Implementing Trusted AI with MLOps

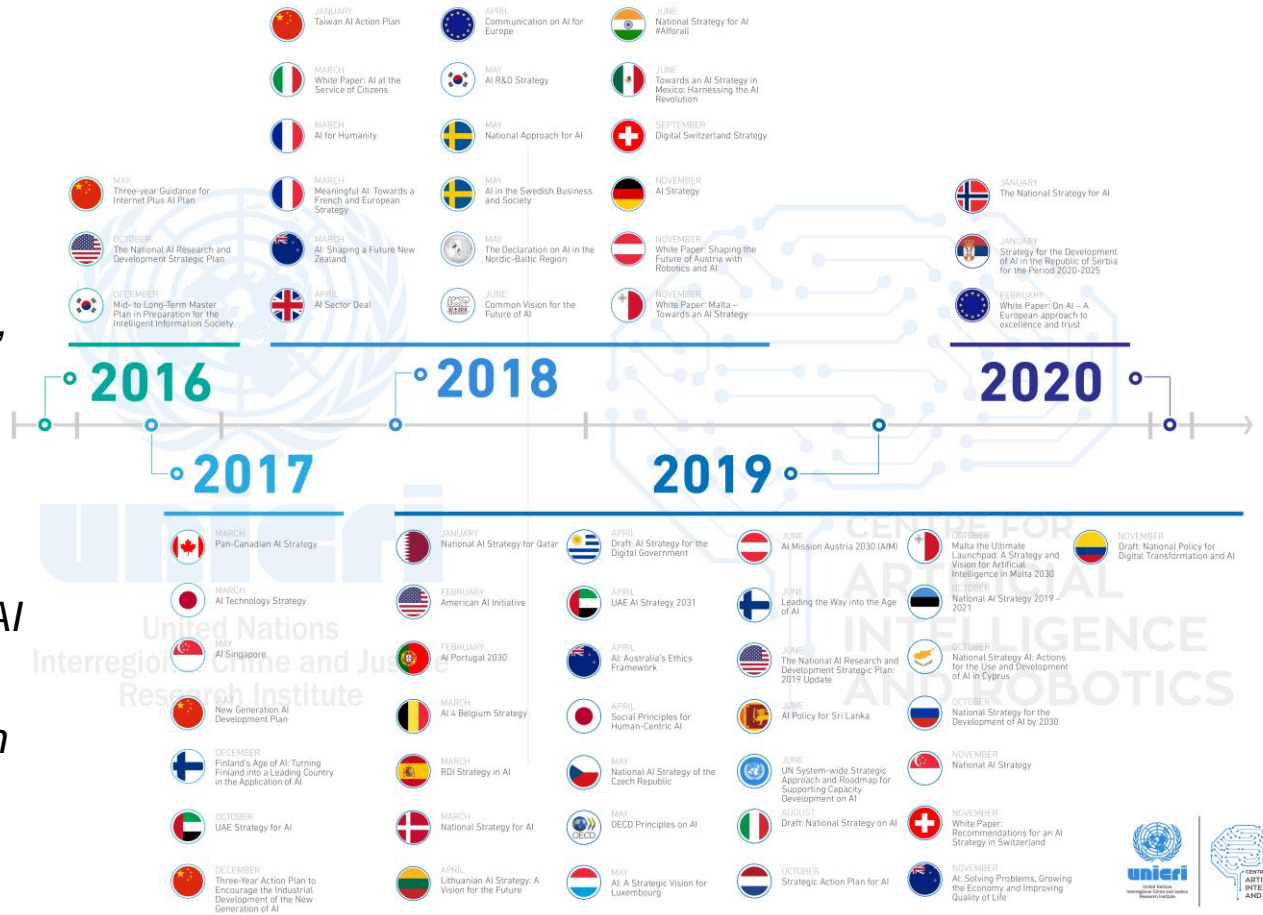
***Dr. Philip Slingerland** (philip.c.sligerland@aero.org)*
***Dr. Max Spolaor** (max.spolaor@aero.org)*
***Lauren Perry** (lauren.h.perry@aero.org)*
***Mike Nemerouf** (michael.k.nemerouf@aero.org)*

DATAWorks 2022
27 April 2022

Approved for Public Release. OTR-2022-00629.

