Aerospace's Trusted AI Framework

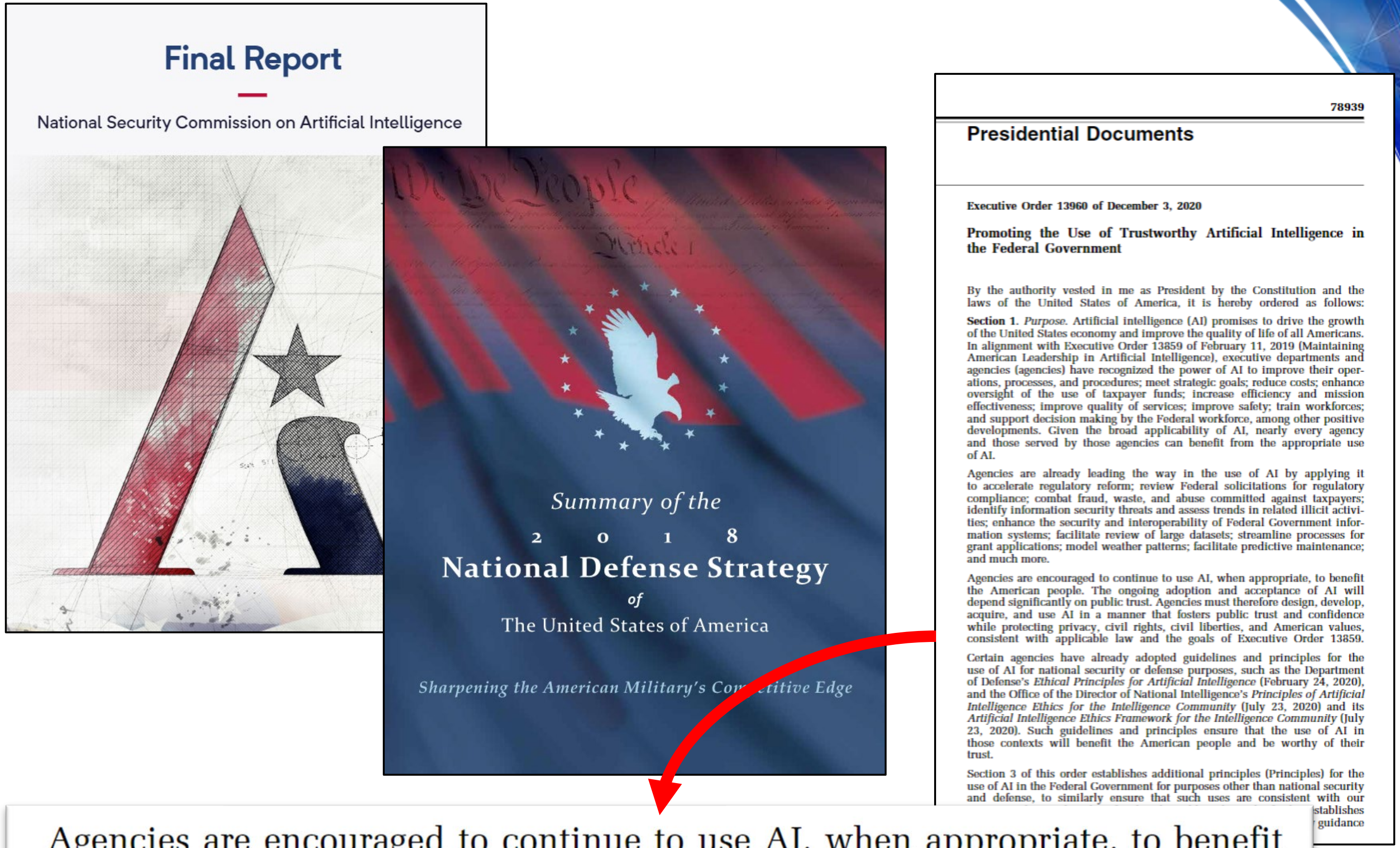
Dr. Phil Slingerland, ESD/RSSAS/MIXD
Lauren Perry, ATD/SRO/SAG

