such guidance when defining goals and objectives can help entities demonstrate a commitment to principles and values that foster public trust in responsible AI use.

1.9 Transparency: Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system.

Documentation and external communication can offer a way for entities to provide transparency. Communication may include disclosing relevant information regarding the use of AI systems, such as (a) what the system is for, (b) what it is not for, (c) how it was designed, and (d) what its

³⁴The Singapore Model AI Governance Framework provides an example of a company which conducts initial risk scoring to determine the risk of the proposed AI activity based on multiple factors—including the impact on individuals from AI decisions. If an AI project has been identified as high risk, it will be referred to the Governance Council for review. Low-risk projects will not be subjected to a review and can proceed to the model development stage. Similarly, the final report of the National Security Commission on Artificial Intelligence also states that “agencies should institute specific oversight and enforcement practices, including auditing and reporting requirements; a mechanism that would allow thorough review of the most sensitive/high-risk AI systems to ensure auditability and compliance with responsible use...” National Security Commission on Artificial Intelligence, *Final Report*.

³⁵IBM Research, *AI Fairness 360*, accessed June 2, 2021, <https://aif360.mybluemix.net/>; IEEE Standards Association, *Raising the Standards in Artificial Intelligence Systems (AIS)*, accessed Mar. 18, 2021, <https://standards.ieee.org/initiatives/artificial-intelligence-systems/index.html>; ODNI, *Artificial Intelligence Ethics Framework*.

limitations are.³⁶ These disclosures should also take into account privacy issues, whether sensitive law enforcement and personally identifiable information is involved, national security issues, and concerns related to other kinds of protected information. In addition, forum participants noted the importance of promoting transparency and explainability, while also protecting individual privacy and the developer's intellectual property rights.³⁷ The text box below provides an example of how documentation can provide increased transparency in AI systems.

Example of Transparency through Documentation

According to the nonprofit organization Partnership for AI, *"transparency requires that the goals, origins, and form of a system be made clear and explicit to users, practitioners, and other impacted stakeholders seeking to understand the scope and limits of its appropriate use. One simple and accessible approach to increasing transparency in [machine learning] lifecycles is through an improvement in both internal and external documentation. This documentation process begins in the machine learning system design and set up stage, including system framing and high-level objective design. This involves contextualizing the motivation for system development and articulating the goals of the system in which this system is deployed, as well as providing a clear statement of team priorities and objectives throughout the system design process."*

Source: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles (ABOUT ML). | GAO-21-519SP

³⁶ODNI, Artificial Intelligence Ethics Framework.

³⁷See key practice 2.8 for additional questions to consider regarding data privacy.

1. Governance Framework

Governance at the Organizational Level

1.1 Clear goals: Define clear goals and objectives for the AI system to ensure intended outcomes are achieved.

 Questions to Consider	 Audit Procedures
• What goals and objectives does the entity expect to achieve by designing, developing, and/or deploying the AI system? • To what extent do stated goals and objectives represent a balanced set of priorities and adequately reflect stated values? • How does the AI system help the entity meet its goals and objectives? • To what extent does the entity communicate its AI strategic goals and objectives to the community of stakeholders? • To what extent does the entity have the necessary resources—funds, personnel, technologies, and time frames—to achieve the goals and objectives outlined for designing, developing and deploying the AI system? • To what extent does the entity consistently measure progress towards stated goals and objectives?	• Collect strategic and implementation plans that describe the goals and objectives for the design, development, and deployment of each AI system. • Review goals and objectives relevant to each AI system to assess whether they are specific, measurable, and clear so that they are understood at all levels of the entity. Assess whether the goals and objectives clearly define what is to be achieved, who is to achieve it, how it will be achieved, and the time frames for achievement. In addition, determine whether the AI system provides functions more effectively, efficiently, economically, ethically, and equitably relative to conventional approaches used by the entity to achieve its goals and objectives. • Interview community of stakeholders—management, program managers, developers, data scientists, legal and policy officers, information technology officers, subject matter experts, civil liberty advocates, and end users, to determine whether goals and objectives are clearly defined and understood

1.2 Roles and responsibilities: Define clear roles, responsibilities, and delegation of authority for the AI system to ensure effective operations, timely corrections, and sustained oversight.

 Questions to Consider	 Audit Procedures
• What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system? • To what extent has the entity clarified the roles, responsibilities, and delegated authorities to relevant stakeholders?	• Collect organizational charts, governance board charters, policies or guidance documents that define roles, responsibilities, and delegated authorities for personnel involved in design, development, deployment, assessment, and monitoring of the AI system. • Review organizational documents that outline roles, responsibilities, and delegation of authority to assess whether they are defined and clarified to relevant stakeholders. In addition, assess whether the operational structure enables timely corrections and oversight. • Interview management, program managers, and relevant stakeholders to determine whether roles, responsibilities, and delegated authorities are defined, clarified, and understood at all levels.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP
Notes: In this table, “Community of stakeholders” can include management, program managers, developers, data scientists, legal and policy officers, information technology officers, subject matter experts, civil liberty advocates, and users.

1. Governance Framework

Governance at the Organizational Level

1.3 Values: Demonstrate a commitment to values and principles established by the entity to foster public trust in responsible use of the AI system.

 Questions to Consider	 Audit Procedures
• How does the entity demonstrate a commitment to stated values and principles? • To what extent has the entity operationalized its stated core values and principles for the AI system? • What policies has the entity developed to ensure the use of the AI system is consistent with its stated values and principles? • To what extent do these policies foster public trust and confidence in the use of the AI system?	• Collect strategic plans, policies, design documents, and mission statements that state values and principles applicable to the AI system. • Review strategic plans, policies, and mission statements to determine whether management has established and prioritized values and principles, including those that ensure sufficient oversight. • Interview management, program managers, legal and policy officers, subject matter experts, civil liberty advocates, and relevant stakeholders to determine whether entity’s use of the AI system reflects a commitment to stated values and principles.

1.4 Workforce: Recruit, develop, and retain personnel with multidisciplinary skills and experiences in design, development, deployment, assessment, and monitoring of AI systems.

 Questions to Consider	 Audit Procedures
• How does the entity determine the necessary skills and experience needed to design, develop, deploy, assess, and monitor the AI system? • What efforts has the entity undertaken to recruit, develop, and retain competent personnel? • What efforts has the entity undertaken to recruit, develop, and retain a workforce with backgrounds, experience, and perspectives that reflect the community impacted by the AI system? • How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?	• Collect position descriptions, recruitment practices, performance reviews, and workforce development documents that describe job responsibilities, qualifications, training, guidance, and resources for personnel involved in the design, development, deployment, assessment, and monitoring the AI system. • Review recruitment/retention practices, training programs, needs assessments, candidate evaluations, performance reviews, certifications (e.g., conferences, training, and learning), and workforce development documents to assess alignment with necessary technical skills, competencies, backgrounds, experiences, and knowledge required. • Interview management, program managers, and human resource specialists to assess recruitment, development, and training practices.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP

1. Governance Framework

Governance at the Organizational Level

1.5 Stakeholder involvement: Include diverse perspectives from a community of stakeholders throughout the AI life cycle to mitigate risks.

 Questions to Consider	 Audit Procedures
• What factors were considered when identifying the community of stakeholders involved throughout the life cycle? • Which stakeholders did the entity include throughout the design, development, deployment, assessment, and monitoring life cycle? • To what extent do the teams responsible for developing and maintaining the AI system reflect diverse opinions, backgrounds, experiences, and perspectives? • What specific perspectives did stakeholders share, and how were they integrated across the design, development, deployment, assessment, and monitoring of the AI system? • To what extent has the entity addressed stakeholder perspectives on the potential negative impacts of the AI system on end users and impacted populations?	• Collect documentation that describes stakeholders and their involvement throughout the life cycle of the AI system. • Review outreach documents and correspondence to assess whether perspectives from a community of stakeholders were collected, assessed, and integrated throughout the life cycle. • Interview community of stakeholders to determine how their perspectives were assessed and integrated, if at all, throughout the life cycle of the AI system.

1.6 Risk management: Implement an AI-specific risk management plan to systematically identify, analyze, and mitigate risks.

 Questions to Consider	 Audit Procedures
• To what extent has the entity developed an AI-specific risk management plan to systematically identify, analyze, and mitigate known and unknown operational, technical, as well as societal risks associated with the AI system? • To what extent has the entity defined its risk tolerance for using the AI system? • To what extent does the plan specifically address risks associated with acquisition, procurement of packaged software from vendors, cybersecurity controls, computational infrastructure, data, data science, deployment mechanics, and system failure?	• Collect risk management plans and evaluation documents that describe the entity’s risk management approach for the AI system. • Review risk management plans to assess how the entity identifies, reviews, and mitigates risks associated with the AI system, including those associated with civil liberties, privacy, and packaged software from vendors. In addition, assess whether risks identified existed prior to the use of the AI system and related to the inherent nature of the issue area. Assess whether the entity has determined if the information submitted and received from vendors is complete, accurate, and valid. • Interview management, program managers, legal and policy officers, subject matter experts, vendors, and relevant stakeholders to assess extent to which risk management plans are established.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP
Notes: In this table, “Community of stakeholders” can include management, program managers, developers, data scientists, legal and policy officers, information technology officers, subject matter experts, civil liberty advocates, and end users.

1. Governance Framework

Governance at the Systems Level

1.7 Specifications: Establish and document technical specifications to ensure the AI system meets its intended purpose.

 Questions to Consider	 Audit Procedures
• What challenge/constraint is the AI system intended to solve? • To what extent has the entity clearly defined technical specifications and requirements for the AI system? • How do the technical specifications and requirements align with the AI system’s goals and objectives? • What justifications, if any, has the entity provided for the assumptions, boundaries, and limitations of the AI system?	• Collect documents on technical specifications and requirements, including architecture diagrams, workflows, data characterization, and test plans. • Review documents to map program goals with technical specifications and requirements, including components and test plans, to ensure the system meets its intended purpose. In addition, identify constraints and assumptions incorporated into the AI system. • Interview program managers, developers, data scientists, subject matter experts, and systems engineers to determine whether the AI system’s technical specifications align with its goals and objectives.

1.8 Compliance: Ensure the AI system complies with relevant laws, regulations, standards, and guidance.

 Questions to Consider	 Audit Procedures
• To what extent has the entity identified the relevant laws, regulations, standards, and guidance, applicable to the AI system’s use? • How does the entity ensure that the AI system complies with relevant laws, regulations, standards, federal guidance, and policies? • To what extent is the AI system in compliance with applicable laws, regulations, standards, federal guidance, and entity policies?	• Collect strategic, design, and implementation documents that describe the AI system’s compliance with relevant laws and regulations, standards, federal guidance, and entity policies. • Review relevant laws, regulation, standards, federal guidance, and entity policies; and assess whether the entity has established controls to ensure the AI system meets these requirements. • Interview management, program managers, legal and policy officers, subject matter experts, and relevant stakeholders to assess compliance with relevant laws and regulations.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP

1. Governance Framework

Governance at the Systems Level

1.9 Transparency: Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system.

 Questions to Consider	 Audit Procedures
• What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system? • How easily accessible and current is the information available to external stakeholders? • To what extent is this information sufficient and appropriate to promote transparency?	• Collect information from entity’s public websites, correspondence, and related outreach efforts on information shared with external stakeholders on use and limitation of the system. • Review the public websites, correspondence, and related documents to assess whether the type of information shared is clear, appropriate, and sufficient to promote awareness of the use and limitation of the AI system. In addition, consider conducting structured interviews, focus groups, and/or surveys of impacted users, regulators, consumers and other individuals affected to assess the extent to which outreach efforts provide the public with necessary information on the use and limitation of the AI system. • Interview community of stakeholders to assess entity’s information communication and outreach efforts.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP

Notes: In this table, “Community of stakeholders” can include management, program managers, developers, data scientists, legal and policy officers, information technology officers, subject matter experts, civil liberty advocates, and end users.



FRAMEWORK
PRINCIPLE

2.

DATA

Key Data Practices

Data Principle:

Ensure quality, reliability, and representativeness of data sources, origins, and processing.

Source: GAO. | GAO-21-519SP

Data used to train, test, and validate AI systems should be of sufficient quality and appropriate for the intended purpose to ensure the system produces consistent and accurate results.³⁸ Management should provide reasonable assurance of the quality, reliability, and representativeness of the data included in the system, from its development stage to system operation. GAO developed eight key practices for this principle, divided into two categories:

- **Data used for model development:** This category refers to training data used in developing a probabilistic component, such as a machine learning model for use in an AI system, as well as data sets used to test and validate the model.
- **Data used for system operations:** This category refers to the various data streams that have been integrated into the operation of an AI system, which may include multiple models.³⁹

Data Used for Model Development

AI based on machine learning models begins with data and infers rules or decision procedures that predict outcomes. Quality, reliability, and representativeness of the data therefore may affect the accuracy, precision, recall, and biases of the predicted outcomes of the AI models. We have identified five key practices to help entities use appropriate data for developing AI models.

³⁸Machine learning systems learn from data, known as the training set, in order to perform a task. A testing set is a subset of the data used to test the trained model. The testing set should be large enough to yield statistically relevant results and be representative of the data set as a whole.

³⁹M. Arnold, R.K.E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, D. Reimer, A. Olteanu, D. Piorkowski, J. Tsay, and K. R. Varshney, "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity," *IBM Journal of Research and Development*, vol. 63, no. 4/5 (2019): pp. 6:1-6:13.

2.1 Sources: Document sources and origins of data used to develop the models underpinning the AI system.

The *Federal Internal Control Standards* describe the importance of documentation to ensure effective design, implementation, and operational effectiveness.⁴⁰ According to our review of literature and forum participants, documenting the provenance⁴¹ and use of data in AI models can ensure data quality and enable third-party assessments.⁴² In addition to documenting how the data were collected, entities should document how they were curated, and used to increase transparency and accountability. Forum participants and literature identified several key sources of evidence that auditors may collect to assess data used for AI models, including documentation on: 1) how data have been collected, prepared, labeled, and maintained; and 2) how data are monitored on a continual basis. According to OECD, when data used in the AI system are well documented and traceable, it enables analysis of the system's outcomes and ensures they are consistent and appropriate for the use case.⁴³

⁴⁰[GAO-14-704G](#).

⁴¹The term “data provenance” refers to a record that accounts for the origin of a piece of data (in a database, document, or repository), together with an explanation of how and why it got to the present place. A provenance record will document the history for each piece of data.

⁴²European Union Agency for Fundamental Rights (EU-FRA), *Data Quality and Artificial Intelligence – Mitigating Bias and Error to Protect Fundamental Rights* (Vienna, Austria: June 11, 2019); S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, *The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards*, <https://arxiv.org/pdf/1805.03677>.

⁴³OECD, *Recommendation of the Council on Artificial Intelligence*.

2.2 Reliability: Assess reliability of data used to develop the models.

GAO Guidance on Data Reliability

In an audit environment, reliability of data means that data are applicable for audit purposes and are sufficiently complete and accurate.

Applicability for audit purpose refers to whether the data, as collected, are valid measures of the underlying concepts being addressed in the objectives.

Completeness refers to the extent to which relevant data records and fields are present and sufficiently populated.

Accuracy refers to the extent to which recorded data reflect the actual underlying information.

