



Trustworthy AI (TAI) Playbook

U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES

SEPTEMBER 2021

Table of Contents

| CHAPTER | SUB-CHAPTERS | QUESTIONS ADDRESSED | SLIDE # |
|--|--|--|--------------|
| I. Introduction | i. Message from HHS Chief AI Officer ii. Background iii. Trustworthy AI (TAI) Playbook Overview | <ul style="list-style-type: none"> • <i>Why is trustworthy AI important?</i> • <i>What is the purpose of the playbook?</i> • <i>Who is the intended audience?</i> | 3-8 |
| II. AI Building Blocks | i. AI Definition ii. AI Building Blocks | <ul style="list-style-type: none"> • <i>What is AI?</i> • <i>What are the components of AI solutions?</i> | 9-13 |
| III. Principles for Use of Trustworthy AI in Government | i. Fair / Impartial ii. Transparent / Explainable iii. Responsible / Accountable iv. Safe / Secure v. Privacy vi. Robust / Reliable | <ul style="list-style-type: none"> • <i>What makes AI solutions trustworthy?</i> • <i>What do the principles look like in practice?</i> • <i>What are the key considerations for each principle?</i> • <i>How do the principles align to federal guidance?</i> | 14-23 |
| IV. Internal AI Deployment Considerations | i. AI Lifecycle Overview ii. Initiation and Concept Overview iii. Research and Design iv. Develop, Train, and Deploy v. Operate and Maintain | <ul style="list-style-type: none"> • <i>What are the AI lifecycle phases?</i> • <i>What are recommended activities for applying the principles during each lifecycle phase?</i> • <i>How do the activities apply to sample AI use cases?</i> • <i>How can business leaders know they have been successful?</i> | 24-67 |
| V. External AI Considerations | i. Regulatory ii. Non-Regulatory | <ul style="list-style-type: none"> • <i>How can OpDivs and StaffDivs promote trustworthy AI development externally?</i> | 68-71 |

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 laws¹

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.

Overview

1 Adhere to Principles for Use of AI in Government

The EO outlines **nine principles that agencies must follow** when designing, developing, acquiring, and using AI in the federal government.

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.

What This Means For You

Op/StaffDivs are encouraged to...

- **Assess existing AI solutions** to ensure they adhere to the principles described in the Playbook
- **Carefully review the Playbook** before implementing new AI solutions

2 Create an Agency Inventory of AI Use Cases

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.

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.

Op/StaffDivs are encouraged to...

- **Provide a list of applicable AI use cases** in accordance with forthcoming OCAIO guidance
- **Use the inventory to connect with colleagues and share knowledge** about AI applications, technologies, processes, and best practices

For [more information on federal actions related to trustworthy AI](#), refer to the Appendix.

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 This Playbook

Chapters 2-3 provide high-level information for leaders interested in gaining a basic fluency in TAI, while Chapters 4-5 provide more granular guidance to support development of trustworthy AI solutions.



Chapter 2

AI Building Blocks



Chapter 3

Principles for Use of Trustworthy AI in Government



High-Level Information about TAI

Chapters 2-3 are designed to equip leaders with a basic understanding of the building blocks of AI solutions and the principles that underpin TAI. ***Readers should use these chapters to gain a high-level understanding of TAI regardless of where their current AI solutions are in the deployment lifecycle.***



Chapter 4

Internal AI Deployment Considerations



Chapter 5

External AI Considerations



Detailed Guidance for Leadership Teams

Chapters 4-5 include detailed recommendations for designing TAI solutions and fostering TAI innovation. ***Readers should become familiar with these chapters prior to beginning an AI project and continue to reference specific sections based on where an AI project is in the deployment lifecycle.***

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?*

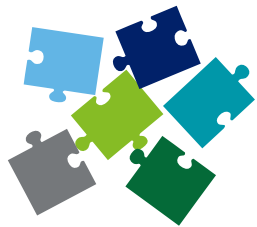
These considerations, while not all-encompassing, serve as a guide in determining whether a solution constitutes AI and whether TAI principles need to be applied

**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)



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

AI Solutions

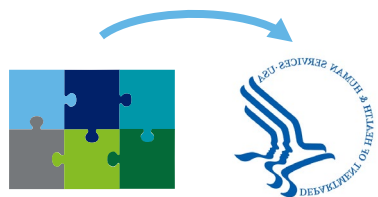


AI Solutions are made up of one or more AI methods and, after assessing problems and needs, are developed to carry out a specific function, purpose, or role (e.g., Call Center Bot)



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

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)



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

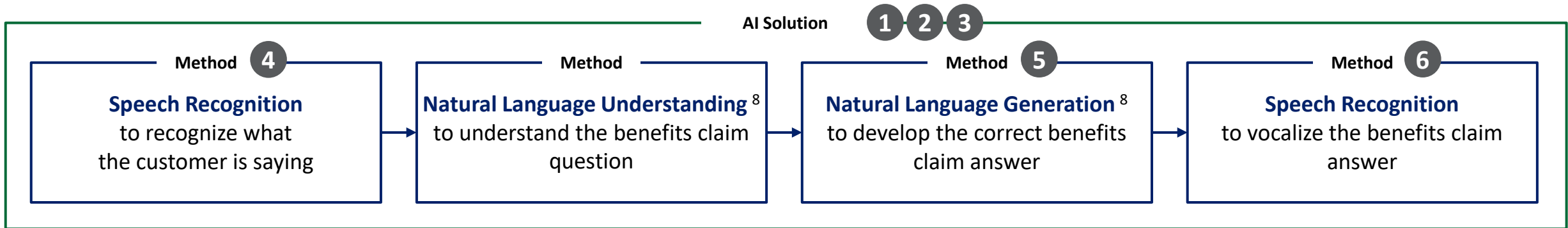
| | DEFINITION | SAMPLE TAI IMPLICATIONS |
|--|--|--|
| Machine Learning (ML) | <i>“A subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed” – MIT⁵ Includes probabilistic methods⁵ and can support predictive analytics⁶</i> | Machine learning should be bias-free and incorporate relevant shifts in healthcare demographics |
| Natural Language Processing (NLP) | <i>“Machines learn to understand natural language as spoken and written by humans” – MIT⁷ and includes both Natural Language Generation (NLG) and Natural Language Understanding (NLU) – IBM⁸</i> | NLP models should be understandable to users to prevent incorrect interpretations that could negatively impact affected individuals |
| Speech Recognition | <i>“Systems [that] interpret human speech and translate it into text or commands.” – Gartner⁹</i> | Voice and speech should be inclusive of a broad range of languages, dialects, and accents |
| Computer Vision | <i>“Intelligent algorithms that perform important visual perception tasks such as object recognition, scene categorization, integrative scene understanding, human motion recognition, material recognition, etc.” – Stanford¹⁰</i> | Computer vision models should be trained with data representative of the patient populations that will use them to support unbiased results |
| Intelligent Automation | <i>“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– IBM¹¹</i> | Intelligent automation solutions should have a human sponsor that is responsible for ensuring protected information (e.g., patient data) is not accessible |

Blockchain is another innovative technology that, while not AI, can be used to support AI solutions. [Learn more about blockchain and AI here.](#)

AI Building Blocks & TAI | Benefits Claims Center Bot Use Case

AI Use Case: Benefits Claims Call Center Bot

An HHS agency currently uses an automated messaging system to route citizens calling with benefits claims questions. To improve customer experience, the agency deploys an AI Benefit Claims Center Bot to receive, understand, and respond to citizen inquiries more quickly. The AI solution and methods are outlined below, along with sample TAI points of failure.



Sample TAI Points of Failure

- 1 The **AI solution** uses historic data based on a policy that was recently changed due to legislation, limiting the accuracy of responses
- 2 The **AI solution** was subject to an attacker with malicious intent, and without proper security protocols, agency privacy data was leaked
- 3 The **AI use case** is intended to reduce the number of claims paid out, raising ethical concerns given the population served
- 4 The **Speech Recognition method** was not designed to incorporate dialect for non-native speakers
- 5 The **Natural Language Generation (NLG) method** does not provide sufficient explanation to the benefit holder, causing frustration
- 6 The **Speech Recognition method** was not designed for hard-of-hearing individuals, impacting fairness

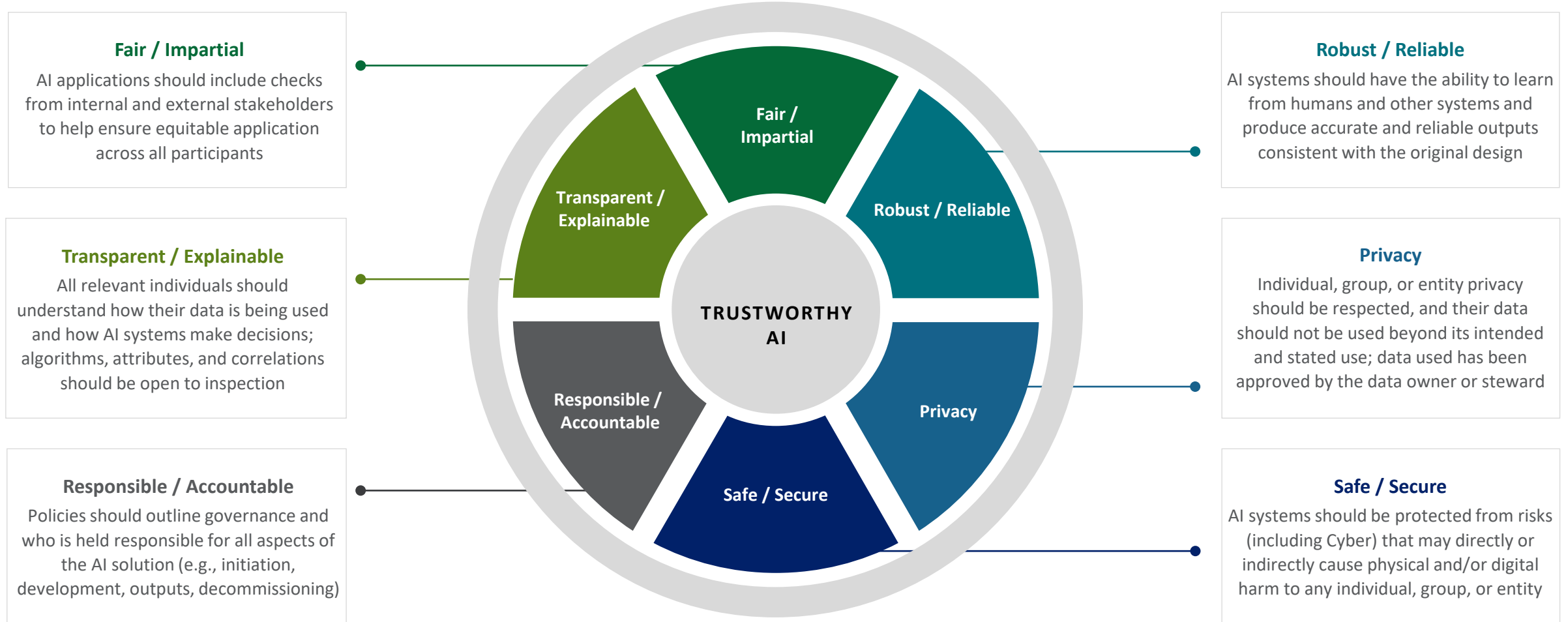
A trustworthy Call Center Bot will enable the agency to protect its reputation and improve operational efficiency.

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.



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 8. Transparent | 2. Public Participation 8. Disclosure and Transparency |
| Responsible / Accountable | 6. Responsible and Traceable 7. Regularly Monitored 9. Accountable | 5. Benefits and Costs |
| Safe / Secure | 4. Safe, Secure, and Resilient | 4. Risk Assessment and Management 9. Safety and Security |
| Privacy | 4. Safe, Secure, and Resilient | 9. Safety and Security |
| Robust / Reliable | 2. Purposeful and Performance-Driven 3. Accurate, Reliable, and Effective | 3. Scientific Integrity and Information Quality |

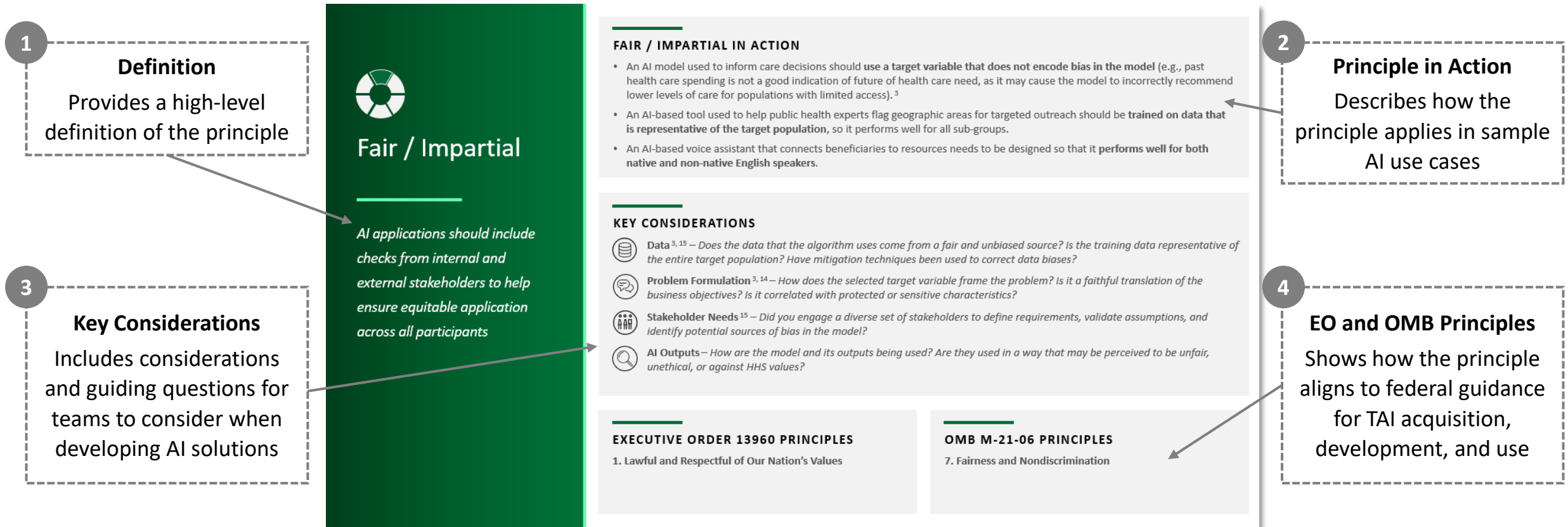
*Refer to the Appendix for more information on the [EO 13960](#) and [OMB M-21-06](#) Principles.

Additional Cross-Cutting Principles:

- 1. Public Trust
- 6. Flexibility
- 10. Interagency Coordination

How to Use This Section

The following slides provide an overview of each principle to equip business leaders with a basic understanding of the components of TAI solutions.



*This section will help you be **conversationally ready** to discuss the core components of TAI.*







Fair / Impartial

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

FAIR / IMPARTIAL IN ACTION

- An AI model used to inform care decisions should **use a target variable that does not encode bias in the model** (e.g., past health care spending is not a good indication of future of health care need, as it may cause the model to incorrectly recommend lower levels of care for populations with limited access).³
- An AI-based tool used to help public health experts flag geographic areas for targeted outreach should be **trained on data that is representative of the target population**, so it performs well for all sub-groups.
- An AI-based voice assistant that connects beneficiaries to resources needs to be designed so that it **performs well for both native and non-native English speakers**.

KEY CONSIDERATIONS

-  **Data**^{3, 15} – *Does the data that the algorithm uses come from a fair and unbiased source? Is the training data representative of the entire target population? Have mitigation techniques been used to correct data biases?*
-  **Problem Formulation**^{3, 14} – *How does the selected target variable frame the problem? Is it a faithful translation of the business objectives? Is it correlated with protected or sensitive characteristics?*
-  **Stakeholder Needs**¹⁵ – *Did you engage a diverse set of stakeholders to define requirements, validate assumptions, and identify potential sources of bias in the model?*
-  **AI Outputs** – *How are the model and its outputs being used? Are they used in a way that may be perceived to be unfair, unethical, or against HHS values?*

EXECUTIVE ORDER 13960 PRINCIPLES

1. Lawful and Respectful of Our Nation's Values

OMB M-21-06 PRINCIPLES

7. Fairness and Nondiscrimination



Transparent / Explainable

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

TRANSPARENT / EXPLAINABLE IN ACTION ¹⁶

- If an organization uses AI to match patients to clinical trials, the organization needs to be able to **explain the solution's decision-making process** to patients, providers, and researchers.
- If an AI model that predicts fraud in medical billing incorrectly flags a claim as fraudulent, the team overseeing the model should **record false identifications and share model performance metrics** with stakeholders.
- An AI-based tool that helps staff monitor child support obligations by flagging specific cases for review should **clearly communicate why specific cases are flagged** so that staff can understand and explain the rationale to stakeholders.

KEY CONSIDERATIONS



Technical and Functional Design – *Have you documented the AI logic and model inputs? Have you consulted with stakeholders to ensure that the technical and functional design documentation is understandable?*



AI Outputs ^{15, 17} – *Are the solution outputs sufficiently clear and comprehensible to end users such that they can take appropriate action? Have you tested the outputs to ensure they meet established explainability requirements?*



Stakeholder Needs ¹⁸ – *What people or groups (e.g., regulatory bodies) have an interest in the outputs of your AI solution? Have you engaged them to understand what they need to know about the model to trust the outputs (e.g., decision-making criteria)?*



Security Risks ¹⁹ – *How are you communicating information about the model to stakeholders? Could adversarial groups use the explanations to engineer attacks on the model?*

EXECUTIVE ORDER 13960 PRINCIPLES

- 5. Understandable
- 8. Transparent

OMB M-21-06 PRINCIPLES

- 2. Public Participation
- 8. Disclosure and Transparency







Responsible / Accountable

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

RESPONSIBLE / ACCOUNTABLE IN ACTION

- If an AI medical device fails to identify macular degeneration in a patient who later develops vision problems, there should be **clear roles and responsibilities in place to respond to the issue.** ¹⁶
- If an AI-based public health surveillance tool produces incorrect predictions about disease surges that lead to misallocation of resources, responsible parties need to be able to **retrace the solution's steps** to determine what went wrong and what corrective action should be taken. ¹⁶
- If stakeholders observe that an AI grants processing tool used to aid grant administrators is inappropriately accessing system data, they should **be able to quickly identify the human custodian** to resolve the issue.

KEY CONSIDERATIONS

-  **Roles and Responsibilities** – *Is there appropriate governance for your AI solution? Are there clear roles and responsibilities for continuously monitoring solution outputs? Have you obtained all necessary approvals?*
-  **Digital Identity Management** ²⁰ – *Have you considered the potential adverse impact level of the AI solution and created a digital identity if needed?*
-  **Traceability** ¹⁸ – *Can you retrace how your AI solution arrived at a given decision? Do you understand the decision-making processes and factors?*
-  **Auditability** ^{15, 18} – *Would a third-party auditor be able to assess the appropriateness of the decision-making processes and factors with the documentation that you have?*

EXECUTIVE ORDER 13960 PRINCIPLES

- 6. Responsible and Traceable
- 7. Regularly Monitored
- 9. Accountable

OMB M-21-06 PRINCIPLES

- 5. Benefits and Costs



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

SAFE / SECURE IN ACTION ¹⁶

- An NLP-based AI solution that interprets handwritten medical records needs to have **data encryption, user authentication, and other applicable security controls** to prevent hackers from stealing records.
- If a team is using open-source code to build an AI model that identifies opportunities for drug development, the team should **use code from whitelisted libraries** to avoid inadvertently downloading malware onto their computers.
- If a healthcare research agency is using an off-the-shelf AI software that scans patient records and identifies trial candidates, they should **obtain and assess vendor documentation of security controls**.

KEY CONSIDERATIONS



Security Risks²¹ – *What types of security risks (e.g., adversarial attacks) may impact your AI solution? What is the potential impact of each risk?*



Vulnerability Level – *How vulnerable is the AI solution in all levels of its stack (e.g., software level, algorithm level)? Are there controls in place to mitigate identified vulnerabilities?*



Access²¹ – *What are the access controls and training requirements for those operating at different stages of the AI lifecycle (e.g., developers, program managers)? Who has access, and what type of access do they have?*



Authorization – *Has the AI solution received proper clearance? Is there an active Authority to Operate (ATO)?*

EXECUTIVE ORDER 13960 PRINCIPLES

4. Safe, Secure, and Resilient

OMB M-21-06 PRINCIPLES

4. Risk Assessment and Management

9. Safety and Security



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

PRIVACY IN ACTION

- A research organization is studying the use of mobile technology to track symptoms of Parkinson’s disease. If a team uses the mobile data to build an AI model that provides tailored recommendations to users, they first need to **obtain consent from patients**.¹⁶
- If an agency is working with a third party to build an AI-based population health management tool, the agency should **implement appropriate privacy protections** (e.g., limiting the transfer of PHI to third party server).¹⁶
- If an agency rolls out an AI solution that helps users identify their colon cancer risk factors, it is important that all **PII (e.g., name, email, IP address) collection is minimized and stripped** prior to use by the AI system.