GSAW 2023

Aerospace's Trusted AI Framework

Motivation

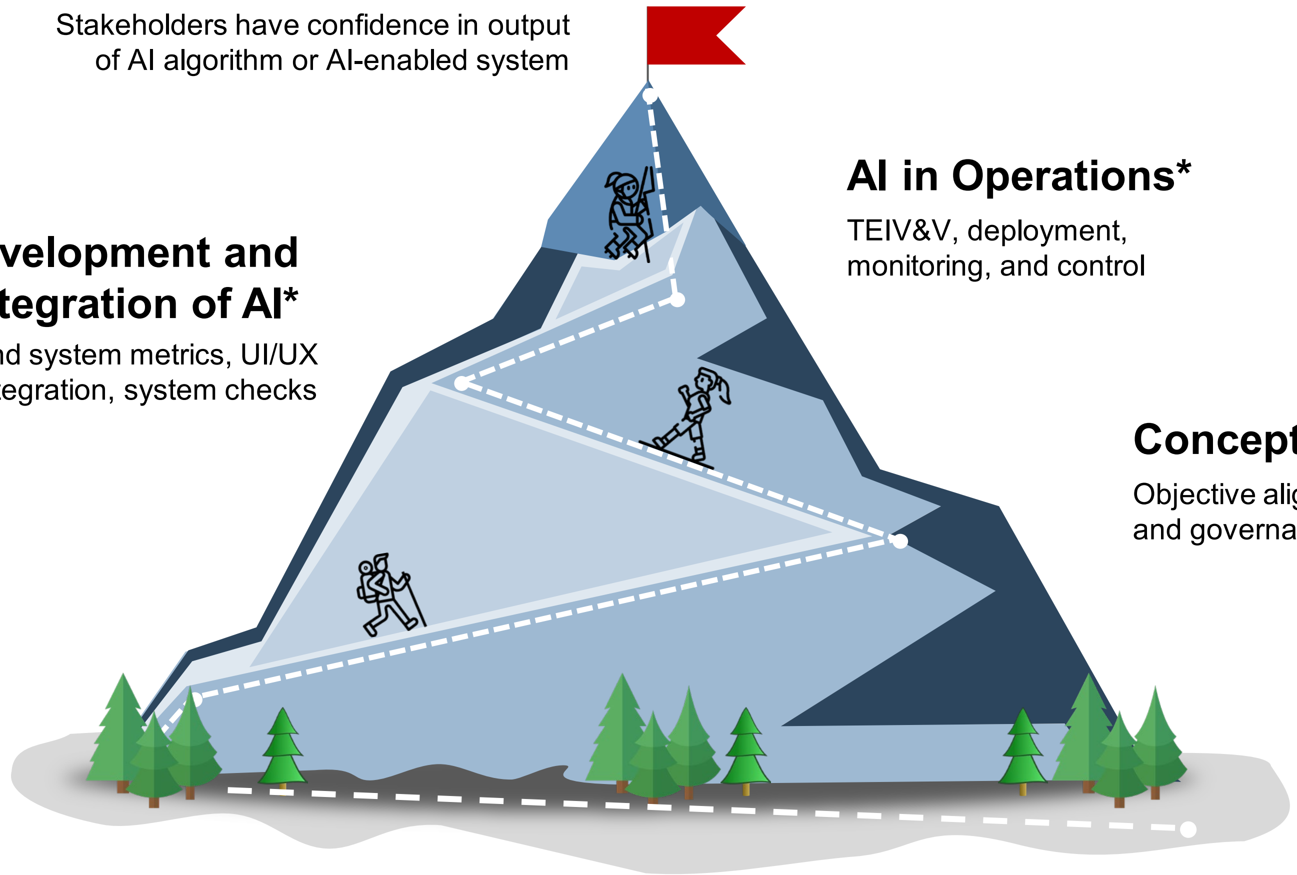
- AI has proven success in wide array of applications, leading to:
  - *Increasing role of automated decision making*
  - *Pervasive use of AI-enabled systems*
- Consequence of increasing presence of AI:
  - *Risk of impact to all stakeholders (business leaders, public, governments)*
  - *Increased awareness of responsibility (financial, legal, ethical)*
- Regulations
  - *Global wave of guidelines and regulation targeting AI*
  - *DoD and IC published recommended policies*
  - *Parallel developments in industry-specific regulation (financial, pharmaceutical, autonomous vehicles, etc.)*



Agencies are encouraged to continue to use AI, when appropriate, to benefit the American people. The ongoing adoption and acceptance of AI will depend significantly on public trust. 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, consistent with applicable law and the goals of Executive Order 13859.

Aerospace Developed A Trusted AI Framework to assist customers in design, implementation, and assessment of AI-based algorithms, with an emphasis on high consequence environments

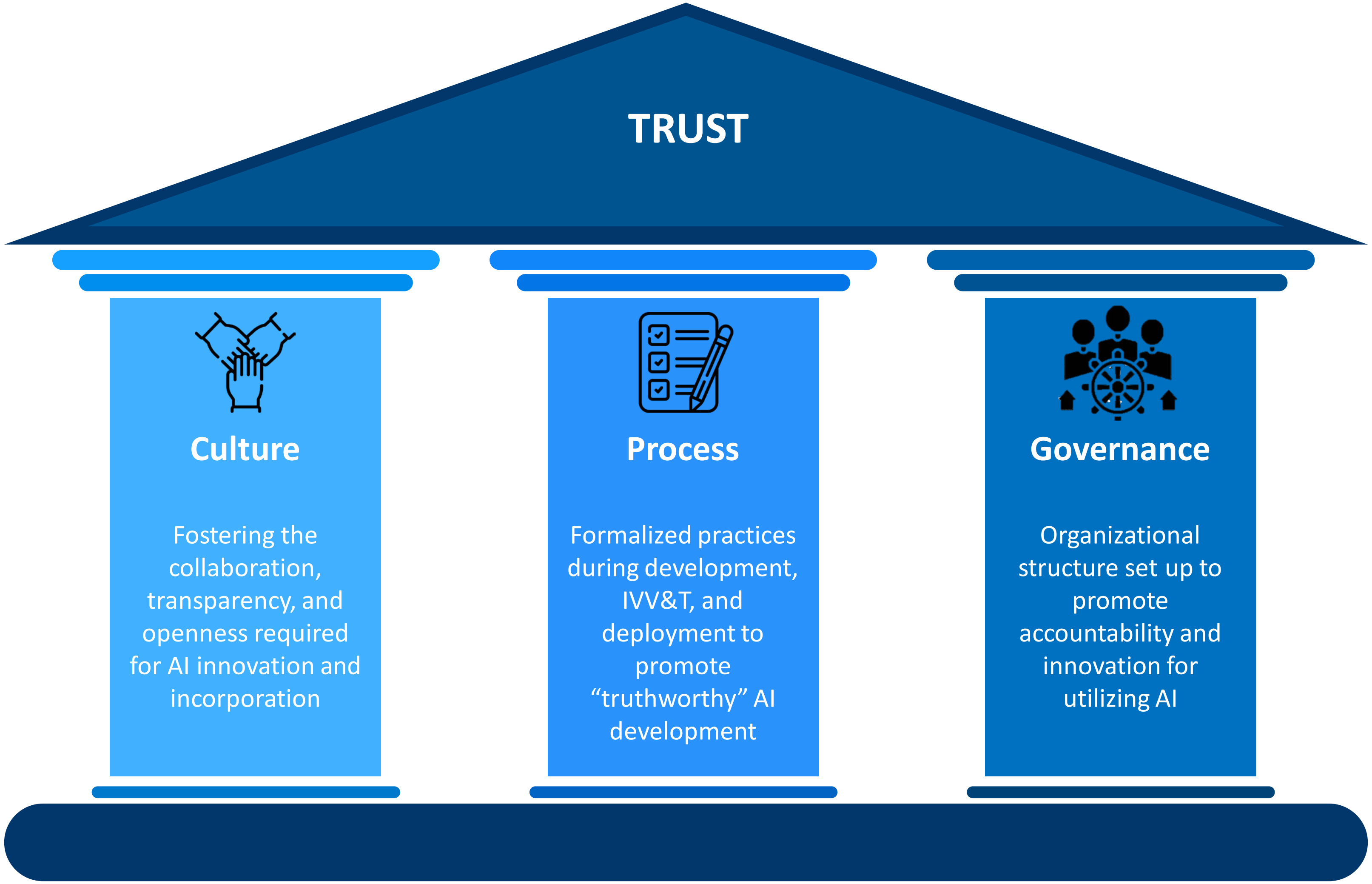
Journey of Trust



**Supported by Trusted AI Framework*

Definition and journey to trusting an AI-enabled system is in the eye of the beholder