Source: Adapted from [GAO-20-283G](#). | GAO-21-519SP

Entities should also ensure that data used to develop the models in the AI models are reliable because data reliability affects the accuracy of model predictions, according to literature and forum participants.⁴⁴ For auditors and entities being audited, GAO provided guidance on what data reliability means (see sidebar). According to GAO's *Federal Information System Controls Audit Manual*, "The entity should implement procedures to reasonably assure that (1) data are inputted in a controlled manner, (2) data inputted into the application are complete, accurate, and valid, (3) any incorrect information is identified, rejected, and corrected for subsequent processing, and (4) the confidentiality of the data is adequately protected. Inadequate controls can result in incomplete, inaccurate, and/or invalid records in the application data or unauthorized disclosure of application data."⁴⁵ Entities should also assess the extent to which the data accurately and verifiably represent constituent populations served by the AI system. AI models trained on data that are not representative of the target population may produce biased results, in that the model performs well for the characteristics that are well represented and poorly for the characteristics that are underrepresented, according to literature we reviewed and one forum participant.⁴⁶

2.3 Categorization: Assess attributes used to categorize data.

Entities should document their rationale for organizing the data and how they are segregated into training, development, and testing sets prior to model development. The documentation can allow a third-party to determine whether the data segregation is appropriate. Inadequate separation between the training and testing data could result in overfitting—that is, the model may perform well during testing but

⁴⁴EU-FRA, *Data Quality and Artificial Intelligence*.

⁴⁵GAO, *Federal Information System Controls Audit Manual (FISCAM)*, [GAO-09-232G](#), (Washington, D.C.: Feb. 2, 2009).

⁴⁶The Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK, *Auditing machine learning algorithms* (Oct. 14, 2020).

underperform in the operational environment when it encounters unfamiliar data.⁴⁷

2.4 Variable Selection: Assess data variables used in the AI component models.

Entities should assess the data variables used in the model to ensure appropriateness, according to forum participants. In particular, entities should document how they select or discard sensitive⁴⁸ variables for modeling processes. In addition, entities should minimize the amount of sensitive data they collect or process and ensure that they are adequate, relevant, and not excessive to the purpose, according to a forum participant. Entities may assess the use of variables to avoid overcomplicating the model, according to literature we reviewed.⁴⁹

⁴⁷According to the Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway, and the UK, inadequate separation between the training and testing data could result in “overfitting” of the model. When overfitting occurs, the performance results can be inflated when some of the data used for training the model are also used for testing the model. Overfitting could lead to degraded model performance in the operational environment when it encounters unfamiliar data.

⁴⁸Sensitive information, as defined by the National Institute of Standards and Technology, is information where the loss, misuse, or unauthorized access or modification could adversely affect the national interest or the conduct of federal programs, or the privacy to which individuals are entitled under 5 U.S.C. § 552a (the Privacy Act); that has not been specifically authorized under criteria established by an Executive Order or an Act of Congress to be kept classified in the interest of national defense or foreign policy.

⁴⁹Complicated models are prone to overfitting; The Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK, *Auditing machine learning algorithms*. “A model overfits the training data when it describes features that arise from noise or variance in the data, rather than the underlying distribution from which the data were drawn. Overfitting usually leads to loss of accuracy on out-of-sample data,” according to C. Sammut and G.I. Webb, *Encyclopedia of Machine Learning*, 2010 ed. (Boston, Mass.: Springer).

2.5 Enhancement: Assess the use of synthetic, imputed, and/or augmented data.

What are Synthetic, Imputed, and Augmented Data?

Synthetic data are artificially produced data that are intended to mirror the features of real data. They provide an approach to preserve privacy when systems use sensitive or personally identifiable information. Synthetic data can serve as a practical replacement for the original sensitive data.

Imputed data are the substituted values for missing data to preserve the viability of the overall data set.

Data may be **augmented or enhanced** using other approaches, such as matching the data on hand against a larger database to complete the desired missing data fields.

Source: GAO summary of multiple sources. | GAO-21-519SP

If synthetic, imputed, and augmented or otherwise enhanced data are used (see sidebar), entities should assess the quality of the data and representativeness of the population served by the AI model. To ensure the AI model meets its intended outcome and is reproducible, entities should be able to explain why and how the data have been produced or manipulated to create the model. Preserving documentation on how the data sets have been enhanced can support testing, as well as traceability throughout the life cycle of the AI model.

Data Used for System Operations

We have also identified three key practices for assessing data streams and models as they are integrated into a system.

2.6 Dependency: Assess interconnectivities and dependencies of data streams that operationalize the AI system.

Once the data streams and models have been connected to operationalize the AI system, entities should check for unexpected correlations across data streams and ensure data are representative. This could be one way to identify biases in data sets.⁵⁰

2.7 Bias: Assess reliability, quality, and representativeness of all the data used in the system's operation, including any potential biases, inequities, and other societal concerns associated with the AI system's data.

Bias is not specific to AI, but AI tools may magnify existing biases.⁵¹ Datasets can be biased for several reasons, according to literature. Data points may be sparse or not exist for certain groups resulting in their underrepresentation in the dataset. Even if data exist, they may reflect

⁵⁰Holland et al., *The Dataset Nutrition Label*; J Stoyanovich and B Howe, "Nutritional Labels for Data and Models," *IEEE Computer Society Technical Committee on Data Engineering*, accessed Jul. 1, 2020, <http://sites.computer.org/debull/A19sept/p13.pdf>.

⁵¹T. Gebru, J. Morgenstern, B. Vecchione, J. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for Datasets," accessed June 22, 2021, <https://arxiv.org/pdf/1803.09010>.

historic biases and inequities.⁵² Therefore, entities should engage stakeholders to proactively identify potential biases, inequities, and other negative societal impacts associated with the AI system's data so that those concerns can be mitigated.⁵³ One forum participant noted that data should be assessed against the organization's own definition of bias. The definition should reflect the kind of impacts the system might have, once deployed. To better understand these potential impacts, entities should work across stakeholder groups to obtain their perspectives.

2.8 Security and Privacy: Assess data security and privacy for the AI system.

Security and privacy protection challenges are not unique to AI. According to GAO *High-Risk Series: Urgent Actions Are Needed to Address Cybersecurity Challenges Facing the Nation*, protecting privacy and sensitive data is one of the four major cybersecurity challenges that the federal government and other entities face.⁵⁴ In that report, GAO found that advances in technology that facilitate the correlation of information about individuals across large and numerous databases pose challenges to the federal government in protecting privacy and sensitive data. GAO identified actions that the federal government and other entities should take, including 1) improving federal efforts to protect privacy and sensitive data and 2) appropriately limiting the collection and use of personal information and ensuring that it is obtained with appropriate knowledge or consent. Forum participants also emphasized the importance of data security and privacy. Entities that are using or plan to implement AI systems should conduct data security assessments,

⁵²N.T. Lee, P. Resnick, and G. Barton, *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms* (Washington, D.C.: Brookings Institution; May 22, 2019), accessed Apr. 15, 2021, <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>; G. Smith and I. Rustagi, *Mitigating Bias in Artificial Intelligence An Equity Fluent Leadership Playbook* (Berkeley Haas Center for Equity, Gender and Leadership, Calif.: July, 2020).

⁵³One forum participant noted that entities could use Bayesian Improved Surname Geocoding, which combines geography-based and surname-based information into a single proxy probability for race and ethnicity, as a proxy method for classifying unidentified race and ethnicity.

⁵⁴GAO, *High-Risk Series: Urgent Actions Are Needed to Address Cybersecurity Challenges Facing the Nation*, [GAO-18-622](#) (Washington, D.C.: Sept. 6, 2018).

including risk assessments,⁵⁵ have a data security plan, and conduct privacy assessments. Any deficiencies or risks identified in testing for security and privacy should be addressed.

⁵⁵See the Federal Information Security Modernization Act of 2014 (FISMA), 44 U.S.C. § 3554(b)(1) and the *NIST Risk Management Framework*, National Institute of Standards and Technology (NIST), *NIST Risk Management Framework* (Gaithersburg, Md.: May 28, 2021), accessed June 2, 2021, <https://csrc.nist.gov/projects/risk-management/about-rmf>. FISMA's information security requirements apply to federal agencies.

2. Data Framework

Data Used for Model Development

2.1 Sources: Document sources and origins of data used to develop the models underpinning the AI system.

 Questions to Consider	 Audit Procedures
• How has the entity documented the AI system’s data provenance, including sources, origins, transformations, augmentations, labels, dependencies, constraints, and metadata? • What processes exist for data generation, acquisition/collection, ingestion, staging/storage, transformations, security, maintenance, and dissemination? • To what extent are the data appropriate for the intended purpose?	• Collect data management plans and documentation—data provenance records, inventories, and dictionaries—that describe the sources and origins of data and methodology used to collect the data. • Review data management plans and documentation to assess whether they clearly identify the sources and origins of data. In addition, assess whether the data are appropriate for the intended purpose. • Interview data stakeholders—data stewards, data custodians, data scientists, program managers—to determine whether sources and origins of data used to develop the AI system are clearly documented.

2.2 Reliability: Assess reliability of data used to develop the models.

 Questions to Consider	 Audit Procedures
• To what extent are data used to develop the AI system accurate, complete, and valid? • To what extent do the data represent constituent populations served by the AI system? • How does the entity ensure that the data collected are adequate, relevant, and not excessive in relation to the intended purpose? • What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data?	• Collect data reliability assessments, provenance records, corrective actions, and samples of data used for training. • Review samples of data used for training to assess whether they are accurate, complete, and valid. In addition, review whether data samples appropriately represent constituent populations, and conduct statistical analysis on the samples of data used for training to verify the data reliability assessment. Review corrective actions to determine the extent to which they are appropriate. • Interview data stakeholders, information technology officers, legal and policy officers, social scientists, and civil liberty advocates to determine whether the AI system’s data are reliable.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP
Notes: In this table, data stakeholders can include data stewards, data custodians, data scientists, and program managers.

2. Data Framework

Data Used for Model Development

2.3 Categorization: Assess attributes used to categorize data.

 Questions to Consider	 Audit Procedures
• What attributes are used to categorize data? • To what extent are the data attributes accurate, complete, and valid? • What is the method for segregating data into training, validation, and testing sets? • To what extent is the training, validation, and testing data representative of the operational environment? • What assumptions, if any, were made about the operational environment?	• Collect data schemas that describe how data are categorized and organized. • Review data schemas and processes to identify sensitive data sets/fields and assess whether the data are complete, accurate, and valid. • Interview data stakeholders to understand how data attributes are defined.

2.4 Variable Selection: Assess data variables used in the AI component models.

 Questions to Consider	 Audit Procedures
• What is the variable selection and evaluation process? • How were sensitive variables (e.g., demographic and socioeconomic categories) that may be subject to regulatory compliance specifically selected or not selected for modeling purposes?	• Collect original, derived, or filtered data, and processes used or evaluations applied to select variables. • Review the independence and suitability of statistical properties. • Interview data stakeholders to understand data variable selection and assessment process.

2.5 Enhancement: Assess the use of synthetic, imputed, and/or augmented data.

 Questions to Consider	 Audit Procedures
• What is the entity's rationale for using synthetic, imputed, and/or augmented data? • How are synthetic, imputed, and/or augmented data generated, maintained, and integrated? • What assumptions, if any, were made in the process of generating synthetic, imputed, and/or augmented data?	• Collect original, synthetic, and imputed/augmented data, and models used to generate the data. • Review imputed/augmented data to assess the validity of the imputation model. In addition, review a sample of synthetic data to assess whether it is representative of the intended population. Further, identify related parameters and potential quality concerns. • Interview data stakeholders to understand the rationale for the use of synthetic, imputed, and/or augmented data.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP
Notes: In this table, data stakeholders can include data stewards, data custodians, data scientists, and program managers.

2. Data Framework

Data Used for System Operation

2.6 Dependency: Assess interconnectivities and dependencies of data streams that operationalize the AI system.



Questions to Consider

- To what extent does operational data result in additional model training and/or validation?
- To what extent do data streams collectively and appropriately represent constituent populations?
- How are the interconnectivity of data streams evaluated to mitigate performance and societal risks associated with dependencies, sequencing, and aggregation?



Audit Procedures

- Collect data management plan and documentation that describe the training and testing data and assessment on representativeness and applicability of the data to operational conditions. In addition, collect a representative sample of the data at different periods of time and conditions.
- Review the representative sample of data to assess whether the data are of sufficient quality and consistency over time to ensure they meet assumptions and requirements of system components.
- Interview data stakeholders involved in training, testing and verifying data for the AI system.

2.7 Bias: Assess reliability, quality, and representativeness of all the data used in the system's operation, including any potential biases, inequities, and other societal concerns associated with the AI system's data.



Questions to Consider

- To what extent has the entity identified and mitigated potential bias—statistical, contextual, and historical—in the data?
- How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?



Audit Procedures

- Collect data used to develop the model, relevant laws and regulations, standards, and federal and entity guidance.
- Review data to assess whether potential biases may have been introduced through the data generation process or affected during data management activities. In addition, assess whether relevant laws, regulation, standards, federal guidance, and entity policies have applied as system controls and constraints to ensure data are appropriate for the use case.
- Interview data stakeholders, legal and policy officers, social scientists, and civil liberty advocates to determine whether potential biases in data are identified and mitigated.

2.8 Security and Privacy: Assess data security and privacy for the AI system.



Questions to Consider

- What assessments has the entity conducted on data security and privacy impacts associated with the AI system?
- How does the entity identify, assess, and mitigate data security and privacy risks associated with the AI system?



Audit Procedures

- Collect security plans, security assessments, and privacy assessments¹ on the AI system's use of data.
- Review the data security plans, security assessments, and privacy assessments to assess whether the methodology, test plans, and results identify deficiencies and/or risks and the extent to which they are promptly corrected.
- Interview chief data officers, information security officers, data stakeholders, and privacy officers to determine extent of the AI system's data security and privacy protection.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP

Notes: In this table, data stakeholders can include data stewards, data custodians, data scientists, and program managers.

¹ This framework is intended to apply to federal agencies and other entities. With regard to federal agencies, one set of requirements that applies is Privacy Impact Assessments (PIAs). The E-Government Act of 2002 requires that agencies conduct PIAs for systems or collections containing personal information.



FRAMEWORK
PRINCIPLE

3.

PERFORMANCE

Key Performance Practices

Performance Principle:

Produce results that are consistent with program objectives.

Source: GAO. | GAO-21-519SP

Management and those charged with oversight of AI can use performance assessment to improve performance and operations, reduce costs, facilitate decision-making by parties responsible for overseeing or initiating corrective action, and contribute to public accountability.⁵⁶

GAO developed nine key practices for this principle, grouped into two categories:

- **Component level:** Performance assessment at the component level determines whether each component meets its defined objective. The components are technology assets that represent building blocks of an AI system. They include hardware and software that apply mathematical algorithms to data.⁵⁷
- **System level:** Performance assessment of the system determines whether the components work well as integrated whole.

As noted in the *GAO Agile Assessment Guide: Best Practices for Agile Adoption and Implementation*, it is important to trace the requirements for software development (including AI development) from the top-level mission needs down to the system or component that enabled those requirements to be met.

Performance at the Component Level

According to the *Government Auditing Standards*,⁵⁸ audit objectives of a performance assessment may include: 1) determining whether management information, such as performance measures, and public reports are complete, accurate, and consistent to support performance and decision-making and 2) determining whether a program produced

⁵⁶[GAO-21-368G](#).

⁵⁷In addition to standard computer hardware such as central processing units, an AI system may include additional hardware such as graphic processing units or assets in which the AI is embedded, as in the case of advanced robots and autonomous cars. Software in an AI system is a set of programs designed to enable a computer to perform a particular task or series of tasks.

⁵⁸[GAO-21-368G](#).

intended results or produced results that were not consistent with the program objectives. To help entities implement AI systems to meet those auditing requirements at the component level, GAO identified four key practices.

3.1 Documentation: Catalog model and non-model components along with operating specifications and parameters.

An AI system may comprise multiple models trained on many data sets.⁵⁹ Based on our review of literature, agencies and entities should catalog the components of the AI system and document the purpose of the components, including their specifications and requirements. Such documentation provides assurance of the appropriateness of the components selected, enhances transparency, and increases users' and public trust in the AI system.⁶⁰

3.2 Metrics: Define performance metrics that are precise, consistent, and reproducible.

During the development phase, individual components should be tested using quantifiable metrics that are consistent with the program goals and objectives to provide reasonable assurance that the components are achieving their objectives, according to literature and forum participants. The metrics should extend beyond assessing for accuracy, safety, and security and include bias, equity, and other societal considerations.⁶¹ Metrics should be selected and applied to align with goals and objectives and reflect the societal impacts of the AI model once deployed.⁶²

⁵⁹Arnold et al., "FactSheets."