KEY CONSIDERATIONS



Sensitivity of the Data²¹ – *Does the data contain PII, PHI, or other sensitive data? Has the data been de-identified and/or encrypted?*



Individual Privacy Rights – *If your AI solution uses data about individuals, are those individuals aware of how their data is being used? Are those individuals able to opt in or out of sharing their data?*



Legal Requirements – *Is the AI solution in compliance with applicable privacy laws and regulations? Did you publish a System of Record Notice if necessary?*



Data Sharing – *Have you considered the tradeoff between protecting data privacy and releasing data for scientific replication and public good?*

EXECUTIVE ORDER 13960 PRINCIPLES

4. Safe, Secure, and Resilient

OMB M-21-06 PRINCIPLES

9. Safety and Security



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

ROBUST / RELIABLE IN ACTION

- An AI model that predicts the risk of substance use disorder in homeless youth needs to be **trained on representative target population data** (e.g., environmental, psychological, and behavioral data) to generate accurate predictions.¹⁶
- An AI-based precision medicine tool needs to be **regularly monitored to identify data drift**, or changes in the underlying data that negatively affect model performance. Otherwise, it could lead to misdiagnoses.¹⁶
- If new research comes out about the relationship between biomarkers and brain deterioration, an AI model that predicts brain deterioration should be **retrained to reflect the updated research**.

KEY CONSIDERATIONS



Data Quality²² – *Is the training data accurate, representative of real-world settings, and free of noise or outliers? Do you monitor the model for data drift? Is the model protected against data contamination?*



Methodology – *Has the model been checked for conceptual soundness? Have you considered the limitations of the selected AI methods and identified any implicit model assumptions?*



Model Performance and Monitoring¹⁵ – *Do model performance metrics meet desired thresholds? Are the model's outputs sensitive to small variations in the inputs? Is the model continuously monitored to identify changes in performance and/or opportunities for improvement?*



Consistency – *How do you manage new versions of the AI solution? If a new version produces different results, how do you resolve performance issues and communicate this with stakeholders?*

EXECUTIVE ORDER 13960 PRINCIPLES

2. Purposeful and Performance-Driven
3. Accurate, Reliable, and Effective

OMB M-21-06 PRINCIPLES

3. Scientific Integrity and Information Quality

CHAPTER IV

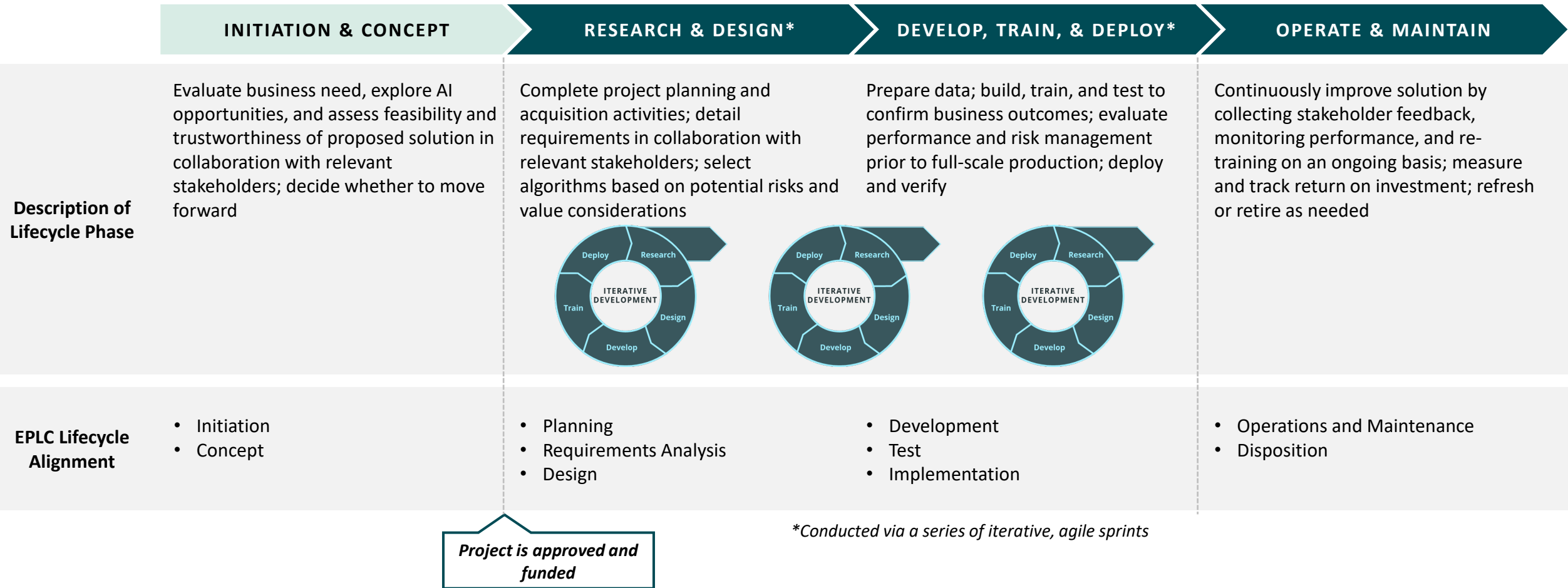
INTERNAL AI DEPLOYMENT CONSIDERATIONS

INTERNAL AI DEPLOYMENT CONSIDERATIONS

AI LIFECYCLE OVERVIEW

AI Lifecycle

There are four phases of a typical AI lifecycle that align to the HHS Enterprise Performance Lifecycle framework.²³ This chapter focuses on the deployment of AI solutions, which begins after an AI concept has been approved and funded.

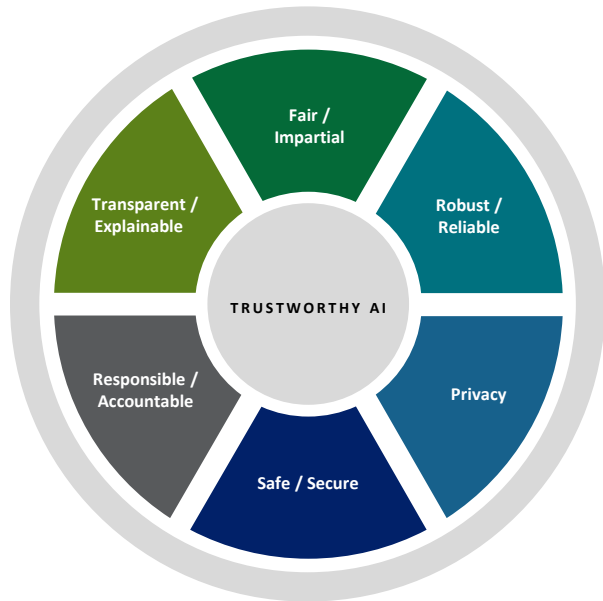


*Leaders must apply the principles **during all stages of the lifecycle** to create TAI solutions.*

Application of TAI Principles Across the AI Lifecycle

Reviewing TAI principles during each phase of the AI lifecycle is critical to effectively creating TAI solutions.

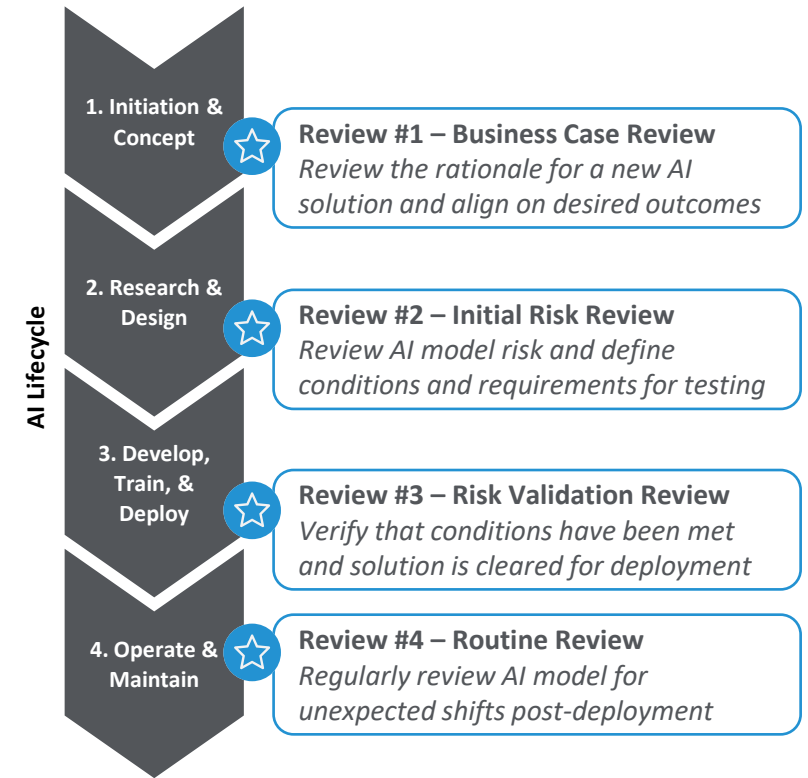
The TAI principles serve as a framework for understanding AI risk.



The TAI principles are applied during every stage of the AI lifecycle.



AI models should undergo reviews during each stage of the AI lifecycle to ensure TAI principles and risks are balanced.



Using the principles to understand and mitigate AI risk throughout the lifecycle supports TAI solutions.

INTERNAL AI DEPLOYMENT CONSIDERATIONS

INITIATION & CONCEPT OVERVIEW

LIFECYCLE PHASE I



How to Use This Section (Lifecycle Phase I)

The following slides will help leaders determine whether to invest in an AI project by evaluating the benefits and TAI risks.

Initiation and Concept | TAI Considerations

Before beginning an AI project, leaders should determine if AI is the right approach for the business problem by conducting a [cost-benefit analysis](#) and considering potential TAI risks.

| TAI Principle | CONSIDERATIONS |
|----------------------------------|--|
| Fair / Impartial | <ul style="list-style-type: none"> Consider how you will translate the business problem into questions that AI algorithms can answer: What are the potential target variables? Are they consistent with protected or sensitive characteristics, measured less accurately for certain subgroups, or more difficult to predict than other variables? (e.g., using “no-shows” as a target may exclude bias in scheduling service for users in more geographically dispersed subgroups) Determine how you will use the available inputs: Will you use the inputs to make resource allocation decisions that could have a disparate impact on affected subgroups? (e.g., providing technical assistance to grantee) Survey the legal and regulatory landscape: What regulations, standards, policies, or laws related to bias and discrimination apply to the proposed solution? (e.g., Social Security Act) |
| Transparent / Explainable | <ul style="list-style-type: none"> Evaluate stakeholder needs: Who will use, be affected by, or have an interest in the solution's output? Are any of those stakeholders external to HHS? What might they want to know about the solution's inputs, outputs, or decision-making process? (e.g., public health professionals may want to know what criteria an AI solution uses to recommend interventions) Consider the modelability of your solution: Will the proposed solution use deep learning, support vector machines, or other AI methods that could increase accuracy but decrease explainability? |
| Responsible / Accountable | <ul style="list-style-type: none"> Complete the IT Acquisition Review (ITAR) Process: 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 to obtain approval to acquire those services? (e.g., ITAR) Use the Digital Services Impact Scenarios Matrix to forecast the solution's potential adverse impact: What type of access will the solution have? Will it be able to act on its own insight? What is the potential impact of the solution's output? |
| Safe / Secure | <ul style="list-style-type: none"> Identify security risks: How might adversarial agents seek to target and compromise the AI solution? (e.g., poisoning data) |
| Privacy | <ul style="list-style-type: none"> Conduct a preliminary Privacy Impact Assessment: Will the proposed solution use sensitive data (e.g., PHI)? If so, why? How will you collect, share, and use that information? |
| Robust / Reliable | <ul style="list-style-type: none"> Consider the likelihood of error: Will the proposed solution require joining and pre-processing data from multiple sources? Are those sources likely to be accurate, appropriate, and representative? How potential target variables been well measured in the past? Are there any dependencies, assumptions, or constraints that may impact the solution? Evaluate the proposed team composition: Will there be more than one data scientist or other team member? Will those who have prior experience with the AI method(s) selected? Will the team include diverse perspectives and programmatic expertise? |

TAI Considerations

For each TAI principle, there is a set of guiding questions to identify risks with a proposed AI solution. Leaders can use that information to understand the resources needed to address TAI risks during development, build a business case, and ultimately decide whether and how to move forward with the project.

Initiation and Concept | Acquisition Approaches

While factors like cost, time, and project objectives may drive the acquisition strategy for an AI project, leaders should also consider the TAI implications of different approaches:

| Buy Approach | Hybrid Approach | Build Approach |
|--|---|--|
| <p>DESCRIPTION</p> <p>Use of pre-trained algorithms developed by a third party, with limited customization by the acquiring organization.</p> <p>THINK: Commercial Off-the-Shelf (COTS) AI Products</p> <p>TAI IMPLICATIONS</p> <ul style="list-style-type: none"> Organizations have less insight into how the vendor trained and tested the algorithm, including what data they used Organizations may need to conduct additional review and validation testing to ensure that the algorithm satisfies all TAI principles Organizations should carefully consider all necessary requirements during procurement, since there is less flexibility after deployment Contract should include provisions for appropriate access to data, model documentation, and test results to enable effective review | <p>DESCRIPTION</p> <p>Use of pre-trained algorithms developed by a third party, with greater customization by the acquiring organization, and some in-house development of custom algorithms</p> <p>THINK: Cloud-Native Machine Learning Apps</p> <p>TAI IMPLICATIONS</p> <ul style="list-style-type: none"> Organizations have some control over how AI project teams build and test the algorithm and some flexibility after deployment Organizations may not have full insight into the training data or model parameters Organizations need to establish and carefully review vendor documentation for pre-trained algorithms, while also providing sufficient oversight for custom development | <p>DESCRIPTION</p> <p>Development of custom algorithms, including fully custom-built code</p> <p>THINK: Custom-Built Solutions</p> <p>TAI IMPLICATIONS</p> <ul style="list-style-type: none"> Organizations have more control over how AI project teams train and test the algorithm Organizations may need to spend more time in the design and development stages of the AI lifecycle to correctly frame the business problem, identify applicable AI methods, and securely code the solution Organizations have more flexibility after deployment to refine and adapt the solution AI project teams should include experienced data scientists who can prepare the training data and write code from scratch |

Regardless of the selected acquisition approach, leaders should determine the optimal [team composition](#) to successfully execute the project and ensure trustworthiness.

Acquisition Approaches

After the principle-specific considerations, there is an overview of the TAI implications for three AI acquisition approaches—build, buy, or a hybrid of the two. Leaders should consider these implications when selecting an approach and developing the project plan.

Initiation and Concept | Risk Review Checklist

Purpose of Review: At the end of the Initiation and Concept phase, there should be a Business Case Review in alignment with the HHS EDC Framework. The purpose of this review is to evaluate that AI is the right approach for the business problem and that the expected benefits of using AI justify the investment costs and potential TAI risks. This review enables leaders to sign up on the identified solutions of an AI solution prior to design and development. It also helps leaders ensure that they have the right resources in place to successfully create an ethical and effective AI solution.

| | |
|---------------------------|---|
| Business Purpose | <ul style="list-style-type: none"> What is the intended business purpose of the AI solution? Does the intended business purpose align with HHS' values? |
| Benefits and Risks | <ul style="list-style-type: none"> Have you identified and researched AI and AI-like approaches to meet the intended business purpose? What are the expected benefits and potential risks associated with each approach? Have you assessed mitigation strategy/tradeoffs? Is the best approach for the intended business purpose? Are there any dependencies, assumptions, or constraints that may impact the solution? Do the expected benefits outweigh the potential risks and justify the authors' investment? |
| Preliminary Design | <ul style="list-style-type: none"> What are the expected inputs to the AI solution? Have sufficient justification for using that data? Will it use or be affected by the solution output? Has a diverse set of relevant stakeholders been consulted? Will what other products and systems will the AI solution need to connect? Are there any barriers to achieving interoperability? Are there any dependencies, assumptions, or constraints that may impact the solution? |
| Project Scope | <ul style="list-style-type: none"> What is the expected timeline for the project? What are the expected resource requirements (e.g., FTEs, technology)? What is the anticipated cost of the effort? Do you plan to procure? (products or services)? Have all applicable acquisitions been reviewed and approved according to the ITAR process? |

| | | | |
|---------------------------|--|---|--|
| POTENTIAL OUTCOMES | <input checked="" type="checkbox"/> Approved There is sufficient justification for using AI for the intended business purpose. The project warrant funding, and the team can proceed with Research and Design. | <input type="checkbox"/> Approved with Conditions The project warrant funding, however, changes need to be made to the solution concept before the next process with Research and Design. | <input type="checkbox"/> Decided There is sufficient justification for using AI for the intended business purpose. The project is placed on hold or abandoned. |
|---------------------------|--|---|--|

Risk Review Checklist

Lastly, there is a checklist to support a Business Case Review. The checklist outlines the purpose of the review, the recommended risks to consider, and the potential outcomes of the review. It will help leaders validate that AI is the right tool to address the business problem and that the expected benefits outweigh the risks.

This section will help you identify TAI risks with an AI solution concept.

Initiation and Concept | TAI Considerations

Before beginning an AI project, leaders should determine if AI is the right approach for the business problem by conducting a [cost-benefit analysis](#) and considering potential TAI risks.

| TAI PRINCIPLE | CONSIDERATIONS |
|---|---|
| <p>Fair / Impartial</p> | <ul style="list-style-type: none"> <input type="checkbox"/> 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, measured less accurately for certain subgroups, or more difficult to predict than other variables? ¹⁴ (e.g., using “no-shows” as a target may encode bias in scheduling since barriers to care are unequally distributed across subgroups) ³ <input type="checkbox"/> 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? (e.g., providing technical assistance to grantees) <input type="checkbox"/> Survey the legal and regulatory landscape: What regulations, standards, policies, or laws related to bias and discrimination apply to the proposed solution? (e.g., Social Security Act) |
| <p>Transparent / Explainable</p> | <ul style="list-style-type: none"> <input type="checkbox"/> Evaluate stakeholder needs: Who will use, be affected by, or have an interest in the solution’s outputs? Are any of those stakeholders external to HHS? What might they want to know about the solution’s inputs, outputs, or decision-making process? ¹⁹ (e.g., public health professionals may want to know what criteria an AI solution uses to recommend interventions) <input type="checkbox"/> 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? |
| <p>Responsible / Accountable</p> | <ul style="list-style-type: none"> <input type="checkbox"/> Complete the IT Acquisition Review (ITAR) Process: ²⁴ 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 to obtain approval in accordance with the HHS Policy? <input type="checkbox"/> Use the Digital Worker Impact Evaluation Matrix to forecast the solution’s potential adverse impact level: ²⁰ What type of access will the solution have? Will it be able to act on its own insights? What is the potential impact of the solution’s outputs? |
| <p>Safe / Secure</p> | <ul style="list-style-type: none"> <input type="checkbox"/> Identify security risks: ^{25, 26} How might adversarial agents seek to target and compromise the AI solution? (e.g., poisoning data) |
| <p>Privacy</p> | <ul style="list-style-type: none"> <input type="checkbox"/> Conduct a preliminary Privacy Impact Assessment: ²⁷ Will the proposed solution use sensitive data (e.g., PHI)? If so, why? How will you collect, share, and use that information? |
| <p>Robust / Reliable</p> | <ul style="list-style-type: none"> <input type="checkbox"/> Consider the likelihood of error: Will the proposed solution require joining and pre-processing data from multiple sources? Are those sources likely to be accurate, appropriate, and representative? ² Have potential target variables been well-measured in the past? ¹⁴ <input type="checkbox"/> 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 programmatic expertise? |

Initiation and Concept | Acquisition Approaches

While factors like cost, time, and project objectives may drive the acquisition strategy for an AI project, leaders should also consider the TAI implications of different approaches.

| Buy Approach | Hybrid Approach | Build Approach |
|--|--|---|
| <p>DESCRIPTION</p> <p>Use of pre-trained algorithms developed by a third party, with limited customization by the acquiring organization</p> <p><i>Think: Commercial Off-the-Shelf (COTS) AI Products</i></p> | <p>DESCRIPTION</p> <p>Use of pre-trained algorithms developed by a third party, with greater customization by the acquiring organization, and some in-house development of custom algorithms</p> <p><i>Think: Cloud-Native Machine Learning Apps</i></p> | <p>DESCRIPTION</p> <p>Development of custom algorithms, including fully custom-built code</p> <p><i>Think: Custom-Built Solutions</i></p> |
| <p>TAI IMPLICATIONS</p> <ul style="list-style-type: none"> Organizations have less insight into how the vendor trained and tested the algorithm, including what data they used Organizations may need to conduct additional reviews and validation testing to ensure that the algorithm satisfies all TAI principles Organizations should carefully consider all necessary requirements during procurement, since there is less flexibility after deployment Contracts should include provisions for appropriate access to data, design documentation, and test results to enable sufficient review² | <p>TAI IMPLICATIONS</p> <ul style="list-style-type: none"> Organizations have some control over how AI project teams build and test the algorithm and some flexibility after deployment Organizations may not have full insight into the training data or model parameters Organizations need to obtain and carefully review vendor documentation for pre-trained algorithms while also providing sufficient oversight for custom development | <p>TAI IMPLICATIONS</p> <ul style="list-style-type: none"> Organizations have more control over how AI project teams train and test the algorithm Organizations may need to spend more time in the design and development stages of the AI lifecycle to correctly frame the business problem, identify applicable AI methods, and securely code the solution Organizations have more flexibility after deployment to refine and adapt the solution AI project teams should include experienced data scientists who can prepare the training data and write code from scratch |

Regardless of the selected acquisition approach, leaders should determine the optimal [team composition](#) to successfully execute the project and ensure trustworthiness.

