WHAT WOULD REVIEW OF AN AI APPLICATION LOOK LIKE?

AI Tabletop Exercise with Argonne and Sandia National Laboratories

Rick Vilim (Argonne National Laboratory)
Art Munson (Sandia National Laboratories)
Matt Dennis (Nuclear Regulatory Commission)

2025-09-24 NRC AI Workshop

SAND2025-11937PE

AI/ML EVALUATION FRAMEWORK

Project Goals:

- Prepare and evaluate a mock submission of an AI/ML safety application.
- Document an evaluation process to inform the regulator and industry.
  - ANL: Representative AI application
  - SNL: Evaluation framework

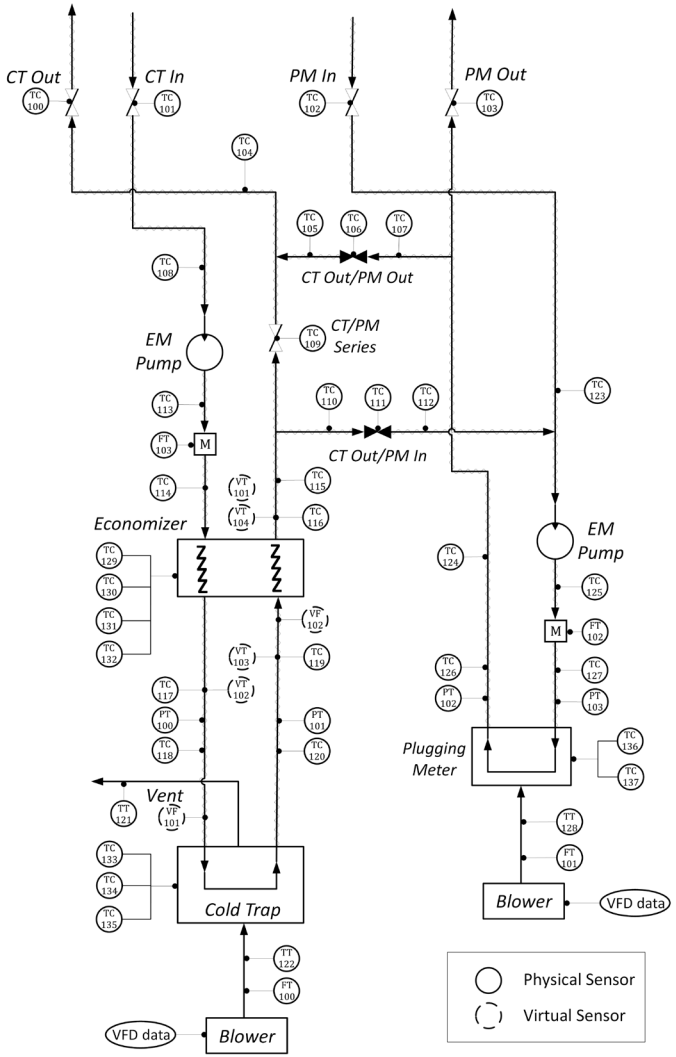
Teams:

- ANL: Akshay Dave, Tim Nguyen, Rick Vilim
- SNL: Art Munson, Mike Smith, Chris Lamb
- NRC: Matt Dennis, Taylor Lamb

AI USE CASE

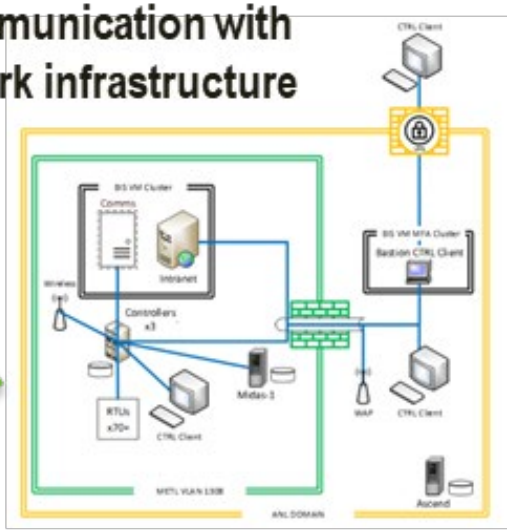
SODIUM PURIFICATION: A SAFETY SIGNIFICANT SYSTEM

