Integrating Responsible AI Principles into Systems Engineering Practices:

A Holistic Approach for Safe and Reliable AI-Enabled Systems

SERC Workshop

September 17, 2024

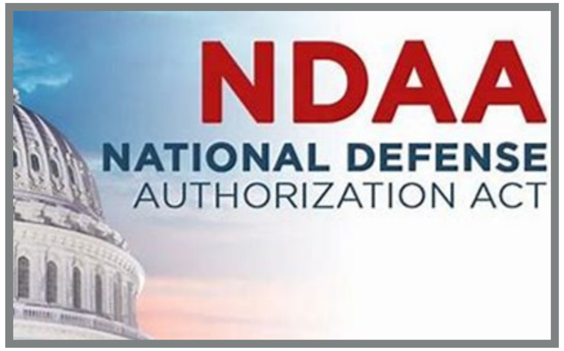
Dr. Rosa R. Heckle
Dr. Michael Hadjimiachel
Dr. Flo Reeder

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Key Policies and Regulations for Implementing Responsible AI (RAI)

White House EO 14110

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence



NDAA of 2024

Advancing AI America Act

DHS Policy Statement 139-06

Acquisition and Use of Artificial Intelligence and Machine Learning Technologies by DHS Components

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

THE DIRECTOR

December 8, 2009

M-10-06

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Peter R. Orszag
Director

SUBJECT: Open Government Directive

OMB M-24-10

Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence

MITRE

The AI Ethics Fog

Navigating the gap between high-level concepts and practical application

Accountable
Safe
Reliable
Transparency
Robust
Fair
Equitable
Explainable
Guidance
Regulations
Laws

Hundreds of AI ethics principles, countless regulations...

But how do we implement them in practice?