Initiation and Concept | Risk Review Checklist

Purpose of Review #1

At the end of the Initiation and Concept phase, there should be a **Business Case Review** in alignment with the HHS EPLC Framework. The purpose of this review is to validate that AI is the right approach for the business problem and that the expected benefits of using AI justify the investment costs and potential TAI risks. This review enables leaders to align on the desired outcomes of an AI solution prior to design and development. It also helps leaders ensure that they have the right resources in place to successfully create an ethical and effective AI solution. ^{23, 26, 28}



Business Purpose

- What is the intended business purpose of the AI solution?
- Does the intended business purpose and AI use align with HHS' values?



Benefits and Risks

- Have you identified and researched AI and non-AI approaches to meet the intended business purpose?
- What are the expected benefits and potential risks associated with each approach? Have you assessed mitigation strategy tradeoffs?
- Is AI the best approach for the intended business purpose?
- Are there existing AI solutions in the Op/StaffDiv AI Use Case Inventory that could apply to this business problem?
- Do the expected benefits outweigh the potential risks and justify the upfront investment?



Preliminary Design

- What are the expected inputs to the AI solution? Is there sufficient justification for using that data?
- Who will use or be affected by the solution outputs? Has a diverse set of relevant stakeholders been consulted?
- With what other products and systems will the AI solution need to connect? Are there any barriers to achieving interoperability?
- Are there any dependencies, assumptions, or constraints that may impact the solution?



Project Scope

- What is the expected timeline for the project?
- What are the expected resource requirements (e.g., FTEs, technology)?
- What is the anticipated cost of the effort?
- Do you plan to procure IT products or services? Have all applicable acquisitions been reviewed and approved according to the ITAR process?

POTENTIAL OUTCOMES



Approved

There is sufficient justification for using AI for the intended business purpose. The project receives funding, and the team can proceed with Research and Design.



Approved with Conditions

The project receives funding. However, changes need to be made to the solution concept before the team proceeds with Research and Design.



Declined

There is not sufficient justification for using AI for the intended business purpose. The project is placed on hold or abandoned.

INTERNAL AI DEPLOYMENT CONSIDERATIONS

HOW TO USE THE DEPLOYMENT CONSIDERATIONS

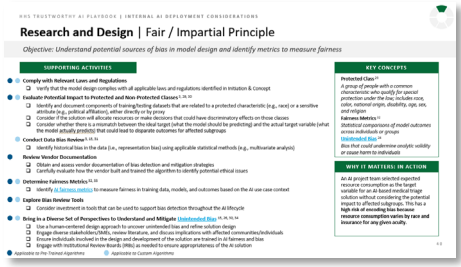
LIFECYCLE PHASES II - IV



How to Use This Section (Lifecycle Phases II-IV)

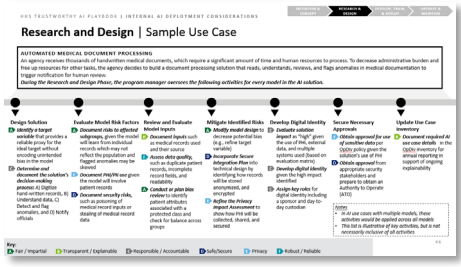
The deployment considerations are organized by the following AI lifecycle phases: Research & Design; Develop, Train & Deploy; and Operate & Maintain.

AI Deployment Lifecycle Phases II-IV Includes Sections for:



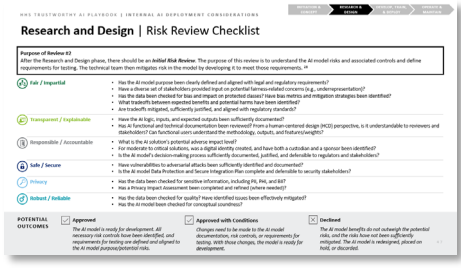
A Considerations for Applying TAI Principles

Within each AI deployment lifecycle phase, there is a set of guidelines and resources to help leaders apply each principle to AI solutions.



B Sample AI Use Case

After the principle-specific considerations, there is a sample AI use case to illustrate how leaders might use the guidance for a real AI solution.



C Risk Review Checklist

At the end of each AI deployment lifecycle phase, there is a risk review checklist to facilitate accountability during each phase. The checklist outlines the purpose of the review, the recommended risks to consider, and the potential outcomes of the review.

This section will help you deliver TAI by identifying the activities relevant for your solution.

Section Overview: A) Considerations for Applying TAI Principles

This section provides guidance for applying TAI principles during each phase of AI solution deployment. The lifecycle phase, use case, and type of AI methods all affect the application of TAI principles.

A

Research and Design | Fair / Impartial Principle

Objective: Understand potential sources of bias in model design and identify metrics to measure fairness

SUPPORTING ACTIVITIES

- **Comply with Relevant Laws and Regulations**
 - Verify that the model design complies with all applicable laws and regulations identified in Initiation & Concept
- **Evaluate Potential Impact to Protected and Non-Protected Classes** ^{3, 29, 30}
 - Identify and document components of training/testing datasets that are related to a protected characteristic (e.g., race) or a sensitive attribute (e.g., political affiliation), either directly or by proxy
 - Consider if the solution will allocate resources or make decisions that could have discriminatory effects on those classes
 - Consider whether there is a mismatch between the ideal target (what the model should be predicting) and the actual target variable (what the model actually predicts) that could lead to disparate outcomes for affected subgroups
- **Conduct Data Bias Review** ^{3, 25, 31}
 - Identify historical bias in the data (i.e., representation bias) using applicable statistical methods (e.g., multivariate analysis)
- **Review Vendor Documentation**
 - Obtain and assess vendor documentation of bias detection and mitigation strategies
 - Carefully evaluate how the vendor built and trained the algorithm to identify potential ethical issues
- **Determine Fairness Metrics** ^{32, 33}
 - Identify [AI fairness metrics](#) to measure fairness in training data, models, and outcomes based on the AI use case context
- **Explore Bias Review Tools**
 - Consider investment in tools that can be used to support bias detection throughout the AI lifecycle
- **Bring in a Diverse Set of Perspectives to Understand and Mitigate Unintended Bias** ^{15, 26, 30, 34}
 - Use a human-centered design approach to uncover unintended bias and refine solution design
 - Engage diverse stakeholders/SMEs, review literature, and discuss implications with affected communities/individuals
 - Ensure individuals involved in the design and development of the solution are trained in AI fairness and bias
 - Engage with Institutional Review Boards (IRBs) as needed to ensure appropriateness of the AI solution

● Applicable to Pre-Trained Algorithms ● Applicable to Custom Algorithms

KEY CONCEPTS

Protected Class ²⁹
A group of people with a common characteristic who qualify for special protection under the law; includes race, color, national origin, disability, age, sex, and religion

Fairness Metrics ³²
Statistical comparisons of model outcomes across individuals or groups

Unintended Bias ²⁸
Bias that could undermine analytic validity or cause harm to individuals

WHY IT MATTERS: IN ACTION

An AI project team selected expected resource consumption as the target variable for an AI-based medical triage solution without considering the potential impact to affected subgroups. This has a **high risk of encoding bias because resource consumption varies by race and insurance for any given acuity.**

4 0

1 Objective
 Describes what business leaders should seek to accomplish when applying the principle during the relevant lifecycle phase

2 Supporting Activities
 Includes recommendations for achieving the objective with links to additional guidance; they are color-coded based on the type of AI model to which they apply

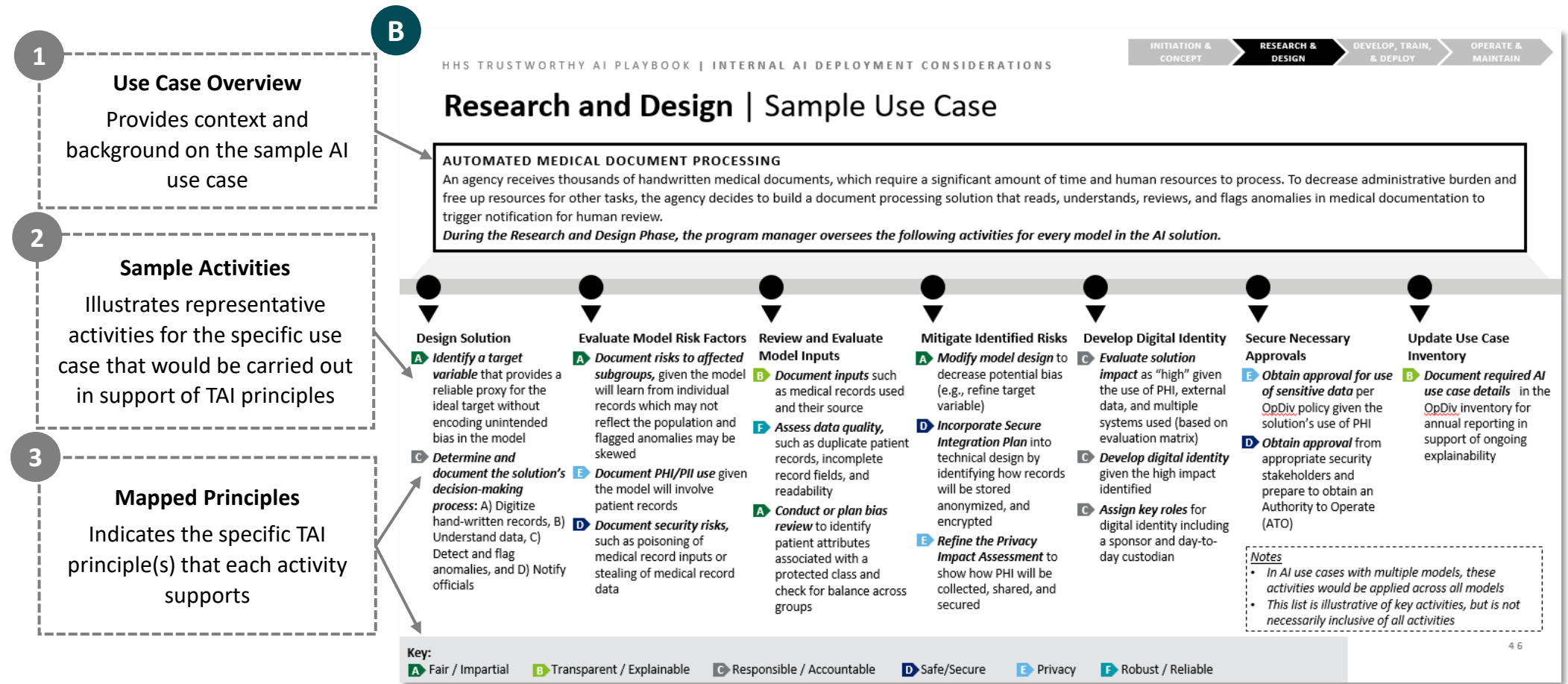
3 Key Concepts
 Provides additional detail on important terms or methodologies described in the supporting activities

4 Why it Matters: In Action
 Highlights the importance of the supporting activities by describing a sample AI project where TAI principles were not fully applied

Leaders should take the specific AI use case into account when determining how to conduct relevant activities.

Section Overview: B) Sample AI Use Case

Not all activities will be relevant for every AI use case, nor will teams conduct all activities in the same way. The sample use cases show how the supporting activities could work in practice.



Section Overview: C) Risk Review Checklist

Before moving from one AI deployment lifecycle phase to the next, leaders should use the risk review checklist as a guide to ensure TAI principles have been applied.









INTERNAL AI DEPLOYMENT CONSIDERATIONS

RESEARCH & DESIGN

LIFECYCLE PHASE II



Research and Design | Overview of How TAI Principles Are Applied

| PRINCIPLE | <p>Fair / Impartial</p>  | <p>Transparent / Explainable</p>  | <p>Responsible / Accountable</p>  | <p>Safe / Secure</p>  | <p>Privacy</p>  | <p>Robust / Reliable</p>  |
|----------------------------------|--|---|--|---|--|---|
| HOW IT'S APPLIED* | <ul style="list-style-type: none"> Comply with relevant laws and regulations Evaluate potential impact to protected/non-protected classes Conduct data bias review Review vendor documentation Determine fairness metrics Explore bias review tools Bring in diverse perspectives | <ul style="list-style-type: none"> Update Op/StaffDiv Use Case Inventory Establish explainability requirements Create a feedback mechanism Document model inputs in design documentation Evaluate and justify model explainability in design documentation | <ul style="list-style-type: none"> Create a digital identity if needed Assign key roles for the AI solution Document the solution's decision-making process | <ul style="list-style-type: none"> Evaluate identified security risks Incorporate a data protection and secure integration plan into technical design documentation Obtain necessary approvals | <ul style="list-style-type: none"> Refine the Privacy Impact Assessment for solutions using sensitive data Obtain approval for the use of sensitive data Identify applicable laws and regulations | <ul style="list-style-type: none"> Assess data quality Select AI methods Perform conceptual soundness check Review vendor documentation of model design decisions |
| SAMPLE STAKEHOLDERS ² | <ul style="list-style-type: none"> Individuals affected by the AI solution Legal Counsel Civil Rights, Ethics, and Minority Health / Health Equity Offices | <ul style="list-style-type: none"> Users Communications and Public Affairs Offices Data and Analytics Offices | <ul style="list-style-type: none"> Op/StaffDiv OCIO System Administrator | <ul style="list-style-type: none"> HHS OCIO Op/StaffDiv OCIO System Owner Database Owner ISSO or CISO | <ul style="list-style-type: none"> Op/StaffDiv Senior Official of Privacy System Owner | <ul style="list-style-type: none"> Users Data and Analytics Offices |

*Note: Activities will be not be applicable for every AI use case



Research and Design | Fair / Impartial Principle

Objective: Understand potential sources of bias in model design and identify metrics to measure fairness

SUPPORTING ACTIVITIES

KEY CONCEPTS

Protected Class²⁹

A group of people with a common characteristic who qualify for special protection under the law; includes race, color, national origin, disability, age, sex, and religion

Fairness Metrics³²

Statistical comparisons of model outcomes across individuals or groups

Unintended Bias²⁸

Bias that could undermine analytic validity or cause harm to individuals

WHY IT MATTERS: IN ACTION

An AI project team selected expected resource consumption as the target variable for an AI-based medical triage solution without considering the potential impact to affected subgroups. This has a **high risk of encoding bias because resource consumption varies by race and insurance for any given acuity.**

● ◆ Comply with Relevant Laws and Regulations

- ☐ Verify that the model design complies with all applicable laws and regulations identified in Initiation & Concept

● ◆ Evaluate Potential Impact to Protected and Non-Protected Classes^{3, 29, 30}

- ☐ Identify and document components of training/testing datasets that are related to a protected characteristic (e.g., race) or a sensitive attribute (e.g., political affiliation), either directly or by proxy
- ☐ Consider if the solution will allocate resources or make decisions that could have discriminatory effects on those classes
- ☐ Consider whether there is a mismatch between the ideal target (what the model should be predicting) and the actual target variable (what the model actually predicts) that could lead to disparate outcomes for affected subgroups

◆ Conduct Data Bias Review^{3, 15, 31}

- ☐ Identify historical bias in the data (i.e., representation bias) using applicable statistical methods (e.g., multivariate analysis)

● Review Vendor Documentation

- ☐ Obtain and assess vendor documentation of bias detection and mitigation strategies
- ☐ Carefully evaluate how the vendor built and trained the algorithm to identify potential ethical issues

● ◆ Determine Fairness Metrics^{32, 33}

- ☐ Identify [AI fairness metrics](#) to measure fairness in training data, models, and outcomes based on the AI use case context

● ◆ Explore Bias Review Tools

- ☐ Consider investment in tools that can be used to support bias detection throughout the AI lifecycle

● ◆ Bring in a Diverse Set of Perspectives to Understand and Mitigate [Unintended Bias](#)^{15, 26, 30, 34}

- ☐ Use a human-centered design approach to uncover unintended bias and refine solution design
- ☐ Engage diverse stakeholders/SMEs, review literature, and discuss implications with affected communities/individuals
- ☐ Ensure individuals involved in the design and development of the solution are trained in AI fairness and bias
- ☐ Engage with Institutional Review Boards (IRBs) as needed to ensure appropriateness of the AI solution



Research and Design | Transparent / Explainable Principle

Objective: Document the AI use case in the Op/StaffDiv inventory and evaluate model explainability

SUPPORTING ACTIVITIES

- ◆ **Update Op/StaffDiv Use Case Inventory (additional guidance to be provided)**
 - ❑ Confirm that the AI use case is non-classified and non-sensitive, and document required information about the AI use case in the Op/StaffDiv inventory for annual reporting
- ◆ **Establish Explainability Requirements** ^{18, 19}
 - ❑ Engage a diverse group of stakeholders to understand what they should know about the model to trust the outputs
 - ❑ Decide how to safely share that information with stakeholders
- ◆ **Create a Feedback Mechanism** ¹⁵
 - ❑ Use a human-centered design approach to develop a mechanism by which stakeholders can regularly provide feedback on the AI solution (e.g., report potential inconsistencies) and, if needed, contest the outcome of the AI solution
- ◆ **Document Model Inputs in Design Documentation** ²
 - ❑ Describe the data source and parameters of each model input
 - ❑ If any input data is synthetic, imputed, or augmented, document the process used to produce the data
 - ❑ For user-sourced inputs, specify the business unit that owns the data and the method by which the input is obtained
 - ❑ If applicable, treat infrequently calibrated and hard-coded parameters as assumptions
 - ❑ Identify AI [interpretation methods](#) (e.g., LIME, SHAP) to use in the model design
- ◆ **Evaluate and Justify Model Explainability in Design Documentation** ²
 - ❑ Assess the explainability of the model, considering the type, number and complexity of model features/methods used; hyperparameters; and interdependencies with other AI models
 - ❑ [Model explainability is often reduced with increasing model accuracy and power](#); identify potential consequences of using a less explainable model and determine if the benefits outweigh the risks
 - ❑ Document the evaluation and provide justification that the level of explainability is appropriate for the use case

KEY CONCEPTS

Explainability¹⁸

Extent to which you can understand and communicate the rationale behind the model's outputs

Model Feature³⁵

Independent input variable used to generate model predictions

Hyperparameter³⁶

Configuration variable used to control the model learning process

WHY IT MATTERS: IN ACTION

When designing an AI-based precision medicine solution, the AI project team did not properly document how and from where the input data was obtained. Later, the team could not answer questions about the input data during a demo with end users, which **raised concerns and delayed deployment.**



Research and Design | Responsible / Accountable Principle

Objective: Understand the potential actions the AI solution can take and establish appropriate oversight