Evolving Landscape of AI Deployment

- 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.)*



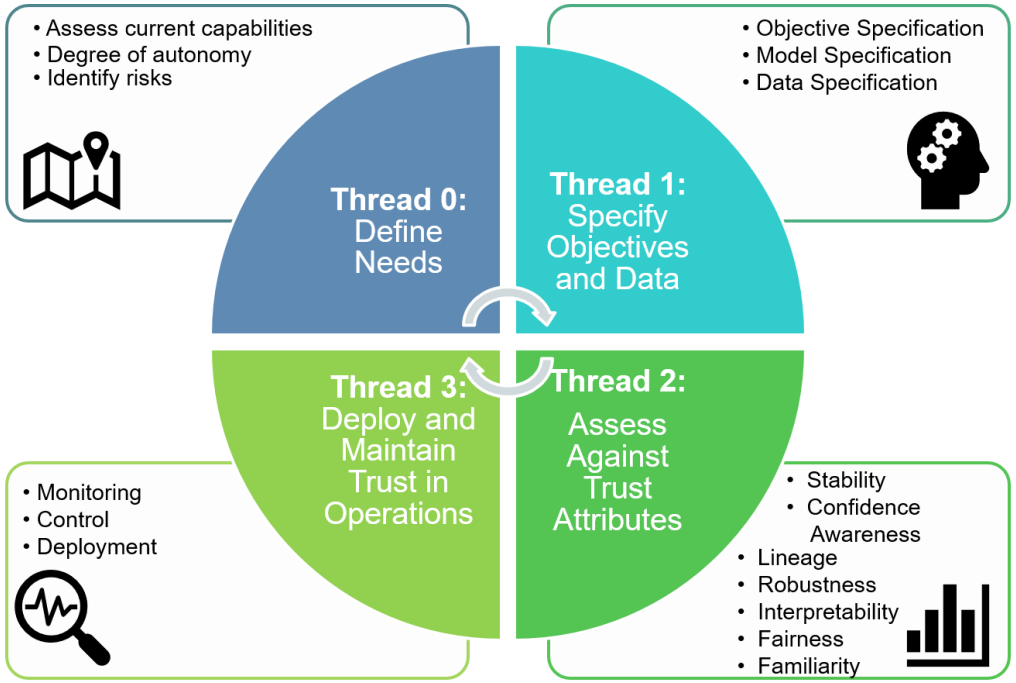
Timeline of strategies, action plans and policy papers setting defining national, regional and international approaches to AI – retrieved from www.unicri.it "UNICRI: United National Interregional Crime and Justice Research Institute"

A Trusted AI Framework can help to navigate through potential impact to new and existing AI projects

Aerospace's Trusted AI Framework

Background

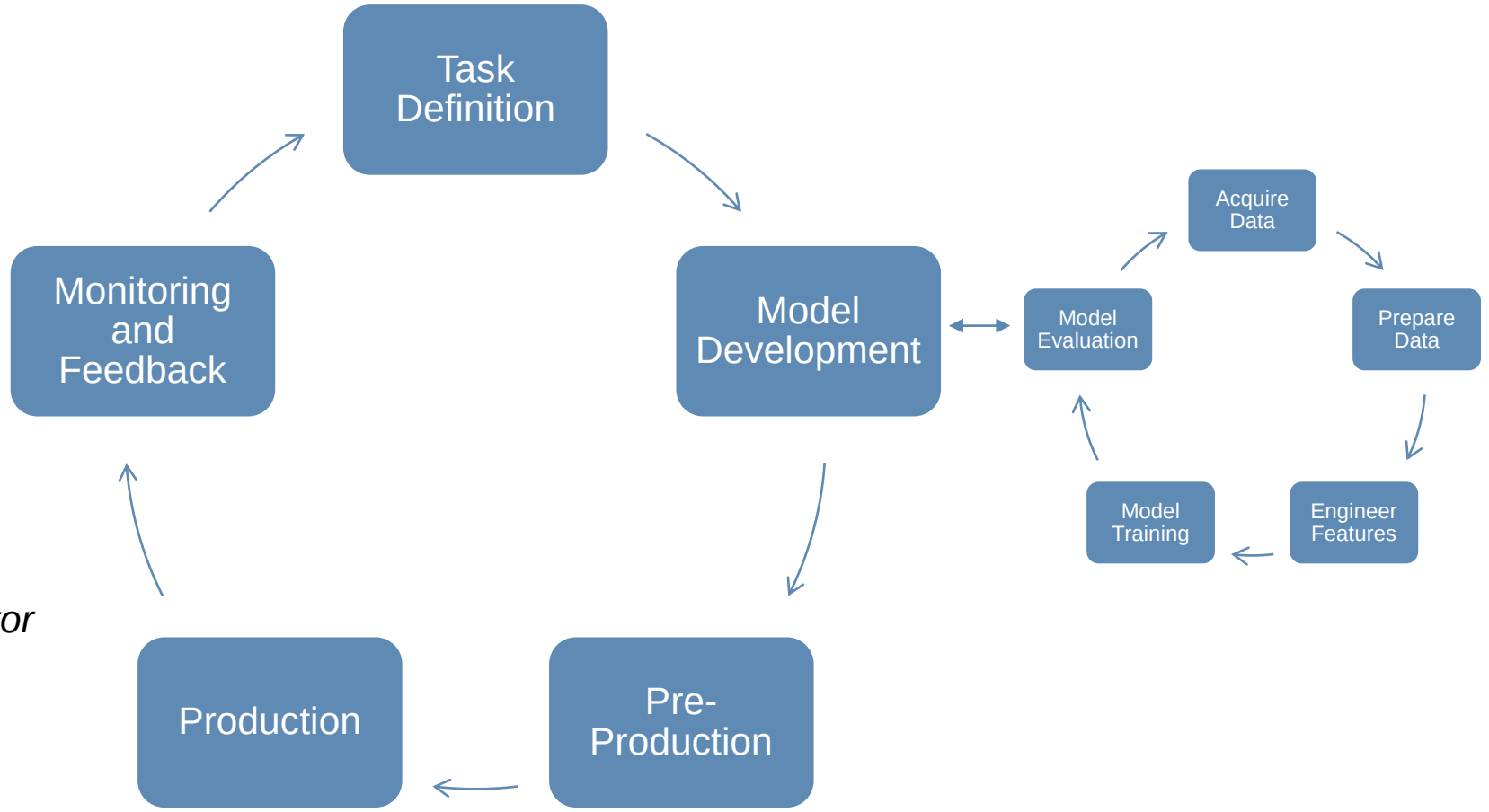
- The Framework was developed to assist Aerospace customers in trustworthy design, implementation, and assessment of AI-based algorithms, with an emphasis on high consequence environments
  - ***Trusted AI**: AI capability that provides sufficient confidence of satisfying user-defined objectives in a proper, interpretable, and safe way over its lifetime*
  - *Threads and attributes of trust relevant to each phase of AI development lifecycle addressed with best practices in mind*