Pillars of Trust



Trust requires a facilitating environment and collaborative team

Trust Is More Than Technology

The People Side of Trust

Values of Trust

Encourage openness and transparency

- Faithfully capture risks inherent to AI capability
- Anticipate challenges of deployment with additional data collection, training, and testing
- Avoid imprecise language of AI hype

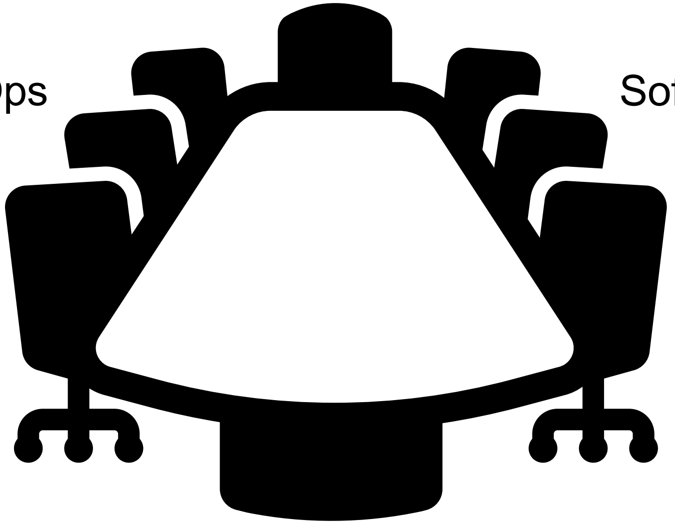
Develop collaboratively with users

- Frame development as journey in building user trust
- Strive for informed users that have input to AI design, function, interpretability, and control

Set high expectations for traceability

- Set high expectations for traceability
- Software, data, and model version control
- Record design decisions and R&D progress
- Log performance metrics throughout development

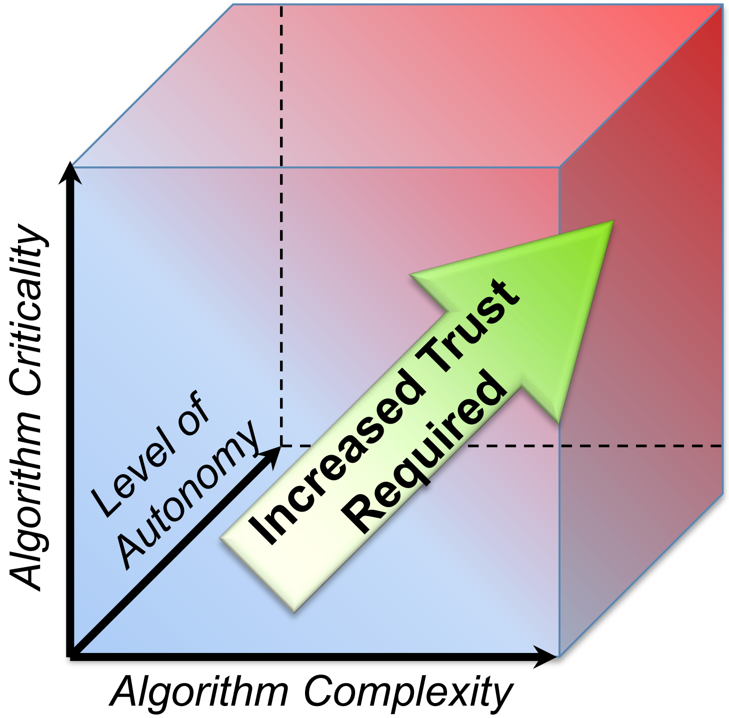
Stakeholders



Without consideration of various stakeholders, trust of an AI-enabled system will never be fully realized

Degree of Trust

How Much Trust Investment is Needed?



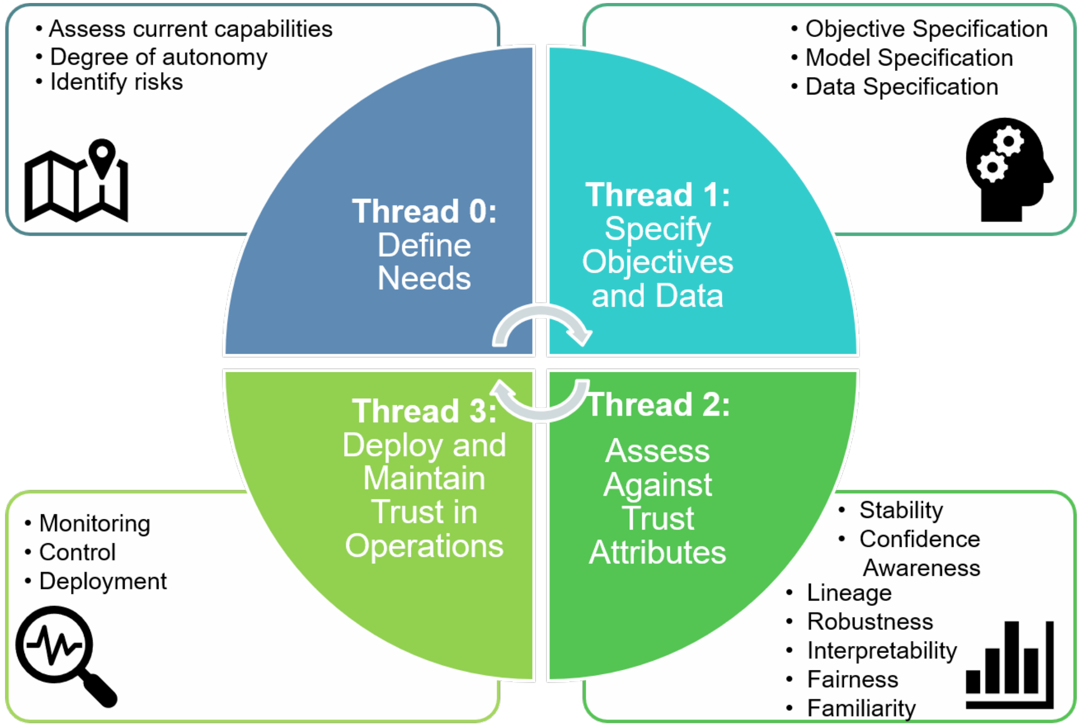
- The amount of trust required is related to risk to mission integrity defined by three dimensions:
  - ***Algorithm Criticality****: impact on mission success*
  - ***Algorithm Complexity****: degree of interpretability*
  - ***Level of Autonomy****: independence from human intervention*
- Assess degree of trust required for application
  - *Engage stakeholders on potential impacts of deployment*
  - *Define benchmarks for success*
  - *Estimate LOE and budget and weigh against expected value*
- Operational risks can be mitigated through DevOps best practices
  - *Prepare for unavailable, drifting, or poorly performing models*
  - *Provide awareness of data, model, and objective alignment throughout model lifecycle*