SUPPORTING ACTIVITIES

- ◆ **Create a Digital Identity if Needed** ²⁰
 - ❑ Use the [ICAM Program Management Guide](#) to reference the latest policies for AI identities
 - ❑ Decide whether to create a digital identity for the solution based on the [potential adverse impact level](#) (low-impact solutions may not require one)
 - ❑ Assign digital workers the lowest level of access required to complete the solution's task
 - ❑ Validate that the digital worker does not create a separation of duty conflict
- ◆ **Assign [Key Roles](#) for the AI Solution** ²⁰
 - ❑ Assign a sponsor, or a federal government employee responsible for solution compliance
 - ❑ Assign a custodian, or a federal government employee or contractor responsible for day-to-day solution management
 - ❑ Determine whether additional [human supervision](#) may be required depending on the risk level and objectives of the AI solution, and assign roles as applicable
- ◆ **Document the Solution's Decision-Making Process** ²
 - ❑ Create functional/technical documentation of data transformation steps and potential actions the solution can take
 - ❑ Provide justification for each step in the process and document all manual components and assumptions
 - ❑ Include calibration steps and any necessary files for model input processing
 - ❑ Formally review the calibration approach to ensure consistency and lack of errors

KEY CONCEPTS

Digital Worker ²⁰

An automated, software-based tool, application, or agent that performs a business task or process similar to a human user and uses AI or other autonomous decision-making capabilities

WHY IT MATTERS: IN ACTION

An AI project team did not assign a sponsor or custodian to an AI solution that identifies individuals at risk of developing Type II diabetes and recommends intervention. After the solution was deployed, an end user noticed a problem with the recommendations, but **there was no responsible party for the user to contact about the issue.**



Research and Design | Safe / Secure Principle

Objective: Categorize and embed necessary security protections in model design

SUPPORTING ACTIVITIES

- ◆ Evaluate Identified [Security Risks](#)^{25, 37}
 - ❑ Evaluate the likelihood and potential impact of risks identified in Initiation & Concept to inform security protections
 - ❑ Consider using [MITRE's ATLAS](#) as a framework to create a threat modelling plan
- ◆ Incorporate a **Data Protection and Secure Integration Plan into the Technical Design Documentation**²¹
 - ❑ Evaluate and document data storage and backup requirements
 - ❑ Identify user authentication mechanisms, session management controls, and other applicable access controls to ensure acceptable data use; document controls in the solution's design documentation
 - ❑ Identify network layer protections, data encryption/cryptographic storage mechanisms, and other secure transmission controls to ensure secure integration with dependent systems for all environments (e.g., Dev, Test, Prod); document controls in the solution's technical design documentation
 - ❑ Identify risk monitoring controls to flag potential security threats during the AI lifecycle for all environments (e.g., Dev, Test, Prod); document controls in the solution's technical design documentation
- ◆ Obtain Necessary Approvals²¹
 - ❑ Engage appropriate security stakeholders for clearance
 - ❑ Prepare to obtain an Authority to Operate (ATO) for the AI solution

KEY CONCEPTS

Security Risk³⁸

The level of impact on agency operations (including mission functions, image, or reputation), agency assets, or individuals resulting from the operation of an information system given the potential impact of a threat and the likelihood of that threat occurring

WHY IT MATTERS: IN ACTION

When developing a data protection and secure integration plan, an AI project team failed to account for all environments. As a result, an **adversarial attacker was able to hack into the development environment and modify the input data.**



Research and Design | Privacy Principle

Objective: Understand sensitivity of the data and refine the Privacy Impact Assessment

SUPPORTING ACTIVITIES

● ◆ Refine the Privacy Impact Assessment for Solutions Using Sensitive Data²⁷

- Confirm if the AI solution will use personally identifiable information (PII), protected health information (PHI), business identifiable information (BII), or other sensitive data
- Review the preliminary Privacy Impact Assessment from Initiation & Concept to ensure it accurately reflects what information is collected, why it is collected, how it is intended to be used, and with whom it will be shared
- Describe what opportunities individuals have to decline to provide information
- Explain how the information will be secured, drawing on the [Data Protection and Secure Integration Plan](#)
- Indicate whether a System of Record is being created under the Privacy Act of 1974
- Describe what potential risks to individuals are posed by the AI solution and what mitigation strategies are in place

● ◆ Obtain Approval for the Use of Sensitive Data

- Provide justification for the use of sensitive data
- Obtain necessary approvals according to applicable Op/StaffDiv data privacy policies

● ◆ Identify Applicable Laws and Regulations

- Identify and document privacy laws and regulations that would be applicable to the AI application
- Validate that the model design is consistent with the identified laws and regulations

KEY CONCEPTS

PII³⁹

Information that can be used to distinguish an individual's identity

PHI⁴⁰

Individually identifiable information relating to the health status of an individual

BII⁴¹

Trade secrets and commercial or financial information obtained from a person that is privileged or confidential

WHY IT MATTERS: IN ACTION

In research and design for an AI solution identifying patients at high-risk for cardiovascular disease, a Privacy Impact Assessment wasn't completed. Consequently, important privacy components were missed during the subsequent development stage and **patients' sensitive privacy information was leaked.**



Research and Design | Robust / Reliable Principle

Objective: Understand data quality and ensure that the model design is conceptually sound

SUPPORTING ACTIVITIES

◆ Assess Data Quality^{2, 26, 42}

- Evaluate the quality and reliability of the training and testing data
- Determine if the data, including any proxy attributes, is appropriate and representative of the AI use case
- Document data quality risks and mitigation measures
- Document data lineage (i.e., source, characteristics, relationships to other data, transformations)
- Create a data flow diagram to show how the data flows through the AI system

◆ Select AI Methods⁴²

- Consider the benefits and limitations of different AI methods
- Select the method(s) that support the desired outcomes and are appropriate for the data quality, type, and size
- Provide justification for selected method(s) and the algorithm's hypothesis
- Identify and document implicit model assumptions (e.g., dependent data)

◆ Perform Conceptual Soundness Check

- Assess academic and industry research when selecting model theory and design
- Determine if theoretical limitations have been identified and mitigated
- Critique other design decisions, such as justification for programming or vendor selections

● Review Vendor Documentation of Model Design Decisions

- Assess whether the vendor has sufficiently completed a conceptual soundness check

KEY CONCEPTS

Factors Affecting Data Quality⁴²

1. Sampling - *Is the sample representative of the use case context? Is it large enough?*
2. Completeness – *Are there any missing or incomplete values?*
3. Duplication – *Are there any repeated or duplicate values?*
4. Noise – *How much variation is there? Are there significant outliers?*
5. Accuracy – *Are there any errors?*

WHY IT MATTERS: IN ACTION

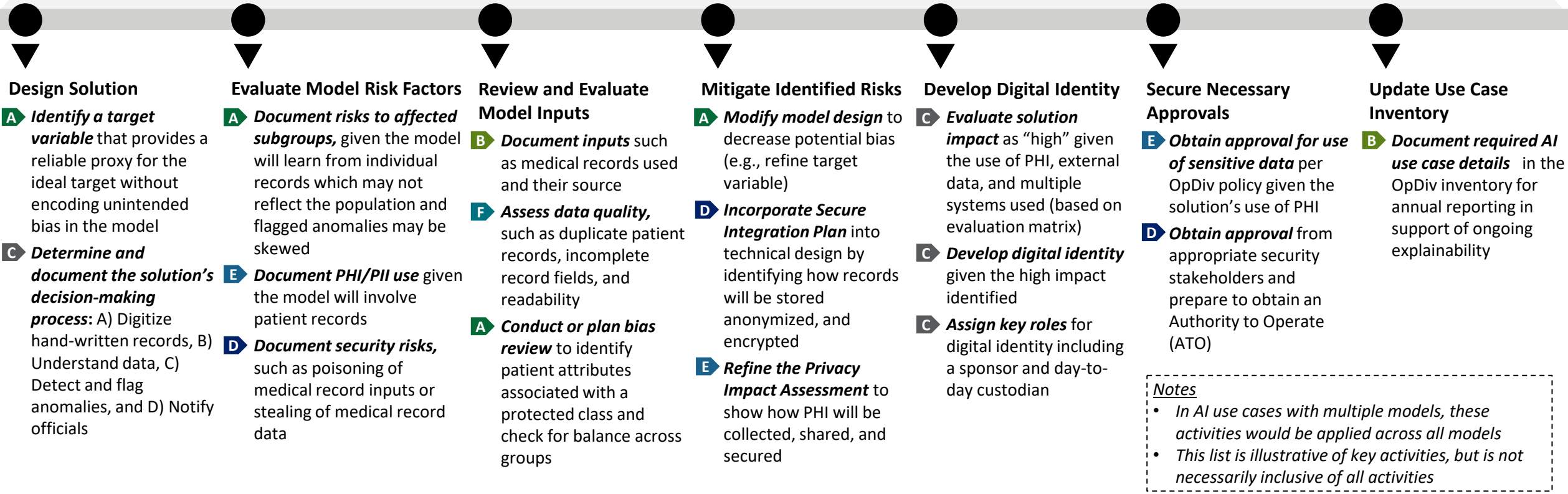
An AI project team obtained a dataset from another business unit to train an AI model. When evaluating the quality of the data, the team did not consider whether the sample was sufficiently large and representative of the use case context. As a result, **the model did not perform well during testing.**

Research and Design | Sample Use Case

AUTOMATED MEDICAL DOCUMENT PROCESSING

An agency receives thousands of handwritten medical documents, which require a significant amount of time and human resources to process. To decrease administrative burden and free up resources for other tasks, the agency decides to build a document processing solution that reads, understands, reviews, and flags anomalies in medical documentation to trigger notification for human review.

During the Research and Design Phase, the program manager oversees the following activities for every model in the AI solution.



Key:

- A** Fair / Impartial
- B** Transparent / Explainable
- C** Responsible / Accountable
- D** Safe/Secure
- E** Privacy
- F** Robust / Reliable

Research and Design | Risk Review Checklist

Purpose of Review #2

After the Research and Design phase, there should be an **Initial Risk Review**. The purpose of this review is to understand the AI model risks and associated controls and define requirements for testing. The technical team then mitigates risk in the model by developing it to meet those requirements. ²⁸



Fair / Impartial

- Has the AI model purpose been clearly defined and aligned with legal and regulatory requirements?
- Have a diverse set of stakeholders provided input on potential fairness-related concerns (e.g., underrepresentation)?
- Has the data been checked for bias and impact on protected classes? Have bias metrics and mitigation strategies been identified?
- What tradeoffs between expected benefits and potential harms have been identified?
- Are tradeoffs mitigated, sufficiently justified, and aligned with regulatory standards?



Transparent / Explainable

- Have the AI logic, inputs, and expected outputs been sufficiently documented?
- Has AI functional and technical documentation been reviewed? From a human-centered design (HCD) perspective, is it understandable to reviewers and stakeholders? Can functional users understand the methodology, outputs, and features/weights?



Responsible / Accountable

- What is the AI solution’s potential adverse impact level?
- For moderate to critical solutions, was a digital identity created, and have both a custodian and a sponsor been identified?
- Is the AI model’s decision-making process sufficiently documented, justified, and defensible to regulators and stakeholders?



Safe / Secure

- Have vulnerabilities to adversarial attacks been sufficiently identified and documented?
- Is the AI model Data Protection and Secure Integration Plan complete and defensible to security stakeholders?



Privacy

- Has the data been checked for sensitive information, including PII, PHI, and BII?
- Has a Privacy Impact Assessment been completed and refined (where needed)?



Robust / Reliable

- Has the data been checked for quality? Have identified issues been effectively mitigated?
- Has the AI model been checked for conceptual soundness?

POTENTIAL OUTCOMES



Approved

The AI model is ready for development. All necessary risk controls have been identified, and requirements for testing are defined and aligned to the AI model purpose/potential risks.



Approved with Conditions

Changes need to be made to the AI model documentation, risk controls, or requirements for testing. With those changes, the model is ready for development.



Declined

The AI model benefits do not outweigh the potential risks, and the risks have not been sufficiently mitigated. The AI model is redesigned, placed on hold, or discarded.







INTERNAL AI DEPLOYMENT CONSIDERATIONS

DEVELOP, TRAIN, & DEPLOY

LIFECYCLE PHASE III



Develop, Train, & Deploy | Overview of How TAI Principles Are Applied

| PRINCIPLE | <p>Fair / Impartial</p>  | <p>Transparent / Explainable</p>  | <p>Responsible / Accountable</p>  | <p>Safe / Secure</p>  | <p>Privacy</p>  | <p>Robust / Reliable</p>  |
|----------------------------------|--|---|--|---|---|---|
| HOW IT'S APPLIED* | <ul style="list-style-type: none"> Continue to comply with applicable laws and regulations Check and mitigate unintended bias in training data, models, and outcomes Incorporate fairness into operational readiness review | <ul style="list-style-type: none"> Use measures and tools to increase explainability Document model outputs Document model performance test plan and results Assess model outputs/explanations Consider independent verification and validation (IV&V) testing | <ul style="list-style-type: none"> Provision digital identity (if applicable) Maintain and use a change access plan during development Consider Independent Verification and Validation (IV&V) testing Obtain approval | <ul style="list-style-type: none"> Employ secure practices for AI configuration and setup Develop and test defenses against adversarial attacks Review vendor documentation Rigorously scan for vulnerabilities Obtain an Authority to Operate (ATO) | <ul style="list-style-type: none"> Finalize the Privacy Impact Assessment (PIA) for solutions using sensitive data Implement privacy protections Test privacy protections Publish a System of Record Notice (if required) | <ul style="list-style-type: none"> Clean the training data Create data quality controls Perform model verification and validation (V&V) testing Establish reliability metrics |
| SAMPLE STAKEHOLDERS ² | <ul style="list-style-type: none"> Individuals affected by the AI solution Legal Counsel Civil Rights, Ethics, and Minority Health / Health Equity Offices | <ul style="list-style-type: none"> Users Third-party Testers Communications and Public Affairs Offices Data and Analytics Offices | <ul style="list-style-type: none"> Op/StaffDiv OCIO System Administrator Third-party Testers | <ul style="list-style-type: none"> HHS OCIO Op/StaffDiv OCIO ISSO or CISO | <ul style="list-style-type: none"> Op/StaffDiv Senior Official of Privacy | <ul style="list-style-type: none"> Users Data and Analytics Offices |

*Note: Activities will be not be applicable for every AI use case



Develop, Train, & Deploy | Fair / Impartial Principle

Objective: Check for and mitigate bias in the AI solution prior to deployment

SUPPORTING ACTIVITIES

● ◆ Continue to Comply with Applicable Laws and Regulations

- ❑ Reference the laws, regulations, policies, and standards identified during the Initiation & Concept phase
- ❑ Ensure model development complies with the identified laws and regulations

◆ Check and Mitigate [Unintended Bias](#) in Training Data, Models, and Outcomes ^{14, 15, 28, 31}

- ❑ Use customized [AI fairness metrics](#) identified in Research & Design to assess bias in training data, models, and outcomes
- ❑ Reevaluate the algorithm's target variable, and reformulate the problem as needed to prevent unfair outcomes
- ❑ Consider and discuss with stakeholders the tradeoffs between fairness and other TAI components (e.g., accuracy)
- ❑ Incorporate bias mitigation [algorithms or techniques](#) (e.g., reweighing inputs) to improve fairness metrics as needed

● ◆ Incorporate Fairness into Operational Readiness Review (ORR) ²

- ❑ Incorporate a bias review as part of the ORR to ensure the model is fair/impartial prior to production
- ❑ Assess dependencies between data and models used to operationalize the AI solution, and check for unexpected correlations that could create bias

KEY CONCEPTS

Areas for Fairness Measurement and Bias Mitigation

- Training Data: Assessing fairness of model inputs (i.e., pre-processing)
- Models: Assessing fairness of how the model works (i.e., in-processing)
- Outcomes: Assessing fairness of model outputs and model learning (i.e., post-processing)

WHY IT MATTERS: IN ACTION

In developing an AI solution for detecting cancer from routine tests, the developers failed to assess the solution against the equal odds fairness metric identified. Consequently, **members of the Pacific Island community were less likely to have cancer cases accurately flagged.**



Develop, Train, & Deploy | Transparent / Explainable Principle

Objective: Create and enhance explanations for the model's outputs

SUPPORTING ACTIVITIES

Use Measures and Tools to Increase Explainability (if applicable)

- Enable intermediate model outputs to better understand model behavior
- Generate explanations for individual model outputs and global model behavior using applicable [interpretation methods](#)

Document Model Outputs

- Describe the model outputs, including how they are used and, if applicable, transformed for the model purpose
- Document how the model is applied to a given scenario to obtain results

Document Model Performance Test Plan and Results

- Create a test plan outlining the test steps (based on Op/StaffDiv testing methodology), the criteria that need to be met to advance to the next step, and the rationale for those criteria
- [Conduct model performance testing](#) according to the test plan, and collect and store all testing evidence

Assess Model Outputs and Explanations¹⁸

- Employ a human centered technology design (HCTD) approach to support the development of model outputs that are sufficiently clear and comprehensible to a diverse group of stakeholders
- Evaluate whether model explanations satisfy the established explainability requirements
- Ensure that model explanations include, at a minimum, the type and source of model input data, the high-level data transformation process, the decision-making criteria and rationale, and risks and mitigation measures
- Identify and mitigate explainability risks using metrics and appropriate documentation

Consider Independent Verification and Validation (IV&V)* Testing⁴³

- Use IV&V testing to validate that the model's actions and outputs are understandable
- Define the IV&V scope, conduct applicable tests, and create a plan to remediate findings

KEY CONCEPTS

Interpretation Methods⁴⁴

Techniques that provide insight into how an algorithm arrived at its conclusion by revealing correlations made within a model

Human-Centered Technology Design⁴⁵

The integration of human sciences with computer science to design computing systems with a human focus to support human activities and needs

WHY IT MATTERS: IN ACTION

After several years in use, an AI benefits fraud detection solution was challenged by a beneficiary in federal court. Due to insufficient documentation of outputs and testing and a lack of external validation, **the agency was held legally liable for biased fraud detection practices.**



Develop, Train, & Deploy | Responsible / Accountable Principle

Objective: Capture identity management information, and track all development activities

SUPPORTING ACTIVITIES

● ◆ Provision Digital Identity (if applicable)²⁰

- Capture identity management information to uniquely identify a digital worker
- Track the last acknowledgement date and recertification date for both the sponsor and the custodian

● ◆ Maintain and Use a Change Access Plan During Development

- Describe the version control mechanisms that are in place to protect the integrity of the model, track which developers wrote each section of code, and track who approved changes
- Provide detailed remediation plans if no process or controls are in place

● ◆ Consider Independent Verification and Validation (IV&V)* Testing⁴³

- As an audit best practice, use IV&V testers to verify and sign off on the model’s performance
- Define the IV&V scope, including fairness and explainability components (e.g., compliance with applicable regulations)
- Conduct applicable tests and create a plan to remediate findings

● ◆ Obtain Approval

- Ensure all necessary parties, including the sponsor, review the test outcomes and provide approval before deployment

KEY CONCEPTS

Last Acknowledgement Date²⁰

The date the sponsor / custodian acknowledged responsibility for the AI solution

Recertification Date²⁰

The date on which the sponsor / custodian must re-acknowledge responsibility for the AI solution

WHY IT MATTERS: IN ACTION

An AI solution malfunctioned and began releasing confidential healthcare organization data via email. Since the agency did not create an appropriate digital identity for the solution, **agency officials were delayed in taking the solution offline to minimize the data leak.**



Develop, Train, & Deploy | Safe / Secure Principle

Objective: Protect the AI model against attacks and scan for vulnerabilities

SUPPORTING ACTIVITIES

◆ Employ Secure Practices for AI Configuration and Setup²¹

- Incorporate strong user authentication, session management, and other access controls per the Data Protection Plan
- Incorporate network layer protections and other secure transmission controls per the Secure Integration Plan
- Encrypt all communications and data at rest and in transit per applicable HHS and Op/StaffDiv policies (e.g., HHS IS2P, Standard for Encryption of Computing Devices and Information)
- Use open-source code from whitelisted libraries only

◆ Develop and Test Defenses Against Adversarial Attacks

- Use whitelisted sources for training data to prevent data poisoning attacks
- Use applicable [techniques and tools](#) to protect the model from adversarial attacks during training and testing
- Conduct penetration testing by simulating adversarial attacks, document results, and improve defenses where needed

● Review Vendor Documentation

- Obtain and assess vendor documentation of security controls
- Verify that the model includes applicable [defense mechanisms](#)

● ◆ Rigorously Scan for Vulnerabilities²¹

- Scan the AI solution for vulnerabilities in all levels of its stack (e.g., software level, algorithm level) according to applicable HHS and Op/StaffDiv policies (e.g., HHS Policy for Vulnerability Management)
- Document and mitigate identified vulnerabilities

● ◆ Obtain an Authority to Operate (ATO)²³

- Prior to implementation, obtain an ATO in compliance with IS2P and other relevant HHS and Op/StaffDiv policies

KEY CONCEPTS

Whitelisting⁴⁶

A list of discrete entities, such as hosts, email addresses, network port numbers, runtime processes, or applications that are authorized to be present or active on a system according to a defined baseline

Adversarial Machine Learning

A set of techniques that attempt to fool models by supplying deceptive inputs

WHY IT MATTERS: IN ACTION

In developing an AI solution for beneficiary payment disbursement, user authentication and session management practices were not followed. Unknown to developers, **an adversary incorporated code that redirected a portion of funds to an offshore bank account**, costing the agency millions of dollars.