METL is a semi-scale advanced reactor sodium facility at ANL



Sodium Purification System

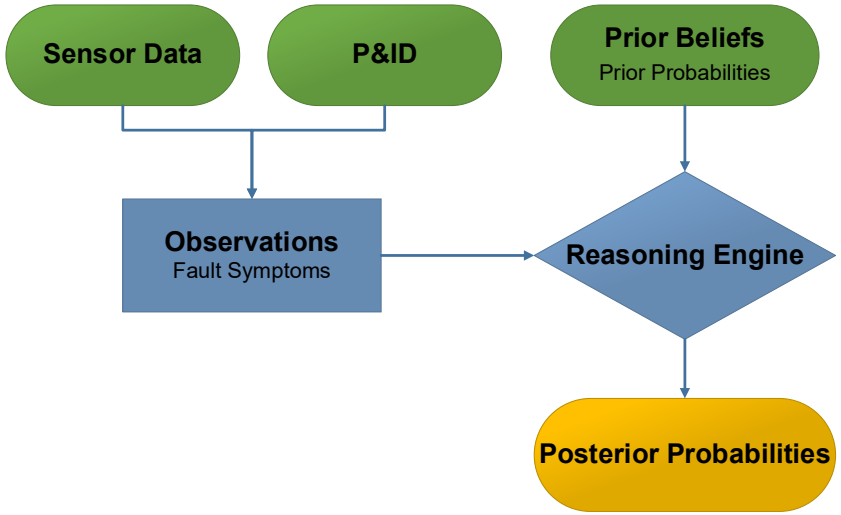
Real-time communication with existing network infrastructure



Argonne NATIONAL LABORATORY

AUTOMATED REASONING FOR ONLINE MONITORING

- Physics-based diagnosis:
  - Enables detection and diagnosis of component and sensor faults
  - Provides robust treatment of uncertainty in the reasoning process



Probabilistic Reasoning Framework in PRO-AID

T. N. Nguyen, et al. "A digital twin approach to system-level fault detection and diagnosis for improved equipment health monitoring." *Annals of nuclear energy* 170 (2022): 109002.

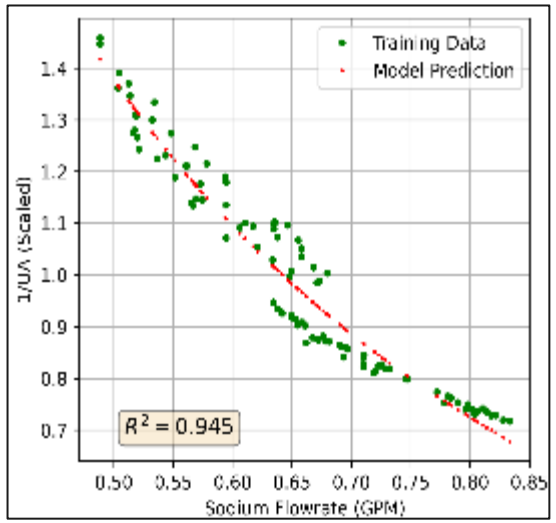
T. N. Nguyen, et al. "A physics-based parametric regression approach for feedwater pump system diagnosis." *Annals of Nuclear Energy* 166 (2022): 108692.

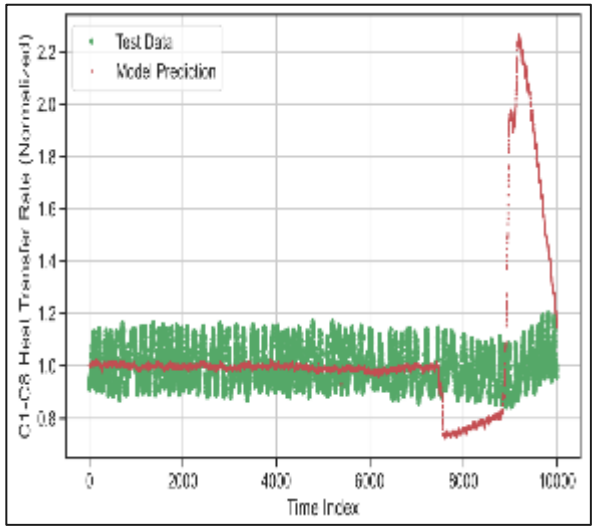
CAPABILITY	DATA DRIVEN	PHYSICS BASED
Robust to operating point change?	N	Y
Diagnosis resolved to specific fault?	N	Y
Rank ordering of likelihood of faults?	N	Y
Designed for engineering systems?	-	Y
Free of need for library of fault signatures?	N	Y
Generates virtual sensors?	N	Y
Adapts upon dropped sensor?	-	Y
Yields component performance index?	N	Y
Supports design of optimal sensor set?	N	Y