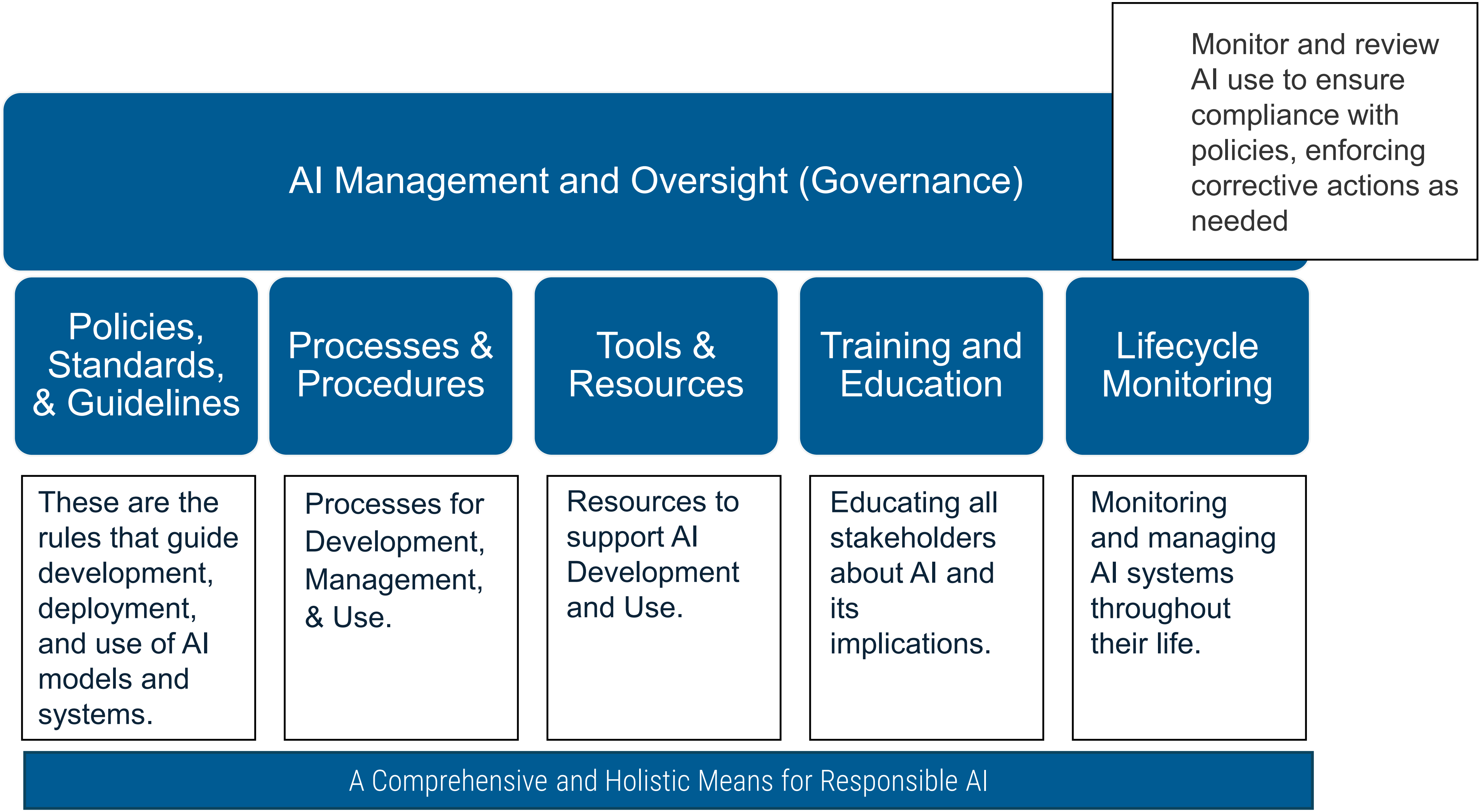
Non-Actionable and Too Late

“What are the correct metrics to assess the AI’s output? Would the margin of error be deemed tolerable by those who use the AI? What is the impact of using inaccurate outputs and how well are these errors communicated to the users?”
(AI Ethics Framework_for_the_Intelligence_Community_10.pdf (odni.gov))

What are the forms of attack to which the AI system is vulnerable? Which of these forms of attack can be mitigated against?”
ai_hleg_draft_ethics_guidelines_32A2C883-DFF3-95CF-73EFE1D88C14A69C_57112.pdf

“During the ***deployment*** phase, assess the potential for algorithmic bias and ensure that the system does not perpetuate or exacerbate existing inequalities.“ (IEEE Ethically aligned Design Guideline.)

Responsible AI Framework



AI Management and Oversight (Governance)

Monitor and review AI use to ensure compliance with policies, enforcing corrective actions as needed

Policies, Standards, & Guidelines

These are the rules that guide development, deployment, and use of AI models and systems.

Processes & Procedures

Processes for Development, Management, & Use.

Tools & Resources

Resources to support AI Development and Use.

Training and Education

Educating all stakeholders about AI and its implications.

