Trusted Artificial Intelligence Challenge for Systems Engineering: Results and Insights

Overview for the SE4AI/AI4SE Workshop

September 17-18, 2025

Principal Investigator: Emma Meno

Presented By: Sami Saliba

Background and Motivation

- Rapid AI advancements can introduce both performance improvements and risks for mission-critical systems, particularly in uncertain/evolving conditions
  - *Under-utilization* → reduced mission effectiveness
  - *Over-reliance* → misplaced confidence & inadequate oversight

- **Trusted AI Challenge for Armaments Systems Engineering** (SE) aimed to develop SE methods capable of enhancing trustworthiness of AI-enabled systems (AIES), particularly in life-critical operational settings
  - **Trustworthiness** = intrinsic system property, demonstrated through attributes like verifiability, reliability, safety, & transparency
  - Build/operate systems with trustworthy behaviors using less trustworthy components

Primary Research Questions:

1. What SE activities and artifacts are best suited to build trust in AIES?
2. What infrastructure is needed to validate trust of AIES?
3. What workforce skills & abilities are required for integrated product teams to develop and manage these systems?

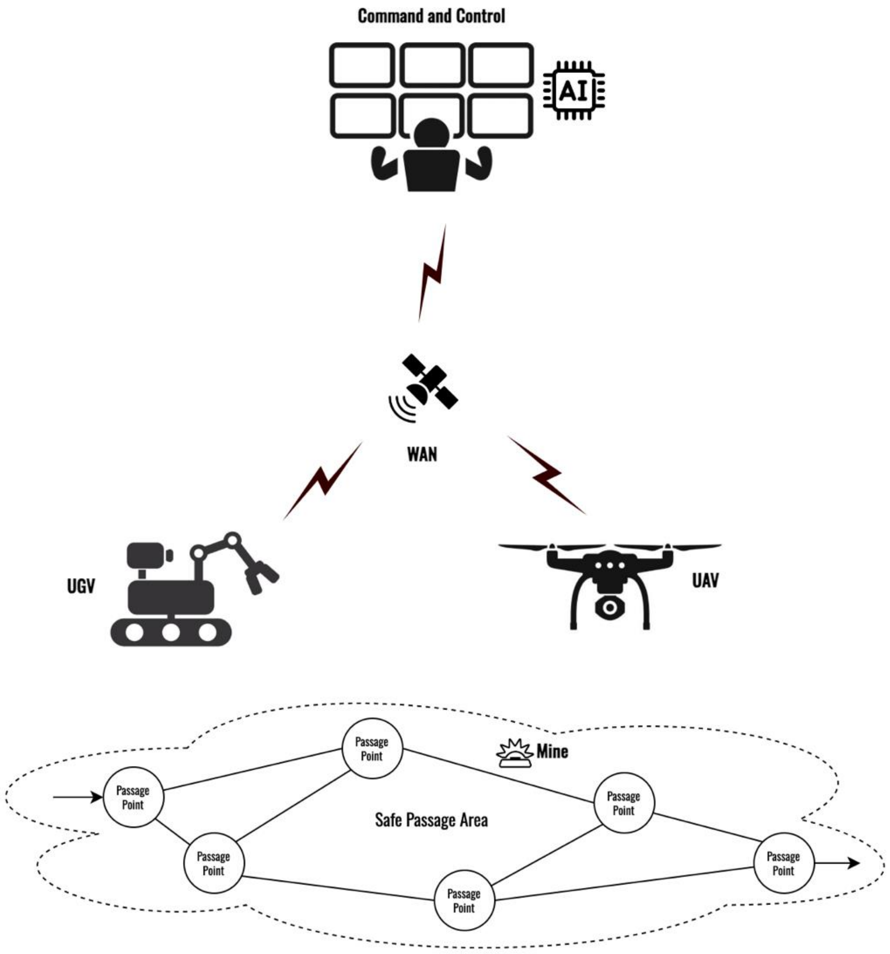
Mission – Operation Safe Passage

Goal: design trusted AI-enabled control system capable of guiding troops through mine-laden terrain

- Determine optimal routing and detection policies that balanced speed, safety, and trustworthiness

Three agents:

- **Unmanned Aerial Vehicles (UAVs)** providing aerial reconnaissance and forwarding mine detection estimates based on environmental scans
- **Mine Detection Systems** with AI-enabled estimator and human SME with varying reliability
- **Unmanned Ground Vehicles** navigating terrain to defuse mines and create path for human troops



Trust integrated by providing mine detection systems ***as-is***

Competition Phases

- **Heat 1: Concept Development**
  - Proposed initial system designs
  - Documented trust assumptions
  - Produced foundational SE artifacts
- **Heat 2: Prototype Demonstration**
  - Developed & deployed prototypes
  - Translated theoretical trust frameworks into operational architectures
- **Heat 3: Lethality Integration**
  - Refined systems to include realistic operational challenges (e.g. complex environmental conditions, additional agents, explicit lethality)

AI Performance Table (% Accuracy)

	Row Index									
Column Index	1	2	3	4	5	6	7	8	9	10
1	0.95	0.95	0.95	0.73	0.95	0.95	0.73	0.94	0.95	0.96
2	0.95	0.95	0.95	0.70	0.95	0.73	0.73	0.95	0.96	0.96
3	0.95	0.95	0.94	0.72	0.70	0.72	0.56	0.95	0.96	0.96
4	0.95	0.56	0.95	0.70	0.68	0.56	0.57	0.95	0.95	0.96
5	0.56	0.57	0.68	0.62	0.94	0.57	0.95	0.96	0.94	0.96
6	0.57	0.57	0.68	0.64	0.94	0.57	0.95	0.96	0.95	0.96
7	0.96	0.57	0.70	0.65	0.94	0.56	0.57	0.95	0.95	0.96
8	0.96	0.96	0.68	0.65	0.94	0.56	0.57	0.94	0.96	0.96
9	0.96	0.96	0.69	0.67	0.70	0.56	0.56	0.96	0.96	0.95
10	0.96	0.96	0.72	0.95	0.72	0.56	0.56	0.96	0.96	0.95

Human Performance Table (% Accuracy)

	Row Index									
Column Index	1	2	3	4	5	6	7	8	9	10
1	0.90	0.90	0.90	0.85	0.90	0.90	0.75	0.90	0.90	0.90
2	0.90	0.90	0.90	0.85	0.90	0.75	0.75	0.90	0.90	0.90
3	0.90	0.90	0.90	0.85	0.85	0.75	0.75	0.90	0.90	0.90
4	0.90	0.56	0.90	0.85	0.70	0.75	0.75	0.90	0.90	0.90
5	0.75	0.75	0.75	0.75	0.90	0.75	0.95	0.90	0.90	0.90
6	0.57	0.75	0.75	0.75	0.90	0.75	0.95	0.90	0.90	0.90
7	0.90	0.75	0.85	0.75	0.90	0.75	0.75	0.90	0.90	0.90
8	0.90	0.90	0.75	0.75	0.90	0.75	0.75	0.90	0.90	0.90
9	0.90	0.90	0.75	0.75	0.85	0.75	0.75	0.90	0.90	0.90
10	0.90	0.90	0.85	0.90	0.85	0.75	0.75	0.90	0.90	0.90

Surface Type Table

	Row Index									
Column Index	1	2	3	4	5	6	7	8	9	10
1	Grassy	Grassy	Grassy	Rocky	Sandy	Sandy	Rocky	Sandy	Sandy	Swampy
2	Grassy	Grassy	Grassy	Rocky	Sandy	Rocky	Rocky	Sandy	Swampy	Swampy
3	Grassy	Grassy	Grassy	Rocky	Rocky	Rocky	Wooded	Sandy	Swampy	Swampy
4	Grassy	Wooded	Grassy	Rocky	Rocky	Wooded	Wooded	Grassy	Grassy	Swampy
5	Wooded	Wooded	Rocky	Rocky	Sandy	Wooded	Grassy	Grassy	Grassy	Grassy
6	Wooded	Wooded	Rocky	Rocky	Sandy	Wooded	Grassy	Grassy	Grassy	Grassy
7	Swampy	Wooded	Rocky	Rocky	Sandy	Wooded	Wooded	Grassy	Grassy	Grassy
8	Grassy	Swampy	Rocky	Rocky	Sandy	Wooded	Wooded	Grassy	Grassy	Grassy
9	Swampy	Swampy	Rocky	Rocky	Rocky	Wooded	Wooded	Swampy	Swampy	Grassy
10	Swampy	Swampy	Rocky	Sandy	Rocky	Wooded	Wooded	Swampy	Swampy	Grassy