⁶⁰Entities could benefit from sharing this documentation with external stakeholders, as appropriate; Partnership on AI (PAI). *Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles (ABOUT ML)* version 0 (San Francisco, Calif.: 2019), accessed Nov. 2020, <https://www.partnershiponai.org/wp-content/uploads/2019/07/ABOUT-ML-v0-Draft-Final.pdf>; Arnold et al., "FactSheets."

⁶¹M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, "Model Cards for Model Reporting," *ACM '19: Conference on Fairness, Accountability, and Transparency*, (Atlanta, GA, Jan. 30, 2019). PAI, *Annotation and Benchmarking of Machine Learning*.

⁶²Mitchell et al., "Model Cards for Model Reporting."

3.3 Assessment: Assess the performance of each component against defined metrics to ensure it functions as intended and is consistent with program goals and objectives.

3.4 Outputs: Assess whether outputs of each component are appropriate for the operational context of the AI system.

Entities should document the test plans for assessing the validity and performance of each individual component to ensure auditors and third-party assessors can reasonably reproduce those results. The documentation should include testing methodology, to help users understand how the AI model(s) were trained and validated for accuracy, confidence, recall, and potential limitations of the model.⁶³ When testing model components, entities should consider whether the discrete outputs are appropriate for the operational context and consistent with values and principles that foster public trust.

Performance at the System Level

Three of the key practices for ensuring performance at the component level are also applicable at the system level (3.5 to 3.7).

3.5 Documentation: Document the methods for assessment, performance metrics, and outcomes of the AI system to provide transparency over its performance.

3.6 Metrics: Define performance metrics that are precise, consistent, and reproducible.

3.7 Assessment: Assess performance against defined metrics to ensure the AI system functions as intended and is sufficiently robust.

As the components are integrated into the AI system, entities should iteratively test the system as a whole to ensure that it performs correctly and reliably across a wide range of operational conditions. The performance metrics depend on the particular use case and should map to the desired outcomes. For example, an AI model for classifying objects would have performance metrics specific to its use case and produce a different type of output than an AI model used to generate a quantitative prediction.⁶⁴ Additionally, some metrics will be domain specific (e.g.,

⁶³PAI, *Annotation and Benchmarking of Machine Learning*.

⁶⁴Mitchell et al., "Model Cards for Model Reporting."

finance, autonomous vehicles, health care).⁶⁵ According to forum participants and literature we reviewed, assessing an AI system may include testing for robustness against any malicious or deliberate attempt to make the system do something other than meeting its intended purpose.⁶⁶ In addition, entities can consider testing the performance of the AI system with data that the system has not encountered before or that have different distributions from data it has encountered.⁶⁷ Entities should document the tests performed, and the corresponding results to ensure transparency and for auditing or third-party assessments.

We identified two additional practices that are applicable at the system level (3.8 and 3.9).

3.8 Bias: Identify potential biases, inequities, and other societal concerns resulting from the AI system.

Differential Impacts of Facial Recognition Technologies

AI-enabled facial recognition has been used by law enforcement to assist with suspect identification. However, National Institute of Standards and Technology tests of facial recognition technology found that it generally performs better on lighter-skinned men than it does on darker-skinned women, and does not perform as well on children and elderly adults as it does on younger adults. These differences could result in more frequent misidentification for individuals within certain demographics.

Source: GAO, *Facial Recognition Technology: Privacy and Accuracy Issues Related to Commercial Uses*, [GAO-20-522](#) (Washington, D.C.: Jul. 13, 2020). | [GAO-21-519SP](#)

Entities should test the AI system for biases, inequities, or other unintended consequences (see sidebar). For example, entities could test the accuracy of the AI system for each demographic group to identify potential biases.⁶⁸ According to literature, even when a model achieves a good overall accuracy, the errors may not be evenly distributed across

⁶⁵Arnold et al., “FactSheets.”

⁶⁶According to one forum participant, current technologies and practices that could be used for performance testing may not provide sufficient confidence to support strong requirements for robustness or trustworthiness. In those cases, entities may consider limiting the reliance on AI, involving a high level of human supervision, or re-engineering the system to better support performance testing.

⁶⁷Arnold et al., “FactSheets”; European Commission Independent High-Level Expert Group on Artificial Intelligence (AI HLEG), *Ethics Guidelines for Trustworthy AI*, (Brussels, Belgium: Apr. 8, 2019).

⁶⁸The Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK, *Auditing machine learning algorithms*.

demographic groups or concentrated on certain groups.⁶⁹ Engaging individuals and communities impacted by use of the AI system and civil liberties advocates on the testing and validation of the AI system could help surface potential societal concerns and mitigate issues.

3.9 Human supervision: Define and develop procedures for human supervision of the AI system to ensure accountability.

Prior to deploying an AI-enabled system, a point of consideration is the level of human supervision of the AI system needed to ensure accountability. This level depends on several factors, including the purpose and potential consequences of the system. For example, a higher level of human supervision may be necessary if the AI output could result in significant consequences, such as impacting individual civil rights and liberties.⁷⁰ Entities should determine the appropriate degree of human supervision and establish procedures accordingly to ensure the system goals are met. The text box below shows three broad approaches to human supervision of AI-enabled systems.

⁶⁹A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *Frontiers in Big Data*, vol. 2, no. 13: <https://doi.org/10.3389/fdata.2019.00013>.

⁷⁰ODNI, *Artificial Intelligence Ethics Framework*.

Three Broad Approaches to Human Supervision of AI systems

The Model Artificial Intelligence Governance Framework, Second Edition outlines three broad approaches to human supervision of AI systems: 1) human-in-the-loop, 2) human-out-of-the-loop, and 3) human-on-the-loop. The extent to which human supervision is needed depends on the objectives of the AI system and a risk assessment, as illustrated by the examples below.

Human-in-the-loop refers to active human oversight of the AI system, “with the human retaining full control and the AI only providing recommendations or input.” A human reviews the output of the AI system and makes the final decision. The GAO report *Artificial Intelligence in Health Care: Benefits and Challenges of Technologies to Augment Patient Care* provided examples of AI in clinical decision support, including tools used to recommend treatments for cancer, sepsis, and stroke, as well as tools to aid the planning of surgical care. In these examples, AI tools could assist provider decision-making with a greater comprehensiveness and speed than would be possible without such tools, but a human expert (i.e., provider) makes the decision for patient care.

Human-out-of-the-loop refers to the lack of human supervision of the execution of decisions, as in the AI system “has full control without the option of human override.” The GAO report *Artificial Intelligence: Emerging Opportunities, Challenges, and Implications* provided an example of an AI-enabled cybersecurity system that can find and patch system vulnerabilities without human intervention. Mayhem, the winning system in the Defense Advanced Research Projects Agency (DARPA) 2016 Cyber Grand Challenge, is designed to protect apps (software) from new attacks by hackers. Mayhem works by hardening applications and simultaneously and continuously looking for new bugs that may be exploited by hackers. When the system finds new bugs, it autonomously produces code to protect the software vulnerability. Mayhem is an expert system that performs prescriptive analytics, where machines detect and interact without human intervention. This is in contrast to traditional signature-based intrusion detection systems, which rely on human intervention in anticipating cybersecurity attacks.

Human-on-the-loop refers to human supervision in which “the human is in a monitoring or supervisory role, with the ability to take over control when the AI model encounters unexpected or undesirable events.” The *Model AI Governance Framework* used a GPS navigation system as an example. The GPS plans the route from point A to point B, offering several options to the driver based on parameters such as shortest distance, shortest time, or avoid toll roads. After a route is selected and navigation is ongoing, the driver can still take control over the GPS and alter the navigational parameters in the case of unforeseen road congestions.

Source: GAO summary of Info-communications Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC), *The Model Artificial Intelligence Governance Framework*, Second Edition (Singapore, Jan. 2020); [GAO-18-142SP](#), and [GAO-21-75P](#). | GAO-21-519SP

3. Performance Framework

Performance at the Component Level

3.1 Documentation: Catalog model and non-model components along with operating specifications and parameters.

 Questions to Consider	 Audit Procedures
• How is each model component solving a defined problem? • How are the operating specifications and parameters of model and non-model components selected, evaluated, and optimized? • How suitable are the components to the available data and operating conditions? • To what extent are the dimension reduction techniques applied appropriate?	• Collect performance documentation that describe the model and non-model components in the AI system. • Review performance documentation to assess whether selected components are appropriate and suitable for the available data. In addition, assess whether the training and optimization process for each component is appropriate. • Interview developers and program managers to determine whether the AI components are appropriate for the system.

3.2 Metrics: Define performance metrics that are precise, consistent, and reproducible.

 Questions to Consider	 Audit Procedures
• What metrics has the entity developed to measure performance of various components? • What is the justification for the metrics selected? • Who is responsible for developing the performance metrics? • To what extent do the metrics provide accurate and useful measure of performance? • To what extent are the metrics consistent with goals, objectives, and constraints?	• Collect performance management plans and documentation on component specifications and metrics. • Review performance management plans and documents to assess whether the metrics are consistent with goals, objectives, and constraints. • Interview program managers and developers to assess whether the performance metrics align with goals and objectives.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP

3. Performance Framework

Performance at the Component Level

3.3 Assessment: Assess the performance of each component against defined metrics to ensure it functions as intended and is consistent with program goals and objectives.

 Questions to Consider	 Audit Procedures
• How is the selection of the mathematics and/or data science techniques, including any ensemble method, documented and assessed? • To what extent is the selection of techniques appropriate? • How appropriate is the training and optimization process for each component within the system?	• Collect documentation, on edge case testing¹, training and optimization procedures, data quality assessments, and internal control documentation related to component performance metrics. In addition, collect predictive models, software codes, operational parameters, variable selection processes, and training methodology. • Review data quality assessments and compare them to the initial test plans to assess the validity of the component and its integration into the system. In addition, run test cases through the models to assess reproducibility and statistical performance (accuracy, precision, recall, statistical errors, and confidence). • Interview AI technical stakeholders—data scientists, data engineers, developers, and program managers—to determine how individual components perform and interact within the AI system.

3.4 Outputs: Assess whether outputs of each component are appropriate for the operational context of the AI system.

 Questions to Consider	 Audit Procedures
• How did the entity determine whether the outputs of each component is suitable for the operational context? • To what extent is the output of each component appropriate for the operational context? • To what extent are the model outputs consistent with the entity’s values and principles to foster public trust and equity?	• Collect model outputs, business rules, and predictive results. • Review outputs, business rules, and predictive results of each component to determine whether they align with program objectives, compliance and values. • Interview legal counsel, compliance officers, and AI technical stakeholders on compliance efforts and mitigation strategies to assess compliance with relevant laws and regulations.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP
Notes: In this table AI technical stakeholders can include data scientists, data engineers, developers, and program managers.
¹ Edge case testing refers to testing computer programs with input values that are at extreme ends of the expected values.

3. Performance Framework

Performance at the System Level

3.5 Documentation: Document the methods for assessment, performance metrics, and outcomes of the AI system to provide transparency over its performance.



Questions to Consider

- To what extent has the entity documented the AI system's development, testing methodology, metrics, and performance outcomes?
- To what extent does the documentation describe test results, limitations, and corrective actions, including efforts to minimize undesired effects in the outcomes?



Audit Procedures

- Collect documentation on performance testing, metrics, and methodology.
- Review documentation to determine whether it clearly defines the performance metrics and testing conducted before deployment of the AI system.
- Interview AI technical stakeholders to understand the methodologies, metrics, and outcomes.

3.6 Metrics: Define performance metrics that are precise, consistent, and reproducible.



Questions to Consider

- What metrics has the entity developed to measure performance of the AI system?
- What is the justification for the metrics selected?
- Who is responsible for developing the performance metrics?
- To what extent are the metrics consistent with system goals, objectives, and constraints, including ethical and compliance considerations?



Audit Procedures

- Collect performance management plans and documentation on system specifications and metrics.
- Review performance management plans and documents to assess whether the metrics are consistent with systems goals, objectives, and constraints and appropriate for the use case.
- Interview program managers and developers to assess whether the performance metrics align with goals and objectives.

3.7 Assessment: Assess performance against defined metrics to ensure the AI system functions as intended and is sufficiently robust.



Questions to Consider

- What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e. adversarial or stress testing)?
- Who is responsible for testing the AI system?
- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?



Audit Procedures

- Collect documentation on testing methodology and data on performance testing.
- Review documentation and data to assess whether the testing was sufficient to ensure system performance, robustness, and detection of unwanted biases or other concerns.
- Interview AI technical stakeholders to assess whether system tests were appropriate and sufficient.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP
Notes: In this table AI technical stakeholders can include data scientists, data engineers, developers, and program managers.

3. Performance Framework

Performance at the System Level

3.8 Bias: Identify potential biases, inequities, and other societal concerns resulting from the AI system.

 Questions to Consider	 Audit Procedures
• To what extent does the AI system perform differently when using different demographics or populations? • Which population(s) does the AI system impact? • Which population(s) does it not impact? • How did the entity address disparate impacts resulting from the AI system, if any? • To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?	• Collect performance management plans and documentation which describes the mitigation techniques to minimize unintentional consequences. In addition, collect performance outcomes, relevant laws and regulations, professional standards, and federal and entity guidance. • Review performance documentation and performance outcomes to assess whether the system performs differently when using different demographics or populations and whether the mitigation techniques to address or reduce bias, inequity, and/or other concerns were effective. • Interview AI technical stakeholders, legal and policy officers, social scientists, and civil liberty advocates to determine whether potential biases in data are identified and mitigated to determine if the mitigation techniques were effective in addressing unintended consequences.

3.9 Human supervision: Define and develop procedures for human supervision of the AI system to ensure accountability.

 Questions to Consider	 Audit Procedures
• How has the entity considered an appropriate degree of human involvement in the automated decision-making processes? • What procedures have been established for human supervision of the AI system? • To what extent has the entity followed its procedures for human supervision to ensure accountability?	• Collect procedures, work load assessments between the humans and the AI system, operational documents, test plans, intermediate results, and user experience descriptions. • Review documentation to assess whether the AI system provides accurate and interpretable information to the human. In addition, assess whether testing included use cases, and compared test results with the use case goals. • Interview developers, user experience designers, program managers, and test users to determine whether the degree of human involvement and supervision is appropriate.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP
Notes: In this table AI technical stakeholders can include data scientists, data engineers, developers, and program managers.



FRAMEWORK
PRINCIPLE

4.

MONITORING

Key Monitoring Practices

Monitoring Principle:

Ensure reliability and relevance over time.

Source: GAO. | GAO-21-519SP

AI systems are dynamic and adaptive, and performance can vary over time. Management should establish a monitoring framework to ensure the AI system maintains its utility and remains aligned with current objectives.

GAO developed five key practices for this principle, grouped into two categories:

- **Continuous monitoring of performance:** This category involves tracking inputs of data, outputs generated from predictive models, and performance parameters to determine whether the results are as expected.⁷¹
- **Assessing sustainment and expanded use:** This category involves examining the utility of the AI system, especially when applicable laws, programmatic objectives, and the operational environment may change over time. In some cases, entities may consider scaling the use of the AI system (across geographic locations, for example) or expanding its use in different operational settings.

Continuous Monitoring of Performance

According to the *Federal Internal Control Standards*,⁷² entities use monitoring to assess the quality of performance over time and promptly resolve deficiencies. Corrective actions are a necessary complement to control activities in order to achieve objectives. Forum participants emphasized the need for continuous monitoring and evaluation of the AI system once it is operationally deployed. As an example, these participants noted that data collected to train a machine learning model may no longer be applicable to the current context. As a result, the predictions generated by the model may not be accurate. GAO identified three key practices to help entities provide assurance that the AI system operates as intended over time.