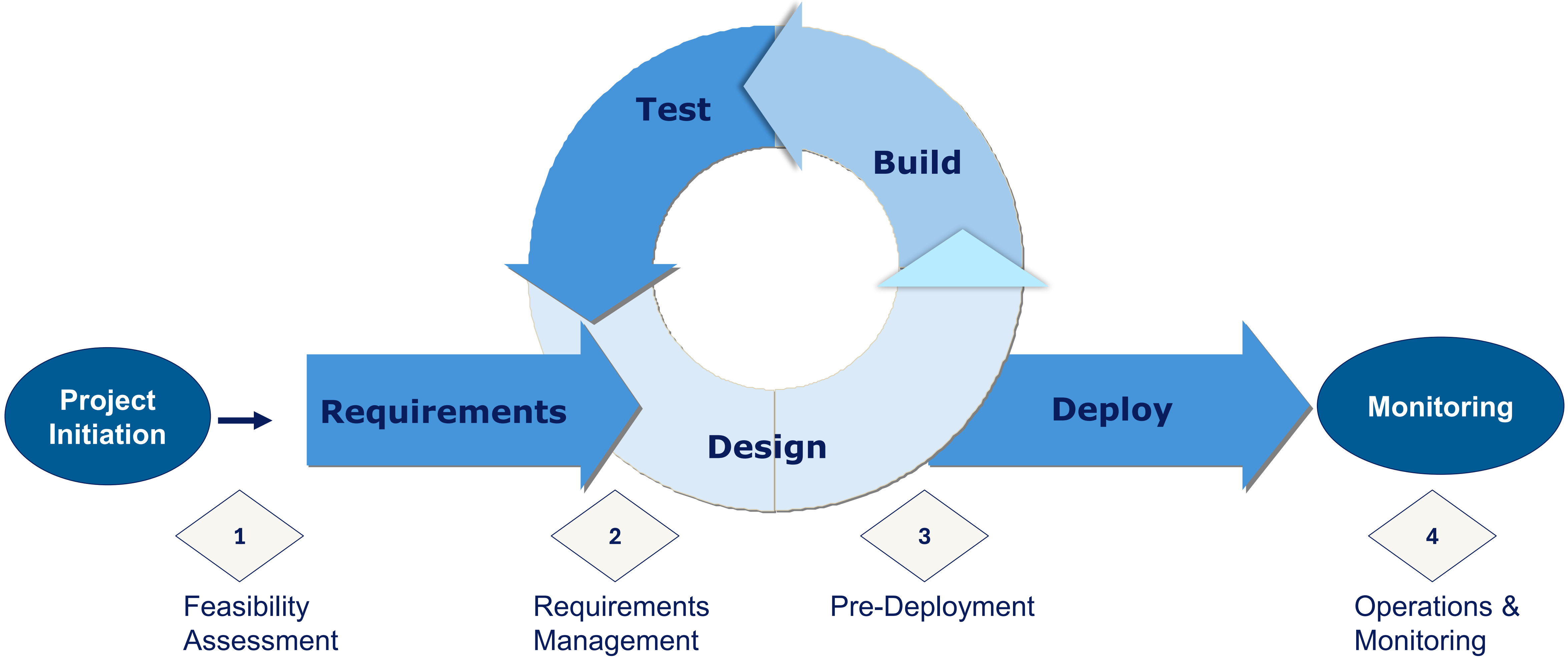
Lifecycle Monitoring

Monitoring and managing AI systems throughout their life.