Trust is not free and should be tailored to intended use

How to Trust AI?

Trusted AI

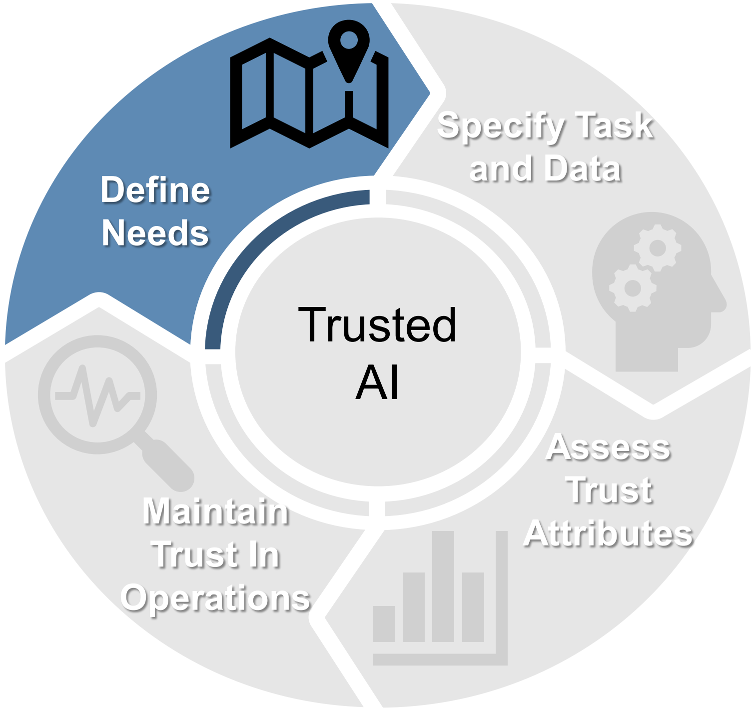
AI capability that provides sufficient confidence of satisfying user-defined objectives in a proper, interpretable, and safe way over its lifetime



Need Definition, Value Proposition and Intent

Thread 0: Definition of Needs

- Define and justify need for AI-based capability
  - *Compare with simplest approach or existing capability*
  - *Conduct literature review to support AI appropriateness*
- Identify desired level of autonomy of AI-based capability
  - *Degree of human intervention desired in operations*
- Reduce and manage risk
  - *Consider appropriate model complexity*
  - *Isolate operation to single function*
  - *Leverage existing systems and codebases*



Developing business case and rationale for bringing AI into a system

Task Specification

Thread 1: Specify Task and Data

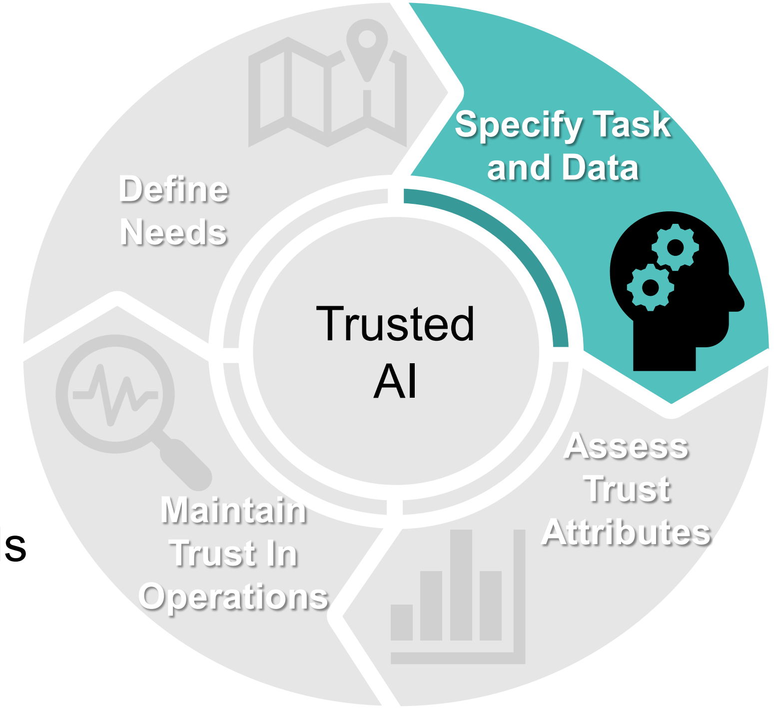
- Translate proposed capability to AI task
  - *Modular description:*
    - Compose task as collection of simple functions
    - Develop requirements for each to assemble trust
  - *Identify performance metrics relevant to target domain*

- Identify potential failure modes and how AI could cause them
  - *Early expectation of potential failures to help mitigate them*
  - *Note hardware and software limitations of deployed system*