⁷¹GAO, *Foreign Assistance: Federal Monitoring and Evaluation Guidelines Incorporate Most but Not All Leading Practices*, [GAO-19-466](#) (Washington, D.C.: Jul. 31, 2019).

⁷²[GAO-14-704G](#).

4.1 Planning: Develop plans for continuous or routine monitoring of the AI system to ensure it performs as intended.

Entities should develop plans to continuously or routinely monitor performance and risks, including the risk of bias and risks to privacy and security. The plan should include a monitoring frequency that is appropriate for each use case. Monitoring activities for some purposes should be more frequent than others and proportional to the impact of an incorrect output, according to literature.⁷³ For example, AI for recognizing images of an object may not need to be monitored as frequently as AI for phishing detection. Patterns of the object's images are unlikely to change quickly,⁷⁴ whereas perpetrators of phishing schemes are likely to adjust their tactics to avoid detection.⁷⁵

4.2 Drift: Establish the range of data and model drift that is acceptable to ensure the AI system produces desired results.

As part of the monitoring plan, entities should decide and document the range of data and model drift that is acceptable and will produce desired results.⁷⁶ Data drift refers to the changes in the statistical properties of the input data in an operational environment, as compared to the training data. Model drift refers to the changes in the relationship between the data inputs and the prediction outputs. Data and model drifts could result in performance degradation.⁷⁷ Entities may need to retrain the components of the AI system if the data or model drift for each component is not within the acceptable range. The range should be

⁷³Information Commissioner's Office (ICO). *Guidance on AI and Data Protection* (Wilmslow, UK: July 30, 2020).

⁷⁴World Economic Forum, *Companion to the Model AI Governance Framework*.

⁷⁵D. F. Engstrom, D. E. Ho, C. M. Sharkey, and M.-F. Cuéllar. *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies* (Washington, D.C.: Administrative Conference of the United States, Feb. 2020); World Economic Forum, *Companion to the Model AI Governance Framework*.

⁷⁶ICO, *Guidance on AI*.

⁷⁷S. Shendre, *Model Drift in Machine Learning*, accessed May 21, 2021, <https://towardsdatascience.com/model-drift-in-machine-learning-models-8f7e7413b563>; Microsoft, *What is Data Drift?*, accessed May 25, 2021, <https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?tabs=python>

established based on the nature, scope, and purpose of the component and the risks it poses.⁷⁸

4.3 Traceability: Document results of monitoring activities and any corrective actions taken to promote traceability and transparency.

Monitoring plans should also include corrective actions to be taken if monitoring reveals a performance deficiency, according to a 2019 GAO report on *Federal Monitoring and Evaluation Guidelines*. If such deficiencies are detected, the entity should correct them in a timely manner. Further, the entity should document the deficiencies, corrective actions, and the effect on the output. Documentation may include how often corrective actions were needed and how they affected performance. These records serve as lessons learned⁷⁹ and enhance traceability and transparency,⁸⁰ according to our review of literature.

Assessing Sustainment and Expanded Use

We identified two key practices that entities should consider as they sustain or expand the use of the AI system.

4.4 Ongoing assessment: Assess the utility of the AI system to ensure its relevance to the current context.

Entities should regularly consider the utility of the AI system to ensure that it is still useful. For example, as one forum participant noted, an AI system trained on traffic patterns in 2019 might not be useful in 2020 because of reduced traffic during the COVID-19 pandemic. In assessing utility, entities should also consider the extent to which the AI system is still needed to address the goals and objectives. In addition, changing laws, operational environments, resource levels, or risks could affect the utility of the AI system compared to other alternatives. Therefore, entities should also consider metrics for determining when to retire the system and the process for doing so.

⁷⁸ICO, *Guidance on AI*.

⁷⁹[GAO-19-466](#).

⁸⁰Arnold et al., "FactSheets."

4.5 Scaling: Identify conditions, if any, under which the AI system may be scaled or expanded beyond its current use.

Some entities may wish to scale up or expand use of an AI system that has been successfully deployed and has demonstrated its utility. Before doing so, entities should identify conditions, if any, under which scaled or expanded use is appropriate. Forum participants cautioned that an AI system successfully deployed in one region may not perform well in another because of differences in context, demographics, or other factors. For example, a 2020 GAO report on AI in medical services found that expanding an AI system across multiple health care systems may be challenging because of differences among institutions and patient populations. This report further noted that many AI systems are initially designed to solve a problem at one health care system, based on the patient population specific to that location and problem. To scale across different settings, the system needs to be able to accept and use data from other sources or locations—the more locations, the more complex the challenge. Bias could be introduced if an AI system was developed based on data from a patient population that may not be representative across health care systems.

4. Monitoring Framework

Continuous Monitoring of Performance

4.1 Planning: Develop plans for continuous or routine monitoring of the AI system to ensure it performs as intended.



Questions to Consider

- What plans has the entity developed to monitor the AI system?
- To what extent do the plans describe processes and procedures to continuously monitor the AI system?
- What is the established frequency for monitoring the AI system?
- To what extent is the frequency feasible and appropriate for effectively managing system performance?



Audit Procedures

- Collect monitoring plans, schedules, and related tracking documents for the AI system.
- Review monitoring plans to assess whether the entity clearly describes its efforts to continually identify, assess, and mitigate differences between the range of data or model drift that is acceptable and the system's performance. In addition, review monitoring plans and schedules to assess whether established frequencies and schedules for monitoring are feasible and appropriate for the specific use case.
- Interview monitoring stakeholders—developers, quality assurance engineers, program managers, and data scientists—to assess whether plans to monitor the AI system perform as intended.

4.2 Drift: Establish the range of data and model drift that is acceptable to ensure the AI system produces desired results.



Questions to Consider

- To what extent has the entity established an acceptable range for the data and model drift?
- To what extent was the acceptable range for the data and model drift established based on a risk assessment, and is it appropriate for its use case?
- What mechanisms have been developed to detect data and model drift?



Audit Procedures

- Collect monitoring plans and other related documents that specify the acceptable range of data and model drift and describe how it was established.
- Review monitoring plans, statistical tests, and other related documents to assess whether the established range of acceptable data and model drift is appropriate for mitigating risks.
- Interview monitoring stakeholders to determine whether range of data and model drift is appropriate.

4.3 Traceability: Document results of monitoring activities and any corrective actions taken to promote traceability and transparency.



Questions to Consider

- To what extent do the monitoring and tuning activities track performance and avoid undesired consequences?
- To what extent did the entity document the frequency and rationale for updating the AI system?
- To what extent did the entity document the results of monitoring activities and corrective actions?



Audit Procedures

- Collect results from monitoring activities and change logs on corrective actions and versions of the AI system.
- Review monitoring documentation to assess whether the established processes are followed, if results show any changes in performance over time, and if corrective actions were applied.
- Interview monitoring stakeholders to determine whether corrective actions, if any, impacted the AI system's performance.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP

Notes: In this table, monitoring stakeholders can include developers, quality assurance and operational engineers, program managers, and data scientists.

4. Monitoring Framework

Assessing Sustainment and Expanded Use

4.4 Ongoing assessment: Assess the utility of the AI system to ensure its relevance to the current context.

 Questions to Consider	 Audit Procedures
• How does the entity determine the ongoing utility of the AI system? • To what extent is the AI system still needed to achieve goals and objectives? • To what extent has the entity identified metrics that will help it determine whether and when to retire the AI system? • To what extent does the entity have established procedures for retiring the AI system, if it is no longer needed?	• Collect any internal and third-party assessments, Inspector General audits, and evaluations of the AI system and information on the context for which it is applied. • Review assessments, audits, and evaluations to determine the utility, relevance, and continuance of using the AI system. • Interview development, quality assurance and operational engineers and program managers to determine the utility of the AI system in its current context.

4.5 Scaling: Identify conditions, if any, under which the AI system may be scaled or expanded beyond its current use.

 Questions to Consider	 Audit Procedures
• What assessments and evaluations have been completed to determine if the AI system can be applied to address other issues/problems? • To what extent did these assessments and/or evaluations identify conditions under which such applications can or cannot take place? • How did the entity use these assessments and/or evaluations to determine if the system can be scaled up? • To what extent can the AI system be applied to different use cases or problems?	• Collect results of any monitoring activities, internal and third-party assessments, Inspector General audits, and evaluations as well as correspondence on the conditions, assumptions, and limitations on scaling up the AI system. • Review assessment, evaluations, and correspondence to determine whether the conditions, assumptions, and limitations under which the system can or cannot scale up were identified and documented. • Interview monitoring stakeholders to determine whether conditions or limitations under which the system can scale up are appropriate.

Source: GAO analysis of U.S. and foreign government, industry, academic, and nonprofit organization documents; interviews; and forum discussion. | GAO-21-519SP

Notes: In this table, monitoring stakeholders can include developers, quality assurance and operational engineers, program managers, and data scientists.

Appendix I: Objectives, Scope, and Methodology

This report identifies key practices in a single framework to help ensure accountability and responsible use of artificial intelligence (AI) by federal agencies and other entities involved in the design, development, deployment, and continuous monitoring of algorithm-based systems. This framework provides auditors with approaches to develop credible assurance assessments of AI systems.

To address this objective, we developed the Framework based on the following sources: (1) information we collected through the Comptroller General (CG) Forum, review of literature, and interviews with subject matter experts; (2) analysis and criteria of key practices; and (3) validation of key practices and technical comments.

Information Collection

Interviews. We conducted 34 interviews with entities with experience in designing, developing, and deploying AI systems. We identified experts and organizations to interview according to criteria including organizational impact, subject matter expertise, literature references, and interview referrals. We interviewed officials from 15 federal agencies, departments, and Offices of Inspector General, as well as one state and one international governmental entity. In addition, we conducted the remaining interviews with experts from the academic, industrial, and nonprofit sectors. Experts interviewed represented software developers, data scientists, privacy and security experts, risk management professionals, attorneys, civil liberties advocates, users, and individuals affected by AI systems.

Comptroller General Forum. We convened a forum, in September of 2020, of experts in industry, government, nonprofits, and academia to discuss factors affecting oversight of AI, including AI governance, sources of evidence, methods to assess implementation of AI systems, and identifying and mitigating potential bias and inequities.

The forum addressed the following objectives:

- What existing standards and governance frameworks can third-party entities or audits apply to assess AI systems?
- According to experts, what tools and practices can third-party entities use to identify, assess, and mitigate bias and ethical concerns in the AI life cycle?
- According to experts, what control activities can be applied to help assure accountability over AI systems in the public sector?

- According to experts, what tools and practices can help third-party entities evaluate an AI system across its life cycle?
- According to experts, what types of risks and challenges exist in applying AI systems in the public sector?

We summarize the views of experts who participated in the Comptroller General's Forum on Artificial Intelligence in Appendix II. To describe statements made by forum participants, we distinguish between issues identified by a single participant and those identified by more than one participant by using the phrase "forum participants." The forum agenda was structured in sessions to address specific topics, see Appendix III, one of which included presentations by participants. Other sessions were moderated discussions with a facilitator from GAO's Applied Research and Methods team.

We selected 23 experts representing 20 organizations to participate in the forum. Participants represented federal program managers and auditors, state auditors, private industry, academia, and nonprofits. These individuals presented a variety of perspectives, including those of software developers, data scientists, privacy/security experts, risk management professionals, legal counsel, civil liberties advocates, users, and individuals affected by AI systems. For a list of forum participants, see Appendix IV. The team's methodology identified experts and organizations according to criteria such as organizational impact, subject matter expertise, literature references, and interview referrals.

In advance of the forum, we prepared a background reading package based on interviews with experts and relevant literature, which we distributed to forum participants. The reading package featured a brief overview of the issues to consider in each of the five sessions.

Literature Review. We conducted a review of literature, which included reports, journal articles, and trade publications related to accountability, governance, equity, and assessments of AI. We reviewed various sources, including (1) publications identified during a formal literature review, aided by a GAO research librarian; and (2) literature recommendations from external experts, entities we interviewed, and discussions with GAO experts. Our literature review included a search for peer-reviewed articles, government reports, and trade publications, among other sources, in databases such as Scopus, Institute of Electrical and Electronics Engineers (IEEE), and ProQuest's science collections. We limited our results to publications from January 1, 2013, to the summer of 2020, when we conducted the search. As a result of this

search, we ultimately identified 58 publications for in-depth review of practices related to governance and accountability of AI systems. In addition to the literature review, we also selected and reviewed 10 publications from a list of sources recommended by external and internal experts we interviewed. In all, we reviewed a total of 68 publications from our literature review and recommended literature. As part of our research, we considered existing frameworks and guides related to AI governance and auditing, including publications by foreign governments and the U.S. government (see table 1 and 2 on the list of frameworks we reviewed). For a list of all publications noted in the product, see Appendix VI.

Criteria for Key Practices

We conducted an analysis on the information we collected from the sources above by first compiling a list of practices from those sources. We also categorized the practices according to two source types—testimonial and documentary—to facilitate analysis and application of the criteria described below. Our testimonial information came from discussions during the Forum. The documentary sources included the 68 publications identified during our literature review as well as other sources recommended to us by individuals we interviewed and who participated in the Forum. We defined a practice as a key practice if at least two independent sources described it as important for implementing an AI system. In addition, we assessed whether the practices were relevant to GAO's *Government Auditing Standards* and *Standards for Internal Control in the Federal Government*.

To determine whether a practice was relevant, we identified at least two independent sources that note the importance of a certain practice in implementing AI systems, as well as applied professional judgment when analyzing practices, and we sought validation of key practices from subject matter experts (see next section for additional details). In addition, we considered other factors when analyzing the importance of a particular practice or concept for effectively addressing oversight of AI, such as whether the statements were broadly applicable or limited to certain circumstances. We mitigated the risk of omitting key practices by seeking validation from external entities, as described below. Moreover, we considered challenges and trade-offs when analyzing the practices, and used professional judgment to develop key practices that give agencies flexibility to handle those challenges.

Validation of Key Practices and Technical Comments

To validate the key practices, we provided the draft of the framework and an outline of the forum findings to all of the forum panel participants (see app. IV). These individuals represented subject matter experts in AI from three federal government agencies and two Offices of Inspector General, one state audit agency, and one international organization, as well as industry, academia, and the nonprofit sectors. Forum participants provided technical comments and edits, which we incorporated, as appropriate. In addition, GAO reviewers who are subject matter experts in AI and other related matters also provided comments and edits.

The framework covers broadly applicable principles, key practices, and audit procedures. It does not cover all possible principles, nor does it specify the exact criteria for evaluation or level of acceptable performance. Many of those aspects are specific to the AI use case or domain. In addition, not all practices apply to every AI system. This framework does not provide specific criteria to assess cybersecurity or privacy risks, nor does it address specific protections or actions that should be taken with regard to the use of AI.

Appendix II: Insights from a Comptroller General Forum on Oversight of Artificial Intelligence

In September 2020, the Comptroller General of the United States (CG) convened a forum of experts in industry, government, nonprofits, and academia to discuss factors affecting oversight of Artificial Intelligence (AI), including AI governance, sources of evidence, methods to assess implementation of AI systems, and identifying and mitigating potential bias and inequities. In fiscal year 2021, the White House requested a nondefense budget of \$1.5 billion for AI, a 54 percent increase over the fiscal year 2020 budget request. However, AI systems pose unique accountability challenges and raise concerns related to civil liberties, bias, and social disparities. The U.S. government, industry, professional associations and others have begun to develop principles and frameworks to address these concerns, but there is limited information on how these will be implemented to enable third-party verifications and assessments of AI systems. The Forum was convened to better understand these issues to ensure effective, efficient, economical, ethical, and equitable implementation of AI in the public sector. Forum participants discussed how to operationalize recent principles and frameworks on the use of AI into practices for managers and supervisors of these systems, as well as mitigation strategies to address challenges in implementing AI in the public sector.

