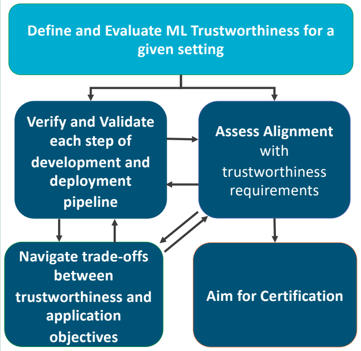
Sandia National Laboratories

TOWARD CERTIFYING TRUSTWORTHY MACHINE LEARNING

Goal: Develop Generalizable Process for ML Certification


Define and Evaluate ML Trustworthiness for a given setting

Verify and Validate each step of development and deployment pipeline

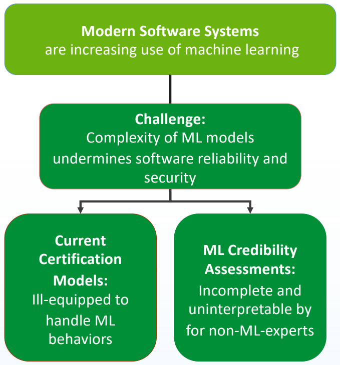
Assess Alignment with trustworthiness requirements

Navigate trade-offs between trustworthiness and application objectives

Aim for Certification

- ➢ Aim to provide criteria to satisfy exacting standards for ML application

The State of Practice


Modern Software Systems are increasing use of machine learning

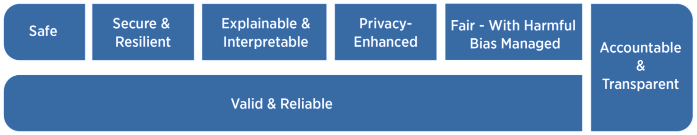
Challenge: Complexity of ML models undermines software reliability and security

Current Certification Models: Ill-equipped to handle ML behaviors

ML Credibility Assessments: Incomplete and uninterpretable by non-ML-experts

- ➢ Current ML credibility approaches are necessary but insufficient for certification

"Trustworthy" is Application Specific


Safe

Secure & Resilient

Explainable & Interpretable

Privacy-Enhanced

Fair - With Harmful Bias Managed

Accountable & Transparent

Valid & Reliable

Characteristics of ML Trustworthiness*

Addressing AI trustworthiness characteristics individually will not ensure AI system trustworthiness; tradeoffs are usually involved, rarely do all characteristics apply in every setting… organizations can face difficult decisions in balancing these characteristics.

*NIST AI Risk Management Framework 1.0

We Want Your Input!

- How to integrate this process into software engineering workflows?
- How to make trustworthiness metrics actionable for developers?
- How to align ML evaluation with existing quality assurance processes?

Share your thoughts!

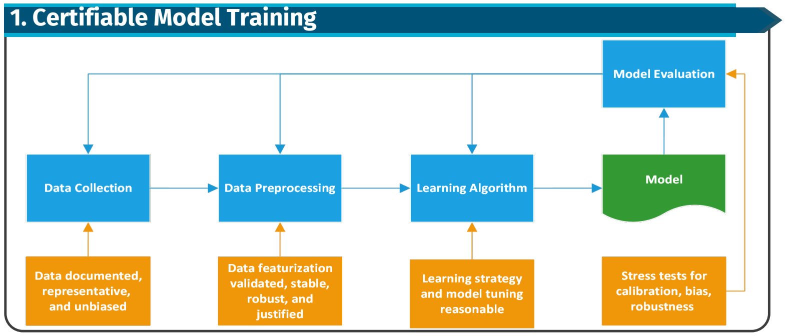
Our Approach

Inspired by multi-tiered software testing strategies, we will:

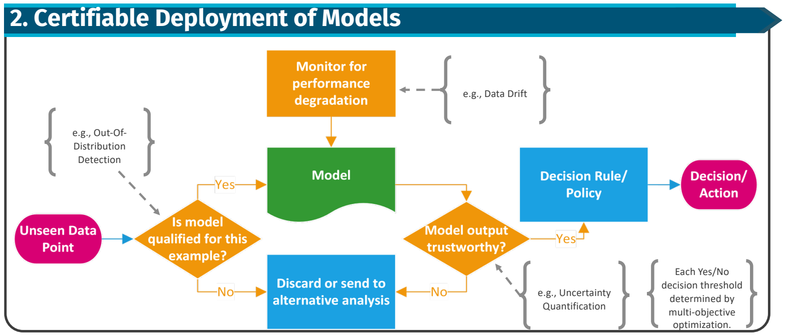
1. **Assess trustworthiness throughout the ML development lifecycle**
2. **Conduct system-level evaluations of trustworthy properties such as transparency, and fairness both pre and post deployment.**
3. **Provide means for stakeholders to navigate trade-off decisions among competing objectives**

We're working toward…

1. Certifiable Model Training


Model Evaluation
Data Collection
Data Preprocessing
Learning Algorithm
Model
Data documented, representative, and unbiased
Data featurization validated, stable, robust, and justified
Learning strategy and model tuning reasonable
Stress tests for calibration, bias, robustness

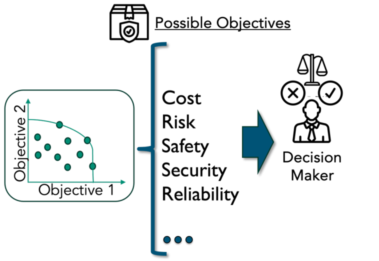
2. Certifiable Deployment of Models


Monitor for performance degradation
e.g., Data Drift
e.g., Out-Of-Distribution Detection
Yes
Model
Decision Rule/ Policy
Decision/ Action
Unseen Data Point
Is model qualified for this example?
Model output trustworthy?
Yes
No
Discard or send to alternative analysis
No
e.g., Uncertainty Quantification
Each Yes/No decision threshold determined by multi-objective optimization.

3. Robust, Multi-Objective Control Over System Characteristics

Given many, sometimes competing, trust objectives, how do we:

- Empower ML system stakeholders to understand and control tradeoffs?
- Determine which solutions are "good enough" for the target application?



Michael C. Darling, Reed M. Milewicz, Erin C.S. Acquesta, Karin M. Butler, Edward R. Carroll, Esha Datta, J. Jake Nichol, Mark A. Smith, Ann E. Speed

Departments: Secure Algorithms, Software Engineering and Research, Computational Data Science, Applied Cognitive Science, Geophysical Detection Programs, Scientific Machine Learning, Advanced Decision Analytics, Cognitive & Emerging Computing

correspondence: michael.darling@sandia.gov

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2024-13731C