# Trustworthy AI (TAI) Playbook: Executive Summary

U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES

SEPTEMBER 2021

# Table of Contents



*Please note that there are links in this document that direct to external sources. These sources are outside of OCAIO's direct control and may not be 508 compliant.*

CHAPTER I

# INTRODUCTION

# Message from HHS Chief AI Officer Oki Mek

> HHS has a significant role to play in strengthening American leadership in Artificial Intelligence (AI). As we use AI to advance the health and wellbeing of the American people, **we must maintain public trust by ensuring that our solutions are ethical, effective, and secure**. The HHS Trustworthy AI (TAI) Playbook is an initial step by the Office of the Chief AI Officer (OCAIO) to support trustworthy AI development across the Department.

# Background | Why is Trustworthy AI (TAI) Important?

Increased AI adoption unlocks new value for agencies, but it also introduces new risks. To achieve the full benefits of AI across the HHS ecosystem, we must mitigate those risks by embedding principles that foster trust in each stage of AI development.

**Trustworthy AI** refers to the design, development, acquisition, and use of AI in a manner that **fosters public trust and confidence** while protecting privacy, civil rights, civil liberties, and American values, consistent with applicable laws [1]

***Trustworthy practices can help agencies achieve mission success with AI by protecting against four key risks…***

## Strategy and Reputation

Loss of public trust and loyalty due to lack of transparency, equitable decision-making, and accountability

*Example: If an AI model uses health care expenses as a proxy for health care needs, it may perpetuate biases that affect Black patients' access to care since Black patients tend to spend less than White patients for the same level of need. In turn, Black patients may lose trust in the health care community.* [2, 3]

## Cyber and Privacy

Security and privacy breaches due to inadequate data protection and improper use of sensitive data

*Example: If an AI model that uses protected health information (PHI) to inform public health interventions is not properly secured, it may be compromised by adversarial attacks. This can cause emotional and financial harm to affected individuals.*

## Legal and Regulatory

Unfair practices, compliance violations, or legal action due to biased data or a lack of explainability

*Example: If an AI-based benefits distribution system discriminates against a protected class due to biased data, the agency may face legal ramifications.*

## Operations

Operational inefficiencies due to disruption in AI systems or inaccurate or inconsistent results

*Example: If a call center bot that answers grantee inquiries about compliance requirements provides inconsistent responses, it may cause confusion among grantees and additional work for agency officials managing compliance.*

# Background | Executive Order 13960 [1]

EO 13960, "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government," outlines two requirements for agencies.

| | 1 Adhere to Principles for Use of AI in Government | 2 Create an Agency Inventory of AI Use Cases |
|---|---|---|
| *Overview* | The EO outlines **nine principles that agencies must follow** when designing, developing, acquiring, and using AI in the federal government. | The EO requires agencies to prepare an inventory of non-classified and non-sensitive **current and planned AI use cases** and update it annually thereafter. Agencies must share their inventories with the public and other agencies, to the extent practicable. |
| *HHS Response* | OCAIO created the **Trustworthy AI (TAI) Playbook** to help Divisions meet this requirement. The Playbook consolidates the EO principles into six TAI principles and reflects the latest Department perspective on TAI adoption. | OCAIO is building upon existing datasets (e.g., PMA data call) to create an **HHS AI Use Case Inventory** that not only satisfies the EO requirements but also increases awareness of and cross-agency collaboration on AI initiatives. |
| *What This Means For You* | Op/StaffDivs are encouraged to…<br>• **Assess existing AI solutions** to ensure they adhere to the principles described in the Playbook<br>• **Carefully review the Playbook** before implementing new AI solutions | Op/StaffDivs are encouraged to…<br>• **Provide a list of applicable AI use cases** in accordance with forthcoming OCAIO guidance<br>• **Use the inventory to connect with colleagues and share knowledge** about AI applications, technologies, processes, and best practices |

# HHS Trustworthy AI (TAI) Playbook Overview

The TAI Playbook is designed to support leaders across the Department in applying TAI principles. It outlines the core components of TAI and helps identify actions to take for different types of AI solutions.

## PLAYBOOK OBJECTIVES

1. **Promote understanding** of the TAI principles outlined in EO 13960
2. **Provide guidance and frameworks** for applying TAI principles throughout the AI lifecycle
3. **Centralize relevant federal and non-federal resources** on TAI
4. **Serve as a framework for future HHS policies** on TAI acquisition, development, and use

The Playbook is not…

- ⊗ A formal policy or standard
- ⊗ An exhaustive guide to building and deploying AI solutions

## INTENDED AUDIENCE

The TAI Playbook is intended for Op/StaffDiv Leadership Teams, including:

### Agency Leadership

*Should use the Playbook to…*

- **Create Op/StaffDiv-specific policies** related to TAI
- **Evaluate TAI risks** associated with new AI investments

### Program/Project Managers

*Should use the Playbook to…*

- **Incorporate TAI principles into the business requirements** for an AI solution
- **Provide guidance to their teams *before* building an AI solution** about what actions to take
- **Oversee AI projects throughout the lifecycle** to ensure solutions adhere to all six TAI principles
- **Identify and mitigate TAI risks** for an AI solution

*While on-the-ground AI users will also need to understand TAI principles, the main audience for this Playbook is Agency Leadership and Program/Project Managers.*

# How to Use The Executive Summary

Leaders should reference Chapters 2-3 to gain baseline fluency in TAI and Chapters 4-5 to understand how to apply TAI principles.