Judging Criteria and Scores

Judges used 7-pt Likert Scale to indicate extent of agreement with twelve positively worded statements

- Highest & lowest scores eliminated
- Scores calculated based only on sponsor & industry judges to remove academic bias

Rank your agreement with the following sentences:

1. This team's deliverables described the **<u>systems engineering activities and artifacts</u>** best suited to build trust in AI-enabled systems.

Strongly Disagree	Disagree	Mildly Disagree	Neutral	Mildly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

Score:__________

Notes:

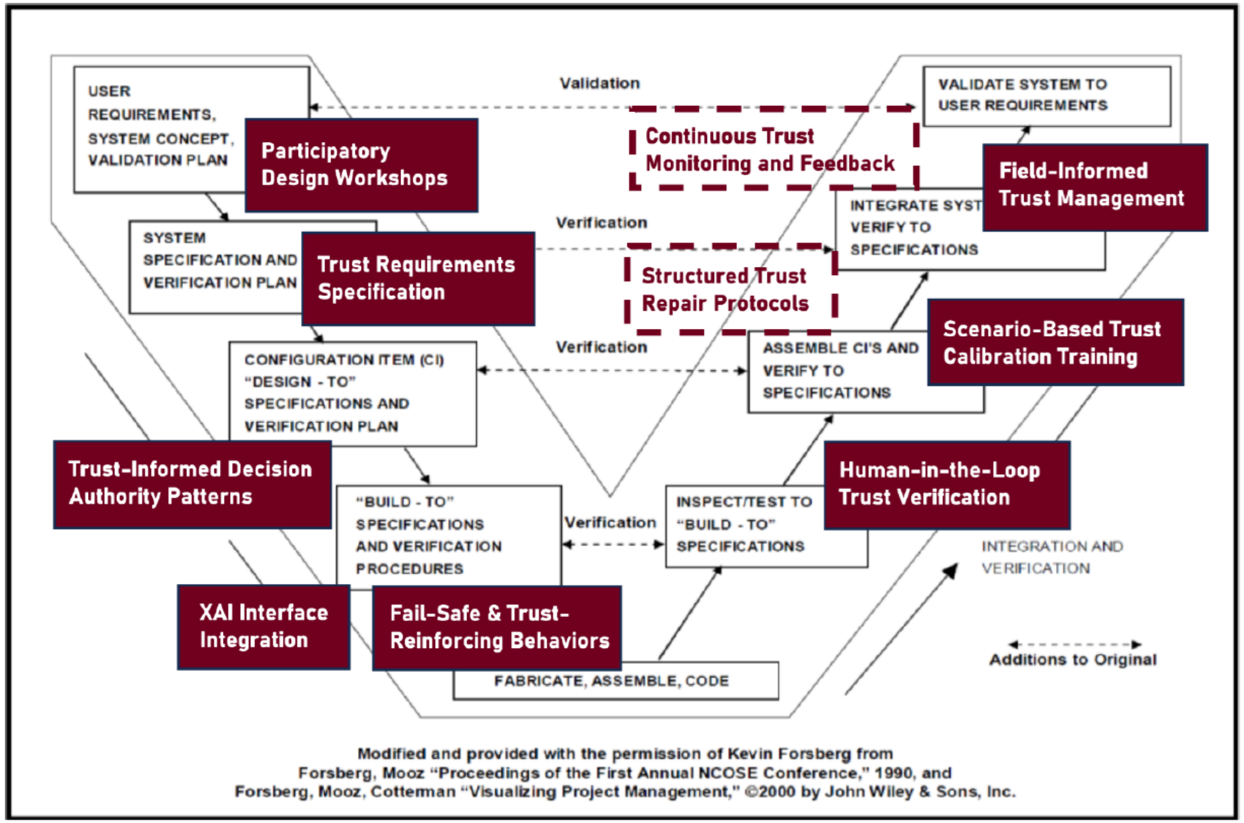
Statements included statements related to research questions as well as design patterns, risk-based monitoring, best practices & novel approaches