- Capture performance and trust requirements; define metrics/TPMs
  - *Tailor metrics to use case*
  - *Algorithm, system, and mission focused metrics*
  - *Include upstream and downstream metrics (for operational monitoring)*

Objective Specification

Clear description of AI objective with metrics that demonstrate task alignment



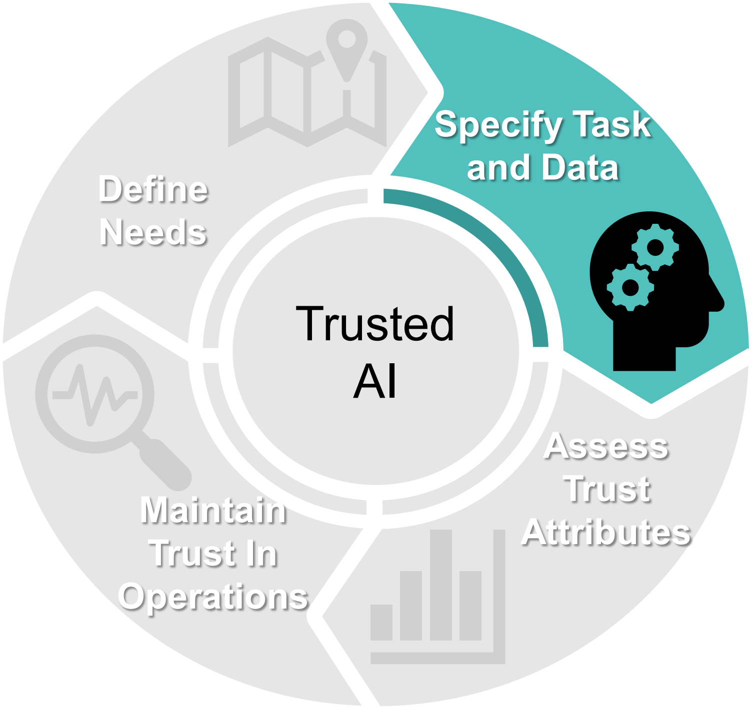
Data Specification

Thread 1

- Identify datasets for each development stage and V&V
  - *Specify how data will be collected and partitioned*
  - *Capture details of sensors, data pre-processing (Traceability)*
- Perform exhaustive exploratory data analysis (EDA):
  - *Identify:*
    - Nominal and out-of-scope data parameters
    - Subgroups and their relative representation (Fairness)
    - Challenging data examples and mitigation plans
    - High spatial or temporal correlations, especially across data splits
- Develop upstream protection to check for out-of-scope data
  - *Re-route out-of-scope data to alternate system*
  - *Tailor confidence based on scope (Pertinence)*
- Provide tools that enhance data pedigree:
  - *Interface to gathering annotations from multiple SMEs, logging annotations from users*
  - *Enable review and disagreement between SMEs*
- Collect and assess representative target data near deployment phase (Pre-deployment)

Data Specification

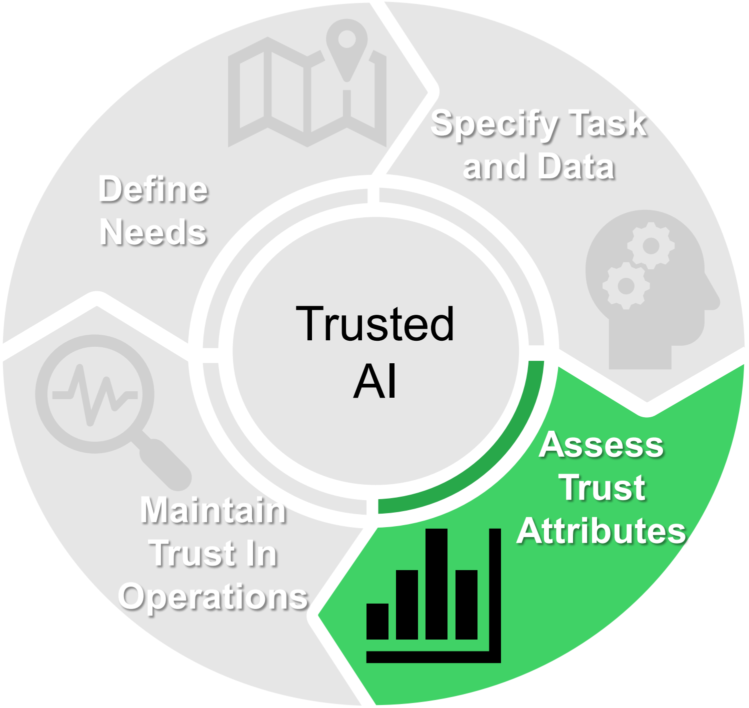
Plan for collection, characterization, annotation, and management of data through AI lifecycle



Assess and Enhance Trust

Thread 2

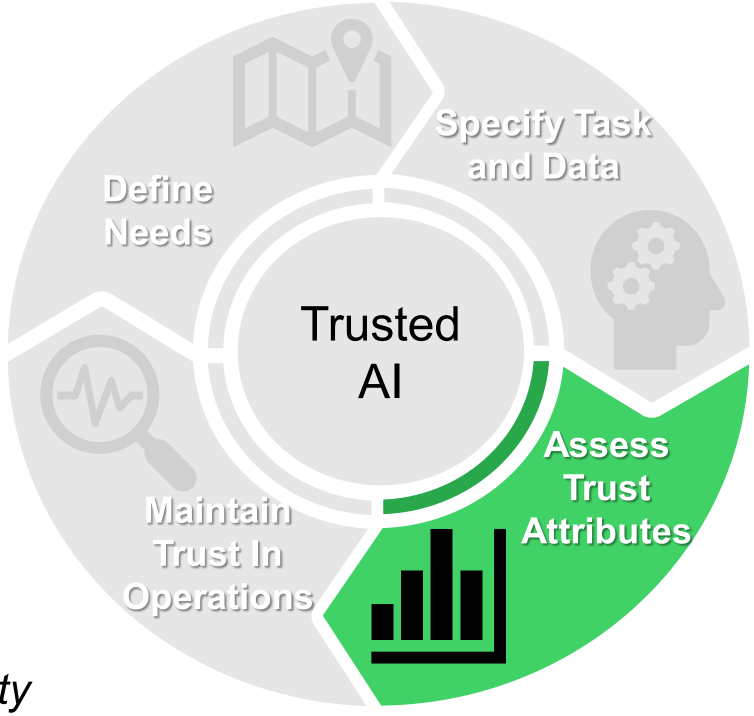
- Evaluate proposed design using attributes of trust:
  - ***Traceability***: *document and maintain artifacts from implementation and evaluation of AI system*
  - ***Stability***: *demonstrate consistency of AI behavior over nominal scope*
  - ***Confidence awareness***: *assess pertinence of inputs and predict uncertainty of output*
  - ***Adversarial resilience:*** *detect and provide stable output when inputs are modified by external processes*
  - ***Interpretability***: *maximize user comprehension of causes leading to AI predictions*
  - ***Fairness***: *demonstrate equitable outcomes to known subgroups and characterize residual biases*
  - ***Familiarity***: *comfort with which a user successfully operates system*