**Chapter 2**

*AI Building Blocks*

**Chapter 3**

*Principles for Use of Trustworthy AI in Government*

**High-Level Information about TAI**

Chapters 2-3 provide an overview of the building blocks of AI solutions and the principles that underpin TAI. The full version of the Playbook* provides additional information about the principles, including examples of how they work in practice.

**Chapter 4**

*Internal AI Deployment Considerations*

**Chapter 5**

*External AI Considerations*

**Guidance for Leadership Teams**

Chapters 4-5 include recommendations for designing TAI solutions and fostering TAI innovation. The full version of the Playbook* provides more detailed guidance and supplementary resources (e.g., sample use cases, risk review checklists) for each phase of the AI lifecycle in Chapter 4.

*The full version of the HHS TAI Playbook can be found on the HHS OCAIO Intranet site.*

CHAPTER II

# AI BUILDING BLOCKS

# AI Definition

To understand whether TAI principles need to be applied to a technology solution, let's first discuss what defines AI.

## To help determine if a use case constitutes AI*, consider whether the solution or system… [1, 4]

A. *…**performs tasks under varying and unpredictable circumstances without significant human oversight**, or **can learn from experience and improve performance** when exposed to data sets?*

B. *…uses computer software, physical hardware, or other technology to **solve tasks that require human-like perception**, thinking, planning, learning, communication, or physical action?*

C. *…thinks or acts like a human, including the use of **cognitive architecture or neural networks** (e.g., developed to mimic the underlying mechanisms of the human mind)?*

D. *…relies on a **set of techniques, including machine learning, to approximate a cognitive task**?*

E. *…is designed to act rationally by utilizing **intelligent software or an embodied robot to achieve goals** using perception, planning, reasoning, learning, communicating, decision-making, and acting?*

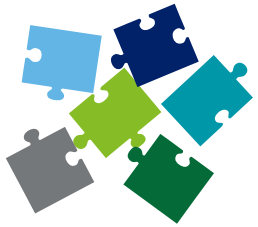
*These considerations, while not all-encompassing, serve as a guide in determining whether a solution constitutes AI and whether TAI principles need to be applied*

**Based on the National Defense Authorization Act for Fiscal Year 2019, Section 238 (g), as utilized in Executive Order (EO) 13960.*

# AI Building Blocks & TAI | Overview

To assure an AI solution is perceived as an enhancement rather than met with mistrust, protocols are needed to ensure trustworthiness across AI methods, the collective AI solution, and how that AI solution is applied for a specific HHS use case.

## AI Methods

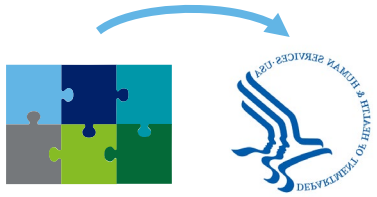


***AI Methods*** are the different types of AI techniques that can be used to perform activities that normally require human intelligence (e.g., Natural Language Processing)

## AI Solutions

***AI Solutions*** are made up of one or more *AI methods* and, after assessing problems and needs, are developed to carry out a specific function, purpose, or role (e.g., Call Center Bot)

## AI Use Cases



***AI Use Cases*** involve how *AI solutions* are used to meet specific HHS mission objectives (e.g., Call Center Bot used to respond to claims benefits inquiries)

### *Trustworthy AI Implications*

Understanding the AI methods used in an AI solution is necessary to determine the TAI techniques to apply in design, development, and testing

Understanding how the comprehensive AI solution functions and the degree of human involvement helps ensure the right level of focus and scrutiny on specific TAI principles

It is important to consider not only whether the solution itself meets TAI guidelines but also whether it is used in a way that upholds HHS' TAI principles

# AI Building Blocks & TAI | AI Methods

AI solutions are built upon one or more AI methods. Recognizing the AI methods that a solution uses is important, as they each have different TAI implications that need to be addressed.