Framework provides principles of trust, but implementation guidance makes them approachable

MLOps Provides Implementation Roadmap

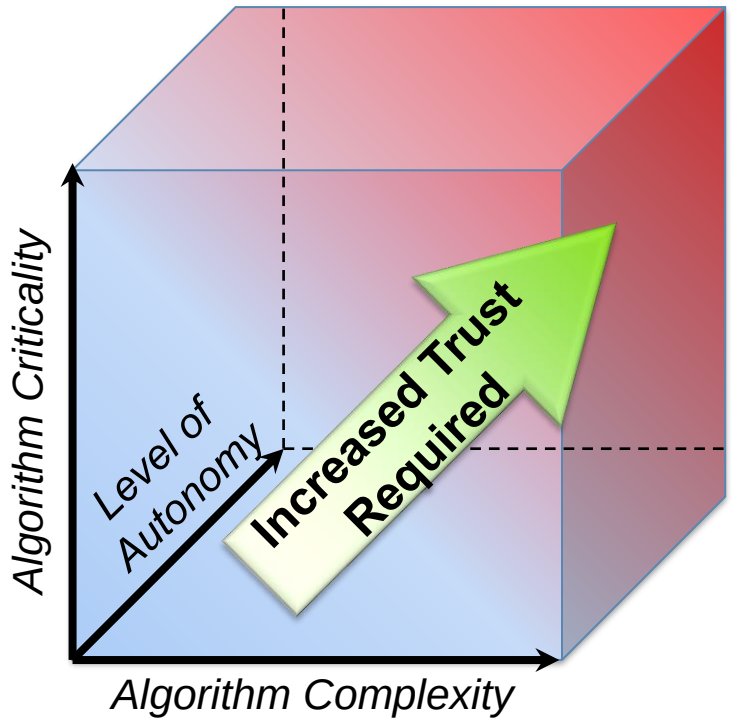
- When deploying AI, MLOps provides maintenance of trust
- Helps realize the value of ML-enabled systems in deployment
  - *Provides mechanism for stakeholder awareness of ML*
  - *Addresses challenges inherent to ML lifecycle*
- Mitigates risk imposed by data driven behavior of systems
  - *Infrastructure and tooling to monitor and track ML performance*
  - *Rapidly address sub-optimal performance in critical situations*



MLOps provides the means through which trust is proven and maintained

Degree of Trust

How Much Trust and MLOps 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 are mitigated through MLOps 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