Attributes of Trust

Thread 2

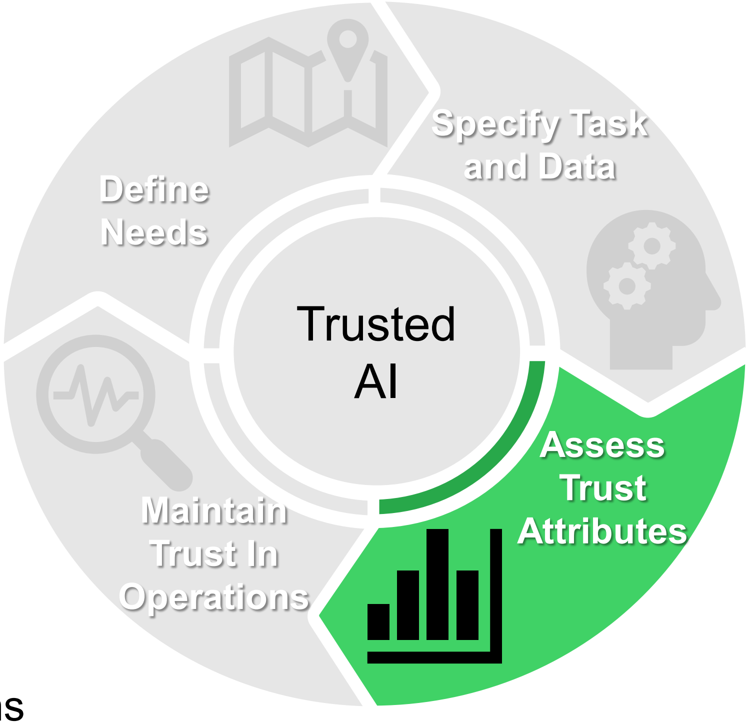
- Stability:
  - *Characterize performance on nominal scope, out-of-scope, and known challenge case data*
  - *Verify consistency of output when input is varied with background noise or inconsequential content*
  - *Consider third-party, blind V&V to assess stability of task and data specification*
  - *Leverage modern development practices to deploy across different platform (MLOps)*
- Confidence awareness:
  - *Determine if inputs are outside nominal, recording anomalies, and lowering prediction confidence (pertinence awareness)*
  - *Provide calibrated estimation of confidence when inputs are nominal*
  - *Provide ability to incorporate and propagate uncertainty from inputs*
- Adversarial resilience:
  - *Consider intents of nefarious actors or processes affecting data integrity*
  - *Evaluate against worst case deployment conditions*
  - *Assess sensitivity to range of attacks and attack strengths and train with them*



Attributes of Trust

Thread 2

- Interpretability:
  - *Consider expertise and training of users when providing interfaces to AI*
  - *Go beyond explanations - demonstrate that additional attributes contribute to user trust*
  - *Prefer algorithms with higher inherent interpretability*
  - *Prefer annotation methods that include concept attribution*
  - *Incorporate user input on relevant features, consider incorporation in model or UX*
  - *Provide evidence for prediction when requested by user:*
    - Display input or feature attributions
    - Display statistics of data and metadata
    - Query influential or similar training examples
  - *Study user engagement and get feedback on application utility:*
    - Identify attribution types that are most relevant
    - Record user expectation of model performance based on attributions (Familiarity)

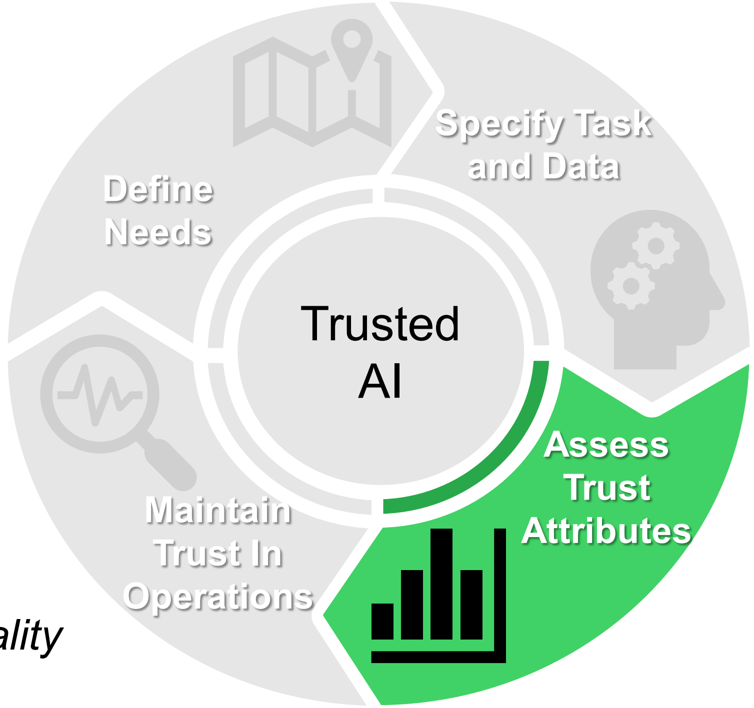


Attributes of Trust

Thread 2

- Fairness:
  - *Leverage EDA to monitor subgroup disparities (Data Specification)*
    - Track performance metrics on subpopulations
    - Note degree of class separability and background diversity
  - *Augment disparities in data and annotation representation between subgroups*
  - *Estimate risk of unresolved biases in data or model*