SAMPLE AI METHODS

| | DEFINITION | SAMPLE TAI IMPLICATIONS |
|---|---|---|
| **Machine Learning (ML)** | *"A subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed" – MIT*[5] *Includes probabilistic methods*[5] *and can support predictive analytics*[6] | Machine learning should be bias-free and incorporate relevant shifts in healthcare demographics |
| **Natural Language Processing (NLP)** | *"Machines learn to understand natural language as spoken and written by humans" – MIT*[7] *and includes both Natural Language Generation (NLG) and Natural Language Understanding (NLU) – IBM*[8] | NLP models should be understandable to users to prevent incorrect interpretations that could negatively impact affected individuals |
| **Speech Recognition** | *"Systems [that] interpret human speech and translate it into text or commands." – Gartner*[9] | Voice and speech should be inclusive of a broad range of languages, dialects, and accents |
| **Computer Vision** | *"Intelligent algorithms that perform important visual perception tasks such as object recognition, scene categorization, integrative scene understanding, human motion recognition, material recognition, etc." – Stanford*[10] | Computer vision models should be trained with data representative of the patient populations that will use them to support unbiased results |
| **Intelligent Automation** | *"The use of automation technologies – artificial intelligence (AI), business process management (BPM), and robotic process automation (RPA) – to streamline and scale decision-making across organizations– IBM*[11] | Intelligent automation solutions should have a human sponsor that is responsible for ensuring protected information (e.g., patient data) is not accessible |

CHAPTER III

# PRINCIPLES FOR USE OF TRUSTWORTHY AI IN GOVERNMENT

# Overview of TAI Principles [12]

By applying these six TAI principles across all phases of an AI project, OpDivs and StaffDivs can promote ethical AI and achieve the full operational and strategic benefits of AI solutions.

**Fair / Impartial**

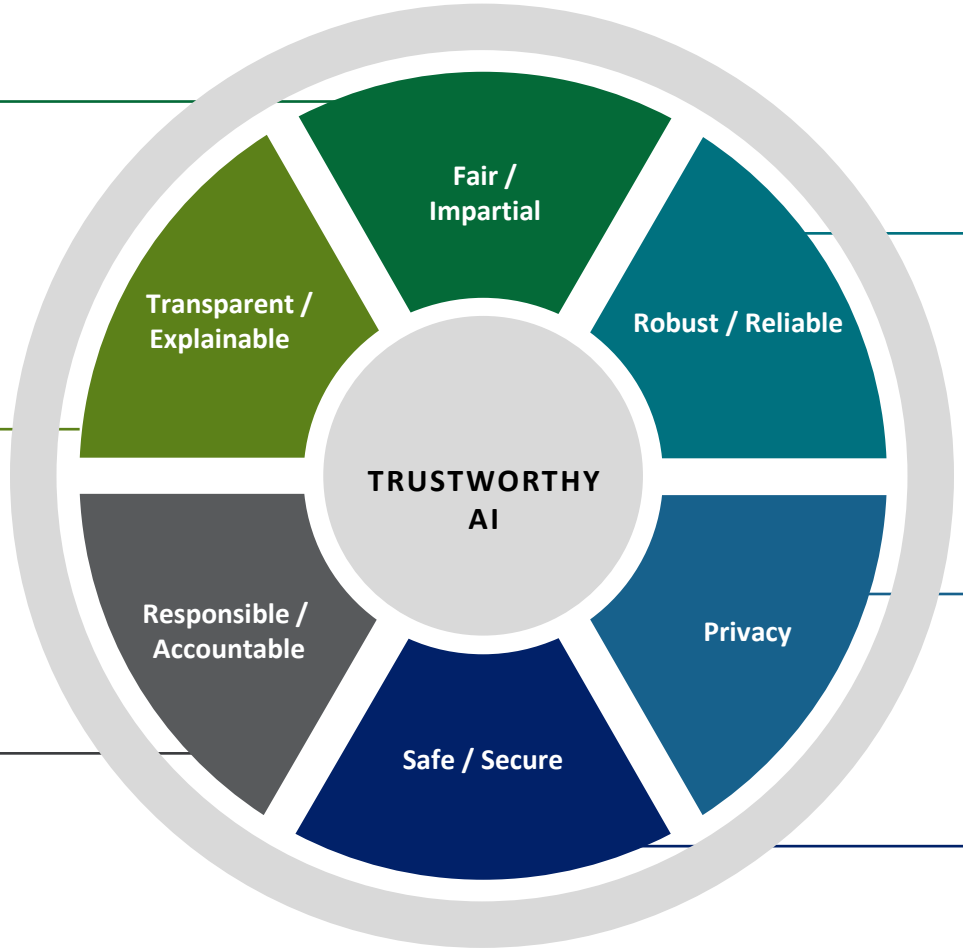
AI applications should include checks from internal and external stakeholders to help ensure equitable application across all participants

**Transparent / Explainable**

All relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations should be open to inspection

**Responsible / Accountable**

Policies should outline governance and who is held responsible for all aspects of the AI solution (e.g., initiation, development, outputs, decommissioning)




**Robust / Reliable**

AI systems should have the ability to learn from humans and other systems and produce accurate and reliable outputs consistent with the original design