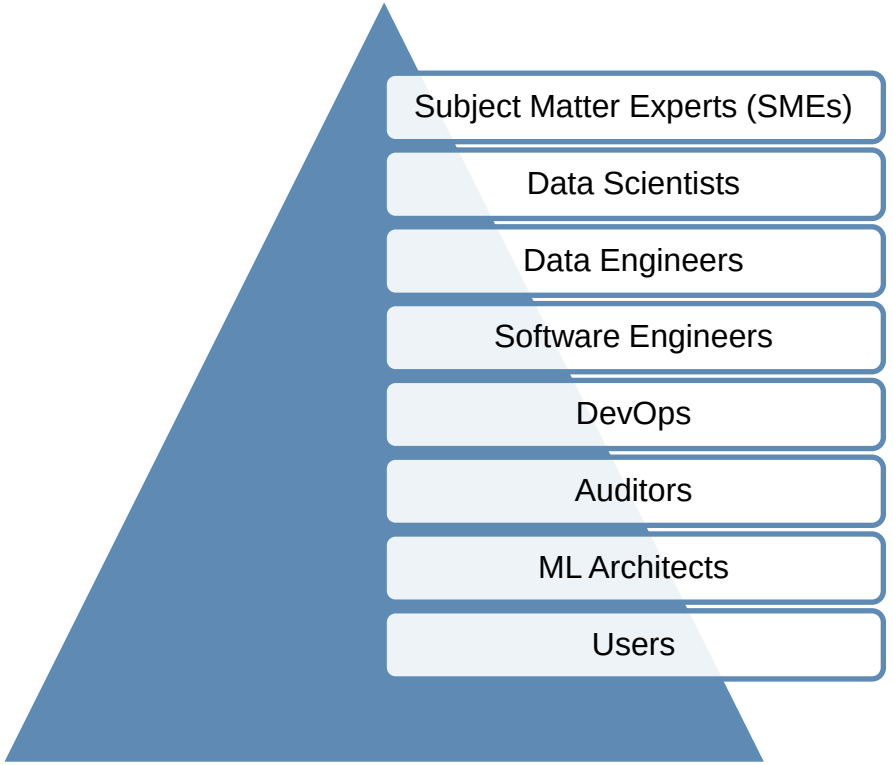
Trust Is More Than Technology

Importance of Cultural Elements

Culture 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
  - *Software, data, and model version control*
  - *Record design decisions and R&D progress*
  - *Log performance metrics throughout development*

Stakeholders

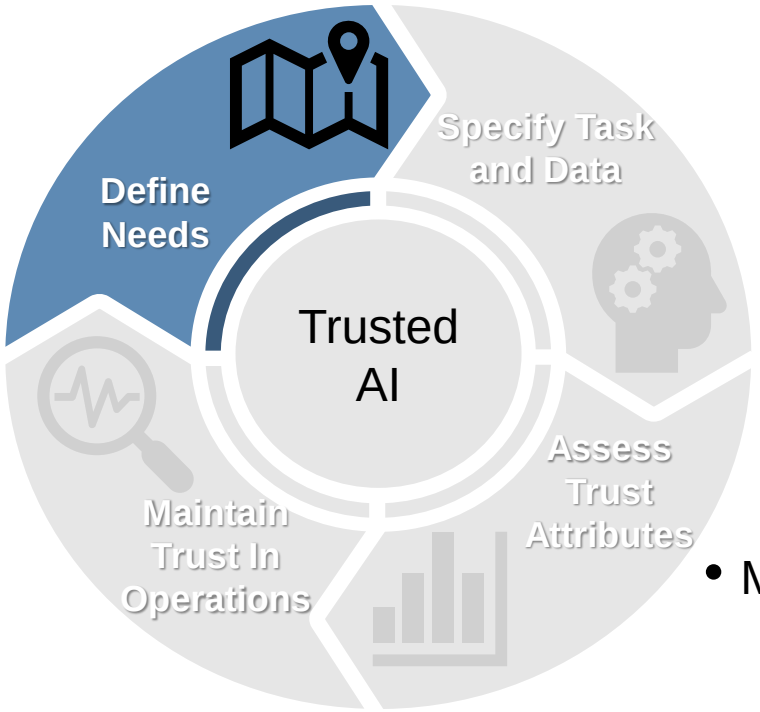


Deliverables: change management plan

Value Proposition and Intent

Prior to Development

- Define and justify need for AI-based capability
  - *Compare with simplest approach or existing capability*
  - *Conduct literature review to support AI appropriateness*
- Create a value proposition
  - *Identify stakeholders*
    - Tailor language to their needs
    - Incorporate perspectives in development plan
  - *Provide metrics of success in target domain*
- Reduce and manage risk
  - *Consider appropriate model complexity*
  - *Isolate operation to single function*
  - *Leverage existing systems and codebases*



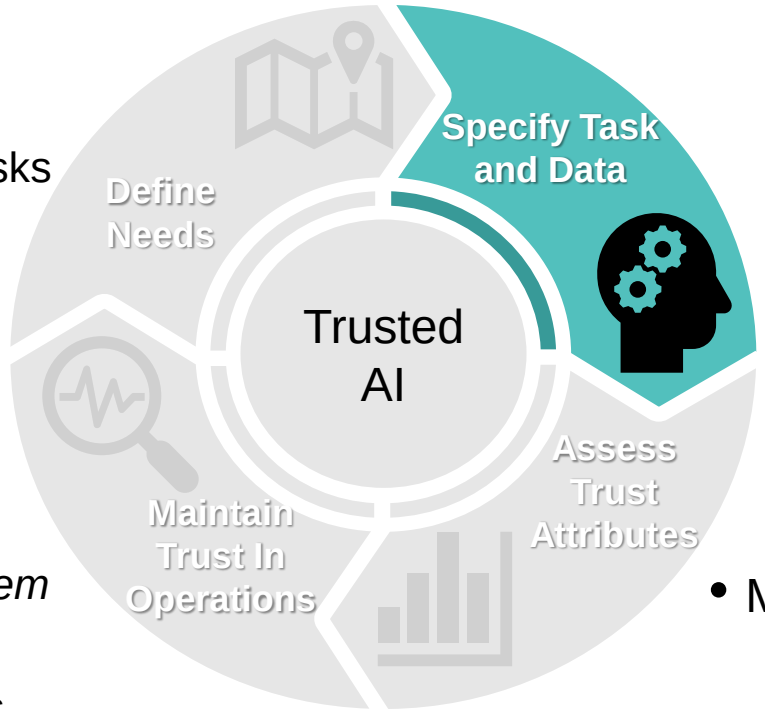
- Stakeholders:
  - *Business Leaders / Tech Directors*
  - *SMEs*
- Metrics:
  - *Performance targets*
  - *Infrastructure, software, and cost constraints*
  - *Key Performance Indicators (KPIs)*