The following are key results from each session of the forum.

Session I: Factors to Consider When Auditing AI Systems

Forum discussions focused on: (1) operationalizing principles on the use of AI systems into implementation practices; (2) establishing governance practices; and (3) developing methods to allow assessments of AI systems. Entities should consider these factors to ensure necessary and sufficient information is available for auditors and third party assessors. Participants noted that there is no one-size-fits-all approach to AI, and thus different organizations may structure these three elements differently.

Operationalizing AI principles on the use of AI. General themes discussed by participants included performance management, data veracity, fairness, transparency, and explainability.¹ A participant noted that an entity's own established principles on the use of its AI system, if they exist, can serve as high-level goals for the entity to review as it evaluates the AI system. Auditors may also use these principles to review the AI system, but only as a starting point. A participant added, however,

¹Explainability refers to methods and techniques in the application of artificial intelligence such that the results of the solution can be understood by humans.

that principles alone will not provide sufficient and necessary criteria to determine if the AI system meets its objectives. Rather, they can serve as a guide and provide high-level characteristics of what the entity wants the AI system to achieve, according to participants. For example, one participant said that, at their organization, trust and transparency serve as a basis for the AI governance model and are considered at each step in the AI life cycle.² At the same time, forum participants highlighted the importance of operationalizing these principles into practices. While each entity—whether it is in the public, private, or nonprofit sector—may develop its own AI principles, a forum participant said that AI principles should be 1) specific to AI, 2) implementable in policy and practice, 3) context-specific, 4) flexible to stand the test of time, and 5) facilitate innovation and trust.

Establishing governance practices. Forum participants provided various examples of governance practices,³ including:

- documenting and defining goals and objectives;
- defining roles and responsibilities of key personnel;
- defining the regulatory environment—including minimum requirements in laws and regulations;
- and establishing policies related to risk management and mitigation.

One participant highlighted the need for entities to identify outcomes they want to achieve when implementing an AI system. Another said embodiment of AI principles within governance practices helps the entity design and build AI systems that align with the entity's values. For example, to demonstrate an entity's commitment to transparency, governance structures may include requirements to document all of the design and development decisions across the phases of the AI life cycle.

Participants also noted the importance of establishing an oversight and ethics committee involving multidisciplinary stakeholders, both internal and external to the entity, throughout an AI system's development to ensure that societal concerns are considered. Some organizations represented on the panel have established committees as a method to

²Transparency includes the balance between the public's right to know and the proprietary of the algorithms used, according to forum participants.

³Governance includes structures, processes, and procedures for directing, managing, and monitoring AI activity, according to a forum participant.

address potential societal concerns. According to participants, these committees should include stakeholders from across the entity—legal, policymakers, developers, as well as top executives from all divisions—to ensure appropriate development, deployment, and application of the AI system.

Testimonial:

"We [built] an AI ethics board that has representatives from all divisions [who] can add ... decision power. We [learned] that this is fundamental ... to give the board enough power to make decisions."

Source: Participant in CG Forum on AI Oversight. | GAO-21-519SP

Participants also noted the need to include external stakeholders during development and deployment of public sector AI systems. They emphasized the need to engage a variety of individuals from the broader community, such as impacted users and subject matter experts in the legal, policy, and civil liberties communities. Participants highlighted several challenges that come with developing governance practices for AI. For example, AI is not a static process, and therefore these practices must be adaptive. In addition, moving from theory (AI principles) to practice (governance structures) can be challenging.

Developing methods to validate AI systems. Forum participants noted the importance of developing tools and/or methods to help validate AI systems. Auditors can review and assess these methods and procedures to determine if the system is working as intended. One participant noted that procedures to validate AI systems are necessary to increase trust in decisions made by AI systems and the resulting outcomes. For example, one such process is using *nutrition labels*—inspired by food nutritional labels—that aim to highlight the key “ingredients” in a data set, such as metadata and populations, as well as unique features regarding distributions and missing data.⁴ In addition, some entities are developing certification systems to evaluate the extent to which AI systems are being implemented in a responsible way. Such methods and procedures can help to initially validate AI systems.

⁴For a description of nutrition labels for datasets, see website link <https://datanutrition.org>, accessed May 10, 2021. See section IV in this appendix for additional information on technical tools to assess AI.

Session II: Criteria for and Challenges Associated with Auditing AI Systems

Forum participants provided several examples of criteria that may be applied when assessing AI systems including 1) the audited entity's internal control system, as well as its own requirements and objectives for the AI system; 2) existing legislation or regulatory guidance; and 3) established frameworks and standards in areas related to AI. For example, auditors can determine the extent to which the system is meeting the entity's own objectives or requirements for the AI system. Federal agencies may have agency-specific requirements for using AI systems and specific objectives they should meet.

In addition, participants noted that existing law or regulatory guidance may serve as criteria for assessing AI systems. In particular, the Office of Management and Budget (OMB) requires federal agencies to ensure government data used in AI systems comply with the 2017 Open, Public, Electronic, and Necessary Government Data Act, among other relevant directives.⁵ OMB also notes that agencies should follow legal and policy requirements on protecting sensitive information and public interest such as privacy, security, and national economic competitiveness.

Furthermore, there are well-established frameworks and standards—such as the National Institute of Standards and Technology's (NIST) cybersecurity framework and the International Organization for Standardization (ISO) data privacy management standards—which could be applied to audits until AI-specific standards are developed and adopted.⁶ Forum participants noted that these frameworks and standards address overlapping concepts in managing data, governance, and security which will be relevant in assessing AI systems.

However, according to participants, existing frameworks and standards may not provide sufficient detail on assessing social and ethical issues

⁵Office of Management and Budget, OMB Memorandum M-21-06, *Guidance for Regulation of Artificial Intelligence Application*, (Nov. 17, 2020) (referring to Title II of the Foundations for Evidence-Based Policymaking Act of 2018, and also referred to as the "OPEN Government Data Act," Pub. L. No. 115-435, tit. II, 132 Stat. 5534 (2019), in addition to other relevant directives).

⁶National Institute of Standards and Technology (NIST). *The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management*, (Gaithersburg, Md.: Jan. 2020), accessed May 5, 2021, <https://www.nist.gov/privacy-framework>. Also see International Organization for Standardization, (ISO/IEC) TR 24028:2020 *Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*, accessed May 5, 2021, <https://www.iso.org/standard/77608.html>.

which may arise from the use of AI systems. Similarly, according to NIST, *“while there is broad agreement that societal and ethical issues, governance, and privacy must factor into AI standards, it is not clear how that should be done and whether there is yet sufficient scientific and technical basis to develop those standards provisions.”*⁷ Based on our review of literature, while several entities (e.g. the Organisation for Economic Co-operation and Development, the European Commission, the U.S. Department of Defense) have adopted or drafted high-level principles for implementing trustworthy and equitable AI, many entities—specifically those in industry and government—are still developing standards for these areas.⁸

Testimonial:

“I don’t think that, right now, there is established AI governance in governmental entities. I think that some people confuse IT governance, which is part of AI governance, and think that that’s basically equivalent. [Those people are] not considering the social and ethical factors.”

Source: Participant in CG Forum on AI Oversight. | GAO-21-519SP

Regardless, forum participants noted that while in many cases there are commonalities in AI principles across entities, there is a lack of consensus on how to define and operationalize AI principles and develop common standards and criteria. In order to translate principles into technical requirements, there needs to be a consensus on how the principles are defined and what constitutes trustworthy AI, participants said. Participants also highlighted that definitions or criteria may not be generalizable and can depend on the social and regulatory environment. For example, fairness may be defined and measured differently depending on the purpose and objective of the program. These issues may raise some challenges in auditing AI systems.

⁷NIST. *FAQ about NIST’s Role in Planning Federal Engagement in AI Standards Development*, accessed May 5, 2021, <https://www.nist.gov/artificial-intelligence/faqs-about-nists-role-planning-federal-engagement-ai-standards-development>.

⁸Organisation for Economic Co-operation and Development, *Recommendation of the Council on Artificial Intelligence OECD/LEGAL/0449* (Paris, France: adopted May 22, 2019) <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>; European Commission Independent High-Level Expert Group on Artificial Intelligence (AI HLEG), *Ethics Guidelines for Trustworthy AI*, (Brussels, Belgium: Apr. 8, 2019); Department of Defense, Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence* (Washington, D.C.: Oct. 2019).

Testimonial:

"I think [a clear and common set of vocabulary with a consensus on definitions] is one of the things that's missing right now. Once we know what it is that we want to measure, then we can work towards the development of the metrics and test it for benchmark validation, verification, performance evaluations, or compliance."

Source: Participant in CG Forum on AI Oversight. | GAO-21-519SP

Session III: Challenges to Using and Auditing AI Systems in the Public Sector

Forum participants identified several challenges to the use and auditing of AI systems within the public sector. These challenges include 1) a need for expertise, 2) limited understanding of how the AI system makes its decisions, and 3) limited access to key information due to commercial procurement of such systems. According to forum participants, there is a greater need for explainability and transparency when using AI systems in the public sector, compared to the private sector. Similarly, an OECD working paper states that governments have to ensure balance between encouraging AI to foster public sector innovation and improved public services, while protecting the public and service users' interests from potential unintended negative consequences.⁹

Public sector managers and subject matter experts need a baseline understanding of the technical aspects of AI systems, according to forum participants. A lack of understanding may pose a challenge, as entities should provide information to a variety of stakeholders, according to participants. Moreover, citizens affected by the AI system including users, impacted communities, and auditors—need to be able to understand how the system makes its decisions. Some AI systems are automated and make decisions based upon the information provided; however, within the public sector, many AI systems are often used in an advisory role, and provide information to a human making decisions, according to participants. For example, when AI-generated criminal risk assessments are used in criminal proceedings, the judge can overrule the AI-generated

⁹Organisation for Economic Co-operation and Development (OECD), Working Papers on Public Governance No. 31. *State of the Art in the Use of Emerging Technologies in the Public Sector*, accessed Apr. 09, 2021, https://www.oecd-ilibrary.org/governance/state-of-the-art-in-the-use-of-emerging-technologies-in-the-public-sector_932780bc-en.

assessment of an offender's likelihood to re-offend.¹⁰ In situations like this, according to one participant, clear explanations are needed on the function of the AI system as well as the role of the judge in the decision-making process. A participant noted that auditors may also lack the expertise needed to audit AI systems. Auditors may understand some content aspects of AI systems—such as security and compliance—but understanding the algorithms and technical aspects of the system requires specific expertise.

Testimonial:

"As cybersecurity auditor[s]... we understand security and compliance and all of those issues. However, to understand the algorithms that go into developing the AI technology, and to understand an outcome, we felt that we needed to work closely with data scientists, the technical experts that [are] actually developing the AI project and have some access to some type of data analytics team so that we can have them interpret for us what those algorithms are, how they're supposed to work, and whether the algorithms are meeting the outcomes that were expected for that project."

Source: Participant in CG Forum on AI Oversight. | GAO-21-519SP

Entities and auditors may also face challenges in understanding how AI systems make decisions due to the inherent lack of explainability of some systems. Such challenges may limit confidence and trust in AI systems. Participants noted that even with full access to the software, data, and trained networks, AI systems are naturally opaque and resistant to direct analysis in most cases, making them difficult to evaluate and verify.¹¹

¹⁰According to the Brookings Institution, criminal risk assessment algorithms are designed to use a range of factors—such as an individual's age and history of misconduct—to predict the likelihood of an individual committing a new crime or failing to appear in court. A decision-making framework translates these risk scores into release-condition recommendations, with higher risk scores corresponding to stricter release conditions. Judges can disregard these recommendations if they seem too strict or too lax. Similar algorithms influence a wide variety of judicial decisions, including sentencing decisions and probation and parole requirements. See Brookings Institution, *Understanding risk assessment instruments in criminal justice*, (Washington, D.C.: June 19, 2020) accessed Apr. 12, 2021, <https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice/#:~:text=One%20class%20of%20algorithmic%20tools,an%20individual%20before%20their%20trial>.

¹¹Modern machine-learning systems are black-box systems, for which users are unable to understand why the system makes a specific decision or recommendation, why a decision may be in error, or how an error can be corrected. The goal of explainable AI is to develop machine-learning systems that provide an explanation for their decisions and recommendations and allow users to know when and why the system will succeed or fail. For additional information, see GAO, *Artificial Intelligence: Emerging Opportunities, Challenges, and Implications*, [GAO-18-142SP](#) (Washington, D.C.: Mar. 28, 2018).

Forum participants also noted that a lack of documentation—such as when model outputs or data are accepted, retained, or discarded—may further limit understanding and auditing of these systems.

Forum participants noted that auditors may have limited access to key information in cases where the procured AI system was commercially developed or licensed.¹² For example, commercial developers may have concerns about providing auditors access to the source code or methods used to develop the model due to concerns about the system being reverse engineered and overall risks to their intellectual property. Additionally, participants noted that there are privacy concerns related to providing access to certain aspects of the AI system—such as training data. For example, according to one forum participant, if the AI system was trained on patient medical records, releasing the training data for review could infringe upon the privacy of the patients whose records were used. Participants noted that these concerns may make it difficult for organizations—including auditing organizations—to obtain access to key information needed to audit the system, especially in commercially developed procured AI systems.

Participants discussed several strategies to mitigate challenges in using or adopting AI systems in the public sector. For example, one way to mitigate the challenges with a lack of expertise is to develop in-house expertise or partnerships with experts, participants noted—not just technical experts, but people who can acquire both subject matter knowledge and competence with AI systems. According to forum participants, entities within the public sector should focus on building expertise—including capacity of program managers and in the auditing community. One participant noted that a number of federal agencies have used mechanisms, such as the Intergovernmental Personnel Act's Mobility Program and collaboration agreements with academic institutions to provide additional topic-specific expertise.¹³ Another stated that within

¹²Entities may procure “off-the-shelf” commercially licensed software for AI systems or enter into a contract with a third-party for commercially-developed custom software for AI systems, according to participants.

¹³The Intergovernmental Personnel Act of 1970, Pub. L. No. 91-648, 84 Stat. 1909 (Jan. 5, 1971), as amended, codified at 5 U.S.C. §§ 3371-3375. The Intergovernmental Personnel Act's Mobility Program provides for the temporary assignment of personnel between the federal government and state and local governments, colleges and universities, Indian tribal governments, federally funded research and development centers, and other eligible organizations.

the public sector there should be a focus on building civil servant literacy about data and algorithms as a means to close the digital divide.

For auditors, participants highlighted the need to partner with experts in the data science fields. In addition, entities could increase understanding of their algorithms, how they work, and whether they meet expected outcomes by developing partnerships with stakeholders. For example, auditors should work closely with data scientists, technical experts developing the AI project, and other data analytics teams to understand the algorithms that go into developing the AI technology or to understand the outcomes.

To understand how AI systems make decisions, participants said, entities should ensure developers, users, and stakeholders coordinate and collaborate throughout the AI life cycle. Participants noted that documentation of key decisions across the AI life cycle should be maintained to improve explainability and help users, auditors, and stakeholders understand the AI system. Involving stakeholders and users throughout the AI life cycle can help build a network of stakeholders who understand the way the AI system functions. A participant noted that there are tools specifically designed to automate explanations and help developers improve explainability across varying use cases.

Finally, to mitigate access issues, forum participants stated, contracts should be transparent regarding access requirements. These should include requirements for entities using AI-enabled systems to have access to system's test results and other key information on the data and model to ensure oversight and improve the ability to audit those systems. For example, the Government of Canada retains the right to access and test automated decision systems, including all released versions of proprietary software components, in cases where it is necessary for a specific audit, investigation, inspection, examination, enforcement action, or judicial proceeding, subject to safeguards against unauthorized disclosure.¹⁴

¹⁴Government of Canada, Treasury Board, *Directive on Automated Decision Making* (Feb. 5, 2019), accessed Apr. 30, 2020, <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592§ion=html>.