Develop, Train, & Deploy | Privacy Principle

Objective: Protect sensitive data during training, testing, and deployment

SUPPORTING ACTIVITIES

- ◆ **Finalize the Privacy Impact Assessment (PIA) for Solutions Using Sensitive Data**²⁷
 - ❑ Obtain approval of the PIA from the Op/StaffDiv Senior Official of Privacy (SOP) or designee
 - ❑ Submit approved PIA form to the HHS Senior Agency Official of Privacy (SAOP)
- ◆ **Implement Privacy Protections**
 - ❑ Strip, mask, and encrypt sensitive data before directly exposing it to generic processes in the AI workflow
 - ❑ Limit sensitive data transfer to third-party environments
 - ❑ Add application-based data pull controls for cloud-based AI solutions
 - ❑ Employ [privacy protection methods](#) (where applicable)
- ◆ **Test Privacy Protections**⁴⁷
 - ❑ Test privacy protections, including those outlined in the technical design documentation
 - ❑ Document associated risks and mitigation measures
- ◆ **Publish a System of Record Notice (if required)**²³
 - ❑ Prepare and publish a System of Record Notice (SORN) in the Federal Register if the PIA concluded that the AI solution would create a System of Record (SOR)

KEY CONCEPTS

System of Record²³

A group of any records under the control of a federal agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual

WHY IT MATTERS: IN ACTION

An agency developed an AI solution to expedite its drug approval process. However, patients' clinical trial data wasn't stripped and/or masked prior to feeding it into the model. **When a data breach occurred, thousands of clinical trial patients' data was released.**



Develop, Train, & Deploy | Robust / Reliable Principle

Objective: Rigorously train and test the model, and establish controls to monitor data quality and performance in production

SUPPORTING ACTIVITIES

◆ Clean the Training Data^{26, 48}

- Prepare the training data using a data sanitation tool or manual procedures
- Validate that the dataset is still sufficiently large and representative of the model scope after cleaning

● ◆ Create Data Quality Controls^{20, 49, 50}

- Develop controls to monitor the existence of data drift and the percentage of missing data, outliers, and wrong types
- Create signals to indicate when an AI solution cannot reliably perform a requested function

● ◆ Perform Model Verification and Validation (V&V) Testing^{42, 48}

- Execute the [test plan](#) using applicable [testing approaches](#) and data that is representative of the real-world setting
- Verify satisfiability and robustness
- Identify and address performance skews (i.e., differences between model performance during training and testing)
- Identify and mitigate potential risks that could decrease model performance in the future
- Benchmark model results against alternative internal and external data and/or models (e.g., alternative assumptions, model parameters) and document any material deviations

● ◆ Establish Reliability Metrics^{2, 26}

- Consider the cost of errors, or inaccurate predictions (e.g., financial, social)
- Select and determine acceptable thresholds for [model performance metrics](#) (e.g., precision, recall) in production
- Create plans to continuously monitor the AI solution against pre-defined thresholds post-deployment
- Consider metrics for determining when to retire the AI solution

KEY CONCEPTS

Data Drift

Changes in the distribution of underlying training or target data that leads to poor model performance

Satisfiability⁴⁸

Achieved when a given input can produce a certain output

Robustness⁴⁸

Achieved when adding noise to an input does not materially change its output

WHY IT MATTERS: IN ACTION

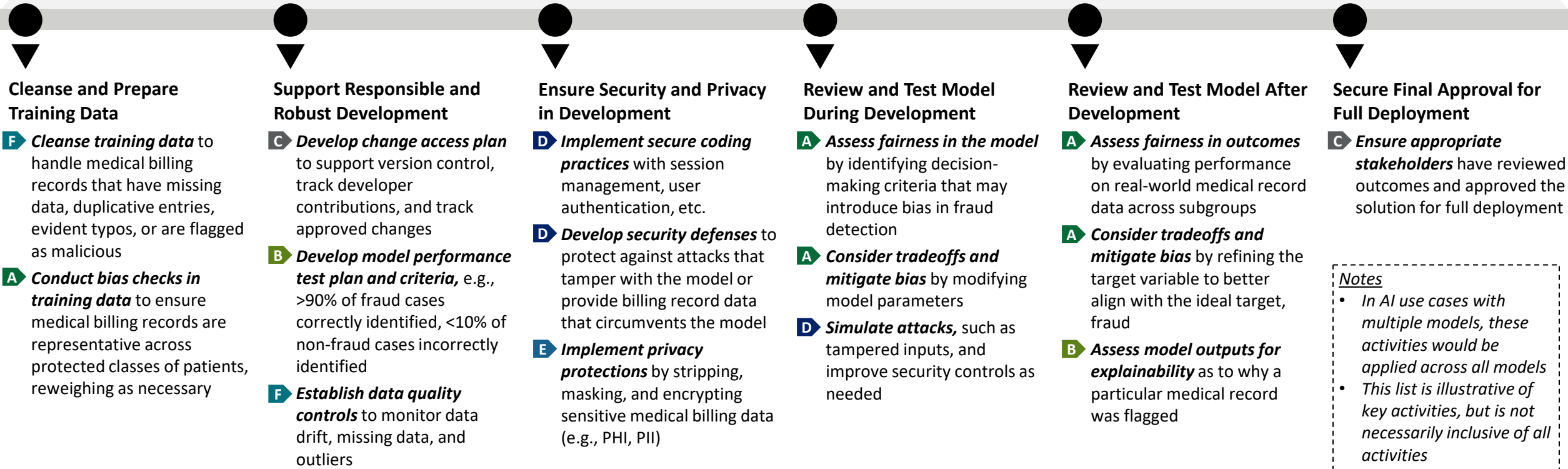
An AI project team did not conduct sufficient model verification and validation testing before deploying an AI-based medical device into production. As a result, the device was **highly sensitive to noise and received criticism from providers.**

Develop, Train, & Deploy | Sample Use Case

FRAUD DETECTION IN MEDICAL BILLING

An agency is responsible for reimbursing billions of dollars for the provision of health services to underserved populations. To mitigate the risk of fraudulent reimbursements, the agency is developing an AI solution to detect and flag potential fraud in medical billing for further investigation by an agency representative.

During the Develop, Train, & Deploy Phase, the program manager oversees the following activities for every model in the AI solution.



Notes

- In AI use cases with multiple models, these activities would be applied across all models*
- This list is illustrative of key activities, but is not necessarily inclusive of all activities*

Key:

- A** Fair / Impartial
- B** Transparent / Explainable
- C** Responsible / Accountable
- D** Safe/Secure
- E** Privacy
- F** Robust / Reliable

Develop, Train, & Deploy | Risk Review Checklist

Purpose of Review #3

Near the end of the Develop, Train, & Deploy phase, there should be a **Risk Validation Review**. The purpose of this review is to validate that all conditions and testing requirements defined in the Initial Risk Review have been met before the model is deployed into production. ²⁸



Fair / Impartial

- Is the model compliant with applicable legal and regulatory standards?
- Has bias in the training data, model, and outcomes been examined, documented, and sufficiently mitigated?
- What are the potential tradeoffs between reducing unintended bias and accuracy? Have decisions to reduce unintended bias at the expense of model accuracy been documented and justified?
- What are the potential consequences of unintended bias? Have decisions to accept bias in the model been documented and justified?



Transparent / Explainable

- Are model outputs and explanations clear and comprehensible to stakeholders? Was IV&V testing considered to validate this?
- Do the explanations for model outputs provide sufficient information about the AI logic, inputs, and limitations?
- Were the model performance test plan and testing results documented?



Responsible / Accountable

- For solutions with a digital identity, have all identity management system data fields been captured?
- Was a change access plan created and used during development? Are changes sufficiently documented?
- Was IV&V testing considered to verify model performance? Have model performance testing results been reviewed and approved?



Safe / Secure

- Have all applicable security controls been implemented according to the Data Protection and Secure Integration Plan?
- Is the AI model sufficiently resilient to adversarial attacks? Have penetration testing results been documented and reviewed?
- Has the solution been scanned for vulnerabilities in all levels of its stack, and have identified vulnerabilities been mitigated?



Privacy

- Were applicable privacy protections used during training and testing?
- Have privacy risks been identified, documented, and mitigated?



Robust / Reliable

- Have applicable data sanitation mechanisms and data quality controls been identified, documented, and implemented?
- Were all relevant performance tests and reviews completed? Do the results satisfy the defined testing outcomes?
- Have performance risks been identified and mitigated, and have reliability metrics been defined for ongoing monitoring?

POTENTIAL OUTCOMES



Approved

The AI model is ready for deployment. All risks have been successfully mitigated, and the model achieved the defined testing requirements.



Approved with Conditions

Additional steps need to be taken to mitigate identified risks. Once those steps are complete, the model is ready for deployment.



Declined

The model cannot satisfy all conditions for risk mitigation or model performance. The model is placed on hold or retired.







INTERNAL AI DEPLOYMENT CONSIDERATIONS

OPERATE & MAINTAIN

LIFECYCLE PHASE IV



Operate & Maintain | Overview of How TAI Principles Are Applied

| PRINCIPLE | <p>Fair / Impartial</p>  | <p>Transparent / Explainable</p>  | <p>Responsible / Accountable</p>  | <p>Safe / Secure</p>  | <p>Privacy</p>  | <p>Robust / Reliable</p>  |
|----------------------------------|---|---|--|---|--|--|
| HOW IT'S APPLIED* | <ul style="list-style-type: none"> Continue to comply with applicable laws and regulations Check and mitigate unintended bias in outcomes | <ul style="list-style-type: none"> Establish a change management process Maintain Op/StaffDiv AI Use Case Inventory Publish and regularly review model performance information | <ul style="list-style-type: none"> Establish an incident management process Recertify key roles Collect third party documentation (if applicable) | <ul style="list-style-type: none"> Develop O&M plans that support the AI solution's safety and security Maintain Authority to Operate (ATO) | <ul style="list-style-type: none"> Routinely evaluate the Privacy Impact Assessment (PIA) for solutions using sensitive data Monitor the storage and privacy of sensitive information Manage data inputs and outputs from a privacy perspective | <ul style="list-style-type: none"> Continuously monitor and improve model performance Retire the AI solution if deemed appropriate |
| SAMPLE STAKEHOLDERS ² | <ul style="list-style-type: none"> Individuals affected by the AI solution Legal Counsel Civil Rights, Ethics, and Minority Health / Health Equity Offices | <ul style="list-style-type: none"> Users Communications and Public Affairs Offices Data and Analytics Offices | <ul style="list-style-type: none"> Op/StaffDiv OCIO System Administrator | <ul style="list-style-type: none"> HHS OCIO Op/StaffDiv OCIO ISSO or CISO | <ul style="list-style-type: none"> Op/StaffDiv Senior Official of Privacy | <ul style="list-style-type: none"> Users Data and Analytics Offices |

*Note: Activities will be not be applicable for every AI use case



Operate & Maintain | Fair / Impartial Principle

Objective: Continue to check for and mitigate bias in the AI solution on an ongoing basis

SUPPORTING ACTIVITIES

● ◆ Continue to Comply with Applicable Laws and Regulations

- ❑ Validate that the regulations, standards, policies, and laws documented in Initiation & Concept are still applicable
- ❑ Identify and document new or upcoming regulations, standards, policies, or laws that may impact use of the AI solution
- ❑ Ensure the AI solution continues to comply with the identified laws and regulations

◆ Check and Mitigate [Unintended Bias](#) in Outcomes

- ❑ Assess AI fairness in outcomes using the customized set of [AI fairness metrics](#) identified in Research & Design
- ❑ Incorporate bias mitigation [algorithms or techniques](#) to improve fairness metrics for outcomes as needed
- ❑ Engage a diverse set of stakeholders to support ongoing bias detection and mitigation

KEY CONCEPTS

Bias Mitigation Algorithms

Set of open-source tools that can be used to address unintended bias in AI models (e.g., systematically less favorable outcomes for individuals in a protected class)

WHY IT MATTERS: IN ACTION

After deploying an AI-based solution into production, the team did not effectively monitor the solution's outcomes for bias. The agency **received negative media attention and criticism from the public** when it was discovered that the solution **negatively impacted rural communities' access to resources.**



Operate & Maintain | Transparent / Explainable Principle

Objective: Provide ongoing explanations of the model and its outputs to stakeholders

SUPPORTING ACTIVITIES

◆ Establish a Change Management Process⁴²

- Create a change control forum to approve or dismiss changes
- Evaluate the differences between the current and proposed model
- Evaluate the impact of the change on system quality and user experience
- Document all changes in a change log containing the change history; model versions; the change(s) made and rationale; and review and approval dates
- Update the design documentation to reflect the change
- Test the performance of the model before and after a change to verify that the change met the requirements

◆ Maintain Op/StaffDiv AI Use Case Inventory (additional guidance to be provided)

- Review and refresh the Op/StaffDiv use case inventory annually (or any time the solution details change)
- Verify that all AI use case updates comply with applicable Op/StaffDiv data sharing policies
- Submit to the HHS OCAIO by the annual reporting deadline, and respond to any inquiries in a timely fashion
- Support other AI efforts across Op/StaffDivs with guidance, lessons learned, and reusable models/code where appropriate

◆ Publish and Regularly Review Model Performance Information^{15, 49}

- Communicate model performance metrics (e.g., accuracy, precision, recall) to stakeholders
- Collect and address stakeholder feedback on model performance and explainability
- Regularly review model performance information to ensure it is complete and accurate
- Where feasible, notify end users when the AI solution makes a mistake or when the solution is enhanced due to performance issues

KEY CONCEPTS

HHS AI Use Case Inventory

HHS is required to maintain an up-to-date inventory of all AI uses cases. It is important that this inventory be updated annually with any post-production changes to the AI solution.

WHY IT MATTERS: IN ACTION

A team managing an AI-based population health management tool made multiple changes to the model parameters but failed to communicate and explain the changes to end users, which **created confusion and caused end users to lose trust in the tool.**



Operate & Maintain | Responsible / Accountable Principle

Objective: Establish governance processes for ongoing monitoring and incident mitigation

SUPPORTING ACTIVITIES

● ◆ Establish an Incident Management Process

- ❑ Assign responsibilities for detecting and responding to incidents and communicating to relevant stakeholders
- ❑ Consider formal Service-Level Agreements (SLAs) between the sponsor and custodian for moderate to critical AI solutions
- ❑ If changes are required to resolve an incident, follow the [change management process](#) for the AI solution

● ◆ Recertify Key Roles²⁰

- ❑ Recertify the sponsor and custodian annually for medium to high adverse impact level solutions or every six months for critical adverse impact level solutions

● Collect Third Party Documentation (if applicable)

- ❑ Obtain and maintain vendor communication logs and data request and receipt logs with the solution's production data

KEY CONCEPTS

Sponsor/Custodian Recertification

Process for re-acknowledging responsibility for the AI solution; While this should take place annually or biannually at a minimum, it should also happen any time a sponsor or custodian leaves the organization or transfers to a new role so that key roles are always in place for the AI solution

WHY IT MATTERS: IN ACTION

An agency uses an AI-based chatbot to address routine questions about its programs. Suddenly, the chatbot began malfunctioning and with no clear responsibilities in place to manage incidents, it **took several days for the agency to identify and resolve the root cause of the problem, disrupting operations.**



Operate & Maintain | Safe / Secure Principle

Objective: Continuously monitor the safety and security of the solution through O&M plans

SUPPORTING ACTIVITIES

◆ Develop O&M Plans that Support the AI Solution's Safety and Security²¹

O&M Plans should include the following components:

Identity and Access Management⁵¹

- Employ a “Least Functionality” approach with a “deny-all, permit by exception” rule
- Reference [NIST SP 800-53](#) for additional guidance on “Least Functionality”

Vulnerability Management

- Vet AI technology for reliability and scan for vulnerabilities per the HHS Policy for Vulnerability Management

Application Whitelisting⁵²

- Maintain the application whitelisting architecture, policies, software, and other solution components
- Reference [NIST SP 800-167](#) for additional guidance on whitelisting

Network Behavior Analysis (NBA)⁵³

- Ensure appropriate network architecture to monitor network traffic for potential threats related to bots and botnets and block malicious bots and botnets
- Reference [NIST SP 800-94](#) for additional guidance on NBA

Automated Security Tools

- Use automated tools to promptly identify malicious source-code modifications and debugging at runtime

◆ Maintain Authority to Operate (ATO)²³

- Complete a periodic Security Authorization
- As needed, obtain a new ATO in compliance with IS2P and other relevant HHS and Op/StaffDiv policies

KEY CONCEPTS

Security Authorization²³

Periodic re-evaluation of the management, operational, and technical information security controls implemented for an information system that is performed during the Operations & Maintenance Phase to ensure that the system is continuing to operate at an acceptable risk level

WHY IT MATTERS: IN ACTION

Without a proper plan in place for Network Behavior Analysis, a team managing an AI solution **did not identify a malicious botnet that caused confidential program information to be compromised.**



Operate & Maintain | Privacy Principle

Objective: Routinely monitor and evaluate privacy practices

SUPPORTING ACTIVITIES

● ◆ Routinely Evaluate the Privacy Impact Assessment (PIA) for Solutions Using Sensitive Data²⁷

- For AI solutions that collect PII or PHI, review and update the PIA at least once every three years
- Review and update the PIA upon any major changes to the solution or the electronic information collected

● ◆ Monitor the Storage and Privacy of Sensitive Information³⁹

- Conduct frequent and random audits at vendor sites handling or storing sensitive data (e.g., PII, PHI)
- Regularly monitor AI use to identify unauthorized attempts to access PII/PHI in the underlying data
- Report suspected or confirmed data breaches of PII or PHI to Op/StaffDiv and/or HHS leadership as soon as possible and without unreasonable delay

● ◆ Manage Data Inputs and Outputs from a Privacy Perspective

- Maintain consent forms and publish privacy notices
- Retain and/or dispose of data in accordance with applicable HHS and Op/StaffDiv policies

KEY CONCEPTS

Data Breach³⁹

Unauthorized use, disclosure, or loss of data; breaches of PII/PHI can cause physical and/or digital harm to affected individuals and result in a loss of public trust with the potential to impede HHS' ability to carry out its mission

WHY IT MATTERS: IN ACTION

The custodian of an AI solution that uses Electronic Health Record (EHR) data noticed unusual activity in the backend of the system. Because the custodian failed to report the unusual activity in a timely manner, it **continued unabated and resulted in stolen EHR data from hundreds of health systems.**



Operate & Maintain | Robust / Reliable Principle

Objective: Ensure the model continues to operate with predictability and accuracy over time

SUPPORTING ACTIVITIES

Continuously Monitor and Improve Model Performance ^{2, 15, 22}

- Assess model performance according to the monitoring plan and mitigate associated risks
- Monitor the model for post-production data drift to determine whether the training data remains relevant for use of the model on production data over time
- Monitor the model for data contamination caused by adversarial attacks
- Assess the potential for model risk to impact downstream models or operational processes
- Confirm the model's output is being used for its intended purpose
- Regularly re-evaluate the model and engage stakeholders to identify opportunities to enhance model performance
- If needed, identify conditions under which scaled or expanded use of the AI solution is appropriate by comparing model performance across use cases

Retire the AI Solution if Deemed Appropriate ^{2, 23}

- 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
- Follow applicable HHS and Op/StaffDiv policies for disposition (e.g., migrating or archiving data)

KEY CONCEPTS

Post-Production Data Drift

Factors that cause changes in the distribution of underlying production data; leads to poor model performance due to differences between current production data and the data that the model was originally trained on

WHY IT MATTERS: IN ACTION

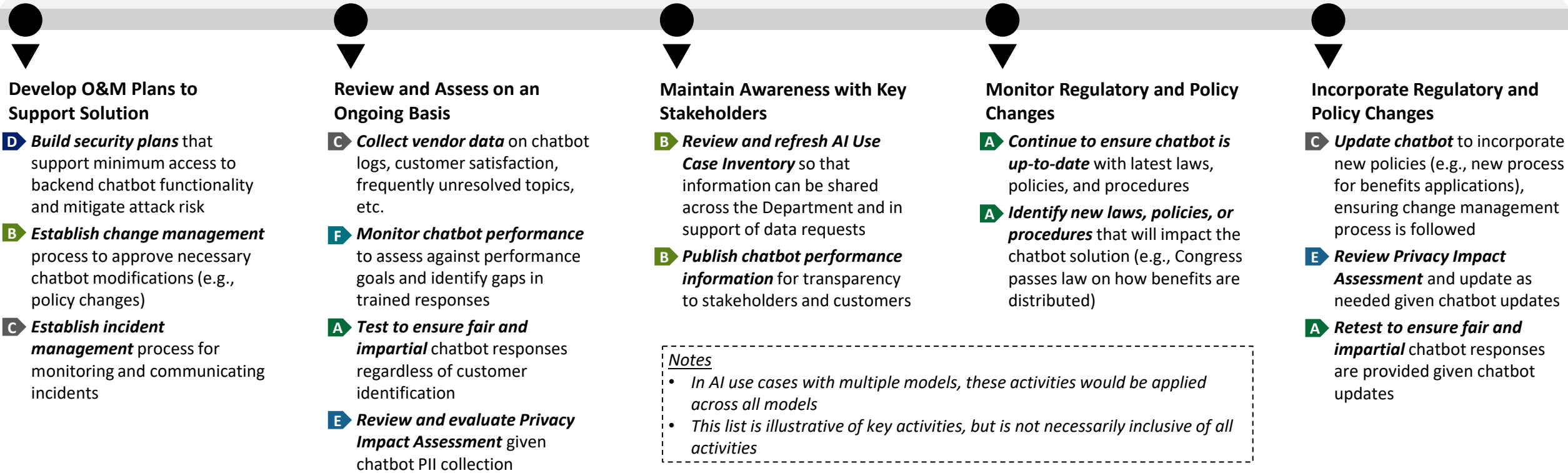
The custodian of an AI solution did not monitor for post-production data drift. As a result, the **solution's predictions became less accurate over time**. The organization had to **pause the solution's use** and reactively address the performance issue by engaging a team to retrain the model on new data.