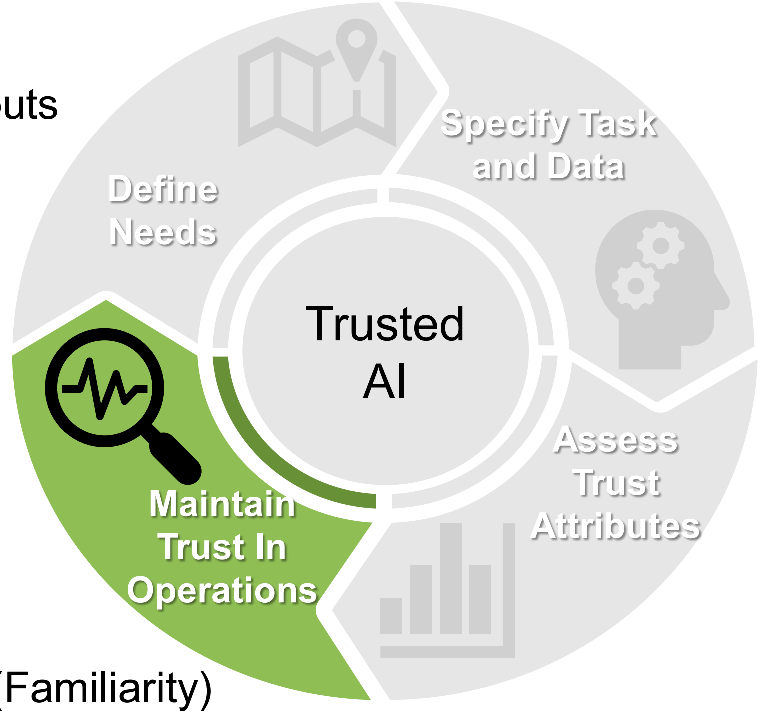
- Familiarity:
  - *Facilitate early and frequent user interaction and incorporate feedback*
  - *Quantify and trend alignment between user actions and AI predictions (Pre-deployment)*
  - *Consider AI validation through tasks that gradually increase task criticality*



Deployment, Monitoring, and Control

Thread 3

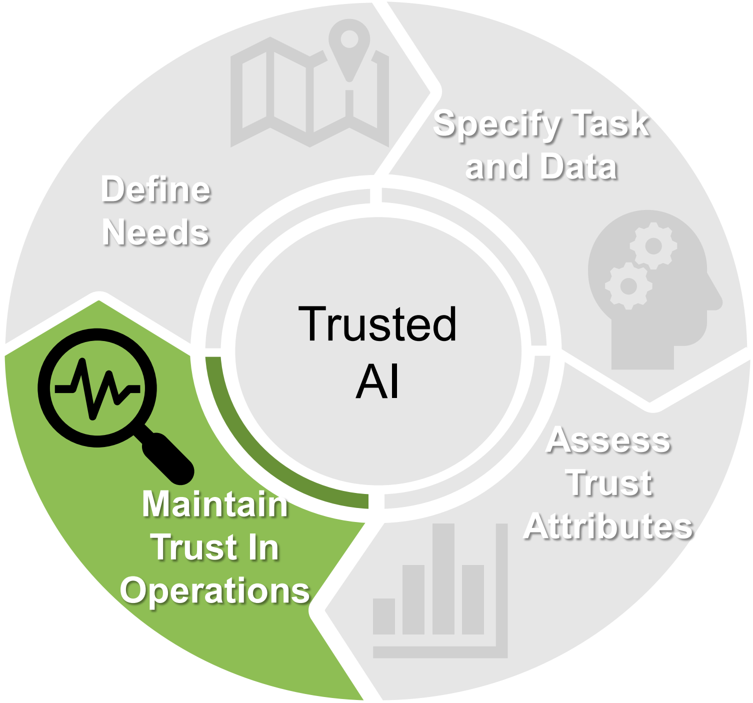
- Structured Deployment
  - *Note differences in target hardware and environment (Traceability)*
  - *Perform limited V&V to reaffirm task alignment and trust attributes:*
    - Verify modular sub-tasks adhere to expected performance
    - Evaluate stability and verify confidence calibration over nominal inputs
    - Record anomalous data and capture SME feedback
  - *Capture data representative of target environment*
    - Assess degree of distribution shift from training data (Data specification)
    - Quantify covariate and prior shift and update nominal / out-of-scope definitions
    - Assess risk for concept, conditional, and sensor shift and estimate impact
  - *Support gradual roll out of AI capability:*
    - Enable shadow mode operation for assessment of user alignment (Familiarity)
    - Support transition from limited to full operations
    - Deploy AI in roles of increasing autonomy and criticality



Deployment, Monitoring, and Control

Thread 3

- Monitoring
  - *Develop upstream and downstream assurance systems to detect undesirable behavior*
  - *Provide metrics that track system degradation or system challenges*
    - Record data and AI prediction statistics over time
    - Identify out-of-scope data and record their frequency
    - Monitor computational trends such as runtime, convergence, memory usage, instrument quality