Deliverables: business case document

Task Specification

Thread 1 - Initial Development

- Clear description of AI objective with metrics that demonstrate task alignment
- Translate proposed capability to AI task
  - *Modular description:*
    - Build complexity from verifiable sub-tasks
    - Define requirements to assemble trust
  - *Identify performance metrics relevant to target domain*
    - Make auditable for SMEs to track adherence to business objective
- AI risk identification
  - *Failure modes and how AI could cause them*
  - *Security and adversarial risks*
  - *Business, legal, safety, reputation impacts*



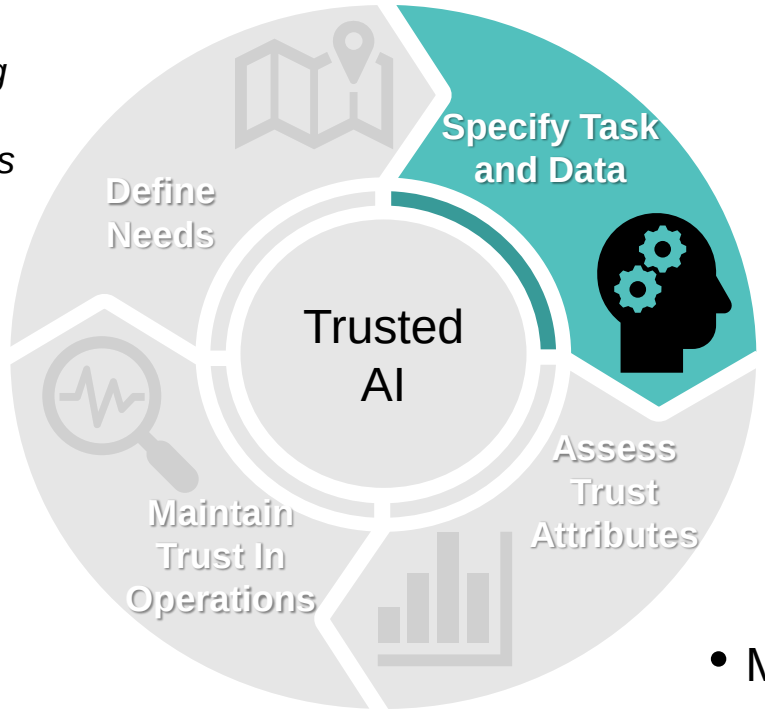
- Stakeholders:
  - *SMEs*
  - *Data Scientists*
- Metrics:
  - *Tailored to use case*
  - *Inform upstream and downstream monitoring*

Deliverables: ADD and deployment plan

Data Specification

Thread 1 - Initial Development

- Plan for collection, characterization, annotation, and management of data through AI lifecycle
- Identify datasets for both development and deployment
  - *Plan for data collection, storage, and partitioning*
  - *Capture details of sensors, data pre-processing*
  - *Capture governance aspects of existing datasets*
- Exploratory data analysis (EDA):
  - *Define nominal and out-of-scope data parameters*
  - *Identify subgroups and their representation*
  - *Identify challenging data examples and mitigation plans*
  - *Perform correlation analysis and select features*
- Develop upstream protection to check for out-of-scope data
  - *Re-route out-of-scope data to alternate system*
  - *Tailor confidence based on scope*
- Provide tools that enhance data pedigree:
  - *Interface to gathering annotations from multiple SMEs, logging annotations from users*
  - *Enable review and disagreement between SMEs*