Operate & Maintain | Sample Use Case

CHATBOT


An agency that is responsible for benefits distribution successfully implemented an AI chatbot solution as a first line of defense in responding to customer inquiries. A year later, Congress passed a law that changed how the agency’s benefits are distributed. The agency noticed a significant increase in customer inquiries and needed to update the chatbot to effectively respond to questions about the new policy.

During the Operate & Maintain Phase, the program manager oversees the following activities for every model in the AI solution.




Operate & Maintain | Risk Review Checklist


Purpose of Review #4
 During the Operate & Maintain phase, there should be **Routine Reviews** at least once per year. The purpose of these reviews is to regularly identify and mitigate new risks that arise after the solution is deployed into production. ²⁸

- 


Fair / Impartial

 - Have new or upcoming regulations, standards, policies, and laws and their potential impact on the AI solution been identified and documented? Have all necessary actions to maintain compliance been identified and assigned to owners?
 - Have bias detection and mitigation activities been completed, documented, and reviewed on a regular basis?
- 


Transparent / Explainable

 - Is there a change management process for the solution? Have all changes been documented, reviewed, and tested?
 - Will it be clear to anyone auditing the AI or consumers of the AI’s outputs which version was in use at any given moment in time? Will it be clear which iteration of a model drew on which data and produced what outputs?
 - Are model explanations and performance metrics communicated to stakeholders? Do they understand and trust the information?
 - Is stakeholder feedback routinely collected and used to continuously improve model performance and explainability?
- 


Responsible / Accountable

 - Is there an incident management process for the solution? Have all incidents been documented and resolved according to SLAs?
 - Were the sponsor and custodian recertified by their respective recertification dates?
- 

Safe / Secure

 - Is the solution continuously monitored for security threats? Are automated security tools used where applicable?
 - Have any changes to the solution introduced new security risks? Were those risks documented and mitigated?
 - Are there robust procedures in place for identity and access management, vulnerability management, application whitelisting, and network behavior analysis? Are those procedures reliably executed?
- 

Privacy

 - Have there been any major changes to the solution? If so, was the PIA reviewed and updated?
 - Have all applicable HHS and Op/StaffDiv policies related to the use, storage, and disposition of sensitive data been followed?
- 

Robust / Reliable

 - Does the solution satisfy the original business intent? Are outputs consistent, reliable, and used for their intended purpose?
 - Is model performance continuously monitored against defined metrics? Have performance issues caused by data drift, changes in the operational/business environment, or other factors been identified and addressed? Is stakeholder feedback incorporated?

POTENTIAL OUTCOMES

Approved
The AI model is still fit for its intended purpose, and all identified risks have been successfully mitigated.

Approved with Conditions
The AI model is still fit for its intended purpose, but additional steps need to be taken to mitigate identified risks.

Declined
The model is no longer fit for its intended purpose, and the potential risks outweigh the benefits. The model is paused or retired.

CHAPTER V

EXTERNAL AI CONSIDERATIONS



How to Use This Section

In addition to modeling TAI development, HHS has an opportunity to foster TAI innovation through AI-related regulatory and non-regulatory actions.

HHS TRUSTWORTHY AI PLAYBOOK | 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.

| | | |
|---|---|---|
| <p>Fair / Impartial</p> <p>Could the use of AI in this area result in discriminatory outcomes? How accessible are AI systems and algorithms in this area to learning and propagating bias?</p> <p>Sample regulatory application: Bias reviews and/or metrics for AI data, models, and outcomes</p> | <p>Transparent / Explainable</p> <p>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?</p> <p>Sample regulatory application: Information disclosure requirements</p> | <p>Responsible / Accountable</p> <p>What are essential outcomes of AI in this area? How should accountable and responsible parties be identified and acknowledged?</p> <p>Sample regulatory application: Traceability and/or accountability requirements</p> |
| <p>Safe / Secure</p> <p>What risks do AI systems and algorithms in this area pose, and what type of control and/or design team might those risks cause?</p> <p>Sample regulatory application: Minimum standard security controls</p> | <p>Privacy</p> <p>Do AI systems and algorithms in this area use sensitive data and generate actions on individuals that could lead to privacy concerns?</p> <p>Sample regulatory application: De-identification requirements for PHI</p> | <p>Robust / Reliable</p> <p>What measures for reliability and consistency do AI systems and algorithms in this area need to meet? How should inaccuracies and unintended outcomes be handled?</p> <p>Sample regulatory application: Model performance thresholds</p> |

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.

*Refer to the Appendix for a list of entities that may authorize Op/StaffDivs to regulate areas that affect AI applications.

Regulatory Considerations

The first set of considerations provides guiding questions to help Op/StaffDivs identify opportunities to incorporate TAI principles into regulations that will affect AI applications.

HHS TRUSTWORTHY AI PLAYBOOK | EXTERNAL AI CONSIDERATIONS

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. 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 | <ul style="list-style-type: none"> Allow pilot programs (i.e., incubators, test sites, challenges, and other pilot programs) to encourage AI innovation Incorporate AI use and TAI principles into grant and research opportunities Use grant reviews, evaluations, and award processes to identify AI applications Collect data on the design, development, deployment, operations, and outcomes of pilot AI applications to better understand AI risks and benefits | <ul style="list-style-type: none"> Issue voluntary, non-regulatory policy statements within existing statutory authority Provide, create, or build upon standards, tools, frameworks, and guidelines to encourage AI understanding and innovation Align standards, frameworks, and tools to TAI principles Leverage private sector conformity assessment programs and related activities before proposing regulations or compliance programs | <ul style="list-style-type: none"> Increase access to government data and models where appropriate Review existing data disclosure protocols and identify systematic ways to do more Explore opportunities to provide granular, anonymized data rather than aggregate data Consider how to balance legal and policy requirements for protecting sensitive data | <ul style="list-style-type: none"> Communicate AI risks and benefits, including how external groups are impacted, to support understanding of and trust in AI Promote non-regulatory consensus standards, frameworks, and guidance Share lessons and lessons learned from pilot programs where appropriate Ensure that AI is subject to AI use informed by agency risk assessments, control specific, and based on sound scientific evidence |
| OTHER RELEVANT GUIDANCE | N/A | OMB Circulars 1319 | Executive Order 13859 National Technology Transfer and Advancement Act Open, Public, Electronic, and Necessary Information (OPEN) Act OMB Circular 5-10 OMB Memorandum 01-13-13 NIST Plan for Federal Assessment in Operations, Technical Standards and Related Tools | Office of Science and Technology Policy Memorandum on Scientific Integrity |

Non-Regulatory Considerations

The second set of considerations summarize four non-regulatory approaches described in OMB M-21-06. Op/StaffDivs can use the considerations to identify policies, investments, or other non-regulatory actions to encourage TAI development.

PERCEPTION OF GOVERNMENT'S ROLE IN PROMOTING TAI ⁵⁴

A recent survey of 250 industry leaders indicated that:



70% of respondents support government investment in fundamental AI research



83% of respondents agree that government investments could enable trustworthy AI innovation to a great or some extent

This section will help Op/StaffDivs determine appropriate regulatory and non-regulatory actions to promote TAI.

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.*

*Refer to the Appendix for a [list of statutes](#) that may authorize Op/StaffDivs to regulate areas that affect AI applications.

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 | <ul style="list-style-type: none"> Allow pilot programs (i.e., hackathons, tech sprints, challenges, and other pilot programs) to encourage AI innovation Incorporate AI use and TAI principles into grant and research opportunities Issue grant waivers, deviations, and exemptions for specific AI applications Collect data on the design, development, deployment, operations, and outcomes of pilot AI applications to better understand AI risks and benefits | <ul style="list-style-type: none"> Issue voluntary, non-regulatory policy statements within existing statutory authority Promote, create, or build upon datasets, tools, frameworks, and guidelines to accelerate AI understanding and innovation Align standards, frameworks, and tools to TAI principles Leverage private-sector conformity assessment programs and related activities before proposing regulations or compliance programs | <ul style="list-style-type: none"> Increase access to government data and models where appropriate Review existing data disclosure protocols and identify systematic ways to share data Explore opportunities to provide granular, anonymized data rather than aggregate data Continue to follow legal and policy requirements for protecting sensitive data | <ul style="list-style-type: none"> Communicate AI risks and benefits, including how external groups are impacted, to support understanding of and trust in AI Promote non-regulatory consensus standards, frameworks, and guidance Share trends and lessons learned from pilot programs where appropriate Ensure that RFIs related to AI are informed by agency risk assessments, context-specific, and based on sound scientific evidence |
| OTHER RELEVANT GUIDANCE | N/A | OMB Circular A-119 | Executive Order 13859 National Technology Transfer and Advancement Act Open, Public, Electronic, and Necessary Government Data Act OMB Circular A-130 OMB Memorandum M-13-13 NIST Plan for Federal Engagement in Developing Technical Standards and Related Tools | Office of Science and Technology Policy Memorandum on Scientific Integrity |

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

A photograph of three medical professionals in a clinical setting. In the center, a man in a white lab coat with a stethoscope around his neck holds a blue folder. To his left, a man in blue scrubs looks down at the folder. To his right, another man in blue scrubs and glasses also looks at the folder. The background shows a hallway with white doors. The image has a blue color overlay.

APPENDIX I

BACKGROUND INFORMATION

Federal Actions Accelerating Trustworthy AI Adoption

Recent federal actions established the development of trustworthy AI solutions as a national priority.

FEBRUARY 11, 2019

Executive Order 13859: Maintaining American Leadership in Artificial Intelligence⁵⁵

Announced the American AI Initiative

*“The United States must foster **public trust and confidence in AI technologies** and protect civil liberties, privacy, and American values in their application in order to fully realize the potential of AI technologies for the American people.”*

FEBRUARY 26, 2020

American AI Initiative: Year One Annual Report⁵⁶

Provided a summary of progress and long-term vision for the American AI Initiative, emphasizing the need to **“embrace trustworthy AI for government services and missions.”**

NOVEMBER 17, 2020

Office of Management and Budget (OMB) Memorandum M-21-06: Guidance for Regulation of Artificial Intelligence Applications¹³

Provided guidance to Federal agencies, including considerations for reducing barriers to AI development and adoption

*“The government’s regulatory and non-regulatory approaches to AI should contribute to public trust in AI by **promoting reliable, robust, and trustworthy AI applications.**”*

DECEMBER 3, 2020

Executive Order 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government¹

Outlined a set of principles and actions to accelerate trustworthy AI use and development

*“Agencies must therefore design, develop, acquire, and use AI in a manner that **fosters public trust and confidence** while protecting privacy, civil rights, civil liberties, and American values”*

Non-AI Building Block | Blockchain ⁵⁷

Blockchain technology is a non-AI building block present in many AI solutions.

What is it?

Blockchain is a **decentralized, distributed ledger of transactions** that record, verify, and maintain information across a network.

Why is it not AI?

While blockchain is an innovative technology, it **relies on a network of devices** rather than cognitive or learned tasks.

How can it be used with AI?

Blockchain can act as a **“master brain” for AI solutions**, serving as a database and system of record across AI solutions.

EO 13960 | Principles for Use of AI in Government ¹

| Principle | Description |
|--|--|
| 1. Lawful and Respectful of Our Nation's Values | Agencies shall design, develop, acquire, and use AI in a manner that exhibits due respect for our Nation's values and is consistent with the Constitution and all other applicable laws and policies, including those addressing privacy, civil rights, and civil liberties. |
| 2. Purposeful and Performance-Driven | Agencies shall seek opportunities for designing, developing, acquiring, and using AI, where the benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed. |
| 3. Accurate, Reliable, and Effective | Agencies shall ensure that their application of AI is consistent with the use cases for which that AI was trained, and such use is accurate, reliable, and effective. |
| 4. Safe, Secure, and Resilient | Agencies shall ensure the safety, security, and resiliency of their AI applications, including resilience when confronted with systematic vulnerabilities, adversarial manipulation, and other malicious exploitation. |
| 5. Understandable | Agencies shall ensure that the operations and outcomes of their AI applications are sufficiently understandable by subject matter experts, users, and others, as appropriate. |
| 6. Responsible and Traceable | Agencies shall ensure that human roles and responsibilities are clearly defined, understood, and appropriately assigned for the design, development, acquisition, and use of AI. Agencies shall ensure that AI is used in a manner consistent with these Principles and the purposes for which each use of AI is intended. The design, development, acquisition, and use of AI, as well as relevant inputs and outputs of particular AI applications, should be well documented and traceable, as appropriate and to the extent practicable. |
| 7. Regularly Monitored | Agencies shall ensure that their AI applications are regularly tested against these Principles. Mechanisms should be maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate performance or outcomes that are inconsistent with their intended use or this order. |
| 8. Transparent | Agencies shall be transparent in disclosing relevant information regarding their use of AI to appropriate stakeholders, including the Congress and the public, to the extent practicable and in accordance with applicable laws and policies, including with respect to the protection of privacy and of sensitive law enforcement, national security, and other protected information. |
| 9. Accountable | Agencies shall be accountable for implementing and enforcing appropriate safeguards for the proper use and functioning of their applications of AI, and shall monitor, audit, and document compliance with those safeguards. Agencies shall provide appropriate training to all agency personnel responsible for the design, development, acquisition, and use of AI. |

OMB M-21-06 | Principles for the Stewardship of AI Applications ¹³

| Principle | Description |
|--|--|
| 1. Public Trust in AI | Since the continued adoption and acceptance of AI will depend significantly on public trust and validation, the government's regulatory and non-regulatory approaches to AI should contribute to public trust in AI by promoting reliable, robust, and trustworthy AI applications. |
| 2. Public Participation | Public participation, especially in those instances where AI uses information about individuals, will improve agency accountability...Agencies must provide ample opportunities for the public to provide information and participate in all stages of the rulemaking process. |
| 3. Scientific Integrity and Information Quality | Agencies should hold information, whether produced by the government or acquired by the government from third parties, that is likely to have a clear and substantial influence on important public policy or private sector decisions (including those made by consumers) to a high standard of quality. |
| 4. Risk Assessment and Management | Regulatory and non-regulatory approaches to AI should be based on a consistent application of risk assessment and risk management across various agencies and various technologies. |
| 5. Benefits and Costs | Agencies should, when consistent with law, carefully consider the full societal costs, benefits, and distributional effects when considering regulations related to the development and deployment of AI applications. |
| 6. Flexibility | When developing regulatory and non-regulatory approaches, agencies should pursue performance-based and flexible approaches that are technology neutral and that do not impose mandates on companies that would harm innovation. |
| 7. Fairness and Non-Discrimination | Agencies should consider, in accordance with law, issues of fairness and nondiscrimination with respect to outcomes and decisions produced by the AI application at issue, as well as whether the AI application at issue may reduce levels of unlawful, unfair, or otherwise unintended discrimination as compared to existing processes. |
| 8. Disclosure and Transparency | Transparency and disclosure can increase public trust and confidence in AI applications...Agencies should carefully consider the sufficiency of existing or evolving legal, policy, and regulatory environments before contemplating additional measures for disclosure and transparency. |
| 9. Safety and Security | Agencies should promote the development of AI systems that are safe, secure, and operate as intended, and encourage the consideration of safety and security issues throughout the AI design, development, deployment, and operation process. |
| 10. Interagency Coordination | Agencies should coordinate with each other to share experiences to ensure consistency and predictability of AI-related policies that advance American innovation and adoption of AI, while appropriately protecting privacy, civil liberties, national security, and American values and allowing sector- and application-specific approaches. |

APPENDIX II

INTERNAL AI DEPLOYMENT CONSIDERATIONS SUPPLEMENTARY MATERIALS

Cost-Benefit Analysis

Leaders can use the below template to evaluate the costs and benefits of a proposed AI solution. This analysis, coupled with an analysis of TAI risks, will help leaders make an informed decision about whether to move forward with an AI project.

ANALYSIS

1. What is the problem that you are trying to solve with AI?

In 1-2 paragraphs, clearly state what problem you are trying to solve with AI and why it is important to solve this problem. Because this is a cost-benefit analysis, it is important to indicate the burdens or costs associated with this problem and which groups bear these burdens or costs.

2. How would an AI solution solve this problem?

In 1 paragraph, clearly state how AI can solve this problem and why it provides a superior solution to alternative methods.

3. What is the value of solving this problem?

In 1-2 paragraphs, indicate how the proposed AI solution would reduce the burdens or costs stated above. AI solutions might also provide other benefits, such as informing policy or improving processes. In these cases, state who would use the AI solution, what it would be used for, how its use would improve operations or inform decision-making, and why these improvements are valuable.