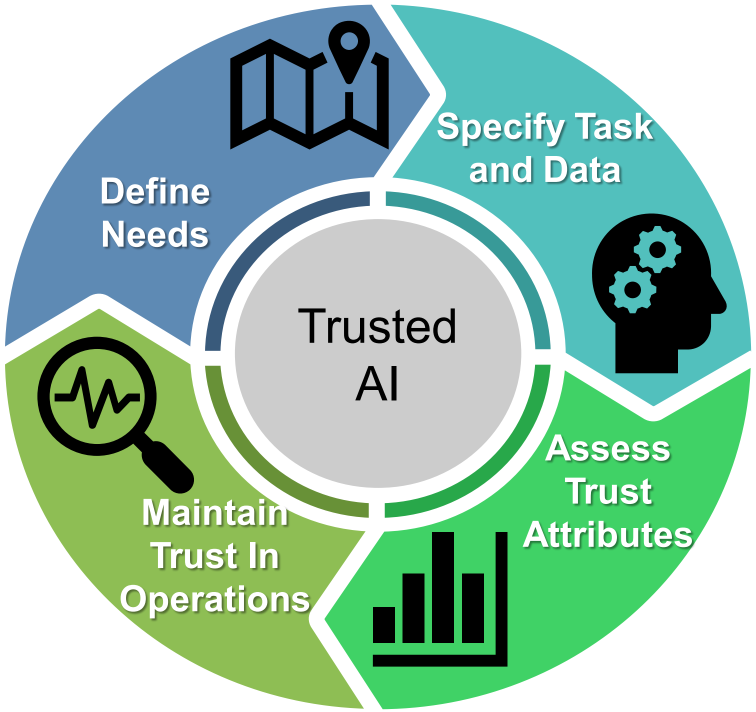


- Control
  - *Provide means for user intervention:*
    - Alert user to system degradations
    - Always obey user request for AI termination
  - *Include fallback systems for when AI termination occurs*
  - *Develop secondary assurance systems to prevent high risk behavior*
  - *Engage with users to consider means for re-fining AI behavior without re-training*

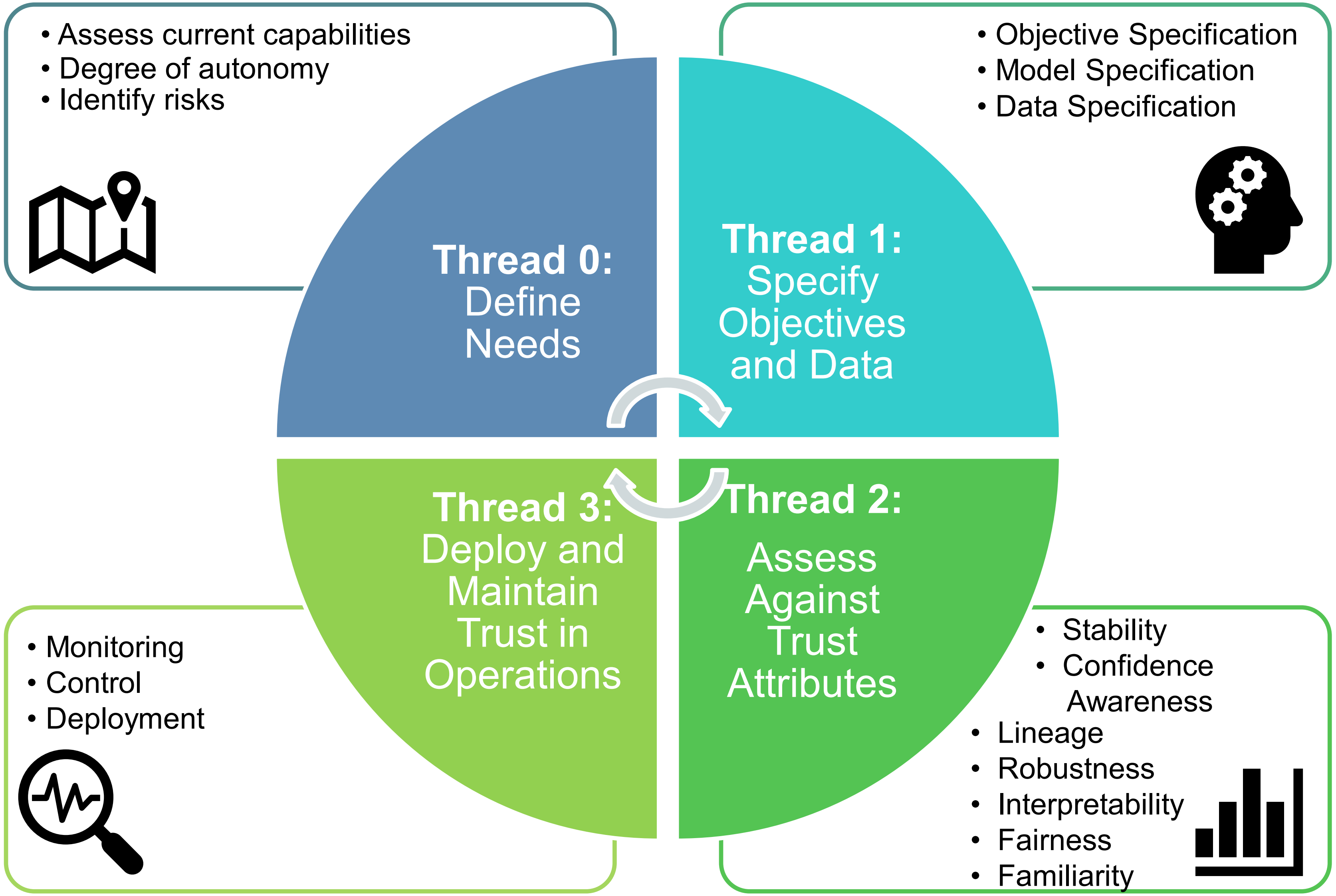
Concluding Remarks

- **The promise of AI**:
  - *AI enabled capabilities are necessary in applications with limited opportunity for communication and control, harsh environments, and where some degree of autonomous discovery is required*

- **Where the Framework fits in**:
  - *Frame AI-based capabilities as trustable*
  - *Identify stakeholders that must be convinced*
  - *Actively manage expectations*
  - *Provide realistic roadmaps for how AI can be implemented and verified*
  - *Mitigate risk of unexpected AI behavior early in the development cycle*
  - *Prepare for the AI monitoring after deployment*



Backup



• Assess current capabilities
• Degree of autonomy
• Identify risks
Thread 0: Define Needs
• Objective Specification
• Model Specification
• Data Specification
Thread 1: Specify Objectives and Data
Thread 3: Deploy and Maintain Trust in Operations
• Monitoring
• Control
• Deployment
Thread 2: Assess Against Trust Attributes
• Stability
• Confidence Awareness
• Lineage
• Robustness
• Interpretability
• Fairness
• Familiarity