**Privacy**

Individual, group, or entity privacy should be respected, and their data should not be used beyond its intended and stated use; data used has been approved by the data owner or steward

**Safe / Secure**

AI systems should be protected from risks (including Cyber) that may directly or indirectly cause physical and/or digital harm to any individual, group, or entity

*TAI principles are not mutually exclusive, and tradeoffs often exist when applying them.*

# Alignment to Federal Guidelines

The six TAI principles map to the principles outlined in Executive Order 13960, "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government," and OMB Memorandum M-21-06, "Guidance for Regulation of Artificial Intelligence Applications."

| TAI Playbook Principles | EO 13960 Principles [1] | OMB M-21-06 Principles [13] |
|---|---|---|
| **Fair / Impartial** | 1. Lawful and Respectful of Our Nation's Values | 7. Fairness and Nondiscrimination |
| **Transparent / Explainable** | 5. Understandable<br>8. Transparent | 2. Public Participation<br>8. Disclosure and Transparency |
| **Responsible / Accountable** | 6. Responsible and Traceable<br>7. Regularly Monitored<br>9. Accountable | 5. Benefits and Costs |
| **Safe / Secure** | 4. Safe, Secure, and Resilient | 4. Risk Assessment and Management<br>9. Safety and Security |
| **Privacy** | 4. Safe, Secure, and Resilient | 9. Safety and Security |
| **Robust / Reliable** | 2. Purposeful and Performance-Driven<br>3. Accurate, Reliable, and Effective | 3. Scientific Integrity and Information Quality |

**Additional Cross-Cutting Principles:**
1. Public Trust
6. Flexibility
10. Interagency Coordination

CHAPTER IV

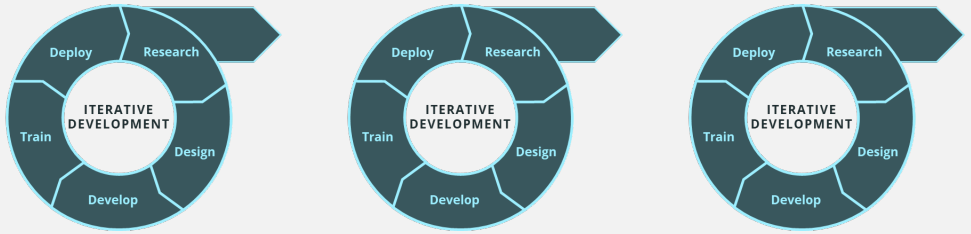
# INTERNAL AI DEPLOYMENT CONSIDERATIONS

*Business Case Development (Pre-Approval/Funding)*
*AI Solution Deployment (Post-Approval/Funding)*

# AI Lifecycle

There are four phases of a typical AI lifecycle that align to the HHS Enterprise Performance Lifecycle framework.[22] This chapter focuses on the deployment of AI solutions, which begins after an AI concept has been approved and funded.

| | INITIATION & CONCEPT | RESEARCH & DESIGN* | DEVELOP, TRAIN, & DEPLOY* | OPERATE & MAINTAIN |
|---|---|---|---|---|
| **Description of Lifecycle Phase** | Evaluate business need, explore AI opportunities, and assess feasibility and trustworthiness of proposed solution in collaboration with relevant stakeholders; decide whether to move forward | Complete project planning and acquisition activities; detail requirements in collaboration with relevant stakeholders; select algorithms based on potential risks and value considerations | Prepare data; build, train, and test to confirm business outcomes; evaluate performance and risk management prior to full-scale production; deploy and verify | Continuously improve solution by collecting stakeholder feedback, monitoring performance, and re-training on an ongoing basis; measure and track return on investment; refresh or retire as needed |
| **EPLC Lifecycle Alignment** | • Initiation<br>• Concept | • Planning<br>• Requirements Analysis<br>• Design | • Development<br>• Test<br>• Implementation | • Operations and Maintenance<br>• Disposition |




***Project is approved and funded***

**Conducted via a series of iterative, agile sprints*

*Leaders must apply the principles **during all stages of the lifecycle** to create TAI solutions.*