A Comprehensive and Holistic Means for Responsible AI

Leveraging Systems Engineering

Inserting AI Requirements into Systems Engineering Methods, Tools, and Processes

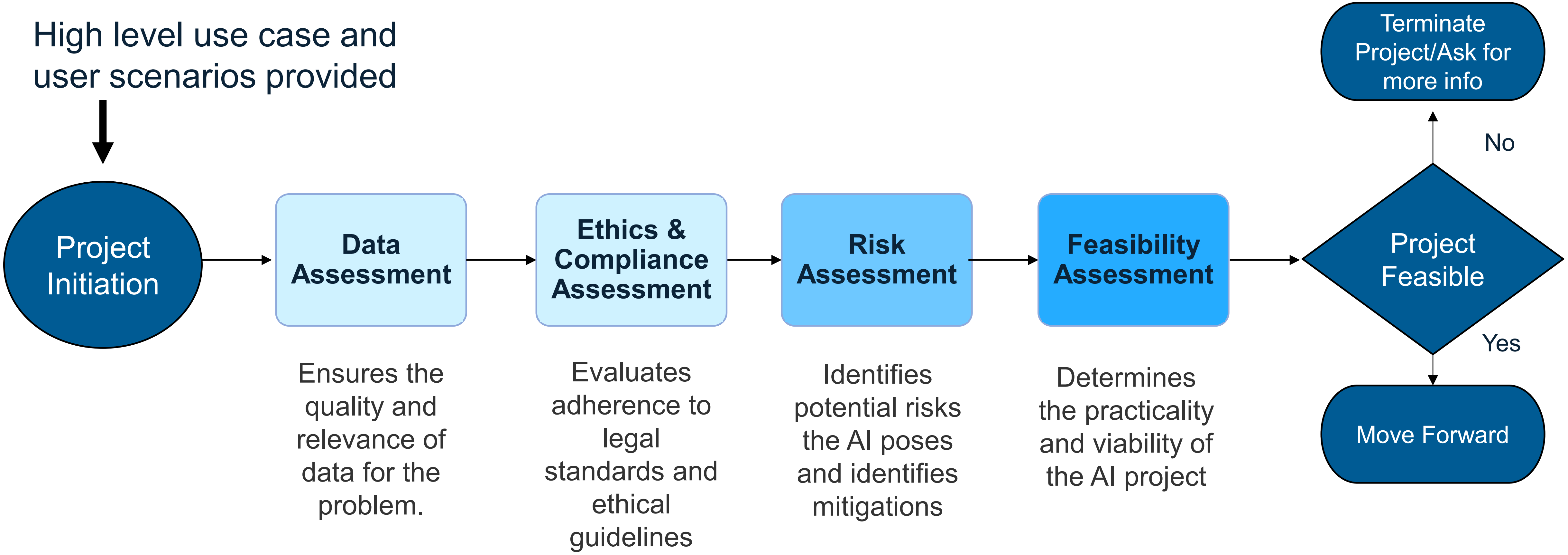


MITRE

Project Initiation Phase

Feasibility Assessment

AI Specific Tasks Within a Project's Initiation or Feasibility Phase


High level use case and user scenarios provided
Project Initiation
Data Assessment
Ethics & Compliance Assessment
Risk Assessment
Feasibility Assessment
Project Feasible
No
Terminate Project/Ask for more info
Yes
Move Forward

Data Assessment: Ensures the quality and relevance of data for the problem.

Ethics & Compliance Assessment: Evaluates adherence to legal standards and ethical guidelines

Risk Assessment: Identifies potential risks the AI poses and identifies mitigations

Feasibility Assessment: Determines the practicality and viability of the AI project

Guides on How to Perform Assessments

1

Feasibility Assessment

Step 1: Complete Assessment Questionnaire

Step 2: Score

Score	0	1	2	3	4
Is there enough RELEVANT data available to train the model?	There is almost no data available.	There is very little data available	There is a moderate amount of data available.	There is a good amount of data available	There is a large amount of data available
Does the data contain sufficient breadth to address the range of real-world inputs the AI might encounter	The data does not contain sufficient breadth and fails to cover the range of real-world inputs	The data contains minimal breadth and fails to cover the range of real-world inputs	The data contains a moderate level of breadth, covering some but not all potential real-world inputs	The data contains good breadth, covering a large range of potential real-world inputs	The data contains excellent breadth, covering nearly all potential real-world inputs

Step 3: Interpret Score

Score	Interpretation
Poor	The data has several issues that could impact the performance or fairness of the AI model.
Fair	The data is of average quality. There may still be some issues, but they are less likely to severely impact the AI model.
Good	The training data is of good quality. There may still be some issues, but they are unlikely to severely impact the AI model.
Good to Excellent	The training data is of high quality. Minor issues may still exist, but overall, the data should be suitable for training the AI model.

Modifying Risk Assessments to Address *AI-Specific* Vulnerabilities

1

Feasibility Assessment

AI has unique vulnerabilities and risks

Vulnerability	Low Risk	Medium Risk	High Risk
Model Source	In-house or reputable provider.	Less experienced or lesser-known provider.	Unknown or untrustworthy source
Model algorithm access	White box access - Transparent and interpretable	Gray box access – partially transparent	Black box access - Opaque and difficult to interpret
Input Data Sensitivity to Drift	High Predictability - Clear patterns reliably anticipated	Moderate Predictability – some unpredictability	Low Predictability - Largely unpredictable
Model Documentation	Comprehensive documentation	Documentation available, but unclear.	No documentation at all provided.