Argonne NATIONAL LABORATORY

AUTOMATED REASONING: ONLINE MONITORING OF COLD TRAP

1. Physics-based model for Sodium Purification System (SPS) is calibrated against training data with sensor and model uncertainties calculated.
2. Fault information is implicit in divergence between physics model and measurements
3. A probabilistic reasoning framework then generates a likelihood ranking of faults for SPS components and sensors.





PROAID monitoring METL

Legend
> 50% probability
> 75% probability
> 90% probability

Fault Description	Probability
Cold Trap Air Side Cooling Fault	79.7%
TC 117 Temperature Sensor Fault	76.1%
Economizer Hot Side Leakage	18.1%
FT 103 Flow Sensor Fault	6.1%
Economizer Fouling	3.6%
FT 100 Flow Sensor Fault	2.4%
Cold Trap Sodium Side Leakage	1.9%
Economizer Cold Side Leakage	1.9%
C7 Temperature Sensor Fault	1.3%

1. Model calibrated with training data

2. Fault symptoms as a divergence between model and measurements

3. Monitoring output: Likelihood of faults, including for Cold Trap

Argonne NATIONAL LABORATORY

EVALUATION

TWO EVALUATION STEPS

QUALITATIVE ACCEPTANCE CRITERIA (QAC)

IS THE SAR READY FOR REVIEW?

	Level 1 (not addressed)	Level 2 (basic)	Level 3 (systematic)	Level 4 (comprehensive)
Performance Characterization	Unknown	Low confidence, tested on broad benchmark	Medium confidence, tested on specific task	High confidence, tested on end user, strong UQ
Bias & Robustness Quantification	Not considered	Some consideration	Significant consideration, communicated to user	Continuous testing
Transparency	Black box	Coarse mental model	Useful mental model	Accurate mental model
Safety & Security	Unknown	Awareness of vulnerability, basic guardrails	Quantified vulnerability, broad guardrails	Confidential with high confidence in integrity
Usability	None	Basic	Intuitive and well-targeted	Intuitive and adaptive to user/task

Steinmetz et al. (2025). The Trust Calibration Maturity Model for Characterizing and Communicating Trustworthiness of AI Systems

AI SYSTEM SAFETY EVALUATION

Pillar	Summary
Performance	How well does the model perform its task? * Dataset quality, model accuracy, prediction uncertainty
Bias & Robustness	Will there be surprises when the model is deployed? * Stability, broken assumptions, retraining, …
Transparency	Is the model correct?
Safety & Security Risks	Can the model or infrastructure be subverted? * Data poisoning, model tampering, secure config files, …
Usability	Potential for unsafe decisions? * Misinterpretation, too much trust in AI, …

System View is Critical

- How does X affect the system's safety?
- Most performant AI model might not be the safest AI-powered system.

THEMES IN REQUESTS FOR ADDITIONAL INFORMATION

Methodology

- How were data divided for train, tune, test?
- What is performance on test data?
- How were physics-based components verified?

Characterize AI Behavior

- How does accuracy vary as function of each input?
- What is impact from removing a faulty sensor?

Deployment & Operations

- What is performance acceptance criteria for deployment?
- How to check if uncertainty est. calibrated?
- How to set detection threshold?
- How to set prior probabilities?
- How to decide if recalibration required during operational use?
- How to combine with standard operating procedures for maintenance?
- What is logic for deciding sensor is faulty?

LESSONS LEARNED

- Opportunity to streamline AI safety evaluations:
  - Consider adopting Qualitative Acceptance Criteria (QAC) as readiness checklist.
  - Knowing regulatory evaluation criteria in advance helps applicant be thorough.
- Safety analysis should carefully evaluate deployment considerations during AI R&D.
  - => Safety evaluations should pay close attention here.
- Future Work:
  - Required AI accuracy should be derived from how it impacts system safety.
- Report publication anticipated in early 2026.

Interested in the project report? Contact matthew.dennis@nrc.gov