- Stakeholders:
  - *SMEs*
  - *Data Scientists*
  - *Data Engineers*

- Many tools becoming available: Druid, DVC, Hopsworks, Pachyderm, Rok, Snorkel

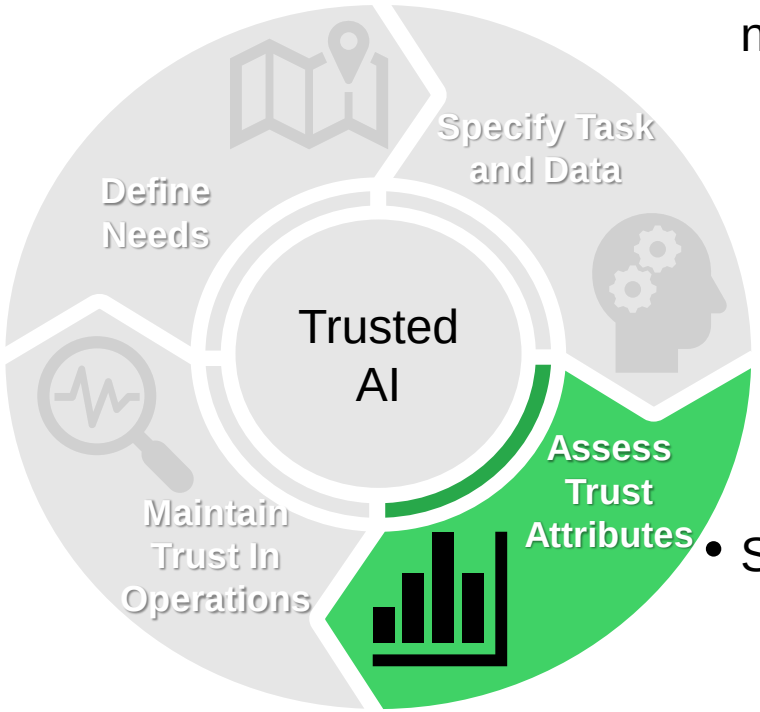
- Metrics:
  - *Resources to prepare and govern data*
  - *Representation of subgroups*

Deliverables: data acquisition, QA, and governance plan

Attributes of Trust

Thread 2 - Building Trust with MLOps

- **Traceability**: document and maintain artifacts from implementation and evaluation of AI system
  - *Record all data preparation, curation, processing steps*
  - *Provide version control to support rapid prototyping*
  - *Document model selection, performance metrics, and R&D trajectory*
- Stakeholders:
  - *SMEs*
  - *Data Scientists*
  - *DevOps*
  - *Auditors*



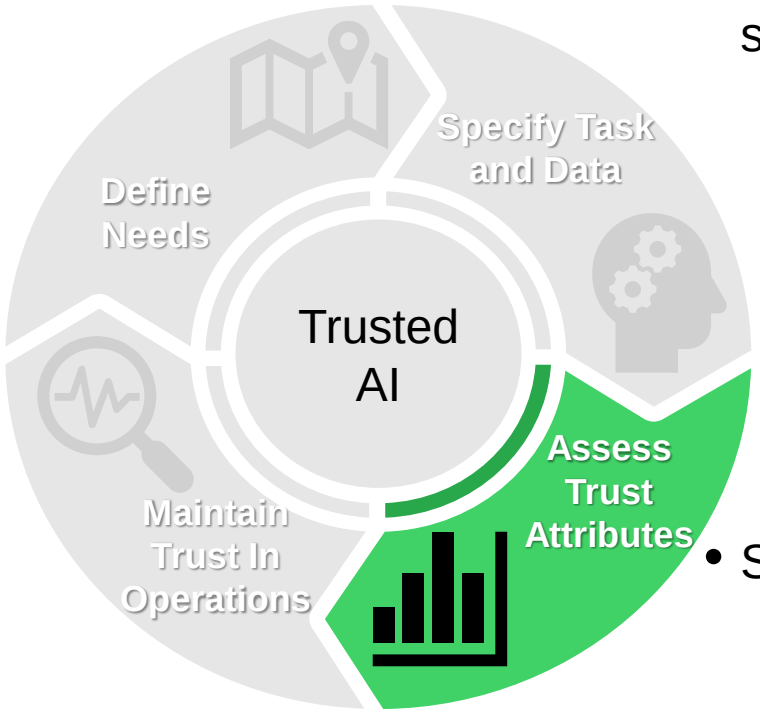
- **Stability**: demonstrate consistency of AI behavior over nominal scope
  - *Characterize performance on nominal scope, out-of-scope, and known challenge case data*
  - *Verify consistency of output over background variations*
  - *Set baseline for DevOps handoff*
  - *Leverage modern processes to deploy across different platform*
- Stakeholders:
  - *Data Scientists*
  - *Software Engineers*
  - *DevOps*

Tools: MLFlow, Neptune, Weights & Biases

Attributes of Trust

Thread 2 - Building Trust with MLOps

- **Confidence awareness**: assess pertinence of inputs and predict uncertainty of output
  - *Determine if inputs are outside nominal, recording anomalies, and lowering confidence*
  - *Provide calibrated estimation of confidence when inputs are nominal*
  - *Provide ability to incorporate and propagate uncertainty from inputs*
- Stakeholders:
  - *SMEs*
  - *Data Scientists*
  - *DevOps*
  - *Auditors*