SUMMARY TABLE

Complete the below table by listing the primary benefits and costs and estimating the dollar value for each.

| Benefits | Estimated Amount (\$) |
|-----------------------|-----------------------|
| Primary Benefit 1 | |
| Primary Benefit 2 | |
| Primary Benefit 3 | |
| Total Benefits | |
| Costs | |
| Primary Cost 1 | |
| Primary Cost 2 | |
| Primary Cost 3 | |
| Total Costs | |
| Net Total | |

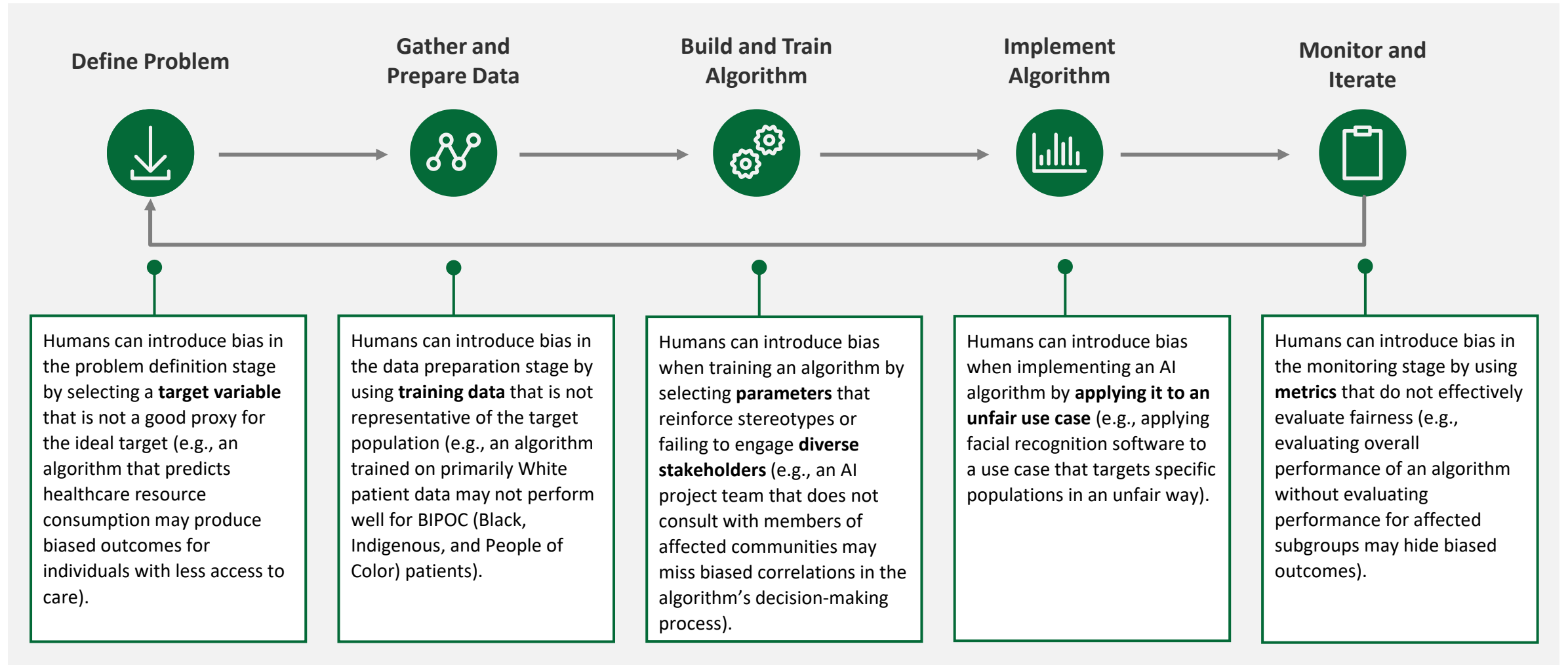
AI Project Team Roles ^{58, 59}

AI project teams often include several of the below roles, but the exact team composition will depend on the AI use case. Leaders should consider what skills they need when planning and issuing solicitations (if applicable) for AI projects.

| Role | Responsibilities |
|---------------------------------|---|
| Project Manager | Plan and lead the execution of an AI project; liaise between project team members; resolve issues |
| Systems Architect | Design and plan the implementation of an AI solution within the existing IT framework and systems |
| DevOps Engineer | Oversee the development, quality assurance, and deployment of an AI solution |
| Full Stack Developer | Write code for both the front end and back end of an AI solution |
| Cloud Engineer | Deploy an AI solution into the cloud; integrate it with existing IT products and services |
| Data Engineer | Integrate data into an AI solution's architecture |
| Data Scientist | Build and train AI algorithms |
| Business Analyst | Translate between business users and technical roles |
| Cybersecurity Specialist | Evaluate and mitigate security risks; monitor, detect, and respond to security incidents |
| User Experience Designer | Ensure that an AI solution is usable, enjoyable, and accessible to end users |
| Agile Coach | Guide the team through agile implementation of an AI solution |

Unintended Bias in Algorithms⁶⁰

Humans can introduce bias at all stages of model development.



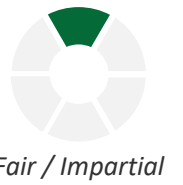
Bias Review Metrics (1 of 2) ³²

There are several metrics that teams can use to measure model fairness, but not all metrics will be relevant for every AI use case.

CLASSIFICATION METRICS

| Metric | Description |
|---|---|
| Statistical Parity Difference | The difference of the rate of favorable outcomes received by the unprivileged group to the privileged group; Suggests that a predictor is unbiased if the prediction is independent of the protected attribute |
| Equal Opportunity Difference | The difference of true positive rates between the unprivileged and privileged groups; A value of 0 implies both groups have equal benefit (i.e., receive the positive outcome at equal rates) |
| Average Odds Difference | The average difference of false positive rates (false positives/negatives) and true positive rates (true positives/positives) between unprivileged and privileged groups; A value of 0 implies both groups have equal benefit |
| Disparate Impact | The ratio of the rate of favorable outcomes for the unprivileged group to that of the privileged group |
| Four Fifths Rule | Heuristic that states that a selection rate for any protected class which is less than four-fifths of the selection rate for the group with the highest selection rate may be an indication of disparate impact |
| Theil Index | Measures the inequality of benefit allocation for individuals via the entropic “distance” the population is away from everyone having the same benefit level |
| False Positive Rate | The probability that a positive result will be given when the true value is negative |
| False Positive Rate Difference | The difference between the false positive rate of the unprivileged group and that of the privileged group |
| False Positive Rate Ratio | The ratio of the false positive rate of the unprivileged group to that of the privileged group |
| False Negative Rate | The probability that a negative result will be given when the true value is positive |
| False Negative Rate Difference | The difference between the false negative rate of the unprivileged group and that of the privileged group |
| False Negative Rate Ratio | The ratio of the false negative rate of the unprivileged group to that of the privileged group |

The links above provide representative examples of tests available from open-source sites. These tests are provided as a non-comprehensive starting point and their inclusion does not represent official endorsement.



Bias Review Metrics (2 of 2) ³²

There are several metrics that teams can use to measure model fairness, but not all metrics will be relevant for every AI use case.

SAMPLE DISTORTION METRICS

| Metric | Description |
|--------------------------------------|--|
| Euclidean Distance | The average Euclidean distance (i.e., distance between two real-valued vectors) between samples from two datasets |
| Mahalanobis Distance | The average Mahalanobis distance (i.e., distance between a point and a distribution) between samples from two datasets |
| Manhattan Distance | The average Manhattan distance (i.e., sum of the absolute differences between two vectors) between samples from two datasets |

OTHER METRICS

| Metric | Description |
|--------------------|--|
| Balance | The mean prediction probability for each protected class, categorized by actual outcomes; Examines whether the average score received by individuals in positive and negative instances are similar regardless of sensitive attributes |
| Calibration | A comparison of actual outcome rates versus the predicted probabilities, by decile; Helps address whether or not the model makes accurate predictions in aggregate for members of each class |

The links above provide representative examples of tests available from open-source sites. These tests are provided as a non-comprehensive starting point and their inclusion does not represent official endorsement.

*Teams should select bias review metrics **based on the context in which the AI model is applied.***



Bias Review Algorithms and Techniques ³²

Bias should be reviewed in training data, models, and outcomes throughout the AI lifecycle. Teams can use the following algorithms and techniques to improve the selected bias review metrics.

Algorithms for checking and mitigating bias in...

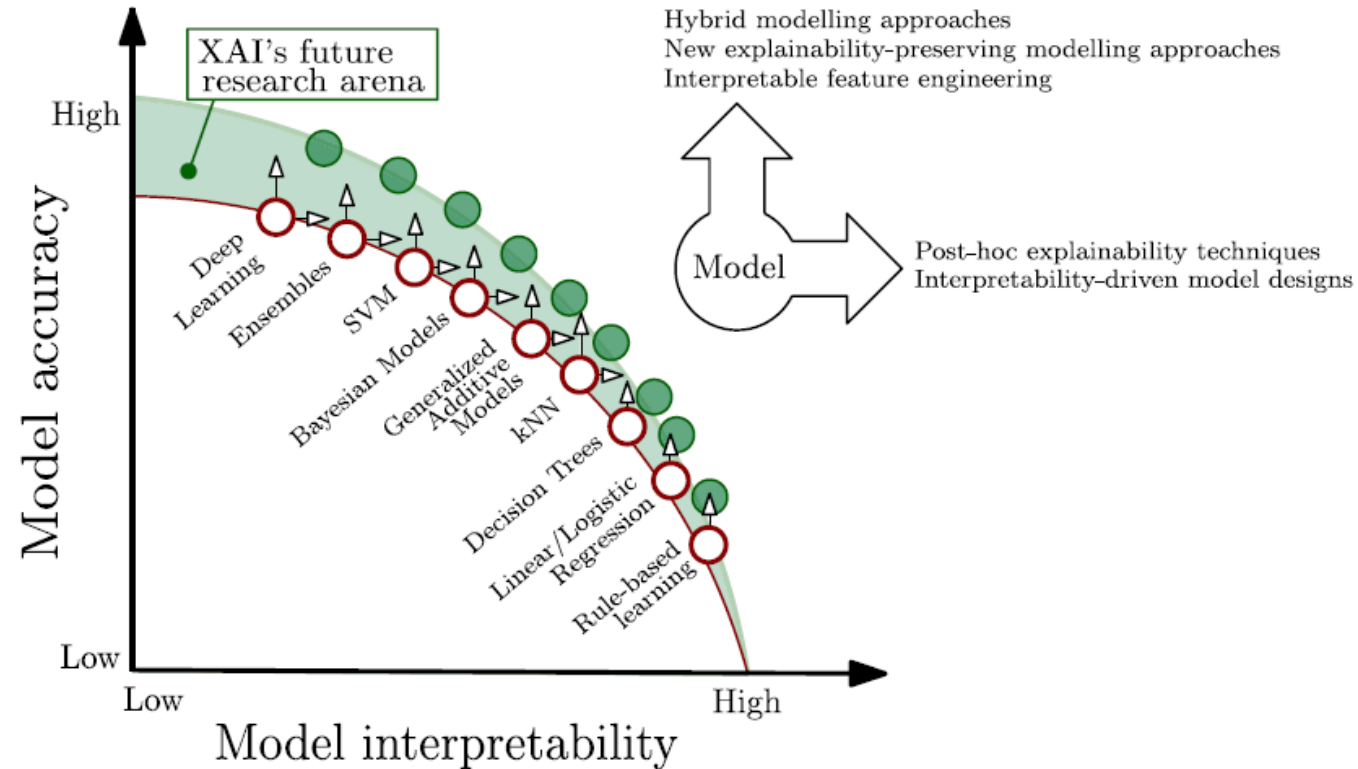
| Training Data | Models | Outcomes |
|---|--|--|
| <p>Disparate Impact Remover Edits feature values to improve groups fairness</p> | <p>Adversarial Debiasing Adversarial techniques maximize accuracy and reduce use of protected attributes</p> | <p>Calibrated Equality of Odds Optimizes over calibrated classifier score outputs that lead to fair output labels</p> |
| <p>Learning Fair Representations Learns fair representations by obfuscating information about protected attributes</p> | <p>Meta Fair Classifier Takes fairness metric as part of the input and returns a classifier optimized for metric</p> | <p>Equality of Odds Modifies predicted labels with optimization scheme to make predictions fairer</p> |
| <p>Optimized Preprocessing Modifies training data features and labels</p> | <p>Prejudice Remover Adds a discrimination-aware regularization term to the learning objective</p> | <p>Reject Object Classification Changes predictions from a classifier to make them fairer</p> |
| <p>Reweighting Modifies the weights of different training examples</p> | <p><i>The links above provide representative examples of tests available from open-source sites. These tests are provided as a non-comprehensive starting point and their inclusion does not represent official endorsement.</i></p> | |

*In applying bias mitigation algorithms to AI models, there is often a **tradeoff between fairness and model accuracy/robustness** that must be carefully considered.*



Explainability-Accuracy Tradeoff ⁶¹

AI models that are easier to interpret tend to have less predictive power. The below diagram illustrates the interpretability and accuracy of common types of AI models.



Note: The importance of model accuracy versus model explainability may vary across use cases depending on the impact of inaccurate model outcomes on affected individuals. For example, an organization may choose to prioritize model accuracy for a public-facing AI solution that directly affects individual's health (e.g., providing medical diagnoses, recommending care decisions), whereas model accuracy may be less critical for an internal solution that streamlines an administrative task.



Interpretation Methods ⁶²

Interpretation methods can provide insight into how an algorithm arrived at its conclusion by revealing correlations made within a model. The below table summarizes five common methods.

| METHOD | DESCRIPTION |
|---|---|
| LIME <i>Local Interpretable Model-Agnostic Explanations</i> | <ul style="list-style-type: none"> Generates model predictions with altered input data Trains interpretable model on new data set and uses it to interpret predictions |
| SHAP <i>Shapley Additive Explanations</i> | <ul style="list-style-type: none"> Uses game theoretically optimal Shapley Values Explains how much an individual feature or variable contributes to the overall prediction or a particular observation's prediction |
| ICE <i>Individual Conditional Expectation</i> | <ul style="list-style-type: none"> Measures marginal effect of a specific instance of a feature (i.e., specific data point) on a predicted output Reveals variance in marginal effects |
| PDP <i>Partial Dependence Plot</i> | <ul style="list-style-type: none"> Measures marginal effect of a feature on a predicted output Average of all lines in an ICE plot Reveals type of relationship between feature and output |
| ALE <i>Accumulated Local Effects</i> | <ul style="list-style-type: none"> Measures marginal effect of a feature on a predicted output over the conditional distribution of the feature Removes interpretation bias when there are highly correlated features |

For more information on interpretation methods, refer to [Captum](#), an open-source library of algorithms to generate model interpretations, including LIME / SHAP values.



Human Supervision of AI Solutions ²

There are three levels of human supervision for an AI solution. The level required depends on the objectives and risks of the AI solution.

- 1 Human-in-the-loop**
The AI solution provides recommendations, and a human reviews the output and makes a final decision
- 2 Human-on-the-loop**
A human monitors the AI solution and takes control when it encounters unexpected or undesirable events
- 3 Human-out-of-the-loop**
The AI solution has full control without the option of human override



Key Roles for Managing AI Solutions ²⁰

All AI solutions require human oversight. Below are two suggested roles and responsibilities for managing AI solutions.

1
SPONSOR

Federal government employee responsible for solution compliance

RESPONSIBILITIES

- Assigns roles and responsibilities to govern the solution
- Fields inquiries from stakeholders
- Oversees who has access to the solution

2
CUSTODIAN

Federal government employee or contractor responsible for day-to-day operational management of the AI solution

RESPONSIBILITIES

- Completes initial and routine training in AI solution management and security
- Maintains access to the AI solution
- Oversees retraining or tuning of underlying model
- Tracks and monitors model data and output
- For AI solutions with digital identities, holds a comparable level of access and rotates digital worker password authenticators



Digital Worker Impact Evaluation Matrix (1 of 2) ²⁰

Score each factor based on the most likely scenario.

| Factor 1 – Is the digital worker attended or unattended? | Score |
|--|-------|
| Attended | 0 |
| Unattended | 10 |
| Factor 2 – What is the highest level of data access by the digital worker? | |
| Data available to the public (either without a user account or with unvetted user account) | 0 |
| Agency operational data, controlled unclassified information (CUI), or data on individuals in low volumes. Doesn't contain PII or PHI | 5 |
| PII and/or PHI | 55 |
| Agency critical operational data or data that could impact life, health, or safety of individuals/systems relied upon for health and safety; or very high volumes of agency operational data | 90 |
| Factor 3 – Does the digital worker have access to internal and/or external networks? | |
| No internal intranet or external internet connection | 0 |
| Either internal intranet access only OR external internet access (not both) | 5 |
| Internal and external network access (i.e., internet and intranet) | 10 |



Digital Worker Impact Evaluation Matrix (2 of 2) ²⁰

Score each factor based on the most likely scenario.

| Factor 4 – What is the impact of the output generated by the digital worker? | Score |
|---|-------|
| Output impacts general internal business operations, but not for critical processes or decisions | 5 |
| Output impacts outside organizations in general business operations or public reporting (e.g., public facing websites or chatbots), but not for critical processes or decisions | 25 |
| Output impacts mission critical organization operations of the agency or other organizations, health or safety of individuals, national economic stability, national security, critical infrastructure, or similarly consequential operations | 90 |
| Factor 5 – What type of system account privileges does the digital worker require? | |
| No system accounts used | 0 |
| Standard system account(s) (roles limited by the business function) | 10 |
| System admin account (privileged access) | 35 |
| Multiple system admin accounts (multiple privileged access roles) | 40 |
| Factor 6 – Does the digital worker act on its own insights? | |
| Digital worker develops insights, but doesn't take action on its insights | 0 |
| Digital worker develops insights and acts on the insights after human review | 5 |
| Digital worker develops insights and acts on the insights without human review or approval before the action is taken | 10 |



Potential Adverse Impact Levels ²⁰

Sum the scores from the Digital Worker Evaluation Matrix to determine the potential adverse impact level of the AI solution.

| Impact Score | Potential Adverse Impact* | Description |
|--------------|---------------------------|---|
| 0-35 | Low | Effects of an error or accident are minimal, resulting in negligible, if any, impacts on organizational operations, finances, assets, individuals, other organizations, or the Nation. |
| 36-55 | Moderate | Effects of an error or accident are limited and may result in minor or temporary impact on organizational missions/business functions, organizational assets, or the Nation. This includes: increased difficulty in performing business operations in a timely manner, with sufficient confidence, or within planned resource constraints; minor damage to agency image, reputation, or trust; minor financial loss to the agency or other organizations; and/or noncompliance with applicable laws or regulations. |
| 56-90 | High | Effects of an error or accident are wide-ranging and could result in serious or long-term impact on organizational missions/business functions, organizational assets, or the Nation. This includes: significant financial losses for the agency; substantially reduced capacity to conduct mission critical business; loss of Personally Identifiable Information (PII), Business Identifiable Information, or Protected Health Information (PHI); and/or damage to agency image or reputation. |
| 91+ | Critical | Effects of an error or accident are extensive and will have severe or catastrophic impact on organizational missions/business functions, assets, or the Nation. This includes: major financial losses for the agency or other organizations; loss of government continuity of operations or ability to conduct mission critical business; life-threatening injury or loss of life; and/or harm to national security. |

**The Potential Adverse Impact Levels are grounded in NIST Special Publication 800-30 and are a recommendation. Agencies may adjust or tailor the levels to fit their individual risk levels or descriptions.*



Types of Security Risks and Defenses ⁶³

Below are four common adversarial techniques that are specific to AI solutions. It is critical to defend AI solutions against these attacks in addition to traditional cybersecurity threats.

| SECURITY RISK | DESCRIPTION | DEFENSE MECHANISMS |
|---------------------|---|---|
| 1. Evasion | Occurs when an attacker modifies input data to trick the model into misclassifying inputs | <ul style="list-style-type: none"> • Adversarial Training – <i>Incorporate adversarial samples into the training stage to improve the model’s robustness</i> • Network Distillation – <i>Use knowledge extracted from a model to further train the model and improve resilience to adversarial samples</i> • Adversarial Detection (e.g., input reconstruction) – <i>Identify adversarial samples in the model inputs</i> |
| 2. Poisoning | Occurs when an attacker feeds contaminated training data to the model to shift the model’s decision boundary in favor of an adversary | <ul style="list-style-type: none"> • Data Filtering – <i>Remove data points that are sufficiently far from the training set</i> • Ensemble Analysis – <i>Use inherently robust learning methods that train multiple algorithms (e.g., bagging)</i> |
| 3. Backdoor | Occurs when an attacker manipulates model components, causing the model to fail on specific inputs while performing well on others | <ul style="list-style-type: none"> • Model Pruning – <i>Remove model components that are activated by the backdoor but not by clean inputs to reduce the backdoor attack’s success</i> • Input Pre-Processing – <i>Train the model on transformed data to make it more difficult for attackers to manipulate the model</i> |
| 4. Stealing | Occurs when an attacker analyzes the input, output, and other external information of an AI system to speculate on the model or the underlying data | <ul style="list-style-type: none"> • Private Aggregation of Teacher Ensembles (PATE) – <i>Train teacher models on disjoint data, noisily aggregate teachers’ answers, and transfer knowledge to a student model</i> • Model Watermarking – <i>Embed content into model as a watermark to enable external verification of ownership and protect intellectual property</i> |



Privacy Protection Methods

There are several methods to enhance data privacy for AI solutions. Common methods and relevant resources are included below.

| METHOD | DESCRIPTION | RESOURCES |
|--|---|--|
| Differential Privacy | Adds noise to a dataset so that it is impossible to reverse-engineer the individual inputs | <ul style="list-style-type: none"> • OpenDP • NIST Differential Privacy Blog |
| Federated Learning | Trains a shared global model across many participating clients that keep their training data locally | <ul style="list-style-type: none"> • Tensor Flow Federated Learning |
| Synthetic Data | Data developed in a digital world as opposed to the real world; synthetic data reflects real-world data as a workaround to healthcare privacy constraints ⁶⁴ | <ul style="list-style-type: none"> • What Is Synthetic Data? • Synthetic Health Data Generation to Accelerate Patient-Centered Outcomes Research HealthIT.gov • Synthea Open-Source Synthetic Patient Generator |
| Secure Multiparty Computation (MPC) | Shares insights from an analysis without sharing the data itself | <ul style="list-style-type: none"> • Open Source MPC Library |
| Homomorphic Encryption | Encrypts data before sharing such that it can be analyzed but not decoded into the original information | <ul style="list-style-type: none"> • Open Source Homomorphic Encryption Library • Homomorphic Encryption Consortium |
| Trusted Execution Environments | Provides input privacy with secure computation through a combination of hardware and software | <ul style="list-style-type: none"> • Confidential Computing Consortium |
| Zero Knowledge Proofs | Proves that encrypted data are within given ranges without revealing any information about the data | <ul style="list-style-type: none"> • ZKProof |



Model Performance Testing Approaches

Technical teams can use the three approaches outlined below to evaluate model performance. Leaders should ensure that model performance testing is sufficiently documented for traceability and approval.

| APPROACH | PURPOSE | RECOMMENDED DOCUMENTATION |
|---|--|---|
| Scenario Analysis / Stress Testing | Assess how the model performs under a variety of conditions including stress scenarios that are outside the range of ordinary expectations | <ul style="list-style-type: none"> • Model outputs under all scenarios • Evidence that the calibration process was considered over a range of input values, including extreme cases • Source and justification for each scenario |
| Sensitivity Analysis | Validate that shocks in input data result in expected and intuitive changes in model outputs | <ul style="list-style-type: none"> • Univariate and multivariate analysis results • Logic for selected shock sizes |
| Outcome Analysis / Backtesting | Assess how the model performs on historical data | <ul style="list-style-type: none"> • Overview of and justification for backtesting framework • Backtesting results and action plan, including vendor source documentation |



Model Performance Metrics

The following metrics can help teams evaluate and improve the model's results before and after deploying the model into production.

| Metric | Description |
|-------------------------------------|---|
| Recall (Sensitivity) | The percentage of correct positive predictions compared to all actual positive classifications; Measures how accurately the positive class is predicted |
| Specificity | The percentage of correct negative predictions compared to all actual negative classifications; Measures how accurately the negative class is predicted |
| Precision | The percentage of correct positive predictions compared to all positive predictions |
| Accuracy | The percentage of correct predictions |
| F1 Score | A harmonic average of recall and precision |
| Area Under Curve (AUC) Score | An aggregate measure of performance across all possible classification thresholds |

Note: *There can be tradeoffs between model performance and both fairness and explainability.* ^{24, 25} A model may have a high percentage of accurate predictions, but the model may be replicating historical biases present in the data. Similarly, a deep learning or other similarly complex model may have strong performance metrics, but it may be more difficult to understand and explain the model's outputs. Leaders should understand and consider these tradeoffs when evaluating model performance.