Session IV: Possible Sources of Evidence for Auditing AI Systems

Forum participants identified several sources of evidence auditors may collect during an audit to assess an entity's governance practices, data sources and documentation, and performance of the AI models and algorithms.¹⁵

Assessing Governance Practices

Forum participants discussed four potential sources of evidence that auditors may collect to assess an entity's AI governance: 1) strategic plans for the AI system, 2) organizational charts, 3) risk management and monitoring plans, and 4) documentation of the appeals process.

According to a forum participant, understanding an organization's AI governing principles is key when comparing the intended goals and established practices for the AI system to actual system operations. Auditors may look for evidence of a strategic plan and its implementation in documents describing the entity's policies and procedures, as well as results of audits internal to the entity. A participant noted that strategic plans may include the audited entity's mission and priorities, as well as governing principles for the AI system.

Participants noted that it is important for entities to document their organizational structure throughout the AI life cycle. Organizational charts promote accountability by establishing roles and responsibilities consistent with the entity's governance priorities, according to a participant. Auditors should gather evidence of how the organization develops these roles and responsibilities, according to participants. Organizational charts that include a layout of the organizational governance structure and identify decision-makers and stakeholders are critical pieces of evidence noted by a participant. In addition, auditors should also collect meeting minutes to corroborate that those responsible within the organizational structure are actively meeting and functioning as intended, according to a forum participant.

¹⁵Models and algorithms are examples of components in an AI system. For additional information on assessing performance of components within AI systems, see component level practices under Performance principle 3.3 in the framework.

Testimonial:

“Meeting minutes [should] document perceived risks related to applications and plans to mitigate them. Does the organization have a risk control matrix to identify risks and control measures?”

Source: Participant in CG Forum on AI Oversight. | GAO-21-519SP

According to forum participants, auditors should collect and assess risk management plans associated with AI systems. To assess risks, participants noted, entities should maintain a risk control matrix, which identifies risks and measures to control them. Auditors can assess these matrices to understand the entity’s perceived risks related to the AI application, as well as plans to mitigate them. For example, mitigation methods can include the entity’s strategies to protect its assets and minimize risks with respect to data protection and privacy, according to a participant. Another participant stated that monitoring the AI system and conducting impact assessments as additional risk mitigation methods.¹⁶ In addition, participants noted that entities may reduce risks by testing the training data for potential biases before their use in the AI system. To assess the entity’s practices, auditors could interview developers about these data and review documentation that demonstrates mitigating factors.

Testimonial:

“A framework should consider organizational risks requiring a consortium [or team] of ... stakeholders impacted. Such teams would be composed of technical developers, data scientists, risk management business, and policy areas. All of them should have ownership in it, and bias can be an unintended consequence and is a serious risk.”

Source: Participant in CG Forum on AI Oversight. | GAO-21-519SP

A final means of auditing governance practices that participants noted was collecting documentation related to the appeals process.¹⁷ Participants stated that auditors should ask for documentation, including roles and responsibilities, rubrics and analyses, and updates to the AI system resulting from the appeals process. If a system is already

¹⁶Similarly, a plan by the National Institute of Standards and Technology advocates that entities conduct research to inform standardization of risk management strategies, including monitoring and mitigating risks. See NIST, *U.S. Leadership In AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools*, Prepared in Response to Executive Order 13,859 (Gaithersburg, Md.: Aug. 9, 2019).

¹⁷The appeals process provides users with options to contest decisions made by the AI system as a means to combat potential inequities or biases in the decision or decision making process.

deployed, auditors may also look at documentation of appeals that have been made, how they have been processed, and the outcomes. Forum participants noted that a lack of an appeals process or one that is not timely in resolutions raises concerns about the organizational commitment to accountability with the public. According to another participant, their organization takes 5 years to resolve disputes, while another participant noted their organization has a 7-year backlog for its appeals process.

Testimonial:

"I think a good source of evidence about governance is about the appeals process . . . If a system is already deployed, you also want to look for sort of the physical evidence of what are the appeals looking like? ... What is coming in? What are the complaints being made? How are they being handled? ... Are they being handled in a timely fashion? And what is the result of the analysis of those things? How is the analysis being fed back into the system to improve the system?"

Source: Participant in CG Forum on AI Oversight. | GAO-21-519SP

Assessing Data

Forum participants identified several key sources of evidence that auditors may collect to assess data used for AI systems, including documentation on: 1) data sources; 2) data labels; and 3) data changes. Auditors should assess the sufficiency and appropriateness of the information in the documentation, regardless of whether the information is provided to auditors or they extract it independently. In *Assessing Data Reliability*, GAO notes that attributing data to its source does not alleviate the need for auditors to assess the reliability of the data.¹⁸

Participants noted that auditors should gather information on the sources of data used in developing and deploying the AI system. Participants stated that auditors should collect documentation, such as provenance records, to understand the data throughout the AI life cycle.¹⁹ Panelists suggested a series of questions for auditors regarding AI system data, including:

- how the data were collected and by whom;

¹⁸GAO, *Assessing Data Reliability*, [GAO-20-283G](#) (Washington, D.C.: Dec. 16, 2019).

¹⁹The term "data provenance" refers to a record that accounts for the origin of a piece of data (in a database, document, or repository), together with an explanation of how and why it got to the present place. A provenance record will document the history for each piece of data.

- the original context in which the data were collected;
- the original method of data verification;
- how the data are being used (i.e. for training, testing, or input);
- how the data are being stored and destroyed; and
- how the data have been manipulated.

Forum participants noted that it is important for auditors to understand how data used in an AI system are labeled and classified. Participants stated that there should be documentation of how the labels were chosen and discarded, as well as other decisions made throughout the labeling process. This process is important to document, as discarding data in the labeling process may introduce bias, according to a participant.²⁰

Data used in AI systems are context-dependent, according to participants. This may mean that the data used to train an AI system may no longer be appropriate or relevant, according to participants. For example, one participant stated that if a model is trained on images from the United States and is then deployed in a foreign country, it may produce inaccurate results in the new location. In addition, data may change or be overwritten as it evolves over time, according to participants. A participant said auditors should assess these changes to determine how and why data were updated, and if the updates were effective. Another participant stated that AI systems can be difficult to audit if changes are not documented. As a result, auditors should look for evidence of continuous monitoring and verification of the appropriateness of the data, according to forum participants.

Testimonial:

“Who made the change? Who reviewed the change? What was the rationale for the change that was made? ... A full change log history can go back, you know, a decade. If the auditor doesn’t have access to that, and it’s very difficult to pick apart what’s going wrong, or figure out how to fix it.”

Source: Participant in CG Forum on AI Oversight. | GAO-21-519SP

²⁰In machine learning models, one or more meaningful and informative labels are applied to raw data—such as images, text files, and videos—to provide context so that it can learn to identify this data in the future; then data labeling typically starts by humans making judgments about a given piece of unlabeled data. A forum participant further commented that if data labelers do not agree or form a consensus on how the raw data should be labeled, they may be discarded and not used to train the model—leading to potential bias.

Assessing Performance

Participants identified several key sources of evidence that auditors can use to assess models and algorithms in AI systems, highlighting the importance of documentation throughout the AI life cycle. This evidence may include: 1) documentation of goals, test results, classification system, and use case parameters; 2) change logs and other documentation describing system updates, parameter tuning, and key decisions; and 3) a copy of the model or algorithm code.

Forum participants noted that the attributes of the current model should be documented. Auditors can collect documentation on an entity's goals, test results, classification system, use case parameters, and other factors of the model. Participants noted that this documentation should be updated throughout the life cycle to reflect the most current model.

Participants provided examples of evidence that can be collected to assess current AI models and algorithms. For example, auditors and other stakeholders can collect documents that provide key information on the model. One such example is model cards, which provide an overview of the model (e.g. inputs, outputs, model architecture) and its performance metrics.²¹ Similarly, fact sheets may contain sections on relevant attributes of an AI system—including intended use, performance, safety, and security.²² Auditors may also conduct tests of the model or review internal documentation of tests performed to determine whether the system meets its intended objectives. For example, auditors may also review documentation of or perform input-output tests, which ensure a system performs within the entity's pre-defined acceptable parameters. Moreover, auditors may also review documentation of or perform model stability tests, such as fuzz testing.²³

According to participants, the model development process and any subsequent updates should be documented, including snapshots of each version of the AI system and its components. Participants stated that

²¹For a description of model cards—including examples, refer to following website <https://modelcards.withgoogle.com/about>.

²²M. Arnold, M. R.K.E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, D. Reimer, A. Olteanu, D. Piorkowski, J. Tsay, and K. R. Varshney, "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity," *IBM Journal of Research and Development* vol. 63, no. 4/5 (2019): pp. 6:1-6:13.

²³Fuzz testing is a process of inputting large amounts of random data, called fuzz, to the test the AI system in an attempt to make it crash.

documentation, such as change logs, should include the training and input data, model parameters, assumptions, and any other information needed to reproduce the outcomes of the model. Participants noted that it is important to reproduce the decisions of the model at a given point in time.

Auditors should also look for evidence to assess model drift and the appropriateness of the model. Relationships between model inputs and outputs may change over time, affecting the performance or accuracy of the model's predictions. Auditors can collect evidence of continuous monitoring of the model and its outputs—including documentation of code and model reviews, according to participants.

A participant stated that an AI model or algorithm's code may also be a source of evidence during an audit. Further, that participant noted that a copy of the model or algorithmic code may provide auditors with the level of detail needed to understand the function of the model. However, participants noted that reviewing the code is time-consuming, and requires a level of expertise that auditors may not have. Code review by auditors may not be an efficient use of public resources, as one participant noted that such audits may take up to 5 years to complete.

Session V: Testing AI Systems for Bias and Equity

An AI system may perform differently for certain demographic groups, potentially leading to adverse outcomes for those groups. Assessing AI systems for bias and inequities may include identifying and evaluating: 1) an entity's established processes designed to address and mitigate potential biases and disparate impacts; 2) its system-level processes to identify and mitigate potential biases; and 3) how it collects and addresses public feedback. Forum participants noted the importance of considering potential biases that could be ingested into an AI system, as well as potential inequities resulting from use of an AI system. Participants noted that issues of bias and equity are not clearly defined and may have differing interpretations across different contexts.

Participants said it is important for entities implementing AI systems to establish processes to address and mitigate equity and anti-discrimination concerns they may have during the AI life cycle.²⁴ Auditors may look for evidence of how an entity defines equity or anti-discrimination issues.

²⁴For example, if an entity is concerned about the potential for discrimination or inequitable outcomes based on protected class variables (e.g., race, sex, and age), the organization should be aware of the potential to mitigate these concerns throughout to AI life cycle.

According to one forum participant, this may include documentation of which data attributes the entity uses to control for or assess potential bias. Another participant noted that entities should document why those attributes were chosen. A third forum participant stated that while proxy variables may be used in place of protected class variables—either due to lack of access to those variables or as a means to reduce potential compliance liabilities—these variables may also contribute to bias.²⁵

Testimonial:

“Something an auditor can look for is whether [the entity has] specifically identified the types of equity and discrimination concerns in advance. ... If you’ve identified that you care about gender discrimination and then you deliberately don’t record anything about gender, then it’s pretty clear that you aren’t doing anything to try to test for it.”

Source: Participant in CG Forum on AI Oversight. | GAO-21-519SP

Forum participants noted potential issues that may result from using or collecting protected class information, such as sex, race, and age, among others. They also noted AI systems should incorporate or otherwise use protected class information. One forum participant stated that entities will need to determine whether the model needs adjustment to reduce bias or to address a disparate impact. Without this information, participants stated that entities cannot know how or if the model is performing differently for different demographic groups. According to a participant, some entities are discouraged from collecting protected class data or taking steps to mitigate bias, because doing so may raise risks associated with anti-discrimination liability. Instead, these entities prefer to remain unaware because they consider this the safest way of proceeding. Participants noted that auditors should be aware of this tension and apprehension, and should take into account an entity’s potential legal concerns around the use of or collection of protected class information.

Entities should check for bias throughout the AI life cycle, according to participants. For example, an entity can report on the demographics of those involved in the design and development of the AI system to capture and communicate potential biases inherent to the development process, according to forum participants. For example, a lack of diversity in the

²⁵For example, forum participants discussed the use of zip codes as a proxy variable for race—a protected class variable—as one method used by entities to potentially alleviate or avoid liability concerns related to use of protected class variables.

individuals involved in data labeling processes may also introduce bias in the classification of data used to train the model.

Auditors may assess system-level processes to determine whether the AI system performs consistently across all demographic groups or attributes of interest, according to forum participants. Measuring performance across groups may help identify potential biases and disparate impacts. One participant noted that the system's model should be assessed against the organization's own definition of bias and should reflect the kind of impacts the system might have in deployment. Furthermore, auditors may conduct tests or collect evidence to verify whether the AI system performs consistently across all demographic groups of interest, including documentation on the entity's mitigation strategies, if any, to address potential concerns and biases. Checks for bias should include assessing whether data classification or development processes for individual models or algorithms reflect biased human decisions, including historical inequities, according to participants.

Forum participants noted that auditors should collect information on the extent to which the entity is responsive to feedback from the public, including those directly impacted by the system. According to a participant, auditors should ask whether the public was informed about government use of AI systems. One participant also stated that before a system is deployed, entities should listen to and address community concerns regarding the development and deployment of the AI system—especially from communities that are at risk of being mistreated. This can help identify and mitigate bias in the inputs into or the decisions made by the AI system, according to a forum participant. Further, this feedback can highlight potential inequities or biases in system outcomes before the organization finds the disparity. Once the AI system is deployed, auditors should ask if impacted users were provided options to appeal such decisions as a means to combat potential inequities or biases.