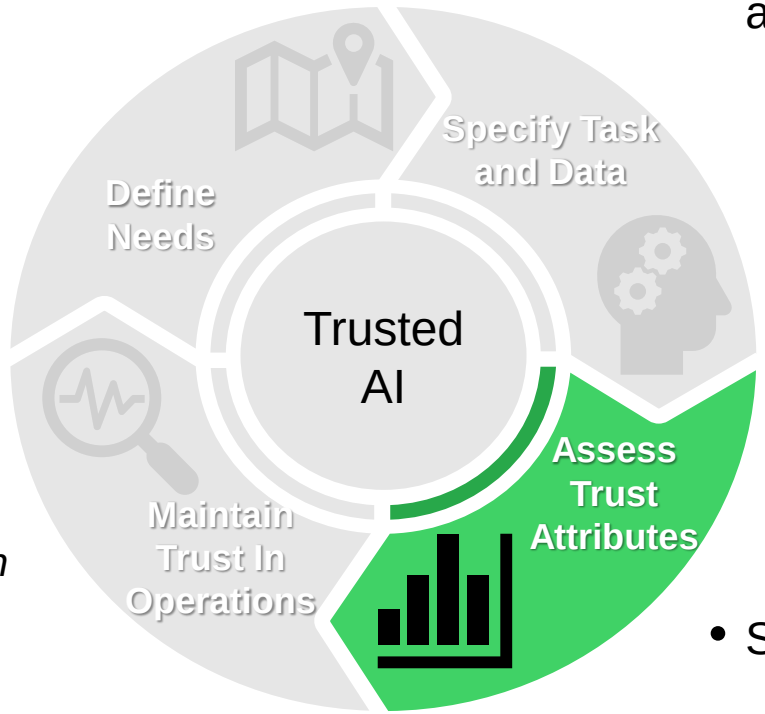
- **Adversarial resilience**: detect when attacks occur and provide stable output
  - *Consider worst case deployment conditions, assuming users want to inflict damage or reverse engineer data*
  - *Assess sensitivity to range of attacks and their strengths*
  - *Consider adversarial training*
- Stakeholders:
  - *SMEs*
  - *DevSecOps*
  - *Auditors*

Tools: Adversarial Robustness Toolbox, ATLAS

Attributes of Trust

Thread 2 - Building Trust with MLOps

- **Interpretability**: maximize user comprehension of causes leading to AI predictions
  - *Consider algorithm interpretability*
  - *Leverage annotations that include concept attribution*
  - *Incorporate user input on 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 training examples
  - *Study user engagement and get feedback on application utility*
- Stakeholders:
  - *SMEs*
  - *Auditors*
  - *Users*



- **Fairness**: seek equitable outcomes to known subgroups and characterize residual biases
  - *Leverage EDA to monitor subgroup disparities*
    - 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*
- Stakeholders:
  - *Business Leaders*
  - *SMEs*
  - *Auditors*

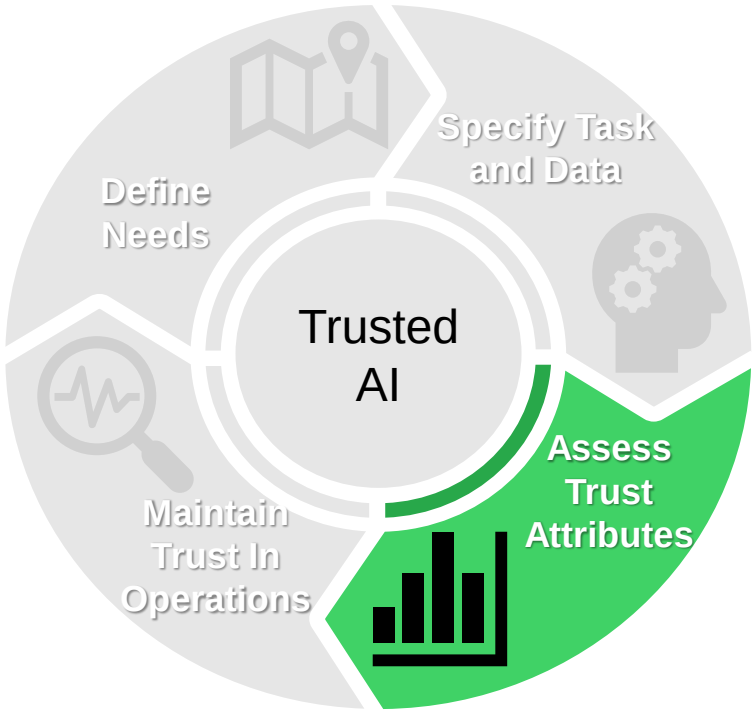
Tools: AI Fairness 360, AI Explainability 360, Explainable AI (GCP)

Attributes of Trust

Thread 2 - Building Trust with MLOps

- **Familiarity**: comfort with which a user successfully operates system
  - *Facilitate early and frequent user interaction and incorporate feedback*
  - *Quantify alignment between user actions and AI predictions and adherence to KPIs*
  - *Consider AI validation through tasks that gradually increase task criticality*