Requirements Management

Incorporating testable ethics and risk mitigation requirements

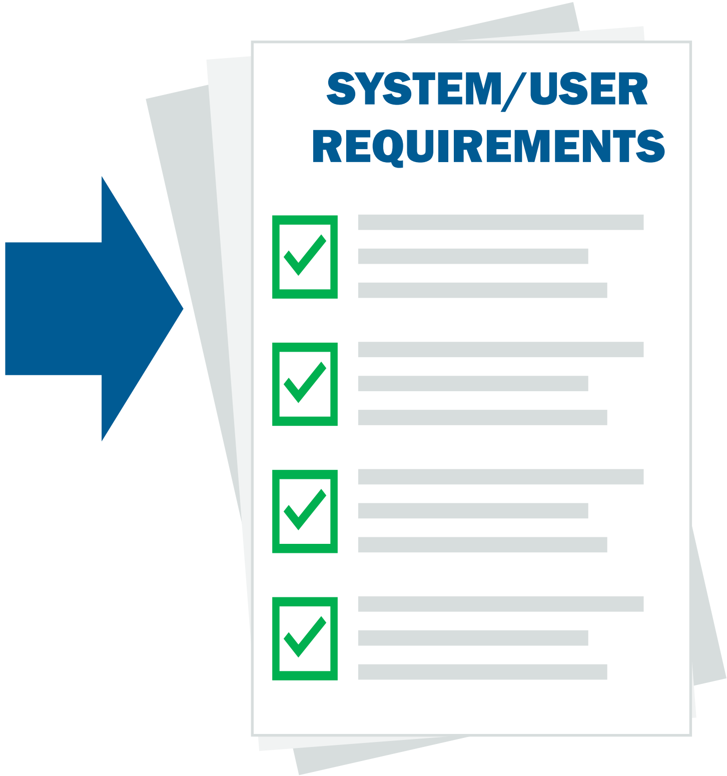
2

Requirements Management

Add required **risk mitigations** identified in risk assessment

Add relevant **foundational AI Ethics** Requirements

** * A set of foundational AI Ethics Requirements was developed by the Organization based on mandates and organizational values.*



Example:

*The AI system shall evaluate missing data, erroneous data, and remove outliers for potential harm to under-represented Group **X** during the data preprocessing stage.*

*The AI system shall ensure that the training data is timely, with all records being no older than **X** days/months/years.*

Test & Evaluation Report Template

3

Pre-Deployment

T&E Report Template to address challenges in documenting critical information and ensure:

Comprehensiveness

Holistic View
Checklist of Required Data

Requirement Validation

Clear Criteria
Traceability

Reproducibility

Scientific Rigor
Easier Replication

Enhanced Decision Making

Data-driven Insights
Easier to derive Recommendations

Traceability

Clear link to requirements
Facilitates auditing and compliance

Consistency

Easier to Read and Compare
Facilitates Repeatability

Operational Documentation and Compliance for ATO

Project documentation for AI transparency, accountability, and robustness

3

Pre-Deployment

Documentation Checklist

- ✓ Requirements
- ✓ Data Sheet
- ✓ Model Card
- ✓ Risk Assessment
- ✓ T&E Report
- ✓ Lifecycle Monitoring Instructions

MITRE

Operations & Monitoring Instructions

How Often

What should be monitored

Things to be monitored include:

- Model Monitoring
- System Monitoring
- Data Monitoring
- Adversarial Monitoring

Example:

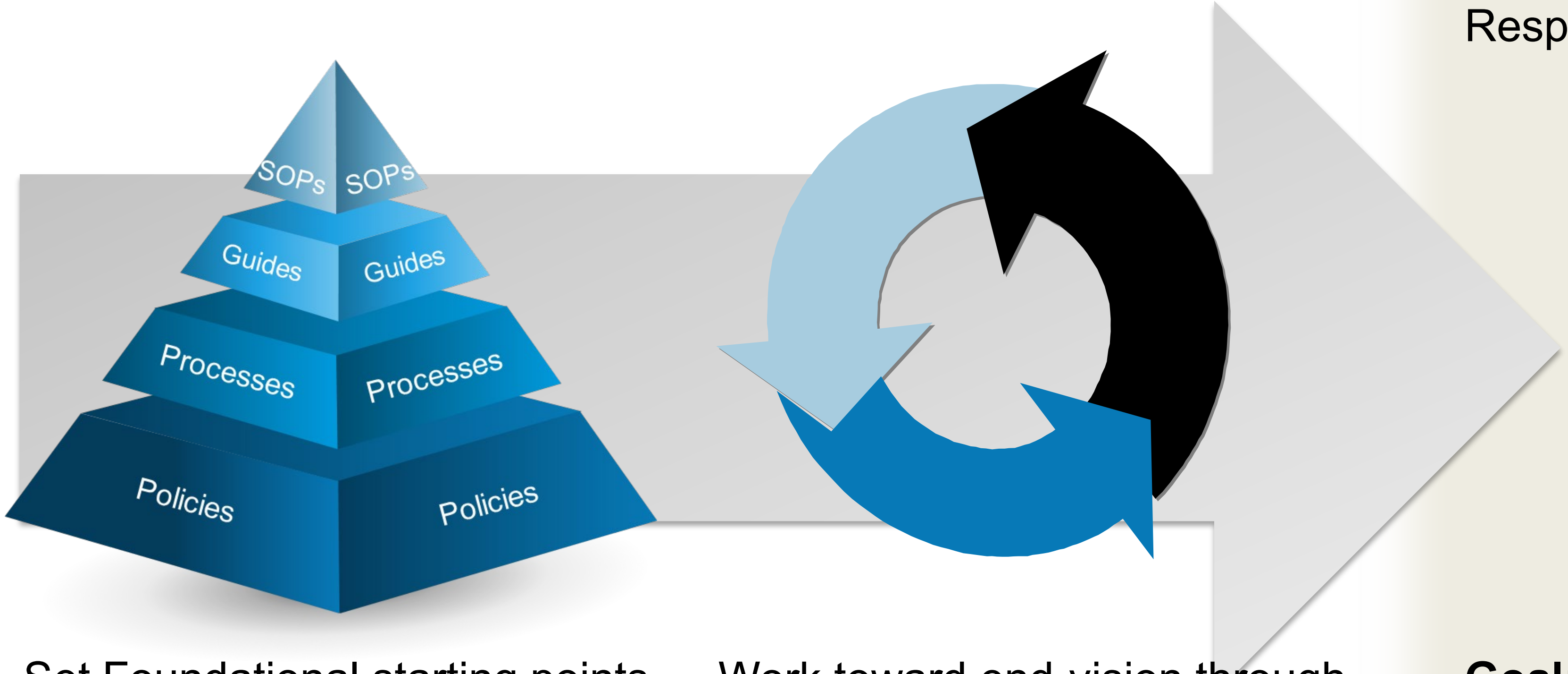
Model Monitoring

Monitoring Frequency	KPI/Metric	Threshold	Severity	Actions to be taken
• Continuous • Daily • Weekly • Monthly • Quarterly • On-Demand • Other	Accuracy	<=.70	Red	Shut down
		<.85	Yellow	Notify PoC
		>=.85	Green	
	Precision			
	Metric X			
	Metric X			

Thresholds

What to do if something goes wrong

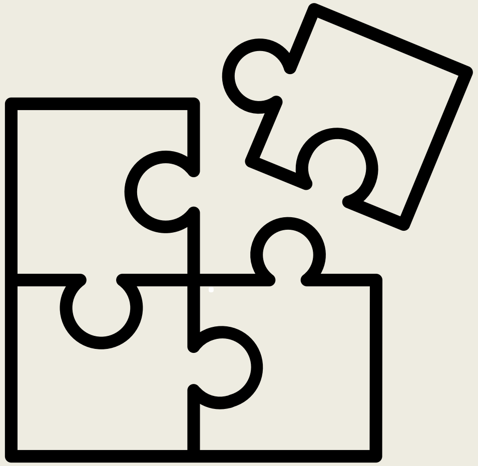
Moving Forward



Set Foundational starting points

Work toward end-vision through iteration and tailoring to Organizational fit.

Vision: A comprehensive TOOL-DRIVEN ecosystem for building & implementing Responsible AI



Goal: To create an environment where Responsible AI is seamlessly integrated, powered by purpose-built tools, automated processes, and adaptive governance frameworks.

MITRE

Questions

The Homeland Security Act of 2002 (Section 305 of PL 107-296, as codified in 6 U.S.C. 185), herein referred to as the "Act," authorizes the Secretary of the Department of Homeland Security (DHS), acting through the Under Secretary for Science and Technology, to establish one or more federally funded research and development centers (FFRDCs) to provide independent analysis of homeland security issues. MITRE Corp. operates the Homeland Security Systems Engineering and Development Institute (HSSEDI) as an FFRDC for DHS under contract HSHQDC-14-D-00006.

The HSSEDI FFRDC provides the government with the necessary systems engineering and development expertise to conduct complex acquisition planning and development; concept exploration, experimentation and evaluation; information technology, communications and cyber security processes, standards, methodologies and protocols; systems architecture and integration; quality and performance review, best practices and performance measures and metrics; and, independent test and evaluation activities. The HSSEDI FFRDC also works with and supports other federal, state, local, tribal, public and private sector organizations that make up the homeland security enterprise. The HSSEDI FFRDC's research is undertaken by mutual consent with DHS and is organized as a set of discrete tasks. This report presents the results of research and analysis conducted under:

70RDA224FR0000003
I&A Mission Services

The results presented in this report do not necessarily reflect official DHS opinion or policy.

Approved for Public Release; Distribution Unlimited.
Case Number 24-2319 / DHS reference number 70RDA224FR-003-02

HSSEDI
Homeland Security Systems Engineering & Development Institute