# Initiation and Concept | TAI Considerations

| TAI PRINCIPLE | CONSIDERATIONS |
|---|---|
| **Fair / Impartial** | ❑ **Consider how you will translate the business problem into questions that AI algorithms can answer:** What are the potential target variables? Are they correlated with protected or sensitive characteristics? [3, 14]<br>❑ **Determine how you will use the solution's outputs:** Will you use the outputs to make resource allocation decisions that could have a disparate impact on affected subgroups?<br>❑ **Survey the legal and regulatory landscape:** What regulations, standards, policies, or laws related to bias and discrimination apply to the proposed solution? |
| **Transparent / Explainable** | ❑ **Evaluate stakeholder needs:** Who will use, be affected by, or have an interest in the solution's outputs? What might they want to know about the solution's inputs, outputs, or decision-making process? [18]<br>❑ **Consider the explainability-accuracy tradeoff:** Will the proposed solution use deep learning, support vector machines, or other AI methods that could increase accuracy but decrease explainability? |
| **Responsible / Accountable** | ❑ **Complete the IT Acquisition Review (ITAR) Process:** [23] Do you plan to acquire IT products or services that meet the minimum criteria for the ITAR process? If so, have you submitted a request in accordance with HHS Policy?<br>❑ **Use the Digital Worker Impact Evaluation Matrix to forecast the solution's potential adverse impact level:** [19] What type of access will the solution have? Will it be able to act on its own insights? |
| **Safe / Secure** | ❑ **Identify security risks:** [24, 25] How might adversarial agents target and seek to compromise the AI solution? |
| **Privacy** | ❑ **Conduct a preliminary Privacy Impact Assessment (PIA):** [26] Will the proposed solution use sensitive data? If so, how will you collect, share, and use that information? |
| **Robust / Reliable** | ❑ **Consider the likelihood of error:** Will the proposed solution require joining and pre-processing data from multiple sources? Have potential target variables been well-measured in the past? [2, 14]<br>❑ **Evaluate the proposed team composition:** Will there be more than one data scientist to peer review code? Will they have prior experience with the AI method(s) selected? Will the team include diverse perspectives and expertise? |

# Research and Design | TAI Considerations

| TAI PRINCIPLE | CONSIDERATIONS |
|---|---|
| **Fair / Impartial** | ❑ **Comply with relevant laws and regulations:** Does the model design comply with identified laws and regulations?<br>❑ **Evaluate the potential impact to protected and non-protected classes:** Are components of training/testing datasets related to a protected or sensitive characteristic? Is there a mismatch between the ideal and actual target variable? [3, 28, 29]<br>❑ **Conduct data bias review, explore tools, and determine metrics:** Have you or the vendor reviewed the data, considered tools to support bias detection and mitigation, and identified metrics to measure fairness? [3, 15, 30, 31, 32]<br>❑ **Bring in diverse perspectives:** Have you discussed model design with stakeholders to uncover unintended bias? [15, 25, 29, 33] |
| **Transparent / Explainable** | ❑ **Update the Op/StaffDiv Use Case Inventory:** Have you documented the use case according to forthcoming guidance?<br>❑ **Establish explainability requirements and create a feedback mechanism:** Have you engaged a diverse set of stakeholders to understand their needs and design a mechanism by which they can provide feedback? [15, 17, 18]<br>❑ **Document model inputs in design documentation:** Have you or the vendor described input source/parameters? [2]<br>❑ **Evaluate model explainability:** Have you considered the model's explainability and the tradeoff with model accuracy? [2] |
| **Responsible / Accountable** | ❑ **Create a digital identity, if needed:** Have you referenced the ICAM Program Management Guide for the latest policies in AI identities and created a digital identify based on the solution's potential adverse impact level? [19]<br>❑ **Assign key roles**: Have you determined what level of human supervision is required for the solution? [19]<br>❑ **Document the solution's decision-making process**: Have you captured data transformation steps? [2] |
| **Safe / Secure** | ❑ **Evaluate identified security risks and create a data protection and secure integration plan:** Have you evaluated the likelihood and potential impact of security risks and determined necessary controls? Have you obtained approval? [20, 24, 34] |
| **Privacy** | ❑ **Refine the PIA:** Does the PIA accurately describe what information will be collected and how it will be protected? [26] |
| **Robust / Reliable** | ❑ **Assess data quality:** Have you or the vendor evaluated the quality and reliability of the training and testing data? [2, 25, 36]<br>❑ **Select AI methods:** Have your or the vendor created a conceptually sound design that supports desired outcomes? [36] |