Appendix III: Forum Agenda

	Wednesday, September 9, 2020 (All times Eastern)
10:30—11:00	OPENING REMARKS: WELCOME AND INTRODUCTION The Honorable Gene Dodaro, Comptroller General of the United States Timothy Persons, Chief Scientist and Managing Director, Science, Technology Assessment, and Analytics James-Christian Blockwood, Former Managing Director, Strategic Planning and External Liaison
11:00—12:00	SESSION #1: GOVERNANCE FACTORS TO CONSIDER IN AUDITING AI SYSTEMS Purpose: Presentations made by selected discussants on factors auditors should consider to assess governance including performance of AI systems. Taka Ariga, Chief Data Scientist and Director of Innovation Lab, Science, Technology Assessment, and Analytics Presenters: Tamara Lilly, Francesca Rossi, Ashley Casovan, Julia Wasserman, Karine Perset, Rayid Ghani (7 minutes each) Open Discussion (15 minutes)
12:00—12:10	BREAK
12:10—12:25	LOGISTICS AND EXPERT INTRODUCTIONS Moderator: Steven Putansu, Senior Social Science Analyst, Applied Research and Methods
12:25—1:25	SESSION #2: IDENTIFYING CRITERIA AUDITORS CAN USE IN ASSESSING AI SYSTEMS Purpose: To highlight common factors discussed in Session #1 and identify criteria that can be used in audits and third-party assessments of AI systems (government standards, industry standards, legislation, or control activities). Primary Discussants: Rayid Ghani, John Havens, Tina Kim, Ashley Casovan, Elham Tabassi, Daniel Ho (40 minutes) Open Discussion (20 minutes)

Appendix III: Forum Agenda

2:10—3:10	SESSION #3: ISSUES AND CHALLENGES IN AUDITING AI SYSTEMS IN THE PUBLIC SECTOR Purpose: To identify challenges in auditing AI systems, including transparency, accountability, and expertise needed to understand the use and operations in the public sector. To identify mitigation strategies to address these challenges (40 minutes) Primary Discussants: Julia Wasserman, Francesca Rossi, Anissa Nash, Daniel Kahn Gillmor, Chris Meserole, Gil Alterovitz, Daniel Ho, Bill Scherlis Open Discussion (20 minutes)
3:10—4:00	Wrap up and adjournment Thursday, September 10, 2020 (All times Eastern)
10:30—10:35	SESSION #4: IDENTIFYING SOURCES OF EVIDENCE AND ASSESSMENT METHODS FOR AI SYSTEMS The objective of this session is to identify tools and practices that can explain the methods, use of data, and model design for third party assessments and audits of AI systems and to identify any other considerations which may have relevance beyond those noted below.
10:35—11:35	Identify Audit Procedures for Assessing Governance in AI Systems Purpose: Identify types of evidence auditors need to collect along the AI life cycle to review governance issues—strategic priorities and system requirements, roles and responsibilities, monitoring and review process; identify assessment methods for the information collected. Primary Discussants: Ashley Casovan, John Elder, Francesca Rossi, Daniel Zimmerman, Karine Perset (40 minutes) Open Discussion (20 minutes)
11:35am – 12:35	<u>Identify Audit Procedures for Assessing Data</u> Purpose: Identify types of evidence auditors should collect on the data used in training, validating, and deploying the AI system. Identify methods on how auditors should assess this information. Primary Discussants: Julia Wasserman, Sorelle Friedler, Ayanna Howard, Caryl Brzymiakiewicz (40 minutes) Open Discussion (20 minutes)
12:35—12:50	BREAK

Appendix III: Forum Agenda

12:50—1:50	<u>Identify Audit Procedures for Assessing Models and Algorithms</u> Purpose: Identify types of evidence auditors should collect on models and algorithms. Identify methods on how auditors should assess this information. Primary Discussants: Jingying Yang, Rayid Ghani, Julia Wasserman, Daniel Kahn Gillmor, Bill Scherlis (40 minutes) Open Discussion (20 minutes)
1:50—2:50	BREAK
2:50—3:50	SESSION #5: TESTING AI SYSTEMS FOR BIAS AND EQUITY Purpose: In addition to assessing the extent to which a model meets its intended goals and objectives, explore indirect or unintended effects related to bias, equity, and ethics. The objective is to identify practices and methods auditors can use to test for sources of bias, equity, and other issues in AI systems. Primary Discussants: Alice Xiang, Ayanna Howard, Daniel Kahn Gillmor, Deborah Raji, John Elder (40 minutes) Open Discussion (20 minutes)
3:50—4:20	WRAP UP AND ADJOURNMENT

Appendix IV: List of Forum Participants

Host	Gene L. Dodaro, Comptroller General of the United States
Participants	Gil Alterovitz, Director, National Artificial Intelligence Institute, U.S. Department of Veterans Affairs, Washington, DC
	Caryl Brzymialkiewicz, Deputy Inspector General, Office of Inspector General, U.S. Department of the Interior, Washington, DC (Formerly Chief Data and Analytics Officer, Department of Health and Human Services Office of Inspector General)
	Ashley Casovan, Executive Director, Responsible AI Institute (Formerly AI Global), Montreal, Canada
	John Elder, Founder, Elder Research, Charlottesville, VA
	Sorelle Friedler, Associate Professor, Computer Science, Haverford College, PA
	Rayid Ghani, Professor, Machine Learning Department and Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, PA
	Daniel Kahn Gillmor, Senior Staff Technologist, ACLU Speech, Privacy, and Technology Project, American Civil Liberties Union, New York, NY
	John Havens, Executive Director, IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Institute of Electrical and Electronics Engineers (IEEE), New York, NY
	Daniel Ho, Associate Director, Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, CA
	Ayanna MacCalla Howard, Dean, College of Engineering, The Ohio State University, Columbus, OH (Formerly Chair, School of Interactive Computing, College of Computing, Georgia Institute of Technology)
	Tina Kim, Deputy Comptroller, Office of the New York State Comptroller, Albany, NY
	Tamara Lilly, Assistant Inspector General for Audit Services, Office of Inspector General, U.S. Department of Health and Human Services, Washington, DC
	Chris Meserole, Director of Research and Policy, Artificial Intelligence and Emerging Technology Initiative, Brookings Institution, Washington, DC
	Anissa Nash, Program Director, Audit Cyberspace Operations Directorate, Office of Inspector General, U.S. Department of Defense, Alexandria, VA
	Karine Perset, Administrator, OECD AI Policy Observatory, Organisation for Economic Cooperation and Development, Paris, France
	Deborah Raji, Fellow, Mozilla, Ottawa, Canada (Formerly Research Fellow, AI Now Institute)
	Francesca Rossi, AI Ethics Global Leader, IBM, Yorktown Heights, NY

Appendix IV: List of Forum Participants

	William Scherlis, Director, Information Innovation Office, U.S. Department of Defense, Defense Advanced Research Projects Agency, Arlington, VA
--	---

	Elham Tabassi, Chief of Staff, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD
--	--

	Julia Wasserman, AI Principles Lead, Product Strategy and Operations, Google Cloud AI & Industry Solutions, San Francisco, CA
--	---

	Alice Xiang, Senior Research Scientist, SonyAI, San Francisco, CA (Formerly Head of Fairness, Transparency, and Accountability Research, Partnership on AI)
--	---

	Jingying Yang, Formerly Program Lead, About ML, Partnership on AI, San Francisco, CA
--	--

	Daniel Zimmerman, Research Analyst, Office of the New York State Comptroller, Albany, NY.
--	---

Appendix V: Information on Auditing Standards, Controls, and Procedures

In this section we describe key auditing standards as outlined in the *Government Auditing Standards*, commonly known as the Yellow Book, and the Standards for Internal Control in the Federal Government, commonly known as the Green Book and hereafter referred to as Federal Internal Control Standards. These two documents provide the foundational principles for auditors and audited entities to ensure accountability. The *Government Auditing Standards* provide a framework for conducting high-quality audits with competence, integrity, objectivity, and independence, while the Federal Internal Control Standards set the standards for an effective internal control system for federal agencies. The objective of both is to ensure economy, effectiveness, efficiency, ethics and equity of federal programs.

These standards can help to ensure accountability and transparency in federal programs and processes—including the governance of AI systems. The discussion below provides a high-level overview of (1) auditing standards; (2) internal controls; (3) sources of evidence; and (4) quality of information. These standards provide the foundational principles we used to develop the AI Accountability Framework.

Auditing Standards

Generally Accepted Government Auditing Standards

Audits provide essential accountability and transparency over government programs and processes. Government auditing is essential in providing accountability to legislators, oversight bodies, others charged with governance, and the public. The federal government's auditing standards are presented in the *Government Auditing Standards*. These standards, commonly referred to as generally accepted government auditing standards (GAGAS) provide the foundation for government auditors to lead by example in the areas of independence, transparency, accountability, and quality throughout the audit process.

Source: GAO. | GAO-21-519SP

The *Government Auditing Standards* contain requirements and guidance to assist both internal and external auditors in obtaining objective, sufficient and appropriate evidence and reporting on the results.

The standards describe how managers should ensure internal governance structures and processes are in place and measure performance of federal programs and processes.

Governance. The *Government Auditing Standards* note that those charged with governance need to know whether (1) management and officials that manage government resources are in compliance with laws and regulations; (2) government programs are achieving their objectives and desired outcomes; and (3) government services are provided effectively, efficiently, economically, equitably and ethically.

Within AI systems, “those charged with governance” may refer to the individuals responsible for overseeing the strategic direction of the system and those with obligations related to the accountability of the system. This includes policy makers, subject matter experts, developers, data scientists, or others responsible for implementing the AI system. In some audited entities, multiple parties may be charged with governance, including oversight bodies, members or staff of legislative committees,

boards of directors, audit committees, or parties contracting for the engagement.

Performance. The *Government Auditing Standards* also provide guidance on assessing performance. Performance is assessed through audits that provide objective analysis, findings, and conclusions. It assists management and those charged with governance with, among other things, improving program performance and operations, reducing costs, facilitating decision-making or initiating corrective action, and contributing to public accountability. For example, an audit of program effectiveness can assess the relative ability of alternative approaches to yield better program performance, whether a program produced the intended results, or the extent to which legislative or regulatory objectives are being achieved.

Within AI, audit objectives that focus on performance assess whether an entity deploying AI technologies (1) measures the extent to which an AI system is achieving its goals and objectives; (2) assesses the relative ability of alternative approaches to achieve the same goals and objectives; and (3) establishes internal controls to provide reasonable assurance that the technologies will perform as intended and initiate corrective action as needed. Control audit objectives may identify risks associated with AI deployment (including bias, inequity, and disparate impacts) and steps to mitigate them.

Internal Controls

In this section, we identify internal controls that can be applied to managing and operating AI systems. Specifically, GAO is interested in identifying internal controls that may help an entity to effectively meet its missions and objectives and adapt to shifting environments, evolving demands, changing risks, and new priorities.

According to the Federal Internal Control Standards, internal control is a process effected by an entity's oversight body, management, and other personnel that provides reasonable assurance that the objectives of an entity will be achieved. Internal control is a series of actions that occur throughout an entity's operations and is recognized as an integral part of the operational processes that management uses to guide its operations. Common internal controls for government programs include agency-level directives, policies, and guidance. While there are different ways to present internal control, the Federal Internal Control Standards approach internal control through a hierarchical structure of five components and 17 principles, along with specific documentation requirements for each (see fig. 4).

Components of an Internal
Control System

The five components represent the highest level of the hierarchy of standards for internal control in the federal government. All components are relevant for establishing an effective internal control system. The five components are:

- **Control environment:** The foundation for an internal control system. It provides the discipline and structure to help an entity achieve its objectives.
- **Risk assessment:** An assessment of the risks the entity faces as it seeks to achieve its objectives. This assessment provides the basis for developing appropriate risk responses.
- **Control activities:** The actions management establishes through policies and procedures to achieve objectives and respond to risks in the internal control system, which includes the entity's information system.
- **Information and communications:** The quality information that management and personnel communicate and use to support the internal control system.
- **Monitoring:** Activities management establishes and operates to assess the quality of performance over time and promptly resolve the findings of audits and reviews.

Figure A shows the 17 principles that support the effective design, implementation, and operation of the associated components. These principles represent requirements for establishing an effective internal control system.

Figure A: The Five Components and 17 Principles of Internal Control

Control Environment

- 1** The oversight body and management should demonstrate a commitment to integrity and ethical values.
- 2** The oversight body should oversee the entity's internal control system.
- 3** Management should establish an organizational structure, assign responsibility, and delegate authority to achieve the entity's objectives.
- 4** Management should demonstrate a commitment to recruit, develop, and retain competent individuals.
- 5** Management should evaluate performance and hold individuals accountable for their internal control responsibilities.

Risk Assessment

- 6** Management should define objectives clearly to enable the identification of risks and define risk tolerances.
- 7** Management should identify, analyze, and respond to risks related to achieving the defined objectives.
- 8** Management should consider the potential for fraud when identifying, analyzing, and responding to risks.
- 9** Management should identify, analyze, and respond to significant changes that could impact the internal control system.

Control Activities

- 10** Management should design control activities to achieve objectives and respond to risks.
- 11** Management should design the entity's information system and related control activities to achieve objectives and respond to risks.
- 12** Management should implement control activities through policies.

Information and Communication

- 13** Management should use quality information to achieve the entity's objectives.
- 14** Management should internally communicate the necessary quality information to achieve the entity's objectives.
- 15** Management should externally communicate the necessary quality information to achieve the entity's objectives.

Monitoring

- 16** Management should establish and operate monitoring activities to monitor the internal control system and evaluate the results.
- 17** Management should remediate identified internal control deficiencies on a timely basis.

Source: GAO. | GAO-21-519SP

Identifying Sources of Evidence

In this section, we identify types of evidence auditors can collect to understand or assess the AI system—including the components, use of data, and system design. The *Government Auditing Standards* provide information on the types of evidence and how evidence is assessed to form a reasonable basis for findings and recommendations.

Evidence Collected for GAO Audits

Examples of Sources of Evidence

The following illustrates examples of the different sources of evidence, as identified in GAO-19-164. In that report, GAO reviewed the Federal Emergency Management Agency's (FEMA) Grants Management Modernization program and found, among other things, that the program needed improvements to strengthen program management and cybersecurity. Evidence collected in support of the findings included:

- **Testimonial Evidence:** GAO conducted interviews with FEMA officials to collect information about their efforts to streamline grants management business processes, collect and incorporate stakeholder input, and manage GMM's IT requirements.
- **Documentary Evidence:** GAO reviewed FEMA documentation of program management business processes, the acquisition program baseline, IT requirements documents, and a concept of operations.
- **Physical Evidence:** GAO observed the program's incremental software development activities and a demonstration of the program's automated requirements management tool at FEMA facilities in Washington, D.C.

Source: Adapted from GAO-19-164. | GAO-21-519SP

GAO categorizes the evidence it collects into the following three categories based on its form and how it is collected:

1. **Testimonial evidence** is elicited from respondents to understand their experience, opinions, knowledge, and behavior. It can be obtained through a variety of methods, including inquiries, interviews, focus groups, expert forums, and questionnaires. Testimonial evidence can be gathered from individuals responding personally based on their own experience or in an official capacity to represent agencies or other entities. Testimonial evidence is evaluated for its objectivity, credibility, and reliability.
2. **Documentary evidence** is existing information, such as letters, contracts, accounting records, invoices, spreadsheets, database extracts, electronically stored information, and management information on performance. Documentary evidence may be used to help verify, support, or challenge testimonial evidence.
3. **Physical evidence** is obtained by direct inspection or observation of people, property, or events. The appropriateness of physical evidence depends on when, where, and how the inspection or observation was made and whether it was recorded in a manner that fairly represents the facts observed. Common considerations for physical evidence include the reliability of site selection, intended analytical approaches, and resource considerations.

Auditors must obtain sufficient and appropriate evidence to provide a reasonable basis for addressing the audit objectives and supporting findings and conclusions.

- **Sufficiency** is a measure of the quantity of evidence used to support findings and conclusions.
- **Appropriateness** measures the quality of evidence and encompasses the relevance, validity, and reliability of evidence.

Auditors perform an overall assessment of the collective evidence, to include any limitations, risks, and source of the evidence (e.g. work of others, computer-processed data).

Assessing the Quality of Information

In this section, we define aspects auditors and third-party assessors can use to assess the quality of information. Auditors and third-party assessor can consider the extent to which the entity uses quality information in regards to the AI systems. According to the Federal Internal Control Standards, quality information is appropriate, current, complete, accurate, accessible, and provided on a timely basis. Quality information should be

used to make informed decisions and evaluate the entity's performance in achieving key objectives and addressing risks.

Within AI, audit objectives that focus on the quality of information may include: (1) the identification of information requirements; (2) assessing data to ensure they relevant and from reliable sources; and (3) assessing how the relevant data are processed into quality information within the entity's information system.

Appendix VI: Bibliography

Arnold, M. and R.K.E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, D. Reimer, A. Olteanu, D. Piorkowski, J. Tsay, and K. R. Varshney, “FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity,” *IBM Journal of Research and Development*, vol. 63, no. 4/5 (2019): pp. 6:1-6:13.

Barredo Arrieta, Alejandro, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera, “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI,” *Information Fusion*, vol. 58 (2020), pp. 82-115.

Brookings Institution, *Understanding risk assessment instruments in criminal justice*, (Washington, D.C.: June 19, 2020), accessed Apr. 12, 2021, <https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice/#:~:text=One%20class%20of%20algorithmic%20tools,an%20individual%20before%20their%20trial>.

Congressional Research Service, *Overview of Artificial Intelligence*, IF 10608, ver. 3 (Washington, D.C.: 2017).

Department of Defense, Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* (Washington, D.C.: Oct. 2019).

Department of Defense, *Ethical Principles for Artificial Intelligence* (Washington, D.C.: Feb. 24, 2020), accessed June 2, 2021, https://www.ai.mil/docs/Ethical_Principles_for_Artificial_Intelligence.pdf.