	Team 1								Team 2							
Factor	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8
SE activities	6	4	6	6	6	7	1	4	4	4	7	6	3	4	1	1
Trust infrastructure	7	6	7	7	4	6	3	4	4	4	7	7	3	6	3	3
Key workforce skills/abilities	6	4	6	4	3	6	6	2	4	4	7	6	6	7	6	6
Added sensing	5	2	5	5	4	2	2	6	5	2	4	4	5	2	6	5
Lethality	6	2	4	5	6	2	3	6	5	2	6	5	3	2	4	3
Design patterns	5	7	7	6	6	5	7	4	6	7	6	6	6	7	7	4
Risk-based monitoring/mgmt	6	5	7	5	3	6	7	3	5	4	6	6	5	2	6	3
Quantitative methods	6	6	7	5	6	6	3	6	5	6	6	7	6	7	4	5
Best practices	5	6	6	5	5	6	4	3	4	7	6	6	4	6	6	3
Novel approaches	6	5	5	6	4	7	5	4	4	6	4	6	6	7	3	4
Future plans	6	4	4	5	4	2	6	2	4	4	4	5	4	5	3	2
Transition	6	5	7	6	5	6	1	4	4	4	7	6	4	6	1	3
Total	89	70	90	82	69	80	58	58	66	66	91	89	67	78	60	52

Final Approaches and Results

Approaches:

- **GWU** – Established structured framework to evaluate human-AI collaboration architectures
- **Purdue** – Extended aspects of the GWU framework, and utilized RL to explore human-AI interaction scenarios
- **Old Dominion** – Developed a modular, transparent simulation framework
- **Stevens** – Employed robust statistical analyses & Monte Carlo simulations
- 2 **Virginia Tech** – Integrated trust into SE V-model with rigorous human-systems integration activities
- 1 **UVA** – Combined RL with explainable statistical methods & risk monitoring and outlined workforce capability requirements
- **Arizona** – *Did not participate in final heat due to unforeseen circumstances*



VT V-Model Definition

Insights and Recommendations

	Near-Term	Intermediate	Future
Systems Engineering (SE) Activities and Artifacts	• Clearly document explicit trust requirements & trust calibration steps • Organize rigorous participatory design sessions, targeting modularization	• Iterate on human-systems integration artifacts • Establish iterative testing cadence and systematic refinement • Structure V& V workflows	• Incorporate comprehensive SE frameworks with explicit trust requirements, iterative validation, and stakeholder engagement
Trust Validation Infrastructure	• Develop human-in-the-loop simulation frameworks • Engineer clear visualization tools	• Employ large-scale modular simulation validated under variety of conditions for operational robustness • Test operational visualization tools	• Establish resilience pipelines and methods integrated in the mission context • Incorporate operational visualization tools into workflows and pipelines
Workforce Skills & Team Competencies	• Prioritize technical proficiency foundational to interdisciplinary skills in hiring and team-building decisions	• Instantiate evaluations and continuous feedback mechanisms • Evaluate cognitive agility and real-time risk assessment skills for operational effectiveness	• Promote continuous learning through training exercises and workshops to stay current with evolving technologies, threats, and best practices

Future Work

- **Human-AI teaming**, emphasizing human-in-the-loop participation in decision-making vs. fully autonomous systems.
- **Operational resilience testing**, including how to identify system failures under disturbance and how to assess trust degradation and recovery capabilities.
- **Systematic risk management** approaches, involving articulating risk indicators for AI subsystem failure or human-AI communication breakdowns.
- Assessing impact of accessing underlying detection methods and **modifying system components on trustworthiness** to enhance system interoperability and adaptability across various operational contexts.

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under Contract HQ0034-19-D-0003. The Systems Engineering Research Center (SERC) is a federally funded University Affiliated Research Center managed by Stevens Institute of Technology.
Any views, opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Department of Defense nor OUSD(R&E).

Thank you

Stay connected with SERC Online:

Email the PI: Emma Meno

emmam99@vt.edu

Email the research team: VT National Security Institute

emmam99@vt.edu

15 YEARS
SYSTEMS ENGINEERING RESEARCH CENTER