# Develop, Train, and Deploy | TAI Considerations

| TAI PRINCIPLE | CONSIDERATIONS |
|---|---|
| Fair / Impartial | ❑ **Comply with laws and regulations:** Is the solution in compliance with identified laws and regulations?<br>❑ **Identify and mitigate unintended bias:** Have you applied fairness metrics and incorporated bias into the Operational Readiness Review? Have you applied bias mitigation techniques and discussed tradeoffs with stakeholders? [14, 15, 27, 30] |
| Transparent / Explainable | ❑ **Use measures and tools to increase explainability:** Did you use interpretation methods to understand the model?<br>❑ **Document model outputs, performance test plan, and test results:** Have you documented how the model is applied to a scenario to obtain results? Have you created a performance test plan and documented testing evidence?<br>❑ **Assess model outputs and explanations:** Do model outputs and explanations satisfy explainability requirements? Have you considered independent verification and validation (IV&V) testing to validate that the outputs are understandable? [17, 37] |
| Responsible / Accountable | ❑ **Provision digital identify (if applicable):** Have you captured identity management information? [19]<br>❑ **Maintain a change access plan during development:** Are there mechanisms in place to track developer actions?<br>❑ **Consider IV&V testing:** Have testers verified model performance, and have necessary parties approved test outcomes? [37] |
| Safe / Secure | ❑ **Employ secure practices for AI configuration/setup:** Have you or the vendor used strong access and data controls? [20]<br>❑ **Develop defenses against adversarial attacks and scan for vulnerabilities:** Have you or the vendor implemented necessary protections? Have you identified and mitigated vulnerabilities in all levels of the solution's stack? [20]<br>❑ **Obtain an Authority to Operate (ATO):** Prior to implementation, did you obtain appropriate clearance? [22] |
| Privacy | ❑ **Finalize the PIA:** Have you obtained approval and submitted the PIA? [26]<br>❑ **Implement and test privacy protections:** Did you encrypt sensitive data, test data controls, and document risks? [38]<br>❑ **Publish a System of Record Notice (SORN):** Will the AI solution create a System of Record? Did you publish a SORN? [22] |
| Robust / Reliable | ❑ **Clean the training data:** Have you or the vendor prepared the training data using applicable tools or procedures? [25, 39]<br>❑ **Create data quality controls:** Have you developed controls to monitor the existence of data drift? [19, 40, 41]<br>❑ **Perform model verification and validation testing:** Have you executed the test plan and addressed any issues? [36, 39]<br>❑ **Establish reliability metrics:** Have you selected and determined thresholds for model performance metrics? [2, 25] |

# Operate and Maintain | TAI Considerations

| TAI PRINCIPLE | CONSIDERATIONS |
|---|---|
| **Fair / Impartial** | ❑ **Continue to comply with laws and regulations:** Are the identified laws and regulations still applicable? Are there any new or upcoming laws or regulations that may impact the solution?<br>❑ **Identify and mitigate unintended bias:** Do you routinely evaluate bias using defined fairness metrics, engage a diverse set of stakeholders to support bias detection, and apply bias mitigation techniques as needed? |
| **Transparent / Explainable** | ❑ **Establish a change management process:** Do you review, test, and document changes to the solution? [36]<br>❑ **Maintain Op/StaffDiv Use Case Inventory:** Do you update the inventory when the solution details change?<br>❑ **Publish and regularly review model performance information:** Do you communicate model performance metrics to stakeholders and collect and address feedback? [15, 40] |
| **Responsible / Accountable** | ❑ **Establish an incident management process:** Have you assigned responsibilities for responding to incidents?<br>❑ **Recertify key roles:** Do you recertify key roles on an annual or semiannual basis? [19]<br>❑ **Collect third party documentation (if applicable):** Do you maintain vendor communication and data logs? |
| **Safe / Secure** | ❑ **Develop O&M plans:** Do your O&M plans include the following components: identity and access management, vulnerability management, application whitelisting, network behavior analysis, automated security tools? [20]<br>❑ **Maintain ATO:** Do you complete a periodic security authorization and obtain a new ATO as needed? [22] |
| **Privacy** | ❑ **Routinely evaluate the PIA:** Do you review and update the PIA every three years? [26]<br>❑ **Monitor the storage and privacy of sensitive information:** Do you regularly monitor AI use to identify unauthorized attempts to access sensitive data and report suspected or confirmed breaches? [35]<br>❑ **Manage data inputs and outputs:** Do you retain and/or dispose of data in accordance with applicable policies? |
| **Robust / Reliable** | ❑ **Continuously monitor and improve model performance:** Do you monitor the model for data contamination and post-production data drift? Do you regularly re-evaluate the model and identify opportunities to enhance performance? [2, 15, 21]<br>❑ **Retire the AI solution when appropriate:** Will you retire the AI solution from production if model performance indicates that the solution is no longer relevant to the use case context or cost-effective for the agency? [2, 22] |

CHAPTER V

# EXTERNAL AI CONSIDERATIONS

# Regulatory Considerations

Op/StaffDivs should consider whether and how to regulate areas within their statutory authority that affect AI applications. In cases where regulatory action is necessary, Op/StaffDivs should apply the TAI principles by considering the below questions.

### Fair / Impartial

*Could the use of AI in this area result in discriminatory outcomes? How susceptible are AI systems and algorithms in this area to learning and propagating bias?*

Sample regulatory application: Bias reviews and/or metrics for AI data, models, and outcomes

### Transparent / Explainable

*To what extent should AI systems and algorithms in this area be open to inspection? How should the use of AI that uses individual data be explained to impacted individuals?*

Sample regulatory application: Information disclosure requirements

### Responsible / Accountable

*What are unintended outcomes of AI in this area? How should accountable and responsible parties be identified and acknowledged?*

Sample regulatory application: Traceability and/or credentialing requirements

### Safe / Secure

*What risks do AI systems and algorithms in this area face, and what type of physical and/or digital harm might those risks cause?*

Sample regulatory application: Minimum standard security controls

### Privacy

*Do AI systems and algorithms in this area use sensitive data and generate actions for individuals that could lead to privacy concerns?*