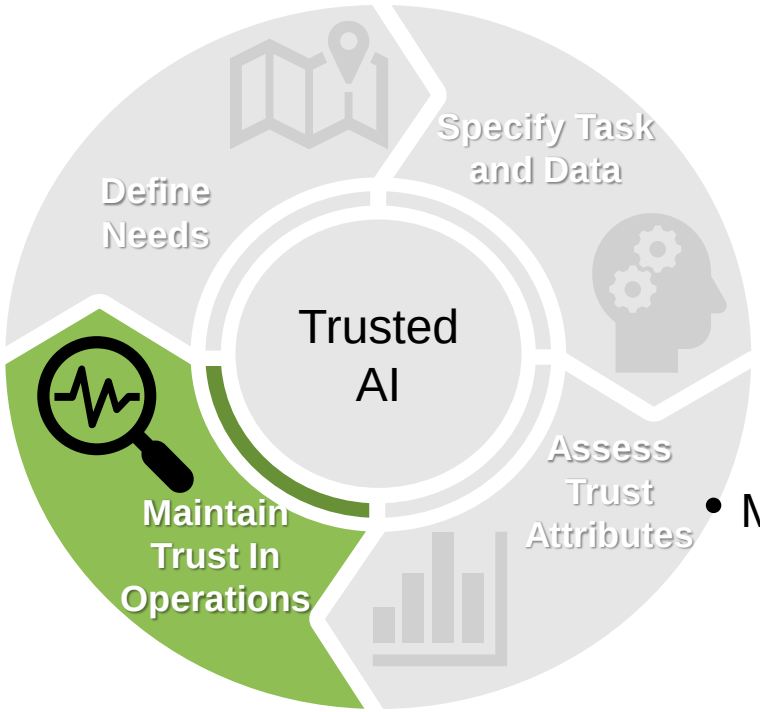
- Stakeholders:
  - *SMEs*
  - *Users*



Pre-deployment, Monitoring, and Control

Thread 3 - Deployment

- Pre-deployment
  - *Transition to target hardware and environment*
  - *Perform limited V&V to reaffirm task alignment and trust attributes*
  - *Capture data representative of target environment*
    - Assess degree of distribution shift
    - 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
    - Deploy AI in roles of increasing scale, autonomy, and criticality



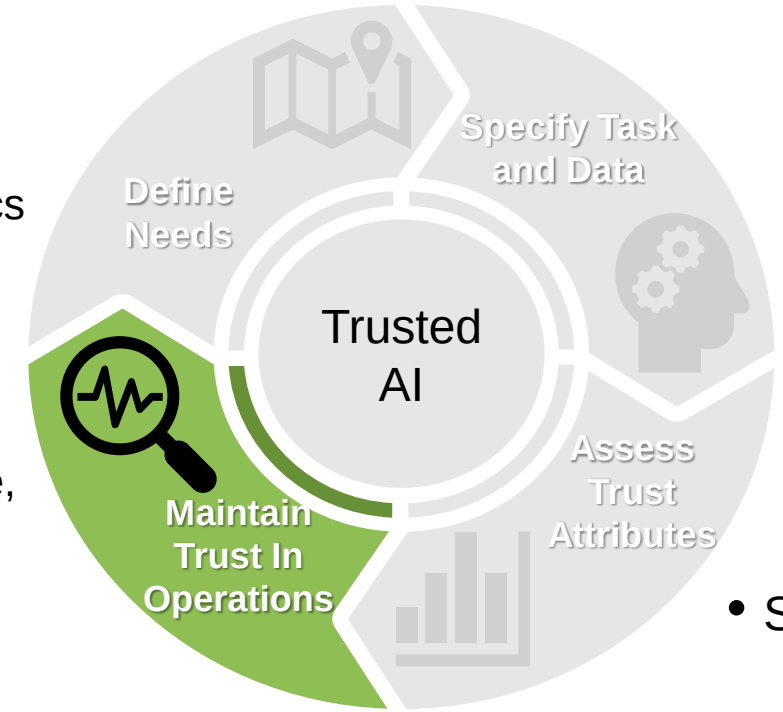
- Stakeholders:
  - *SMEs*
  - *DevOps*
  - *Software Engineers*
  - *Data Scientists*
- Metrics:
  - *Adherence of modular sub-tasks to expected KPIs*
  - *Stability and confidence calibration over nominal inputs*
  - *Covariate / prior shift and reassess nominal / out-of-scope*

Tools: Triton

Pre-deployment, Monitoring, and Control

Thread 3 - Deployment

- **Monitoring**
  - *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 frequency
    - Monitor computational trends such as runtime, convergence, memory usage, instrument quality
- Stakeholders:
  - *SMEs*
  - *Auditors*



- **Control**
  - *Provide means for user intervention:*
    - Alert to system degradations
    - Enable AI termination
  - *Include fallback systems for when AI termination occurs*
  - *Develop secondary assurance systems to prevent failures*
  - *Engage with users to consider means for re-fining AI behavior without re-training*
- Stakeholders:
  - *Users*
  - *Auditors*

Concluding Remarks

- Summary
  - *Organizations developing AI strategy have opportunity to build culture and infrastructure supporting trust*
  - *When proposing AI solutions for safety critical applications, trust will be required*
  - *When preparing to deploy AI, MLOps will implement and maintain proof of trust*
- Where the Framework and MLOps fit 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 and execute AI monitoring*
- For further discussion:
  - *References for Aerospace's Trusted AI Framework can be provided upon request*
  - *Forthcoming paper will go into greater detail on alignment between Trusted AI and MLOps with use case application to satellite pose estimation*