APPENDIX III

STATUTORY AUTHORITIES

Statutory Authorities (1 of 4)

The following statutes authorize HHS to issue regulations on the development and use of AI applications in the private sector

| Statute | Description |
|--|---|
| <p>Statute: Civil Rights Act of 1964 Citation/Codification: Title VI (42 USC 2000d et seq.; 45 CFR Part 80)</p> | <p>The Civil Rights Act of 1964 is an Act that outlaws discrimination based on race, color, or national origin. The law prohibits unequal application of voter registration requirements, and racial segregation in schools, employment, and public accommodations. Title VI of the act prevents discrimination by programs and activities that receive federal funds, including hospitals and other health care facilities.</p> <p>AI applications can result in discriminatory outcomes that negatively impact individuals protected by federal civil rights law. HHS has authority to enforce this statute in the context of AI to the extent such applications result in unlawful discrimination prohibited by the statute.</p> |
| <p>Statute: Rehabilitation Act of 1973 Citation/Codification: Section 504 (29 USC 794; 45 CFR Part 84 (HHS federally assisted programs or activities); 45 CFR Part 85 (HHS federally conducted programs or activities))</p> | <p>The Rehabilitation Act of 1973 prohibits discrimination against people with disabilities in programs that receive federal financial assistance. Section 504 works together with the Americans with Disabilities Act (ADA) and Individuals with Disabilities Education Act (IDEA) to protect children and adults with disabilities from exclusion, and unequal treatment in schools, jobs and the community.</p> <p>AI applications can result in discriminatory outcomes that negatively impact individuals protected by federal civil rights law. HHS has authority to enforce this statute in the context of AI to the extent such applications result in unlawful discrimination prohibited by the statute.</p> |
| <p>Statute: Education Amendments of 1972 Citation/Codification: Title IX (20 USC 1681 et seq.; 45 CFR Part 86)</p> | <p>Title IX of the Education Amendments of 1972 (Title IX) prohibits sex discrimination in any education program or activity receiving federal financial assistance.</p> <p>AI applications can result in discriminatory outcomes that negatively impact individuals protected by federal civil rights law. HHS has authority to enforce this statute in the context of AI to the extent such applications result in unlawful discrimination prohibited by the statute.</p> |
| <p>Statute: The Age Discrimination Act of 1975 Citation/Codification: 42 USC 6101 et seq.; 45 CFR Part 90 (federally assisted programs or activities); 45 CFR Part 91 (HHS federally assisted programs or activities).</p> | <p>The Age Discrimination Act of 1975 (Age Act), prohibits discrimination on the basis of age in HHS-funded programs and activities. Under the Age Act, recipients may not exclude, deny, or limit services to, or otherwise discriminate against, persons on the basis of age.</p> <p>AI applications can result in discriminatory outcomes that negatively impact individuals protected by federal civil rights law. HHS has authority to enforce this statute in the context of AI to the extent such applications result in unlawful discrimination prohibited by the statute.</p> |

Statutory Authorities (2 of 4)

The following statutes authorize HHS to issue regulations on the development and use of AI applications in the private sector

| Statute | Description |
|---|--|
| <p>Statute: Section 1557 of the Patient Protection and Affordable Care Act Citation/Codification: Section 1557. (42 USC 18116; 45 CFR Part 92)</p> | <p>This statute prohibits discrimination on the basis of race, color, national origin, sex (including sexual orientation and gender identity), age, and disability in certain health programs or activities, including health programs or activities in the private sector that receive financial assistance from HHS.</p> <p>AI applications can result in discriminatory outcomes that negatively impact individuals protected by federal civil rights law. HHS has authority to enforce this statute in the context of AI to the extent such applications result in unlawful discrimination prohibited by the statute.</p> |
| <p>Statute: Public Health Service Act Citation/Codification: § 3001(c)(1) PHSA [Standards and Certification Criteria Review] § 3001(c)(2)(A) PHSA § 3001(c)(5)(A) and (B) PHSA [Certification Program] § 3004 PHSA [Process for adoption of endorsed recommendations; adoption of initial set of standards, implementation specifications, and certification criteria]</p> | <p>The Public Health Service Act (PHSA) provides the legal authority for HHS, among other things, to respond to public health emergencies. The act authorizes the HHS secretary to lead federal public health and medical response to public health emergencies, determine that a public health emergency exists, and assist states in their response activities.</p> <p>Specifically, the Office of the National Coordinator for Health Information Technology (ONC) has the authority to review and endorse standards, implementation specifications, and certification criteria for the electronic exchange and use of health information. To the extent an AI application is involved in the exchange or use of health information, HHS has the ability to regulate it. ONC also has the authority to keep or recognize a voluntary certification program to provide for the certification of health IT.</p> |
| <p>Statute: Education Amendments of 1972 Citation/Codification: Title IX (20 USC 1681 et seq.; 45 CFR Part 86)</p> | <p>The United States Federal Food, Drug, and Cosmetic Act (FD&C Act) is a set of laws giving authority to the U.S. Food and Drug Administration to oversee the safety of food, drugs, medical devices, and cosmetics.</p> <p>The FD&C Act contains provisions (or regulatory requirements) that define The Food & Drug Administration (FDA)'s level of control over several products. To fulfill the provisions of the FD&C Act that apply to medical devices and radiation-emitting products, FDA has the authority to regulate such devices or products that use AI and Machine Learning (ML) algorithms that can result serious adverse health consequences, are expected to have significant use in pediatric populations, are intended to be implanted in the body for more than one year, or are intended to be a life-sustaining or life-supporting device used outside a device user facility.</p> |

Statutory Authorities (3 of 4)

The following statutes authorize HHS to issue regulations on the development and use of AI applications in the private sector

| Statute | Description |
|--|---|
| <p>Statute: Safe Medical Devices Act (SMDA) of 1990 Citation/Codification: H.R.3095 — 101st Congress</p> | <p>Safe Medical Device Amendments of 1990 (or Safe Medical Devices Act) sanctioned progressive reporting and tracking rules for medical devices classified by the Medical Device Regulation Act. The Act mandates reporting requirements by medical device manufacturers regarding adverse safety events and product effectiveness of devices classified as substantially equivalent to Class III medical devices. The United States Statute established the HHS Office of International Relations and an FDA office for regulatory activities concerning healthcare products which are considered a combinational biological, device, or drug product.</p> <p>To carry out the reporting provisions of SMDA, FDA has the authority to direct AI or ML algorithms that user facilities or manufacturers use to monitor products after their clearance to market and track devices for maintaining traceability.</p> |
| <p>Statute: Mammography Quality Standards Act (MQSA) of 1992 Citation/Codification: 21 CFR Part 900</p> | <p>MQSA required HHS to develop standards that would be enforced through strict accreditation, certification and inspection of equipment and personnel at mammography facilities. The FDA was tasked with implementing the federal regulations to establish and enforce such procedures.</p> <p>Mammography technologies that use AI or ML are required to meet standards enforced by the MQSA statute. Therefore, HHS has the authority to manage or direct such algorithms to ensure standards are adhered to.</p> |

Statutory Authorities (4 of 4)

Given that AI is still an emerging field, there are other statutes that authorize HHS to regulate health data or health technology but may not directly reference AI. The following statutes may indirectly give HHS an authority to regulate AI, given that AI algorithms and solutions are already used as components in many health technology solutions

| Statute | Description |
|--|---|
| <p>Statute: The Confidential Information Protection and Statistical Efficiency Act (known as the E-Government Act of 2002) Citation/Codification: 116 STAT. 2962 Public Law 107-347 107th Congress</p> | <p>Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)’s primary purposes are to protect information collected for statistical purposes from improper disclosure and to ensure that the information is not used for nonstatistical purposes. To achieve its purposes, CIPSEA establishes limitations on the use and disclosure of statistical data or information. As stated in CIPSEA section 512, data or information acquired under a pledge of confidentiality and for exclusively statistical purposes shall be used for exclusively statistical purposes and shall not be disclosed for a nonstatistical purpose, except with the informed consent of the respondent.</p> <p>CIPSEA gives National Center for Health Statistics (CDC-NCHS) broad statutory authority to protect the confidentiality of information they collect. Therefore, NCHS has the authority to restrict the application of AI/ML techniques applied to their data in order to protect confidentiality.</p> |
| <p>Statute: Health Information Technology for Economic and Clinical Health (HITECH) Act Citation/Codification: Title XIII STAT. 115 Public Law 111-5 111th Congress</p> | <p>The HITECH Act, enacted as part of the American Recovery and Reinvestment Act of 2009 (ARRA), established requirements for Health Insurance Portability and Accountability Act of 1996 (HIPAA) regulated entities to provide notice of breaches of protected health information and increased penalties for non-compliance with the HIPAA Rules.</p> <p>The HITECH Act indirectly authorizes HHS to regulate AI applications by establishing requirements for the safeguarding and notification of a breach of protected health information which may occur through use of an AI application by a HIPAA regulated entity.</p> |

APPENDIX IV

ACRONYMS

Acronyms

| Acronym | Meaning |
|-----------------|--|
| AI | Artificial Intelligence |
| ALE | Accumulated Local Effects |
| ATLAS | Adversarial Threat Landscape for Artificial-Intelligence Systems |
| ATO | Authority to Operate |
| BII | Business Identifiable Information |
| BIPOC | Black, Indigenous, People of Color |
| BPM | Business Process Management |
| CAIO | Chief AI Officer |
| CISO | Chief Information Security Officer |
| COTS | Commercial Off-the-Shelf |
| CUI | Controlled Unclassified Information |
| EPLC | Enterprise Performance Lifecycle |
| HCD | Human-Centered Design |
| HCTD | Human-Centered Technology Design |
| ICAM | Identity, Credential, and Access Management |
| ICE | Individual Conditional Expectation |
| IRB | Institutional Review Board |
| ISSO | Information System Security Officer |
| ITAR | IT Acquisition Review |
| IV&V | Independent Verification and Validation |
| LIME | Local Interpretable Model-Agnostic Explanations |
| ML | Machine Learning |

| Acronym | Meaning |
|----------------|-------------------------------------|
| MPC | Multiparty Computation |
| NLG | Natural Language Generation |
| NLP | Natural Language Processing |
| NLU | Natural Language Understanding |
| O&M | Operations and Maintenance |
| PDP | Partial Dependence Plot |
| PHI | Protected Health Information |
| PIA | Privacy Impact Assessment |
| PII | Personally Identifiable Information |
| RPA | Robotic Process Automation |
| SAOP | Senior Agency Official of Privacy |
| SHAP | Shapley Additive Explanations |
| SLA | Service-Level Agreements |
| SOP | Senior Official of Privacy |
| SOR | System of Record |
| SORN | System of Record Notice |
| TAI | Trustworthy AI |

APPENDIX V

REFERENCES

References (1 of 5)

- 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>

References (2 of 5)

- ¹⁴ Samir Passi, Solon Barocas, *Problem Formulation and Fairness*, January 29, 2019. Available at: <https://dl.acm.org/doi/10.1145/3287560.3287567>
- 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 The Center for Open Data Enterprise, *Sharing and Utilizing Health Data for AI Applications*, 2019. Available at: <https://www.hhs.gov/sites/default/files/sharing-and-utilizing-health-data-for-ai-applications.pdf>
- ¹⁷ 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>
- ¹⁸ Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, Aurelia Tamò-Larriex, *Towards Transparency by Design for Artificial Intelligence*, November 16, 2020. Available at: <https://link.springer.com/article/10.1007/s11948-020-00276-4>
- ¹⁹ 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>
- ²⁰ General Services Administration, *The Digital Worker Identity Playbook*, January 5, 2021. Available at: <https://playbooks.idmanagement.gov/docs/playbook-digital-worker.pdf>
- ²¹ *HHS Policy for Securing AI Technology*, Draft.
- ²² Suchi Saria, Adarsh Subbaswamy, *Safe and Reliable Machine Learning*, April 15, 2019. Available at: <https://arxiv.org/pdf/1904.07204.pdf>
- ²³ *HHS Enterprise Performance Life Cycle Framework*, July 18, 2012. Available at: <https://www.hhs.gov/sites/default/files/ocio/eplc-lifecycle-framework.pdf>
- ²⁴ *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>
- ²⁵ 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
- ²⁶ 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>

References (3 of 5)

- ²⁷ 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>
- ²⁸ 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](https://www.dni.gov/files/ODNI/documents/AI_Ethics_Framework_for_the_Intelligence_Community_10.pdf)
- ²⁹ 42 U.S.C. § 18116(a)
- ³⁰ IT Modernization Centers of Excellence, *Guide to AI Ethics*. Available at: <https://coe.gsa.gov/docs/CoE%20Guide%20to%20AI%20Ethics.pdf>
- ³¹ 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
- ³² IBM, *AI Fairness 360*. Available at: <https://aif360.mybluemix.net/>
- ³³ 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>
- ³⁴ Deon, *An Ethics Checklist for Data Scientists*. Available at: <https://deon.drivendata.org/>
- ³⁵ Google, *Machine Learning Crash Course*. Available at: <https://developers.google.com/machine-learning/crash-course/framing/ml-terminology>
- ³⁶ DeepAI, *Hyperparameter*. Available at: <https://deepai.org/machine-learning-glossary-and-terms/hyperparameter>
- ³⁷ MITRE, *Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS)*. Available at: <https://atlas.mitre.org/matrix/>
- ³⁸ NIST Computer Science Resource Center (CSRC) Glossary. Available at: <https://csrc.nist.gov/glossary>
- ³⁹ 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>
- ⁴⁰ HIPAA Journal, *What is Considered Protected Health Information Under HIPAA?*, March 2, 2021. Available at: <https://www.hipaajournal.com/what-is-considered-protected-health-information-under-hipaa/>

References (4 of 5)

- ⁴¹ U.S. Census Bureau, *DS022: Personally Identifiable Information (PII) Breach Policy*, June 29, 2018. Available at: https://www2.census.gov/foia/ds_policies/ds022.pdf
- ⁴² Google, *Responsible AI Practices*. Available at: <https://ai.google/responsibilities/responsible-ai-practices/>
- ⁴³ 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>
- ⁴⁴ Xiang Zhou, *Interpretability Methods in Machine Learning: A Brief Survey*. Available at: <https://www.twosigma.com/articles/interpretability-methods-in-machine-learning-a-brief-survey/>
- ⁴⁵ Michigan State University, *Human-Centered Technology Design*. Available at: <https://comartsci.msu.edu/research-creative-work/current-research/thematic-research-areas/human-centered-technology>
- ⁴⁶ NIST SP 800-128, *Guide for Security-Focused Configuration Management of Information Systems*, August 2011. Available at: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-128.pdf>
- ⁴⁷ 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>
- ⁴⁸ 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>
- ⁴⁹ 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
- ⁵⁰ 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>
- ⁵¹ NIST SP 800-53 Revision 5, *Security and Privacy Controls for Information Systems and Organizations*, September 2020. Available at: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf>
- ⁵² NIST SP 800-167, *Guide to Application Whitelisting*, October 2015. Available at: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-167.pdf>

References (5 of 5)

- ⁵³ NIST SP 800-94, *Guide to Intrusion Detection and Prevention Systems (IDPS)*, February 2007. Available at: <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-94.pdf>
- ⁵⁴ Chamber Technology Engagement Center, Deloitte AI Institute, *Investing in Trustworthy AI*, July 2021. Available at: <https://www2.deloitte.com/us/en/pages/consulting/articles/investing-in-ai-trust.html?nc=1>
- ⁵⁵ *Executive Order on Maintaining American Leadership in Artificial Intelligence*, February 11, 2019. Available at: <https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence>
- ⁵⁶ Office of Science and Technology Policy, *American Artificial Intelligence Initiative: Year One Annual Report*, February 2020. Available at: <https://www.nitrd.gov/nitrdgroups/images/c/c1/American-AI-Initiative-One-Year-Annual-Report.pdf>
- ⁵⁷ Ron Schmelzer, *AI and Blockchain: Double the Hype or Double the Value?*, October 24, 2019. Available at: <https://www.forbes.com/sites/cognitiveworld/2019/10/24/ai-and-blockchain-double-the-hype-or-double-the-value/?sh=3d8af045eb41>
- ⁵⁸ Evan Knopp, *Building Your AI Team: The Roles Your Enterprise Needs*, September 17, 2018. Available at: <https://www.ibm.com/blogs/systems/building-your-ai-team-the-roles-your-enterprise-needs/>
- ⁵⁹ Maria Korolov, *8 Key Roles of Successful AI Projects*, March 12, 2019. Available at: <https://www.cio.com/article/3356818/8-key-roles-of-successful-ai-projects.html>
- ⁶⁰ *Responsible AI with TensorFlow*, TensorFlow Dev Summit '20, March 11, 2020. Available at: <https://www.youtube.com/watch?v=UEECKh6PLhI>
- ⁶¹ David Gunning, *Explainable Artificial Intelligence (XAI)*, DARPA. Available at: [https://www.cc.gatech.edu/~alanwags/DLAI2016/\(Gunning\)%20IJCAI-16%20DLAI%20WS.pdf](https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf)
- ⁶² Fabricio Pretto, *Uncovering the Magic: Interpreting Machine Learning Black-box Models*, July 28, 2020. Available at: <https://towardsdatascience.com/uncovering-the-magic-interpreting-machine-learning-black-box-models-3154fb8ed01a>
- ⁶³ NISTIR 8269, *A Taxonomy and Terminology of Adversarial Machine Learning*, Draft, October 2019. Available at: <https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf>
- ⁶⁴ Gerard Andrews, *What is Synthetic Data?*, June 8, 2021. Available at: <https://blogs.nvidia.com/blog/2021/06/08/what-is-synthetic-data/>