Sample regulatory application: De-identification requirements for PHI

### Robust / Reliable

*What measures for reliability and consistency do AI systems and algorithms in this area need to meet? How should inconsistencies and unintended outcomes be handled?*

Sample regulatory application: Model performance thresholds

*It is recommended that Op/StaffDivs **share AI-related regulatory priorities with the HHS OCAIO** to support communication with the White House, Congress, and other stakeholders.*

# Non-Regulatory Considerations

OMB Memorandum M-21-06, "Guidance for Regulation of Artificial Intelligence Applications," includes four non-regulatory approaches to reduce barriers to AI deployment and use. [13] The below table summarizes each approach.

| | Pilot Programs and Experiments Support | Non-Regulatory Consensus Standards | Access to Federal Data and Models | Public Communications |
|---|---|---|---|---|
| **OMB M-21-06 GUIDANCE** | • Allow pilot programs (i.e., hackathons, tech sprints, challenges, and other piloting programs) to encourage AI innovation<br>• Incorporate AI use and TAI principles into grant and research opportunities<br>• Issue grant waivers, deviations, and exemptions for specific AI applications<br>• Collect data on the design, development, deployment, operations, and outcomes of pilot AI applications to better understand AI risks and benefits | • Issue voluntary, non-regulatory policy statements within existing statutory authority<br>• Promote, create, or build upon datasets, tools, frameworks, and guidelines to accelerate AI understanding and innovation<br>• Align standards, frameworks, and tools to TAI principles<br>• Leverage private-sector conformity assessment programs and related activities before proposing regulations or compliance programs | • Increase access to government data and models where appropriate<br>• Review existing data disclosure protocols and identify systematic ways to share data<br>• Explore opportunities to provide granular, anonymized data rather than aggregate data<br>• Continue to follow legal and policy requirements for protecting sensitive data | • Communicate AI risks and benefits, including how external groups are impacted, to support understanding of and trust in AI<br>• Promote non-regulatory consensus standards, frameworks, and guidance<br>• Share trends and lessons learned from pilot programs where appropriate<br>• Ensure that RFIs related to AI are informed by agency risk assessments, context-specific, and based on sound scientific evidence |

## CONTACT INFORMATION

**Questions or comments about the HHS Trustworthy AI Playbook?**

**Please reach out to HHS.CAIO@HHS.GOV.**

## CONTRIBUTORS

**Thank you to the following Op/StaffDiv representatives who supported the development and review of the Trustworthy AI Playbook.**

- Joshua Williams (ACF)
- Alan Sim (CDC)
- Brian Lee (CDC)
- Andrés Colón (CMS)
- Rick Lee (CMS)
- Andreas Schick (FDA)
- Satish Gorrela (HRSA)
- Daniel Duplantier (HRSA)
- Renata Miskell (OIG)
- Stephen Konya (ONC)
- Kathryn Marchesini (ONC)

# APPENDIX

# References (1 of 3)

[1] *Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*, December 3, 2020. Available at: https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government

[2] U.S. Government Accountability Office, *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, June 30, 2021. Available at: https://www.gao.gov/products/gao-21-519sp

[3] Ziad Obermeyer, Rebecca Nissan, Michael Stern, Stephanie Eaneff, Emily J. Bebeneck, Sendhil Mullainathan, *Algorithmic Bias Playbook*, The Center for Applied Artificial Intelligence, June 2021. Available at: https://www.chicagobooth.edu/research/center-for-applied-artificial-intelligence/research/algorithmic-bias/playbook

[4] National Defense Authorization Act for Fiscal Year 2019, August, 13, 2018. Available at: https://www.congress.gov/bill/115th-congress/house-bill/5515/text

[5] Zoubin Ghahramani, *Probabilistic Machine Learning and AI*, 2017. Available at: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/03/Ghahramani.pdf

[6] IBM, *Predictive Analytics*. Available at: https://www.ibm.com/analytics/predictive-analytics

[7] Sara Brown, *Machine Learning, Explained*, MIT Sloan, April 21, 2021. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

[8] Eda Kovlakoglu, *NLP vs. NLU vs. NLG: The Differences Between Three Natural Language Processing Concepts*, November 12, 2020. Available at: https://www.ibm.com/blogs/watson/2020/11/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts/

[9] Gartner Glossary, *Speech Recognition*. Available at: https://www.gartner.com/en/information-technology/glossary/speech-recognition

[10] Stanford Computer Vision Lab. Available at: http://vision.stanford.edu/

[11] IBM Cloud Education, *Intelligent Automation*, March 5, 2021. Available at: https://www.ibm.com/cloud/learn/intelligent-automation

[12] Deloitte, *Trustworthy AI: Bridging the Ethics Gap Surrounding AI*. Available at: https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html

[13] Office of Management and Budget, *M-21-06: Guidance for Regulation of Artificial Intelligence Applications*, November 17, 2020. Available at: https://trumpwhitehouse.archives.gov/wp-content/uploads/2020/11/M-21-06.pdf