Department of Defense, Office of the Department of Defense Chief Information Officer, *Artificial Intelligence Governance Plan, Version 1.0*, (May 2020).

Department of Defense Office of the Inspector General, *Audit of Governance and Protection of Department of Defense Artificial Intelligence Data and Technology*, DODIG-2020-098. (Alexandria, Va.: June 29, 2020), available at <https://media.defense.gov/2020/Jul/01/2002347967/-1/-1/1/DODIG-2020-098.PDF>.

Diakopoulos, Nicholas, and Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H. V. Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, Christo Wilson, Cong Yu, Bendert Zevenbergen,. *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms* Fairness, Accountability, and Transparency in Machine Learning, accessed July 6, 2020, <https://www.fatml.org/resources/principles-for-accountable-algorithms>.

El Boucheffry, Khadija, and Rafael S. de Souza, *Learning in Big Data: Introduction to Machine Learning*, Science Direct accessed Mar. 29, 2021, www.sciencedirect.com/topics/computer-science/dimensionality-reduction

Engstrom, David Freeman, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies* (Washington, D.C.: Administrative Conference of the United States, Feb. 2020).

European Commission Independent High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, (Brussels, Belgium: Apr. 8, 2019).

European Union Agency for Fundamental Rights, *Data Quality and Artificial Intelligence – Mitigating Bias and Error to Protect Fundamental Rights* (Vienna, Austria: June 11, 2019), <https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect>.

Exec. Order No. 13,960, *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government* (Dec. 3, 2020), 85 Fed. Reg. 78,939, (Dec. 8, 2020).

GAO, *Agile Assessment Guide: Best Practices for Agile Adoption and Implementation*, [GAO-20-590G](#) (Washington, D.C.: Sept. 28, 2020).

GAO, *Artificial Intelligence: Emerging Opportunities, Challenges, and Implications*, [GAO-18-142SP](#) (Washington, D.C.: Mar. 28, 2018).

GAO, *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development*, [GAO-20-215SP](#) (Washington, D.C.: Dec. 20, 2019, reissued Jan. 31, 2020).

GAO, *Artificial Intelligence in Health Care: Benefits and Challenges of Technologies to Augment Patient Care*, [GAO-21-7SP](#) (Washington, D.C.: Nov. 30, 2020).

GAO, *Assessing Data Reliability*, [GAO-20-283G](#) (Washington, D.C.: Dec. 16, 2019).

GAO, *Disaster Resilience Framework: Principles for Analyzing Federal Efforts to Facilitate and Promote Resilience to Natural Disasters*, [GAO-20-100SP](#) (Washington, D.C.: Oct. 23, 2019).

GAO, *Federal Information System Controls Audit Manual (FISCAM)*, [GAO-09-232G](#), (Washington, D.C.: Feb. 2, 2009).

GAO, *Foreign Assistance: Federal Monitoring and Evaluation Guidelines Incorporate Most but Not All Leading Practices*, [GAO-19-466](#) (Washington, D.C.: July 31, 2019).

GAO, *A Framework for Managing Fraud Risks in Federal Programs*, [GAO-15-593SP](#) (Washington D.C.: July 28, 2015).

GAO, *Government Auditing Standards 2018 Revision Technical Update April 2021*, [GAO-21-368G](#) (Washington, D.C.: Apr. 14, 2021).

GAO, High-Risk Series: Urgent Actions Are Needed to Address Cybersecurity Challenges Facing the Nation, [GAO-18-622](#) (Washington, D.C.: Sept. 6, 2018).

GAO, *Standards for Internal Control in the Federal Government*, [GAO-14-704G](#) (Washington, D.C.: Sept. 10, 2014).

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hannah Wallach, Hal Daumé III, and Kate Crawford, “Datasheets for Datasets”, (Mar. 19, 2020), accessed on June 22, 2021, <https://arxiv.org/pdf/1803.09010>.

Government of Canada, Treasury Board, *Directive on Automated Decision Making* (Feb. 5, 2019), accessed Apr. 30, 2020, <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592§ion=html>.

Holland, Sarah, and Ahmed Hosny, Sarah Newman, Joshua Joseph, Kasia Chmielinski. *The Dataset Nutrition Label: A Framework to Drive*

Higher Data Quality Standards, (May 2018).
<https://arxiv.org/pdf/1805.03677>.

Horneman, Angela, Andrew Mellinger, and Ipek Ozkaya, *AI Engineering: 11 Foundational Practices*, Carnegie Mellon University Software Engineering Institute (Sept. 2019).

IBM Research, *AI Fairness 360*, accessed June 2, 2021,
<https://aif360.mybluemix.net/>.

IEEE Standards Association, *Raising the Standards in Artificial Intelligence Systems (AIS)*, accessed June 2, 2021,
<https://standards.ieee.org/initiatives/artificial-intelligence-systems/index.html>.

Info-comm Media Development Authority and Personal Data Protection Commission, *Model Artificial Intelligence Governance Framework, Second Edition* (Singapore: Jan. 21, 2020).

Information Commissioner's Office, *Big data, artificial intelligence, machine learning and data protection*, Version: 2.2 (Wilmslow, UK: Sept. 4, 2017).

Information Commissioner's Office, *Guidance on AI and Data Protection* (Wilmslow, UK: July 30, 2020),
<https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-ai-and-data-protection/>.

Institute of Internal Auditors, Global, *The IIA's Artificial Intelligence Auditing Framework, Practical Applications, Part A Special Edition*. (Lake Mary, Fla.: Dec. 20, 2017).

International Organization for Standardization TR 24028:2020 *Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*, accessed on May 5, 2021,
<https://www.iso.org/standard/77608.html>.

Jobin, Anna, Marcello Lenca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence*, vol. 1 (2019): pp. 389-399.

Kaminski, Margot E., and Gianclaudio Malgieri, "Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations" *International Data Privacy Law*, Dec. 6, 2020.

Kaminski, Margot E., "The Right to Explanation, Explained", 34 *Berkeley Tech. L.J.* 189 (2019), available at <https://scholar.law.colorado.edu/articles/1227>.

Lee, Nicol Turner, Paul Resnick, and Genie Barton, *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms* (Washington, D.C.: Brookings Institution, May 22, 2019), accessed Apr. 15, 2021, <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>.

Livingston, Morgan "Preventing Racial Bias in Federal AI" *Journal of Science Policy & Governance*, Vol. 16, Issue 2, (May 27, 2020).

Lum, Kristian and William Isaac, "To Predict and Serve?" *Significance Magazine*, vol. 13 (Oct. 7, 2016).

Microsoft, *What is Data Drift?*, accessed May 25, 2021, <https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?tabs=python>.

Ministry of Economy, Trade and Industry (METI) of Japan, *Contract Guidelines on Utilization of AI and Data*, (June 2018) accessed Mar. 19, 2021, https://www.meti.go.jp/english/press/2019/0404_001.html. "Trans. provided by METI."

Mitchell, Margaret, and Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru "Model Cards for Model Reporting," *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, (Association for Computing Machinery, New York, N.Y.: Jan. 2019), pp. 220–229.

Mulligan, Deirdre K. and Bamberger, Kenneth A. "Procurement As Policy: Administrative Process for Machine Learning" *Berkeley Technology Law Journal*, Vol. 34 (Dec. 8, 2019).

National Institute of Standards and Technology. *FAQ about NIST's Role in Planning Federal Engagement in AI Standards Development*, accessed May 5, 2021, <https://www.nist.gov/artificial-intelligence/faqs-about-nists-role-planning-federal-engagement-ai-standards-development>.

National Institute of Standards and Technology. *The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management*. (Gaithersburg, Md.: Jan. 2020), accessed on May 5, 2021, <https://www.nist.gov/privacy-framework>.

National Institute of Standards and Technology, *NIST Risk Management Framework* (Gaithersburg, Md.: May 28, 2021), accessed June 2, 2021, <https://csrc.nist.gov/projects/risk-management/about-rmf>.

National Institute of Standards and Technology, *Robustness*. (Gaithersburg, Md.: January 2020). Accessed May 21, 2021, <https://csrc.nist.gov/glossary/term/robustness>.

National Institute of Standards in Technology, *U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools*, Prepared in Response to Executive Order 13,859 (Gaithersburg, Md.: Aug. 9, 2019).

National Security Commission on Artificial Intelligence, *Final Report*. (Arlington, Va.: Mar. 1, 2021).

National Security Council, Exec. Office of the President, *Interim National Security Strategic Guidance*, (Mar. 2021).

Office of the Director of National Intelligence, *Artificial Intelligence Ethics Framework for the Intelligence Community version 1.0* (Washington, D.C.: June 2020), accessed June 22, 2021, <https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community>.

Office of the Director of National Intelligence, *Principles of Artificial Intelligence Ethics for the Intelligence Community*, (Washington, D.C.: July 23, 2020). June 22, 2021, <https://www.intelligence.gov/principles-of-artificial-intelligence-ethics-for-the-intelligence-community>.

Office of Management and Budget, OMB Memorandum M-21-06, *Guidance for Regulation of Artificial Intelligence Application* (Washington, D.C.: Nov. 17, 2020).

Office of Science and Technology Policy, Exec. Office of the President, *American Artificial Intelligence Initiative: Year One Annual Report* (Feb. 2020).

Olteanu, Alexandra, and Carlos Castillo, Fernando Diaz, Emre Kiciman, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *Frontiers in Big Data*, vol.2, no. 13: <https://doi.org/10.3389/fdata.2019.00013>.

Organisation for Economic Co-operation and Development, *Artificial Intelligence in Society* (OECD Publishing: Paris, revised Aug. 2019), accessed June 22, 2021, <https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm>.

Organisation for Economic Co-operation and Development, *Recommendation of the Council on Artificial Intelligence* OECD/LEGAL/0449 (Paris, France: adopted May 22, 2019), <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

Organisation for Economic Co-operation and Development, Working Papers on Public Governance No. 31. *State of the Art in the Use of Emerging Technologies in the Public Sector*, accessed Apr. 09, 2021, https://www.oecd-ilibrary.org/governance/state-of-the-art-in-the-use-of-emerging-technologies-in-the-public-sector_932780bc-en.

Partnership on AI. *Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles *ABOUT ML*) version 0 (San Francisco, Calif.: 2019), accessed Nov. 2020, <https://www.partnershiponai.org/wp-content/uploads/2019/07/ABOUT-ML-v0-Draft-Final.pdf>.

Preece, Alun, and Dan Harborne, Dave Braines, Richard Tomsett, Supriyo Chakraborty "Stakeholders in Explainable AI" (Sept. 29, 2018), accessed on June 22, 2021, <https://arxiv.org/pdf/1810.00184>.

Raji, Inioluwa Deborah, and Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, Parker Barnes, "Closing the AI Accountability Gap: Defining an

End-to-End Framework for Internal Algorithmic Auditing” *Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain, Jan. 28, 2020).

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. AI Now Institute New York University, Apr. 2018.

Sammut, Claude, and Geoffrey I. Webb, *Encyclopedia of Machine Learning*, 2010 ed. (Boston, Mass.: Springer).

Schelter, Sebastian, and Yuxuan He, Jatin Khilnani, Julia Stoyanovich FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions, *23rd International Conference on Extending Database Technology (EDBT)* (Mar. 30, 2020).

Selbst, Andrew D., Danah Boyd, Sorelle Friedler, Suresh Venkatasubramanian, and Janet Vertesi, “Fairness and Abstraction in Sociotechnical Systems”, *FAT*19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, New York, N.Y.: Jan, 2019), pp. 59–68.

Selbst, Andrew "Disparate Impact in Big Data Policing." *Georgia Law Review* Vol. 52, Issue 1, (2017).

Select Committee on Artificial Intelligence of the National Science and Technology Council, *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update* (Washington, D.C.: June 2019).

Shendre, Sushrut, *Model Drift in Machine Learning*, accessed May 21, 2021, <https://towardsdatascience.com/model-drift-in-machine-learning-models-8f7e7413b563>.

Smith, Genevieve and Ishita Rustagi, *Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook* (Berkeley Haas Center for Equity, Gender and Leadership, Calif.: July, 2020).

Stoyanovich, Julia and Howe, Bill, *Follow the Data! Algorithmic Transparency Starts with Data Transparency*, Shorenstein Center on Media, Politics, and Public Policy, (Nov. 18, 2018).

Stoyanovich, Julia and Howe, Bill, "Nutritional Labels for Data and Models", *IEEE Computer Society Technical Committee on Data Engineering*, accessed July 1, 2020, <http://sites.computer.org/debull/A19sept/p13.pdf>.

The Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK, *Auditing machine learning algorithms* (Oct. 14, 2020).

Winfield, Alan, "Ethical standards in robotics and AI," *Nature Electronics*, vol. 2 (2019): pp. 46–48.

World Economic Forum, *Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations*, (Cologny/Geneva, Switzerland: Jan. 2020)

Appendix VII: GAO Contacts and Staff Acknowledgments

GAO Contacts

Taka Ariga, (202) 512-6888 or ArigaT@gao.gov
Timothy M. Persons, (202) 512-6888 or PersonsT@gao.gov
Stephen Sanford, (202) 512-4707 or SanfordS@gao.gov

Staff Acknowledgments:

In addition to the contact named above, Farahnaaz Khakoo-Mausel (Assistant Director), Dennis Mayo (Analyst-in-Charge from November 10, 2020), Sean Manzano (Analyst-in-Charge until November 10, 2020), Evonne Tang, Kristen Pinnock, Ben Frey, Samuel Huang, Steven Putansu, Jehan Chase, Ben Shouse and Anika McMillon made key contributions to this publication. Also contributing to this publication were Elise Beisecker, Andrew Kurtzman, Martin Skorczynski, Nicole Jarvis, Colleen Candrl, Brian Mazanec, Suzanne Kaasa, Jim Dalkin, Michael Bingham, Jennifer Beddor, Kerry Burgott, Kendall Childers, Jon Menaster, Rebecca Parkhurst, Bethann E. Ritter Snyder, Robert Rivas, and Monica Perez-Nelson.

Image Sources

This section contains credit and copyright information for images and graphics in this product, as appropriate, when that information was not listed adjacent to the image or graphic.

Front cover: theromb/stock.adobe.com (Multicolored 3D cubes), GAO (icon illustrations, background).

Section dividers: theromb/stock.adobe.com (3D cubes), GAO (background).

Section headers: GAO.

GAO's Mission

The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through our website. Each weekday afternoon, GAO posts on its [website](#) newly released reports, testimony, and correspondence. You can also [subscribe](#) to GAO's email updates to receive notification of newly posted products.

Order by Phone

The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's website, <https://www.gao.gov/ordering.htm>.

Place orders by calling (202) 512-6000, toll free (866) 801-7077, or TDD (202) 512-2537.

Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.

Connect with GAO

Connect with GAO on [Facebook](#), [Flickr](#), [Twitter](#), and [YouTube](#).
Subscribe to our [RSS Feeds](#) or [Email Updates](#). Listen to our [Podcasts](#).
Visit GAO on the web at <https://www.gao.gov>.

To Report Fraud, Waste, and Abuse in Federal Programs

Contact FraudNet:

Website: <https://www.gao.gov/about/what-gao-does/fraudnet>

Automated answering system: (800) 424-5454 or (202) 512-7700

Congressional Relations

Orice Williams Brown, Managing Director, WilliamsO@gao.gov, (202) 512-4400,
U.S. Government Accountability Office, 441 G Street NW, Room 7125,
Washington, DC 20548

Public Affairs

Chuck Young, Managing Director, youngc1@gao.gov, (202) 512-4800
U.S. Government Accountability Office, 441 G Street NW, Room 7149
Washington, DC 20548

Strategic Planning and External Liaison

Stephen Sanford, Managing Director, spel@gao.gov, (202) 512-4707
U.S. Government Accountability Office, 441 G Street NW, Room 7814,
Washington, DC 20548



Please Print on Recycled Paper.