[14] Samir Passi, Solon Barocas, *Problem Formulation and Fairness*, January 29, 2019. Available at: https://dl.acm.org/doi/10.1145/3287560.3287567

# References (2 of 3)

[15] American Council for Technology-Industry Advisory Council, *Ethical Application of Artificial Intelligence Framework*, October 8, 2020. Available at: https://www.actiac.org/documents/act-iac-white-paper-ethical-application-ai-framework

[16] NISTIR 8312, *Four Principles of Explainable Artificial Intelligence*, Draft, August 2020. Available at: https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8312-draft.pdf

[17] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, Aurelia Tamò-Larrieux, *Towards Transparency by Design for Artificial Intelligence*, November 16, 2020. Available at: https://link.springer.com/article/10.1007/s11948-020-00276-4

[18] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, Peter Eckersley, *Explainable Machine Learning in Deployment*, July 10, 2020. Available at: https://arxiv.org/pdf/1909.06342.pdf

[19] General Services Administration, *The Digital Worker Identity Playbook*, January 5, 2021. Available at: https://playbooks.idmanagement.gov/docs/playbook-digital-worker.pdf

[20] *HHS Policy for Securing AI Technology*, Draft.

[21] Suchi Saria, Adarsh Subbaswamy, *Safe and Reliable Machine Learning*, April 15, 2019. Available at: https://arxiv.org/pdf/1904.07204.pdf

[22] *HHS Enterprise Performance Life Cycle Framework*, July 18, 2012. Available at: https://www.hhs.gov/sites/default/files/ocio/eplc-lifecycle-framework.pdf

[23] *HHS Policy for Information Technology Acquisition Reviews (ITAR)*, June 2020. Available at: https://www.hhs.gov/web/governance/digital-strategy/it-policy-archive/hhs-ocio-policy-for-information-technology-acquisition-reviews-itar.html

[24] U.S. Department of Homeland Security, *Artificial Intelligence: Using Standards to Mitigate Risks*. Available at: https://www.dhs.gov/sites/default/files/publications/2018_AEP_Artificial_Intelligence.pdf

[25] Amy Paul, Craig Jolley, Aubra Anthony, *Reflecting the Past, Shaping the Future: Making AI Work for International Development*, U.S. Agency for International Development, September 5, 2018. Available at: https://www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf

[26] *HHS Policy for Privacy Impact Assessments (PIA)*, June 4, 2019. Available at: https://www.hhs.gov/web/governance/digital-strategy/it-policy-archive/policy-for-privacy-impact-assessments.html

[27] Office of the Director of National Intelligence, *Artificial Intelligence Ethics Framework for the Intelligence Community*, Version 1.0, June 2020. Available at: https://www.dni.gov/files/ODNI/documents/AI_Ethics_Framework_for_the_Intelligence_Community_10.pdf

# References (3 of 3)

[28] 42 U.S.C. § 18116(a)

[29] IT Modernization Centers of Excellence, *Guide to AI Ethics*. Available at: https://coe.gsa.gov/docs/CoE%20Guide%20to%20AI%20Ethics.pdf

[30] Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, October 31, 2019. Available at: https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

[31] IBM, *AI Fairness 360*. Available at: https://aif360.mybluemix.net/

[32] Michelle Seng Ah Lee, Luciano Floridi, Jatinder Singh, *Formalising Trade-offs Beyond Algorithmic Fairness: Lessons from Ethical Philosophy and Welfare Economics*, June 12, 2021. Available at: https://link.springer.com/article/10.1007/s43681-021-00067-y

[33] Deon, *An Ethics Checklist for Data Scientists*. Available at: https://deon.drivendata.org/

[34] MITRE, *Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS)*. Available at: https://atlas.mitre.org/matrix/

[35] *HHS Policy for Preparing for and Responding to a Breach of Personally Identifiable Information (PII)*, Version 2.0, May 2020. Available at: https://www.hhs.gov/web/governance/digital-strategy/it-policy-archive/hhs-policy-preparing-and-responding-breach.html

[36] Google, *Responsible AI Practices*. Available at: https://ai.google/responsibilities/responsible-ai-practices/

[37] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes, *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, January 3, 2020. Available at: https://arxiv.org/pdf/2001.00973.pdf

[38] NIST, *Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0*, January 16, 2020. Available at: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf

[39] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, *Robustness and Explainability of Artificial Intelligence*, European Commission Joint Research Centre (JRC), 2020. Available at: https://publications.jrc.ec.europa.eu/repository/handle/JRC119336

[40] Microsoft, *Responsible Bots: 10 Guidelines for Developers of Conversational AI*, November 4, 2018. Available at: https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf

[41] Andrew Smith, *Using Artificial Intelligence and Algorithms*, Federal Trade Commission, April 8, 2020. Available at: https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms