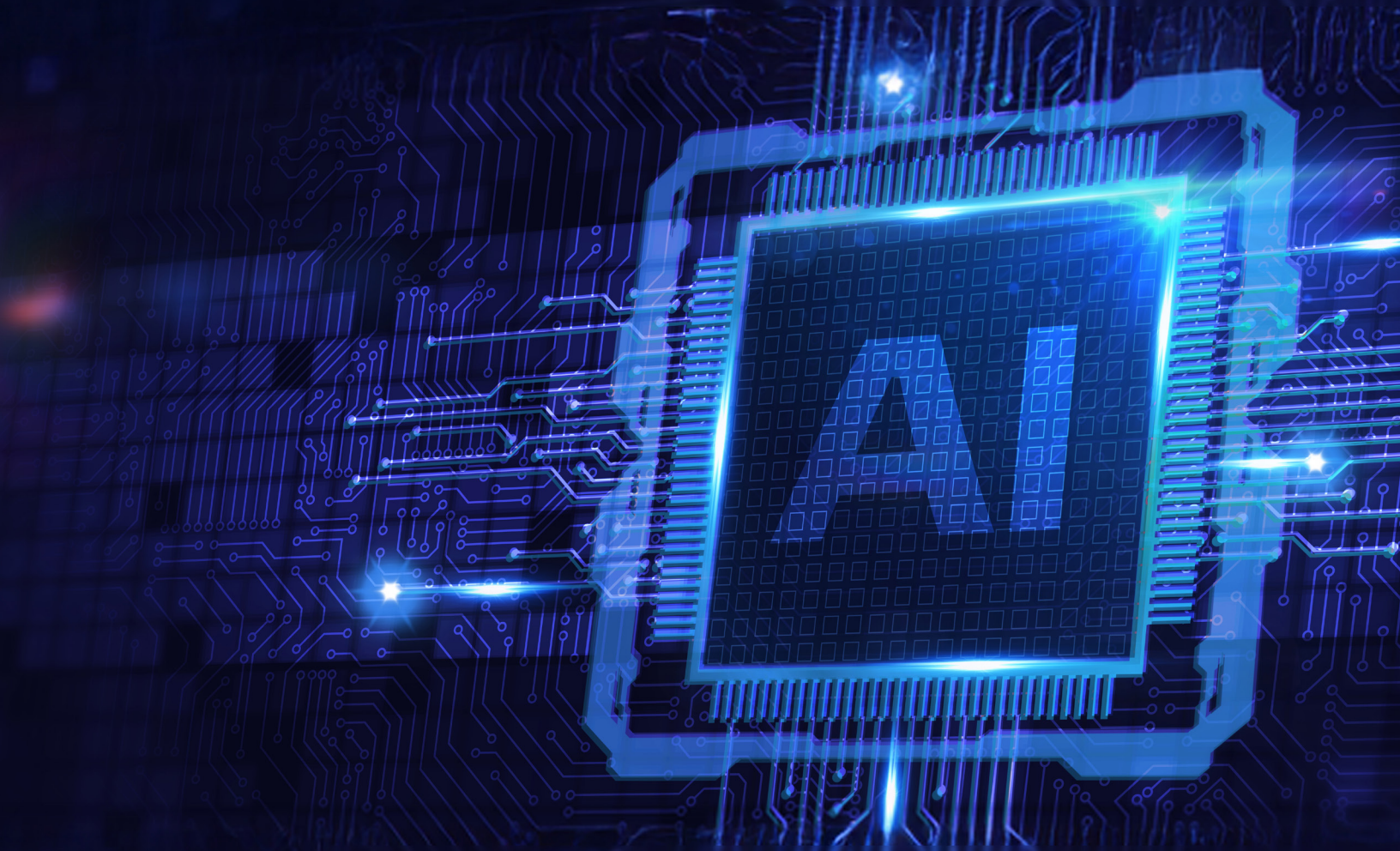
SYSTEMS ENGINEERING RESEARCH CENTER | U.S. ARMY | DEVCOM ARMAMENTS CENTER | GWTAI TRUSTWORTHY AI INITIATIVE

6TH ANNUAL

AI4SE & SE4AI

RESEARCH AND APPLICATION WORKSHOP

September 17-18, 2025 | Washington, DC & Virtual



EXECUTIVE SUMMARY

OBJECTIVE

The U.S. Army DEVCOM Armaments Center (AC) Systems Engineering Directorate (SED) and the Systems Engineering Research Center (SERC), a University Affiliated Research Center (UARC) for the Department of Defense (DoD), jointly sponsored the sixth Artificial Intelligence for Systems Engineering & Systems Engineering for Artificial Intelligence (AI4SE & SE4AI) Research and Application Workshop on September 17-18, 2025. The sold-out, two-day hybrid event was hosted by The George Washington University Trustworthy AI Initiative in Washington, DC, and was attended in-person and virtually by more than 250 people representing government and military, academia, industry, and Federally Funded Research and Development Centers (FFRDCs). Attendees gained insights from leaders using AI in this space, shared ideas, and further explored outcomes that resulted from the previous AI4SE & SE4AI workshops.

EVENT SUMMARY

The conference theme, “Systems Engineering AI That Works: Assuring Transformative Capabilities and Enabling a Digital Transformation,” aimed to foster discussions and insights on how systems engineering (SE) can support the development of trustworthy AI systems, and how AI tools can transform the practice of SE and shape the workforce. The agenda included two keynotes, two plenary panels, 60 submitted technical presentations as well as two in-person interactive sessions, a new addition to the workshop format. Contributions represented perspectives from industry, academia and government.

In the spirit of the workshop, the following AI4SE and SE4AI key points were partially generated by Google’s Gemini generative AI tools as a summary of the full report:

AI4SE Key Points:

Systems engineering is undergoing a digital engineering transformation and AI is right in the middle of it. Digital engineering is foundational to leveraging AI in the SE discipline.

- **AI augments the SE process by shifting it from a manual, documentation-heavy discipline to a rapid, model-driven one where models exist as structured data sources contributing directly to enterprise analytics.** Generative AI accelerates initial engineering work, such as requirements synthesis, architecture generation, and test case creation, often leading to faster iteration and measurable time savings.
- **AI's value lies in transforming unstructured, implicit system knowledge into structured, queryable knowledge bases.** This involves leveraging large language models (LLMs) and Retrieval-Augmented Generation (RAG) to convert dense technical documents and legacy data into dynamic knowledge systems that support traceability, semantic search, and continuous learning.
- **Successful AI integration requires anchoring generalized models with structured, domain-specific knowledge and strict engineering rigor.** LLMs must be guided by semantic scaffolds (ontologies) and modular prompt design to move beyond imprecise output, ensuring generated designs are coherent, traceable, and grounded in engineering truths.
- **Effective AI integration into engineering processes requires human-AI teaming and new human-machine workflows.** AI adoption is also an organizational change and workforce training opportunity.
- The rise of AI necessitates a **shift in the systems engineer's role from model constructor to orchestrator of human-AI collaboration**. AI tools function best as reliable co-pilots within modular, human-in-the-loop systems, where human oversight and cognitive checkpoints remain essential to validate outputs and maintain accountability.
- Trust in AI assistants is determined by relational factors (such as empathy and clear communication) as much as technical accuracy. **Designing for "Responsivity"—how well AI adapts, collaborates, and communicates—is crucial for increasing user engagement and improving performance**, especially in creative or ambiguous design tasks.
- Transforming SE requires **embedding AI directly into existing workflows, governance, and organizational culture, not merely treating it as a standalone tool**. Strategic adoption involves integrating AI tools into existing Systems of Record to ensure scalability, manage risks like technical debt, and overcome institutional inertia in high-regulation industries. How and where AI is embedded into workflows with humans can significantly impact overall outcomes and is an area that requires greater depth of study.

SE4AI Key Points:

Systems engineering in an AI-augmented world requires that we formalize properties like governance, trustworthiness, and resilience as systems properties. New test and evaluation regimes are required, both in scope and timing. There is opportunity for public-private partnerships, particularly around live-virtual-constructive testbeds for AI-enabled systems. Data management is a critical aspect of AI governance.

- **AI must be treated as an engineering discipline for safety**. AI systems must require application of the same safety and certification rigor applied to traditional mission-critical capabilities. The foundational goal of AI assurance is building confidence—in the models, the data, and the people who use them—rather than just ensuring compliance.
- **Assurance requires continuous, layered guardrails**. Assurance is achieved through a multi-layered safety architecture (or redundant assurance model) that integrates technical controls, organizational oversight (like internal review committees), and regulatory compliance as "guardrails". This layered approach treats testing as a continuous, lifecycle process rather than a one-time event.
- **T&E must shift to proactive stress testing**: Testing and Evaluation (T&E) must move beyond static performance metrics to proactive stress-testing and optimization-based falsification. This shift is essential to deliberately search for and disprove safety assumptions, uncovering hidden vulnerabilities, non-determinism, and failure thresholds, particularly for safety-critical machine learning systems.
- **Governance artifacts must be traceable and embedded**. SE4AI frameworks treat AI monitoring and governance artifacts (such as fallbacks, alerts, and model disablement) as traceable, reusable, and auditable SE elements. This integration provides lineage traceability of decisions and the necessary systematized evidence trail to satisfy assurance requirements.
- **System design must preserve human judgment**. Designing and testing AI systems requires Human Systems Integration (HSI) to ensure appropriate levels of human judgment are preserved. Systems must incorporate cognitive checkpoints and in-application reminders to reinforce user responsibility, as hybrid human–AI architectures often exhibit unpredictable emergent behaviors that must be discovered through specific T&E regimes.
- **Formal qualification is needed for adaptive agents**. Rigorous, structured qualification processes (like the Reinforcement Learning Qualification Process—RLQP) are necessary to bridge research-level models to trusted operational use. These frameworks simulate real-world adversities (such as sensor failures) to establish robustness, safety bounds, and operational limits for adaptive learning agents before deployment in critical environments.

****

The AI4SE & SE4AI Research and Application Workshop has grown in scope in its six years and each year has allowed an exchange on progress, challenges, and goals. This year's event highlighted that trustworthy, efficient, and human-aligned AI is no longer a hypothetical, but technically feasible and strategically urgent. Discussions and presentations illustrated a maturing AI ecosystem capable of augmenting, not replacing, human expertise across high-stakes domains and a future where AI doesn't just support technical tasks but meaningfully collaborates across the entire engineering continuum. There was a continued emphasis on the importance of organizational culture and supporting the workforce to ensure individuals understand their role within the larger, interconnected digital ecosystem focused on delivering reliable capabilities to the warfighter.

Based upon this year's registration information, 39% of attendees were from academia, 38% from industry, 16% from government and military, and 7% from FFRDCs. Within this mix of sectors is where answers and solutions can be developed. The workshop organizers and participants look forward to a seventh gathering in 2026 and continued guidance on evolving efficiently and effectively into the future.

TABLE OF CONTENTS

INTRODUCTION

This was the sixth Artificial Intelligence for Systems Engineering & Systems Engineering for Artificial Intelligence (AI4SE & SE4AI) Research and Application Workshop, jointly sponsored by the U.S. Army DEVCOM Armaments Center (AC) Systems Engineering Directorate (SED) and the Systems Engineering Research Center (SERC), a University Affiliated Research Center (UARC) for the Department of Defense (DoD). The two-day hybrid event was hosted by The George Washington University Trustworthy AI Initiative in Washington, DC. The conference theme, "Systems Engineering AI That Works: Assuring Transformative Capabilities and Enabling a Digital Transformation," aimed to foster discussions and insights on how systems engineering (SE) can support the development of trustworthy AI systems, and how AI tools can transform the practice of SE and shape the workforce.

WORKSHOP AGENDA STRUCTURE AND AUDIENCE

This year's keynote speakers were leaders within entities and organizations at the forefront of AI development and exploring the technology's far-reaching impacts. Dr. Matthew Kuan Johnson, Chief of Responsible AI, Chief of U.S. Digital and Artificial Intelligence Office, gave a government perspective as lead of the team that supports operationalization and implementation of Responsible AI and AI Ethical Principles. Dr. Morgan Dwyer, Head of Policy Operations, Open AI, who leads the development and coordination of the company's public policy positions, provided industry insights. Both speakers noted the importance of keeping pace with AI technology and of collaboration across disciplines to increase accountability and safety.

The two-day hybrid event was sold-out and in-person and virtual registration exceeded 250 people representing academia, industry, government and military, and Federally Funded Research and Development Centers (FFRDCs). The workshop agenda was structured into the following two tracks with presentations on corresponding relevant topics:

AI4SE TRACK

- Improving SE
- AI to Manage Complexity
- Cognitive Assistants
- AI Tools Workflows and Training
- Evolving Role of Digital Engineering

SE4AI TRACK

- Trust and Bidirectionality
- System Design Processes That Support AI Across the Lifecycle
- Safety, Reliability, and Ethics
- Evolving Role of Digital Engineering
- Test and Evaluation

Each track was moderated with an interactive discussion and Q&A at the end.

New to this year's agenda were two in-person interactive sessions, offered in addition to the traditional abstract presentations. Mr. Sebastian Völkl, Dalus, led "Practical Approaches to AI-Driven Model Generation in SysMLv2 Environments," which included use cases and a hands-on session for in-person participants. Dr. Deri Draper-Amason, Old Dominion University Center for Mission Engineering, led "Operationalizing Mission Engineering with AI: Leveraging the IDMP User Story Framework for Digital Decision Superiority."

Presentation materials for the entire workshop are available via the event page on the SERC website.

WORKSHOP KEYNOTES, PANELS, AND PERSPECTIVES

DAY 1 | KEYNOTE

Dr. Matthew Kuan Johnson, *Chief of Responsible AI, U.S. Department of Defense*

Dr. Johnson described the Department's current efforts to formalize, automate, and scale AI assurance and evaluation processes across defense agencies using a strategy built on interdisciplinary validation, empirical grounding, and scalable assurance. His team's vision is to align policy, tooling, and mission execution so AI-enabled systems (AIES) can be developed and trusted at the speed of relevance while maintaining the ethical, legal, and safety guardrails expected of federal innovation.

AI governance is now a whole-of-government effort involving collaboration across departments, academia, and industry. Dr. Johnson emphasized the need for interdisciplinary approaches to benchmark design and assurance validation. For example, bringing in social scientists has helped ensure that testing captures the intended phenomena.

The rise of generative AI has democratized red teaming and vulnerability testing. This accessibility reinforces the need for standardized tradecraft and alignment testing. Dr. Johnson's team's focus is on expanding the use of crowdsourced methodologies, automation pipelines, and interdisciplinary expertise to improve alignment testing and safety assurance. He noted the need to streamline processes and identify gaps in these new approaches, and using tools to inform policies and ensure they are empirically grounded.

His office supports integration of assurance documentation and verification artifacts into development workflows and leverages large language models (LLMs) as *judges* for benchmark validation and scoring. He noted that these models are powerful accelerators, but introduce methodological risks if used as evaluation surrogates.

Dr. Johnson's office chairs the White House Working Group on AI Assurance, currently tasked with scaling shared resources and aligning with broader national AI management requirements. His office also develops AI tools and policies in tandem, allowing developers to assemble assurance cases and validate compliance within a "sandbox" environment. He noted that this approach has proven effective in aligning oversight with operational realities, yet emphasized that technology development and workforce skills and staffing must be addressed to achieve High-impact AI for the Government.

Highlights

- *Ethical and Legal Trade-offs***.** AI improves overall system performance but also raises ethical and legal questions that are both engineering and moral in nature, and not easily legislated away.
- *Zero-Trust and Data-Sharing.* The Department is working toward a model that preserves data visibility, yet keeps access secret by design—a theme expected to recur throughout modernization.
- *Evaluation Ownership and Model Contamination.* His office is developing internal testbeds to help services generate, run, and interpret benchmarks tailored to their operational domains.
- *Open-Source Models and Export Controls.* The National Institute of Standards and Technology (NIST) and the federal ecosystem are assessing the advantages and security risks of openness.
- *Policy Development and Acquisition Reform.* The Department's "guardrails and guidelines" approach interprets existing cybersecurity and safety requirements through an AI lens rather than introducing new rules, balancing local mission flexibility with enterprise alignment. His office has partnered with the Defense Innovation Unit and USD(R&E) to create an AI sub-pathway that enables faster acquisition cycles without sacrificing compliance or assurance.
- *Data Foundries and Specialized Research*. The Department is working to establish AI data foundries capable of labeling, generating, and synthesizing mission-specific datasets, easing data burdens for certain use cases. For high-complexity domains, Dr. Johnson's office supports synthetic data generation and labeling services.
- *Human Judgment and User Responsibility.* Training will reinforce appropriate use patterns and ensure users apply proper judgment as AI tools become embedded in daily workflows.
- *Risk Ownership and Assurance Communication*. His team has developed persona-based assurance templates that address transparency and traceability of risk decisions and tailoring assurance documentation to each stakeholder's expertise.

DAY 2 | KEYNOTE

Dr. Morgan Dwyer, *Head of Policy Operations, OpenAI*

Dr. Dwyer emphasized that AI's rate of change demands an equal evolution in systems engineering (SE) thinking for responsible AI integration. The systems approach provides the translation layer between highly technical AI development and real-world policy and governance. The field must integrate agility, transparency, and continuous safety validation as core design principles, and Dr. Dwyer linked SE principles to the governance and safety of advanced AI models. As engineers, educators, and policymakers work to harness AI's potential, attention also needs to be paid to ensure its development remains human-centered, democratically governed, and systemically safe.

In advocating for a systems approach to AI safety, Dr. Dwyer discussed OpenAI's SE-based safety system. It uses a layered architecture of assurance that combines technical controls, human oversight, and transparent public engagement with each layer functioning as a fail-safe. She likened this to classical systems redundancy – distributing risk management across diverse methodologies to maintain assurance despite rising complexity. Safety is integrated into every stage of model development, ensuring models produce accurate, safe outputs. Each new model undergoes thousands of hours of red teaming by more than 400 external experts across domains, which contributes to model resiliency.

Dr. Dwyer discussed practices and approaches used by OpenAI, such as working with national and international partners to conduct independent model evaluations. The industry is widely emulating OpenAI's practice of publishing safety assessments online, providing access to documented model capabilities, risks, and mitigations. In the areas of oversight and governance, OpenAI conducts an internal evaluation of each model at key lifecycle stages and compliance with international regulatory regimes to enforce accountability.

Highlights

- *The Acceleration of AI and Its Engineering Implications*. The three key forces that drive AI acceleration—scaling laws, falling cost of intelligence, and shortening innovation cycles—will reshape SE and challenge the community to reconsider the objective function it is optimizing for.
- *Democratic Imperatives*. Whoever scales compute, talent, and deployment first will set global standards for safety, privacy, and growth, making it essential that democracies lead.
- *AI as a Tool for Systems Risk Analysis*. Such capability could empower analysts to assess vulnerabilities across legacy defense systems quickly, rather than the previously required months of manual tracing.
- *Ethics, Trade-offs, and Future Challenges*. Dr. Dwyer acknowledged the tension between openness and safety (balancing innovation with security) and emphasized transparency as a market incentive for responsible AI behavior. On verification and validation, traditional static testing cannot capture the complexity of general-purpose AI, highlighting the need for continuous, iterative safety evaluation akin to agile SE.

WORKSHOP PANELS

Industry Executive Panel | The Need for Sociotechnical System Testbeds for AI-Enabled Systems
Moderator: **Dr. Dinesh Verma**, *SERC*
Panelists: **Ms. Irene Helley**, Lockheed Martin; **Dr. Dimitrios Lymberopoulos**, Palantir; **Dr. Jay Meil**, SAIC

The panel explored how *augmented intelligence* is reshaping SE for defense and complex missions, from designing components to orchestrating systems-of-systems (SoS) that can sense, decide, and adapt in real time. The panel agreed the future of AI in defense lies not in building ever-larger models, but in engineering trustworthy, interoperable, explainable systems built for mission reality, not lab idealism. Trust needs to be engineered as carefully as technology. Success will hinge on composability, explainability, and speed of learning.

Key Themes

- *From concept to capability,* AI needs to be treated as an engineering discipline, applying safety and certification rigor similar to traditional systems. Integration efforts illustrate how legacy platforms can adopt advanced data-fusion and mission-system intelligence. True progress depends on partnerships spanning academia, industry, and operators.
- *Ruggedizing AI for the battlefield* means defense AI must be hardened, explainable, and capable of operating with limited compute in adversarial conditions. The defense community must invest in small-model optimization, mission-specific evaluation metrics, and workflow adaptation. Iterative, field-driven testing is more valuable than perfection in simulation.
- *Focus on* reducing cognitive load, accelerating decision speed, and building human-machine trust.
- *Testbeds help humanize AI*—seeing where it succeeds, where it fails, and how it can be trusted. Secure, hands-on environments allow for evaluation of AI under real mission constraints. The example of AI fight clubs demonstrated bringing different systems together to maximize capability and highlighted that the best missions are collaborative.
- *Academia's role beyond that of Transformer* includes exploring next-generation reasoning architectures and causal world models, explainability and trust metrics for operational systems, lightweight, embedded AI for constrained platforms, and shared data frameworks to connect research and mission application.
- *The New Systems Engineering* results from AI shifting the focus from individual systems to SoS orchestration. Success requires breaking down data silos and decoupling data from applications; designing for composability, traceability, and reuse; and maintaining SE rigor as the backbone of digital transformation.

PLENARY PANEL: HUMANS AND AI SYSTEMS

Moderator: Dr. Bryan Mesmer*, The University of Alabama in Huntsville*
Panelists: Mr. Eric Mortin, *U.S. Army DEVCOM Analysis Center*; Dr. Valerie Sitterle, *SERC*; Dr. Harrison Hyung Min Kim, *U.S. National Science Foundation, University of Illinois Urbana-Champaign*; Dr. Matt Gaston, *Carnegie Mellon University*

The panelists contributed their views on central issues and themes related to the common perspective that engineering for systems incorporating AI is a more difficult and complex task compared to engineering for systems that do not incorporate AI. Specifically, the panel discussion focused on the comparison between thinking about systems with AI and engineering for humans that are typically in the "role" of the AI. Overall, there was agreement on viewing AI not as a replacement for human decision makers but as able to interface with and augment human capabilities.

The panelists did not consider "engineering for" humans or AI an entirely solved problem and emphasized that design and engineering must account for both together. The inherent sociotechnical aspects present when there are AI agents and human users in a system change failure modes. The community must rethink how tasking relationships are architected in teaming workflows, highlighting comments made in a separate panel by SERC's Chief Scientist, Dr. Zoe Sjanfarber. The roles of humans and AI in systems is dependent on and determined by context, e.g., the role of AI in the context of daily work activities is different than its role in a fielded system. AI will work faster than a person in all contexts, requiring a focus on the effective integration of both within a workflow. Panelists emphasized the necessary shift from traditional deterministic thinking to probabilistic collaboration styles, which will impact skill development needs for the workforce and where and how we place protocols for decision transparency and authority within workflows. Also, differences in AI "reasoning" compared to that of humans were discussed as they related to maintaining and transferring context effectively across AI-AI and human-AI interactions. "Correctness" becomes more nuanced when workflows with humans and agentic AI need steps for iterative refinement, multiple output samples, and a different understanding of confidence levels and uncertainty related to use of agentic AI.

Panelists noted there is much tacit judgement in how humans apply information and the nature of instructions for AI need to be specific and cannot rely on context interpretation. The dynamism of other fields needs to be brought in to help the engineering community address current and future challenges including understanding how humans work, as well as data quality, bias and hallucination, and safety engineering specific to AI. Performance measurement concepts from cognitive science may be extensible foundations for evaluating effectiveness of human-AI teams. Panelists cited recent work from Stanford that found emergence in AI systems could appear due to a researcher's choice of metric rather than fundamental changes in LLM model behavior with scale. Additionally, the panelists discussed pragmatic issues such as maintenance. AI systems will require periodic updates, retraining, and quality assurance for their underlying models and data. Specialized AI maintenance teams may be required to ensure safety and effectiveness.

The panelists concluded by talking about the need for research in human-AI "engineering" for both pre-fielding analysis and development activities and that for humans and AI-enabled cyberphysical systems in theater. Especially for the latter, it will be critical to understand the boundary of what is described as the system in order to add the complexity necessary to represent the system aspects needed for design—yet not adding so extensively that the systems with AI appear more complex than necessary.

U.S. Army DEVCOM Armaments Center Perspective

Mr. Edward W. Bauer, *Director of the Systems Engineering Directorate (SED), U.S. Army DEVCOM Armaments Center (AC)*

Mr. Bauer discussed how the Army is driving modernization through digital engineering (DE), AI, and trusted systems assurance. He framed this work within the Army's three transformation pillars—delivering critical warfighting capabilities, optimizing force structure, and eliminating obsolete programs—underscoring that AI must enhance lethality and efficiency as well as safety, reliability, and ethical integrity. Mr. Bauer outlined how AI and model-based systems engineering (MBSE) are embedded throughout the acquisition lifecycle, supported by three integrated strategies (Digital Engineering, Data, and AI) that each reinforce the others through common ontologies, system modeling (SysML v2), and interoperable digital ecosystems. The AC's transformation rests on four lines of effort: strengthening the digital foundation, developing enterprise data architecture, deploying integrated toolchains such as the IOIF (Interoperability and Integration Framework), and building workforce literacy to sustain adoption.

Mr. Bauer illustrated how AI accelerates analysis while maintaining engineering rigor through four applied AI-driven case studies: a machine learning–based counter–UAS (unmanned autonomous system) performance analysis that reduced months of modeling to days; automated radiographic defect detection for munitions; predictive modeling for next-generation propellants; and robotics-enabled smart manufacturing cells that improve safety and quality assurance. The Army's broader challenge, however, is ensuring measurable confidence in systems that influence battlefield decisions. DEVCOM-AC is contributing to department-wide trust metrics, calibrated assurance evaluations, and human-machine testbeds to integrate human factors early in development. Collaboration across government, industry, and academia through the Army's Technology Transfer Program highlights that digital transformation is about automation but also cultivating trust.

SERC Perspective

Dr. Zoe Szajnfarber, *Chief Scientist/Professor, SERC/The George Washington University*

Dr. Szajnfarber discussed the evolving relationship between Artificial Intelligence for Systems Engineering (AI4SE) and Systems Engineering for AI Systems (SE4AI), framing them as complementary paradigms driving the modernization of defense acquisition and engineering practice.

In the AI4SE context, Dr. Szajnfarber emphasized where AI tools already offer measurable gains in traditional lifecycle functions such as contracting and compliance language generation and enterprise-scale project tracking. She highlighted examples where cognitive assistants and generative models accelerate the development of systems models, suggest design alternatives, and train new engineers by interpreting complex trade spaces. In parallel, reinforcement learning and autonomous test generation are redefining testing and evaluation, while AI-based decision dashboards support continuous oversight in large, distributed programs. These advances signal a shift in the systems engineer's role from model constructor to orchestrator of human-AI collaboration, ensuring that model outputs, trade-offs, and automation remain mission-aligned.

SE4AI applies SE principles to the architecture and assurance of AI systems themselves, emphasizing reliability, explainability, and calibrated trust in non-deterministic, data-driven components. Two ongoing research thrusts were presented: (1) an LLM-enabled architectural generator, capable of transforming text and imagery into systems modeling language (SysML) models, translating between SysML versions, and embedding SE expertise; and (2) GenAI application in contract management to assist clause harmonization, compliance analysis, and traceability. These exemplify AI's growing ability to translate across representational languages, scale repetitive modeling work, and identify patterns across multidimensional data sets, while reinforcing the need for structured human oversight.

The integration of AI into SE is as much an organizational transformation as it is a technical one. AI is an *enabler* of new capabilities, expanding the scope of decision-making and enterprise effectiveness across defense missions. Dr. Szajnfarber identified opportunities at the task level (AI can automate translation and pattern recognition within human-supervised workflows), the individual level (engineers rebalance cognitive workloads between humans and machines), the system level (AI allows organizations to rethink structures once limited by human attention), and the workforce level (the rise of human-AI teaming introduces new skills and new training paradigms, where LLMs themselves can aid in workforce development).

Academic environments provide a unique testing ground for exploring the human-AI interface at both cognitive and organizational levels. Universities serve as neutral sandboxes where researchers, students, and practitioners collaboratively probe how delegation, trust, and shared cognitive load function in hybrid human-AI systems. Academic teams can experiment openly with questions of agency distribution, error tolerance, and adaptive oversight, developing new methods to evaluate where humans should intervene, and where AI can safely operate autonomously. These small-scale, iterative testbeds offer a powerful model for designing trustworthy, resilient systems that reflect the dynamic, co-adaptive realities of future engineering and defense operations.

Army and SERC Perspectives: Summary of Key Points

The following key points were partially generated by Google's Gemini generative AI tools.

- **AI must cultivate trust, not just automation**. Mr. Bauer emphasized that the Army's digital transformation is centered on cultivating trust, assuring that the systems built, and the people who depend on them, can trust the intelligence every time it acts.
- **AI must enhance ethical integrity and safety**. Mr. Bauer outlined that AI must be integrated to enhance safety, reliability, and ethical integrity as well as lethality and efficiency within the Army's three transformation pillars.
- **SE4AI and AI4SE are complementary paradigms**. Dr. Szajnfarber framed the modernization of defense acquisition as driven by AI4SE, which applies AI to enhance traditional SE workflows, and SE4AI, which ensures AI-enabled systems are designed, tested, and governed with mission-critical rigor.
- **SE requires orchestration, not just construction**. In the AI4SE context, the systems engineer's role is shifting from model constructor to orchestrator of human-AI collaboration, ensuring model outputs and automation remain mission-aligned.
- **SE4AI ensures reliability and explainability**. SE4AI applies classic engineering principles to the architecture and assurance of AI systems, specifically emphasizing reliability, explainability, and calibrated trust in non-deterministic, data-driven components.
- **Focus on architecting the human–AI interface**. Dr. Szajnfarber identified this as the future research frontier, using academic testbeds to explore how delegation, shared cognitive load, and trust function in hybrid systems.

INTERACTIVE SESSIONS

Session 1 | Practical Approaches to AI-Driven Model Generation in SysMLv2 Environments
Presenter: **Sebastian Völkl**, *Dalus, Inc.*

Mr. Völkl walked participants through a workflow to generate SysML v2 models from natural language design descriptions using the Dalus MBSE platform and the open SysML v2 pilot implementation. Attendees started with a general overview of the concept and how Large Language Models (LLMs) can facilitate the workflow when model quality improving techniques are used as part of the process. Attendees used a provided sample Design Specification Document (DSD), which they uploaded into browser-based Dalus accounts created for the session. Participants watched an LLM extract over 30 requirements from the DSD and then inspect the auto-created structure within Dalus' graphical editor. While a dashboard made cycle time and acceptance rates visible, participants were guided through exercises to tweak prompts and system policies to discover the impacts to architecture granularity, naming schemes, and constraint solvers. Each participant group then exported the resulting SysML v2 code and ran it through Dalus' Pilot Implementation validator, which made grammar violations visible. Participants left with a hands-on understanding of where and why LLMs can add value in SysML v2 model generation when using best practice prompt engineering and policy guard techniques to improve overall model quality.

Session 2 | Operationalizing Mission Engineering with AI: Leveraging the IDMP User Story Framework for Digital Decision Superiority
Presenter: **Dr. Darryl Draper-Amason**, *Old Dominion University, Center for Mission Engineering*

Dr. Draper-Amason introduced participants to the Integrated Digital Maturity Pathway (IDMP) User Story Framework and its integration with the Mission Readiness Model, an interactive decision-support software developed by ODU's Center for Mission Engineering. The focus was on understanding how AI-supported tools can evaluate readiness across interconnected systems and teams. The IDMP framework provides human-centered, structured maturity modeling for sharing and utilization of technical data using a visual storytelling approach to ensure data is Visible, Accessible, Trustworthy, Interoperable, and Secure (VAULTIS). The Mission Readiness Model dynamically models the interdependencies between system capabilities, constraints, and their impact on mission outcomes. The session focused on the coupling of IDMP and the Mission Readiness model to demonstrate how stakeholders could understand, prioritize, and justify AI-driven decisions across acquisition, sustainment, and operational timelines.

Participants were shown a walkthrough of the IDMP user story-to-maturity mapping process followed by a demonstration of the Mission Readiness Model, showing how user story-driven metrics can inform mission-level tradeoffs and resourcing decisions. Small groups of participants were then walked through co-development of a role-based user story related to digital twin integration, supply chain resilience, or AI-enabled sustainment using the IDMP framework. Participants were provided an IDMP template as a reusable user story framework, sample maturity dashboards, and an understanding of how a digitally informed, stakeholder-centered design process can support decision making about AI-enabled systems in a mission context.

WORKSHOP PRESENTATIONS

AI4SE: Improving SE

Axiom Me a Question!

Erin Smith Crabb, *Leidos, Inc.*

Systems engineering (SE) often relies on complex diagrams difficult for non-experts to interpret, creating barriers to understanding the discipline's value. This research explored whether large language models (LLMs) and visual language models (VLMs) could lower that barrier by translating state machine diagrams into axioms—simple, logical statements that describe the system's flow and behavior. Using only static images with no additional context, the team tested models like GPT-4o and Claude Sonnet to see how effectively these could capture and convey information embedded in the diagrams.

Findings showed that VLMs generated axioms with promising accuracy, with Claude producing outputs more closely aligned with systems engineers' expectations. While the models struggled with certain features such as decision gates and reversed transitions, engineers found the generated axioms useful in reducing interpretation effort and making diagrams more accessible. This pilot suggests a path toward leveraging AI to help non-experts engage with SE artifacts, modernize legacy models, and support broader adoption of systems thinking across domains.

Elevating Digital Engineering Competency

Dr. Heidi Davidz, *MANTECH International*

Digital engineering (DE) faces persistent gaps in workforce capability and systemic failure patterns in SE. The research used AI-driven metadata analysis to highlight critical challenges: technical gaps in MBSE, analytics, cybersecurity, and AI; foundational skill deficits in communication and change management; and organizational barriers such as an aging workforce and limited leadership buy-in. At the systems level, recurring dysfunctions span design flaws, communication breakdowns, inadequate verification, and cultural resistance. These issues remain entrenched and costly, underscoring the need for a more holistic approach.

AI-enabled enterprise architecture can bridge roles, competencies, interventions, and organizational dynamics. By mapping workforce gaps, systemic failure patterns, and team-level influences, leaders can identify interventions that improve value delivery. Elevating DE competency will require moving toward an adaptive, AI-supported strategy that continuously captures influences, identifies dysfunction, and aligns development initiatives. Optimizing workforce utilization through AI and visualization offers a path to build resilience and deliver stronger outcomes at enterprise scale.

Agentic AI for Lifecycle Traceability: A Digital Thread Architecture for Adaptive, Context-Aware Systems Engineering

Dr. Kristen Jaskie, *Prime Solutions Group*

This research introduced a modular digital thread architecture enhanced with agentic AI, addressing a persistent problem in SE: how to achieve lifecycle traceability, explainability, and adaptability without demanding tool expertise at every stage. By integrating a GraphRAG pipeline grounded in structured knowledge graphs and supported by the emerging CaSCADE standard, the proposed system enables LLMs to answer complex lifecycle queries with source-linked, policy-compliant precision, offering a compelling alternative to black-box AI.

The strategic importance rests in reduced anomaly resolution time and design latency by allowing engineers to interact naturally with deeply technical data, democratizing system insight without compromising rigor. Unlike traditional digital thread solutions, this approach does not require replacing legacy tools; it wraps existing artifacts in a framework that supports scalable, AI-augmented reasoning across engineering domains. The introduction of agentic AI further shifts the paradigm by embedding adaptive decision-making capabilities into the workflow, ultimately paving the way for proactive, autonomous engineering support systems. This lays the groundwork for trustworthy, standards-based AI integration in mission-critical environments, where traceability isn't optional and operational complexity demands more than static models.

Generative AI, Visualization, and the Future of Managing Megaprojects

Tom McDermott, *Stevens Institute of Technology*

The presentation explored how emerging technologies, particularly generative AI and advanced visualization, can transform leadership intuition into a more systematic, data-enriched capability for navigating uncertainty in megaprojects. AIVis was introduced, a prototype environment that combines AI, visualization, and learning to track evolving uncertainties and their causes over time. AIVis aims to surface the "hidden data" that leaders can sense but not always quantify, sharpening their intuition with richer situational awareness. Early DoD prototypes underscored the potential and the ethical challenge of responsibly capturing and protecting sensitive data. The research found that megaproject success may hinge less on perfect predictive models and more on augmenting leadership judgment with AI-driven insights and adaptive visualization, creating learning environments where managers can see, question, and act on uncertainty dynamically.

Guiding Large Language Models Through the Engineering Design Process Using Domain-Specific Knowledge Bases

Dr. Jitesh Panchal, *Purdue University*

The presentation explored how embedding structured, domain-specific knowledge, particularly from patent-based knowledge graphs, into LLM prompts can enhance the diversity, relevance, and system-level coherence of AI-generated engineering design solutions. This research envisioned engineering design where LLMs reason through ideas using structured, domain-specific knowledge. This signals that the variety and quality of AI-generated designs are less about model size and more about the precision and alignment of the knowledge embedded in prompts. When LLMs are guided by hierarchically structured knowledge graphs from functionally aligned patent domains, they produce more varied solutions and also begin to surface assumptions and technical constraints in ways that mimic human engineering reasoning.

This suggested a future where engineers may design the structure of knowledge itself, selecting, curating, and shaping domain inputs, as a core part of the creative process. While the study exposed current limitations in subsystem compatibility, it also highlighted a promising path: that modular, graph-based knowledge frameworks can reduce design incoherence and improve system-level integration. This approach opens the door to AI-augmented design environments where LLMs can support early-phase exploration, cross-domain innovation, and collaborative ideation across technical specialties. The key opportunity lies in developing reusable knowledge infrastructure that enhances LLM outputs and also transforms them into usable, system-aware design proposals.

Autonomous Agents for Complex Simulations: A Compute-Efficient Imitation Learning Approach

Dr. Julian Togelius, *New York University*

The research advances a practical and scalable method for training human-like AI agents in complex, real-time environments by leveraging imitation learning from limited, high-quality human gameplay data. It addresses two often competing industry demands, authenticity and compute efficiency, by abstracting sensory input to a low-dimensional space, enabling deployment even on systems without GPUs and within millisecond inference constraints. The ability to fine-tune agents for different skill levels and play styles from under 50 hours of player data demonstrates a cost-effective training pipeline and also a modular design that supports rapid adaptation to new contexts.

This work redefines “realistic” agent behavior as not just optimal performance, but strategic believability that can deceive or engage human players naturally, such as defusing bombs under pressure or adapting to dynamic threats. Implications for industries range from defense to corporate training, where human-like interaction is critical but human availability is limited. Quality-over-quantity data collection, paired with architectural efficiency, enables scalable, believable AI agents that are production-ready, not just lab-bound.

AI4SE: AI to Manage Complexity

A Lockheed Martin Case Study in Accelerating Systems Engineering with Generative AI

Melissa Bradshaw, *Lockheed Martin*

The presentation highlighted that generative AI isn't simply for automating tasks, but can be embedded within a broader digital transformation strategy that aligns platforms, normalizes data, and connects stakeholders through digital thread principles. The integration of tools like LM Navigator, Lockheed's internal generative AI assistant, demonstrates AI can front-load engineering work, like requirement generation or test case creation, giving engineers a head start and enabling faster iteration with greater consistency. This work represents a strategic reframing of SE, from a traditionally manual, documentation-heavy process to an AI-augmented, model-driven, and production-integrated discipline.

The presentation noted AI is not replacing human expertise but scaling and supporting it across the lifecycle, from initial requirements to ABAP code generation and test planning. The phased approach created momentum and buy-in. Notably, cost savings, time reductions, and increased quality were measured and shown to be real, important because buy-in is influenced by strategic leadership backing and grassroots experimentation, overcoming institutional inertia in a high-regulation industry.

A takeaway was the use of hackathons as living labs to co-develop with end users, accelerating adoption and continuous refinement. Tying AI integration to Systems of Record and governance frameworks like SEMPL ensured scalability and compliance. Ultimately, this initiative underscored that transforming SE with AI isn't just about technology, it's about culture, leadership, and embedding AI into the core rhythms of product delivery.

Automated Knowledge Base Generation from Technical Documents Using Open-Source Large Language Models and Retrieval-Augmented Generation

Charles Collard *and* **Dr. Mark Blackburn**, *Systems Engineering Research Center*

This presentation demonstrated how LLMs and retrieval-augmented generation (RAG) can automate the transformation of complex engineering documents into interactive, graph-based knowledge systems. The research team developed an open-source pipeline that consumes technical texts, identifies key topics, and generates interconnected wiki pages linked through knowledge graphs and visual artifacts. The result is an AI-assisted knowledge transfer system that makes dense, static textbooks accessible as dynamic, navigable learning resources. Building on earlier LLM-based research, this effort uses prompt-driven topic extraction, hierarchical structuring, and multimodal embeddings to preserve conceptual relationships while improving usability for practitioners.

The second phase explored "LLM as a judge," introducing metrics for evaluating generative accuracy, with factual consistency identified as the most critical measure. The team also discussed lessons learned in managing model quality for defense-related applications where data trust and model contamination are major concerns. Context window management, model architecture differences, and the use of vetted open-source models (such as Qwen 3-Instruct) were highlighted as key technical considerations. The approach points toward a future of scalable, AI-driven knowledge ecosystems that can guide digital transformation and sustain continuous learning through interactive, trustworthy knowledge bases and follow-on tools such as an AI-powered helpdesk.*

**presented virtually*

Operation Design in the Fifth Industrial Revolution

Avi Harel, *Ergolight*

The Fifth Industrial Revolution (5IR) reframes technology's role by emphasizing human–machine synergy and sustainability, moving beyond automation to collaboration. This research introduced a model-based approach to designing operations that anticipate and prevent exceptional events. Drawing on case studies such as MX981, Proton M, and PL603, the research demonstrated how LLMs can help capture, analyze, and formalize lessons from past incidents into reusable operational rules. The goal is to design systems that inherently detect and prevent failures.

The presentation contrasted the traditional view of failure, which tends to focus on performance and post-hoc analysis, with a proactive, engineering-centered view emphasizing quality assurance. Borrowing from concepts like poka-yoke (error-proofing), this approach uses AI to model exceptional events, identify error-prone conditions, and embed rule-based safeguards into system design. By coupling human judgment with machine learning, organizations can accelerate learning from incidents, improve operational resilience, and make safety-by-design a core feature.*

Onto-Graph: AI-driven Framework for Ontology-Guided Clustering and Hierarchical Structuring in System-of-Systems Engineering

Yinchien Huang, *Purdue University*

The authors strategically combined ontology-driven semantic normalization with LLM capabilities to form a hybrid pipeline that automates knowledge extraction, clustering, and multi-level system representation. Onto-Graph introduces a scalable and flexible method for dynamically constructing system architectures from raw text, reducing modeling burden and improving traceability.

Onto-Graph shifts SoS modeling from a brittle, manual exercise to a living, AI-augmented knowledge system that is domain-adaptable, semantically coherent, and responsive to evolving requirements. Ontologies serve as semantic scaffolds, anchoring free-form LLM outputs to consistent, domain-specific concepts, resolving issues of redundancy, ambiguity, and lack of interoperability. The semi-supervised clustering step, using the ROPE framework (Resources, Operations, Policies, Economy) allows stakeholders to see SoS architectures in layered, semantically meaningful groupings.

This approach supports stakeholder-specific querying, contradiction detection, and temporal evolution analysis, enabling live system navigation and oversight. In dynamic domains like Urban Air Mobility (UAM), where multi-actor coordination and evolving regulations complicate SE, Onto-Graph provides a tool for integration and agility. It is a vision of SE where understanding scales with complexity, and where the architecture is continuously interpreted, queried, and evolved.

Information Synthesis Workflows with Large Language Models: A Review of Current Practices, Opportunities, and Challenges in Literature Review Tasks

Bryce Huffman, *The George Washington University*

While purpose-built tools often advertise autonomous capabilities across tasks like proposal writing or literature synthesis, this research revealed that LLMs are most reliably effective when part of human-in-the-loop systems, especially in article screening, the task most frequently studied and showing early promise. This underscores a growing misalignment between commercial AI tool claims and the empirical realities of how LLMs currently perform in unstructured text synthesis workflows. Tasks like full knowledge synthesis or end-to-end workflow integration remain largely unexplored, underperforming, or poorly evaluated, meaning marketed capabilities often outpace supporting evidence.

The disconnect matters because organizations may invest in tools promising generalized automation when success is highly task-specific, methodologically inconsistent, and dependent on human oversight. The absence of shared evaluation benchmarks or clear definitions of "acceptable performance" makes comparing tools or studies nearly impossible. Findings suggest that workflows in policy analysis, market research, and proposal development can benefit from a granular, modular view of AI capabilities rather than treating LLMs as monolithic solutions. To move forward, researchers and developers must co-design evaluation frameworks, prompt strategies, and teaming models reflecting actual LLM strengths and limitations rather than aspirational marketing.

Generating Simulation Metadata with Large Language Models

Jayaprakash Kambhampaty, *Georgia Institute of Technology*

This presentation addressed an emerging DE challenge—managing and contextualizing vast amounts of engineering data—and proposed LLMs for generating meaningful simulation metadata. As modeling environments evolve toward integrated, multi-step computational workflows, metadata becomes critical for trust, decision support, and automation. Metadata generation remains difficult because it must capture structural details and also the contextual and domain-specific knowledge embedded in engineering documents and models. The presentation reframed modeling as a form of natural language processing that translates human intent, documentation, and implicit assumptions into structured representations that machines can understand and act upon.

Outlined was how AI can serve as the "data plumbing" behind future DE workflows, enabling cleaner and more consistent metadata generation. By automatically extracting, classifying, and enriching model information, LLMs can help preserve the engineering knowledge locked within unstructured documents and accelerate the modeler's OODA loop (observe, orient, decide, act). This emphasized that how knowledge is represented shapes how effectively it can be used, calling for new approaches to AI-driven metadata pipelines that bridge human context with machine reasoning in complex, collaborative engineering environments.*

Leveraging Large Language Models for Logistics Information Extraction: A Case Study on Two International Disasters

Zaid Kbah, *The George Washington University*

This study examined whether LLMs like ChatGPT can meaningfully assist in extracting logistics information from unstructured disaster reports—a task with high stakes for real-time decision-making in humanitarian response. The researchers found that while ChatGPT can contribute to structured information extraction, its effectiveness depends more on the clarity of source documents and prompt design than on model version or settings, challenging common assumptions about performance tuning. The model performed significantly better on structured, logistics-focused documents than on loosely organized reports, revealing LLMs still struggle with ambiguity, which must be addressed for reliable deployment in high-stakes, real-world scenarios. A key insight is that prompt engineering, not model tuning, offers the most leverage for improving extraction quality, especially when examples from the same context are included or when linguistically complex categories are excluded. This suggests that practical performance gains are possible today through strategy rather than technical advancement. Moreover, while the models missed or misclassified some logistics statements, they still successfully identified many others, signaling their potential to reduce cognitive load for information officers, provided there is human oversight for validation. The study's rigorous use of a gold-standard comparison dataset further adds to its credibility and transferability to adjacent fields like contested logistics and military intelligence.

LLMs are not yet plug-and-play tools for critical operations, but can still meaningfully augment human decision-making when used within the right structural constraints. This work offers a tactical blueprint for how humanitarian and other operational sectors can build semi-automated pipelines, not to replace human expertise, but to scale it efficiently under pressure.

Maximizing the Usability of Publicly Available Information Through Improved USPI Filtering

Dr. Donald Koban, *United States Military Academy*

This research reveals a critical inflection point in how the defense entities and similar institutions might evolve their use of Publicly Available Information (PAI): moving from rigid, brittle rule-based systems to adaptable, explainable AI models that scale to the operational demands of modern information warfare. LLMs such as Llama 3.2 dramatically outperform traditional Regex filters, even when limited to sparse metadata, which suggests a path toward more nuanced, compliant, and trustworthy USPI filtering at scale. This enables structured, high-confidence intelligence from chaotic, open-source data streams, without compromising privacy regulations or operational accountability.

The study also uncovered design trade-offs that matter for real-world deployment, precision versus recall, interpretability versus compute cost, and stability versus adaptability. Its insight that LLMs are sensitive to location fields underscores a vulnerability and a roadmap for building more robust classifiers in adversarial or obfuscated data environments. By demonstrating that temperature tuning and ensemble strategies can significantly stabilize outputs, the research offers practical knobs for tuning LLMs in safety-critical systems. This work paves the way for integrating LLMs into hybrid analytic pipelines, combining their contextual reasoning with the speed of Regex and enabling real-time filtering and deeper strategic insight from open-source platforms.

AI-Enhanced Requirements Traceability Using MBSE and Large Language Models for Complex Systems
Henock Legesse, *National Aeronautics and Space Administration*

This research demonstrated how AI, when thoughtfully integrated into established SE workflows, can address a long-standing, high-cost challenge: maintaining traceability across complex, evolving requirement sets. Poor traceability is an administrative burden and a risk factor that can result in missed requirements, late-phase design changes, and system failures or costly rework. Automating traceability gap detection using a layered LLM approach enriched with context and embedded validation offers a path to proactive requirements management. The broader implication is that as systems grow in complexity, traditional SE processes will increasingly require augmentation, not replacement, by AI tools engineered to respect domain constraints, human oversight, and evolving standards.

This research also modeled what responsible LLM deployment looks like: confidence scoring, hallucination filtering, and integration with tools like MagicDraw create a blueprint for trustworthy AI adoption in high-stakes environments. As future LLMs improve in reasoning and reliability, this architecture positions engineering teams to catch traceability gaps earlier and eventually automate higher-order tasks like compliance verification and change impact analysis—unlocking new efficiencies while preserving the integrity of the engineering process.

LLM-Augmented Pipelines for Validated System Model Creation: Extracting and Structuring Complex System Representations from Text
Dr. Carlo Lipizzi, *Stevens Institute of Technology*

This research tackled a central obstacle in DE adoption: how to trust AI-derived system models built from unstructured legacy documents, especially in high-stakes defense contexts. By developing a semi-automated pipeline that uses LLMs to extract knowledge and subject matter experts to validate it, the authors created a process that scales and also embeds trust directly into the model-building workflow. The strategic significance lies in its ability to shift from static documentation to dynamic, queryable, and semantically rich system models, enabling engineers to reason across complex dependencies while maintaining human oversight.

The integration of graph neural networks and retrieval-augmented generation further amplified the utility of these validated models, allowing AI to operate over structured knowledge rather than raw text—key for ensuring explainability and repeatability. This approach directly supports MBSE and digital transformation goals within the Army, showing that AI can be a reliable teammate when its outputs are grounded in mission-critical truths. Operationalizing assurance at every stage offers a scalable and trustworthy bridge between legacy systems and future-ready engineering workflows.

AI-Enhanced DEMA: Transforming Implicit System Knowledge into Intelligent, Compliant, and Documented Processes

Dr. Dan O'Leary *and* Dr. Allison Ledford*, Auburn University*

Traditional data-mapping and process-modeling methods fail to capture the unstructured and implicit knowledge (e.g., conversations, decisions, and tribal expertise) that drive real-world systems. The Data Element Mapping and Analysis (DEMA) methodology introduces a structured approach to uncover and analyze these hidden data flows across an organization's lifecycle. Unlike visual mapping techniques that focus on process steps, DEMA works at the data-element level, integrating human insights with formal system views to reveal the true "as-is" state of complex systems. The authors demonstrated how coupling this methodology with AI can systematically expose and rationalize this implicit knowledge, creating a foundation for digital transformation rather than just digitization. This approach converts tacit knowledge into structured intelligence, reducing cost and cycle time by orders of magnitude—transforming tasks that once took weeks into hours. The key insight—"vessel differences are context, not conflict"—highlights that variation across teams represents valuable perspective rather than inconsistency. With pilot deployments underway, AI-enabled DEMA offers a flexible, scalable path to capture, standardize, and operationalize system knowledge across organizations at a fraction of traditional cost.*

Leveraging AI to Manage Technical Debt During Test and Evaluation in Aerospace Systems Engineering

Zak Ouzzif*, Worcester Polytechnic Institute*

This research reframed technical debt not as a software-only issue but as a systems-level risk factor that, if unmanaged during the test and evaluation phase, can undermine mission assurance in aerospace programs. By targeting verification, test, and documentation debt within complex SE workflows, the AI-driven Technical Debt Management Framework offers a scalable way to catch issues before they propagate downstream, a critical advancement given the scale of modern projects like the F-35. The framework's ability to reduce review time by 45% while maintaining high validation accuracy signals a shift toward AI-enabled early intervention, rather than relying on costly, reactive corrections late in the lifecycle.

Strategically meaningful is the focus on integration: this approach complements existing SE tools and processes, enabling human-AI collaboration without requiring wholesale process redesign. The use of real aerospace artifacts for validation grounds this work in operational relevance, bridging the persistent automation, validation, and integration gaps that have limited AI adoption in SE. As systems grow more interconnected and deadlines more compressed, the ability to continuously monitor and manage technical debt using tools like TDMF may be the difference between routine success and catastrophic failure.

AI4SE: Cognitive Assistants

Open Local AI: An Open-Source Generative AI Approach for Sensitive Information

Dr. Barclay Brown, *Collins Aerospace*

The presentation introduced *Ola Chat (Open, Local AI Chat)*—a secure, open-source generative AI platform purpose-built for SE teams working with classified or export-controlled data. While LLM tools have demonstrated measurable productivity gains in requirements generation and test development, organizations now face strict data-handling and compliance constraints that prohibit cloud-based AI use. *Ola Chat* addresses this by running offline on standard Windows workstations, integrating open-source components such as Ollama for on-device model execution and AnythingLLM for document orchestration. Once installed, it requires no internet connection, making it suitable for air-gapped environments.

The platform delivers powerful capabilities such as unlimited-length document summarization, semantic document comparison, and extensibility through Python, while keeping all embeddings, indexes, and data local to the user. This architecture combines the transparency and adaptability of open source with the confidentiality required for sensitive engineering projects. In pilot deployments, *Ola Chat* reduced approval times dramatically and enabled immediate productivity gains, illustrating how on-premise, open AI solutions can reconcile innovation with security, and may become a cornerstone of future SE workflows.

Managing Complexity for LLM Software Co-Pilots Through Abstraction-Aware Context Modeling

Dr. Brad Dennis, *Valkyrie Enterprises*

The presentation challenged the enticing ease of using LLMs like ChatGPT for software development, warning that their “magic” conceals critical complexity. While LLMs are powerful code generators, they flatten hierarchies, remove boundaries, and eliminate traceability, which are essential for building reliable, mission-critical systems. Software isn’t just about generating code—it’s about managing complexity through abstraction, clear interfaces, semantic cohesion, and structured thinking. When prompts lack these boundaries, LLMs take over too much decision-making, turning precision engineering into imprecise guesswork.

This is especially risky in defense, where traceability and verification are non-negotiable. The solution is reintroducing “abstraction discipline”—modularizing prompts, limiting context to what’s relevant, and preserving a digital thread of decisions. Techniques like progressive disclosure and semantic boundaries help LLMs understand tasks more like expert developers would. Ultimately, LLMs must be guided by software engineering principles, not used as replacements for them.*

Davinci: Transforming Systems Engineering Through Agentic AI Beyond Traditional LLMs

Dr. Chris Helmerich, *Celedon Solutions*

The presentation focused on Davinci, an AI-driven platform designed to accelerate MBSE. Building and maintaining complex SysML system models requires hundreds of manual hours even with current AI assistance. Davinci's vision is to generate, verify, and refine a complete, fully linked system model in under 20 minutes. It was demonstrated how the system transforms lengthy PDFs into structured engineering artifacts, automatically producing traceable requirements, architectures, interfaces, behaviors, risks, and budgets in minutes. Each model element is cross-linked, meaning changes to one parameter (like mass or power) propagate instantly throughout the system and its documentation, which Da Vinci keeps synchronized in real time.

Applications ranging from risk analysis and capability modeling to simulation and document generation were showcased, including examples where Davinci created slide decks directly from model data. This integration—linking digital thread, modeling, and verification—ushers in a new phase of agentic AI-driven engineering, where systems can design, simulate, and validate themselves iteratively under human guidance. While Davinci currently serves as an AI copilot, its future lies in end-to-end automation: AI agents capable of engineering, testing, and manufacturing systems at a pace 1,000 times faster than today. Achieving this vision could transform engineering into a continuous, adaptive, and self-improving process.

Shaping Trust Through System Design: Human Perceptions of AI in Design Ideation

Dr. Ting Liao, *Stevens Institute of Technology*

Testing variations of a chatbot assistant (AIDA) with different levels of appearance, performance, and empathetic behavior yielded a surprising result: empathetic behavior was the most consistent and influential driver of trust, even when performance was low. This reframed trust in AI not just as a technical benchmark, but as a relational and emotional experience.

This challenges the assumption that performance alone earns user trust in AI design tools. Users don't evaluate AI systems in isolation; they make holistic judgments, weighing emotional cues, perceived effort, and system tone alongside functionality. Empathy, expressed through design choices like first-person language and emotional acknowledgment, measurably improved engagement, enjoyment, and usefulness, especially for non-expert users. Designing for emotional resonance isn't just "nice to have": it's a compensatory strategy when performance falls short.

Trust is also user-dependent. Factors like initial trust in automation and user age significantly shaped perceptions, highlighting the need for adaptive, user-aware AI interfaces. The distinctive significance illustrated is the shift from designing AI as a tool to designing it as a cognitive collaborator that supports problem-solving and also the messy, emotional early-stage creativity. For teams building AI in creative or ambiguous domains, trust is built through emotion as much as accuracy, and empathetic design is a strategic lever, not just a UI detail.*

AI-Enabled Mission Engineering

Dr. Michael Pennock, *MITRE*

This research marks a pivotal step in transforming mission engineering by enabling AI to bridge the gap between model complexity and decision-making speed. Manual interrogation of SysML and simulation models sometimes takes weeks to answer simple consistency questions. Building a testbed that integrates fine-tuned LLMs with graph-based retrieval (GraphRAG) aims to create a system where engineers can interact conversationally with complex models, reducing cognitive load and analysis time. The significance is in automation and in enabling human-AI teaming that can improve the quality, adaptability, and trustworthiness of mission engineering outputs.

This opens the door for proactive, iterative exploration of alternatives, which current tools and timelines don't allow. This project lays the foundation for creating a reusable infrastructure for testing, integrating, and scaling AI innovations across the mission engineering community. It offers a forward-looking path where AI augments engineers' ability to deliver faster, more confident decisions in the face of uncertainty and evolving threats. It could fundamentally reshape how the DoD links DE investments to mission outcomes, making future forces more agile, informed, and operationally effective.

Agile Engineering Enabled by Agentic Co-Modelers

Hart Traveller, *SysGit*

This research explored how language models can generate, validate, and manage SysML models with syntactic and semantic integrity. The team focused on benchmarking how LLMs perform when asked to produce SysML code from system documentation, evaluating the models' ability to create syntactically valid and semantically meaningful outputs. Through experiments using Lunar Orbiter mission requirements, the team compared models with and without access to SysML documentation. While documentation-augmented models produced higher-quality syntax, errors and invalid constructs remained prevalent—highlighting the need for a language-model-agnostic validation layer to ensure correctness independent of the model used.

To address this, the team developed a framework that mediates all LLM interactions with SysML models through an interface layer capable of enforcing grammar rules, validating inputs and outputs, and preventing invalid structural edits. This system treats SysML as a graph-based data structure, allowing for parsing, semantic validation, serialization, and unit consistency checks. A Python-based prototype was demonstrated and presented as an approach akin to "Jedi for SysML." The presentation emphasized that the goal is not blind automation but a trustworthy foundation for model generation and manipulation, paving the way for AI-assisted system modeling that guarantees formal validity and, eventually, physical and logical soundness.

Advancing Systems Engineering through a Mathematically-Rigorous Co-Pilot: From GPS Systems to Satellite Constellation Revectoring

Dr. Paul Wach, *Virginia Tech*

The research fundamentally shifts SE from a largely qualitative discipline to one underpinned by rigorous, mathematically grounded theory, addressing a gap that has limited the maturity and reliability of MBSE practices. Quantifying degrees of homomorphism between system models enables engineers to precisely measure abstraction and equivalence rather than rely on subjective judgment, directly improving verification confidence and reducing costly errors in complex system design. The integration of agentic AI and graph-based knowledge representations creates a scalable, interactive co-pilot that can dramatically cut development and analysis time, highlighting the tangible efficiency gains possible when theory meets cutting-edge AI tools.

The significance lies in the potential to transform SE workflows into dynamic, data-driven ecosystems where humans and AI collaborate seamlessly, accelerating decision-making and adaption to evolving system requirements. The shift from descriptive to quantitative SE opens new pathways for handling increasingly complex systems, including national defense, where precision and agility are paramount. Embedding formal mathematical frameworks within accessible AI-enabled platforms enhances current engineering rigor and lays the foundation for future autonomous and trustworthy SE systems. Ultimately, it pushes the field toward a future where SE is more reliable and better equipped to meet the rapid innovation cycles and high-stakes demands of tomorrow's technologies.

AI4SE: AI Tools Workflows and Training

Using LLMs to Accelerate Text-Based SysML Model Requirements Gap Analysis

David Hetherington, *System Strategy*

This research explored how LLMs can accelerate one of SE's most tedious tasks: analyzing requirements in legacy documents and mapping them to SysML models for gap analysis. The research team tested how effectively LLMs could clean and structure unformatted requirement sets, generate short names, and atomize complex statements into model-friendly elements. While LLMs struggled with extracting precise requirements from human-written legacy documents, they performed well at simpler but time-consuming tasks like standardizing structure and identifying potential new requirements or model functions not present in the original data. Performance varied among models, with all tasks requiring human oversight to ensure accuracy.

Results suggest that while this AI-assisted process is not yet a turnkey solution, it holds promise for improving productivity and consistency in requirements engineering. Automation was especially successful in generating short names and developing draft functions to satisfy requirements. Scalability and cost remain limiting factors as API-based implementations are expensive, and human review is still essential. Overall, the study demonstrates encouraging progress toward transforming a manual, research-oriented activity into a repeatable, semi-automated engineering process for future model-based SE workflows.*

The Impact of Subject Matter Background on Generative AI Use and Perception

Stephen Hilton, *The George Washington University*

This study explored how individuals with different academic and professional backgrounds engage with generative AI during engineering design tasks, addressing a growing question among professionals: How will AI change the way I work? Participants were asked to design a robotic arm for the International Space Station using their preferred AI tools, allowing researchers to observe how human-AI collaboration unfolded across disciplines. The analysis revealed two main patterns: AI-augmentative workflows, where humans led the design process with AI providing iterative feedback, and AI-directive workflows, where users relied on AI to generate solutions directly. Contrary to expectations, participants with stronger technical backgrounds relied less on AI's autonomy, preferring to review and validate designs themselves, while others allowed AI greater creative input. Across participants, humans made most key design decisions and development efforts, while AI supported logistical and computational elements. Students reported high satisfaction with the process, noting that AI significantly reduced effort and time, especially for conceptual design tasks. The findings highlight an emerging dynamic: while AI can streamline engineering workflows and enhance productivity, trust and domain expertise continue to anchor human-AI collaboration. Future research will examine how these interaction patterns influence design quality and creativity, shaping best practices for integrating AI into engineering education and professional practice.*

AI in Systems Engineering Education: A Review of Curricula and Training Programs in the U.S.

Dr. Liang Zhang, *Drexel University*

This study examined how AI is being integrated into SE curricula and training programs across U.S. universities and professional institutions. A review of 52 representative programs found that while many universities offer AI-related courses, most focus narrowly on AI fundamentals, such as data foundations, machine learning, and data analytics, without embedding them in SE contexts. Using both qualitative and quantitative analysis, the research team mapped current offerings against SE competencies to identify where AI4SE and SE4AI appear in education.

The work culminated in a proposed framework for curriculum development that spans three tiers: core data and AI foundations, AI integration within SE courses, and advanced AI for systems applications. The recommendation emphasized that future SE programs must move beyond teaching AI as a stand-alone subject and instead weave it into the SE fabric to prepare graduates for the AI-enabled future of engineering practice. By aligning academic and professional training with this integrated model, the study aims to bridge educational gaps and shape a generation of systems engineers capable of designing, managing, and governing intelligent systems.*

AI4SE: Evolving Role of Digital Engineering

Leveraging AI Agents for Adaptive, Data-Driven Requirements in Army Systems

Yvan Christophe, *U.S. Army DEVCOM Armaments Center*

The presentation outlined a novel approach for transforming Army requirements engineering through collaborative AI agents. The research integrated multi-agent Retrieval-Augmented Generation (RAG) with secure, on-premise architectures to automate and improve how requirements are authored, validated, and traced. This effort aligns with Army modernization goals to streamline acquisition cycles and eliminate redundant or outdated requirements. By enabling multiple AI agents to reason collectively across structured and unstructured data, the system can identify gaps, clarify stakeholder intent, and generate new requirements with traceable logic—all while safeguarding Controlled Unclassified Information (CUI).

Technically, this framework relies on a knowledge base, retriever, and language model, operating within a localized vector database that supports semantic search, chunking, and embedding. Single- versus multi-agent architectures were compared, finding that distributed AI reasoning significantly improved precision in requirements interpretation and cross-system traceability. Compact models such as TinyLlama were tested to balance performance and latency for deployability in secure environments. Looking ahead, the aim is to refine the system for adaptive learning and broader Army integration, envisioning a future where AI agents act as collaborative partners, augmenting human analysts in real time while maintaining accountability, data sovereignty, and decision transparency.*

Ethical Issues and Behavior Strategies in Implementing AI-Enabled Digital Twins for Complex Military Systems

Dr. Nate Crews, *Caltech Center for Technology & Management Education*

The presentation provided an overview of how AI, digital twins, and digital threads can be integrated responsibly within complex SE. It was emphasized that AI should remain a subordinate, powerful assistant always under human control and ethical constraint. The presentation discussed how supervised, unsupervised, and reinforcement learning each bring value to engineering applications, from predictive maintenance to nonlinear simulation and automation. Illustrated were how digital twins enable predictive modeling, optimization, and "what-if" analysis for defense, manufacturing, and healthcare systems, while digital threads trace the flow of data and decisions across lifecycles, strengthening collaboration, efficiency, and compliance.

Ethics was framed as the foundation for trustworthy system design. Citing INCOSE and DoD Responsible AI principles, engineers were urged to embed accountability, fairness, and transparency throughout the development lifecycle. Frameworks such as NIST's AI Risk Management Framework can be used to ensure verification, validation, and harm prevention are integrated from concept to deployment. Case studies, from autonomous targeting to military logistics, demonstrated that AI and digital twins, when properly bounded, can accelerate performance without compromising safety or ethics. The next evolution in SE depends not just on intelligent tools but on engineers' disciplined commitment to responsible, interpretable, and ethically aligned innovation.*

GenGroves: A Bridge Between Systems Engineers and Domain Experts

Dr. Paul Wach, *Virginia Tech*

The presentation focused on GenGroves, a two-year research effort aimed at bridging the gap between systems engineers, domain experts, and AI-driven digital engineering tools. The project reflected a call for decisive, integrative leadership in advancing digital transformation. GenGroves was positioned as a framework that uses agentic AI orchestration to translate seamlessly between technical modeling languages like SysML v2 and the natural-language or visual inputs of subject-matter experts. The system allows engineers to generate models from textual descriptions or sketches and, conversely, to summarize complex model content in plain language for decision-makers, creating a two-way bridge between modeling precision and human understanding.

The research team evaluated multiple small and mid-sized LLMs (including Gemma, Mistral, and Llama 3) for use within LangGraph and Model Context Protocol (MCP) architectures, ultimately selecting Gemma as the orchestration model for its explainability and efficiency. Designed for on-premise or edge deployment, GenGroves supports secure, low-latency collaboration without dependence on cloud infrastructure. The goal is automation for collaborative augmentation: empowering systems engineers, project managers, and domain specialists to co-develop and interpret digital models through a shared, AI-mediated interface. GenGroves represents a practical step toward uniting modeling rigor, accessibility, and explainable AI—advancing digital engineering as both a technical and human endeavor.*

SE4AI: Trust & Bidirectionality

Dynamic Risk-Fusion Framework for Human-AI Collaborative Aerial Threat Prioritization

Dr. Nidhal Bouaynaya, *Rowan University*

The research focused on real-time decision-making against swarms of autonomous drones. The framework presented used camera-based sensing, computer vision, and lightweight deep learning models to detect, classify, and prioritize aerial threats based on likelihood and severity of interaction, all deployed on edge devices for field readiness. Lacking existing drone datasets, the research team built synthetic data in Unity and developed an auto-labeling engine to accelerate model training. The resulting AI system can dynamically update intent classifications, highlight the most critical threats to soldiers, and reduce cognitive overload. Tested in live drone-swarm simulations, the model demonstrated effective situational awareness and responsiveness. Emphasized was the goal of building AI that detects and interprets intent, transforming battlefield autonomy into an interpretable, cooperative decision-support tool.

Validating a Contextual Trust Assessment Instrument for AI-Enabled Systems Using the MASTOPIA Testbed

Jessica Lee, *Arizona State University*

The presentation introduced a novel trust evaluation framework based on *Responsivity*, a multidimensional measure of how well AI systems adapt, communicate, and collaborate with humans. The work redefined automation trust beyond static supervisory models by capturing dynamic, context-sensitive relationships between people and intelligent systems. Through both perceptual and pilot studies, it was shown that systems rated high in Responsivity (anticipate user needs, align goals, and communicate decisions clearly) are trusted more than rigid or opaque systems. LLM performance was evaluated using the *MASTOPIA* testbed in complex tasks such as intelligence analysis, comparing “high-responsivity” versus “low-responsivity” AI interfaces. Findings revealed that clarity, contextual awareness, and proactive feedback increase both trust and performance. Responsivity offers a scalable foundation for trust assessment in human–AI collaboration, particularly for adaptive and conversational systems.

Retrieval Augmented Generation for Operations Order Evaluation

Wilmer Maldonado, *U.S. Army DEVCOM Army Research Laboratory*

The presentation focused on *OpOrder-GPT*, a retrieval-augmented generation (RAG) framework designed to help Army planners generate, review, and evaluate Operations Orders (OPORDs). These structured documents guide mission execution but are often dense and manually validated. The highlighted system leverages RAG pipelines, XML schema extraction, and fine-grained context retrieval to preserve the OPORD’s hierarchical structure while answering queries or evaluating completeness. Integrated modules automatically assess sections against doctrinal standards using the Army’s METL rating scale (Trained, Needs Practice, Untrained), offering transparent, source-traceable evaluations with confidence indicators. The system achieved high accuracy in structured tests and significantly reduced review time. Future work will include integrating user feedback loops and benchmarking model performance. Tools like OpOrder-GPT exemplify how trustworthy, explainable AI can enhance decision-making and efficiency in mission-critical systems.

Trusted Artificial Intelligence Challenge for Systems Engineering Results and Insights

Sami Saliba*, Virginia Tech*

Results were presented from the *Trusted Artificial Intelligence Challenge for Systems Engineering*, which explored how to build trust into AI-enabled mission systems. Using a simulated operation called *Safe Passage*, teams were tasked with navigating semi-autonomous vehicles through a minefield while relying on imperfect AI and human subsystems – each with different trade-offs in speed, reliability, and accuracy. The challenge required participants to define, measure, and validate trustworthiness using SE artifacts and metrics. Results showed diverse approaches: some focused on human–AI teaming, others on statistical robustness or explicit trust requirements integrated into extended V-models. Key recommendations emphasized the need to define clear trust requirements early, establish human-in-the-loop validation infrastructure, and build interdisciplinary teams capable of managing trust as a dynamic property of system performance, safety, and transparency.

SE4AI: System Design Processes That Support AI Across the Lifecycle

Risk-Informed Flight Scenario Planning and AI-Enhanced Decision Support in Aviation Domains

Yonas Ayalew, *North Carolina Agricultural and Technical State University*

The researchers developed a framework that combines quantitative risk modeling with natural language AI interfaces, enabling aircraft operators to make informed, mission-specific flight decisions in real-world, often unpredictable conditions. A key insight highlighted the system's ability to identify unsafe airspace states—such as strong winds or energy limitations—and to strategically plan trajectories around these, taking into account FAA regulations and aircraft-specific constraints. This work redefines how AI and risk assessment can be integrated into flight planning for civilian and defense aviation by treating operational safety as a dynamic, data-driven process rather than a static checklist.

Significant is the integration of energy management as a core risk factor, including aircraft battery degradation and reserve requirements, elements often overlooked in traditional planning but critical in electric and advanced aircraft. The framework quantifies operational risk and safety margins, providing flight plans that optimize for safety and performance. Also unique is the use of LLMs to deliver context-aware, natural language decision support, turning the AI into an interactive assistant that interprets operator goals, constraints, and preferences. This human-centered interface enables real-time adaptability and keeps the operator in the loop, crucial for high-stakes military and advanced aviation missions. The use of AI is not to replace pilots or planners but to augment their judgment, enabling safer, more adaptable, and regulation-compliant missions in an increasingly complex airspace.*

A Mission-Centric Resilience Lifecycle and Test Framework for AI-Enabled Systems of Systems

Dr. Tyler Cody, *Virginia Tech, and* Dr. Peter Beling, *University of Virginia*

This presentation redefined resilience for AI-enabled systems by linking component-level AI performance to mission-level outcomes. The authors argued that resilience should not be confined to robustness testing within individual algorithms but should encompass the entire AI pipeline from data collection and model training to deployment and operation. The proposed framework integrated detection, mitigation, and recovery across both AI pipelines and the larger SoS they support. By modeling resilience as a lifecycle property, engineers can identify performance degradation earlier, in particular catching issues at the component or pipeline level before they cascade into mission-level failures.

A key insight is that AI introduces a kind of generic resilience that can be systematically engineered, unlike cyber resilience, which is often threat-specific. The framework's taxonomy spans system, process, and operational levels, showing how resilience mechanisms (such as data validation and fallback configurations) can work across onboard and offboard AI pipelines. The approach bridges the gap between traditional assurance practices and AI-enabled operations, offering a pathway to continuously test, monitor, and adapt AI systems for mission assurance in complex, dynamic environments.

A Modular Agent-Based Architecture for Digital Engineering

Nicole Manno, *ManTech*

The presentation introduced a technically grounded framework for deploying modular, agent-based architectures within digital engineering environments (DEEs), focused on how to make agents interoperable, resilient, and semantically aware. While interest in intelligent agents has surged, most current implementations are brittle: they break when software tools or APIs change, and often lack governance or clear data contracts. The challenge in agentic engineering is not intelligence, but data boundary definition, i.e., how agents access, interpret, and exchange engineering data without cascading errors when environments evolve.

To solve this, a technology-agnostic integration architecture anchored by two design commitments was introduced: (1) agents retrieve and publish data through a semantically linked, ontology-backed digital thread, and (2) every agent is self-contained with explicit input/output contracts, versioning, and permissions. The layered architecture—contract registry, orchestration engine, runtime manager, and observability layer—can all work together to ensure safe composition, schema validation, and automated provenance tracking. This design underpins ManTech Foundry, an operational platform that currently supports tools for obsolescence management, digital twin querying, and capability optimization. Standardizing these agent boundaries and enforcing semantic interoperability will allow engineering teams to focus on ontology and logic development, accelerating the maturity of governed, scalable agent ecosystems in digital engineering.

SE4AI: Safety, Reliability, & Ethics

A Systems Approach for Governable & Sustainable AI

Philip DiBona, *Lockheed Martin*

This presentation addressed a core challenge in defense systems engineering: ensuring safety, explainability, and trust in AI models operating autonomously in dynamic environments. Instead of relying on ad hoc fixes, the OMEGA and Spotlight frameworks combined with the MBSE DevKit create a closed-loop ecosystem that monitors model performance and environmental assumptions in real time, feeding insights directly into MLOps pipelines. This architecture treats AI monitoring and governance artifacts as traceable, reusable, and auditable SE elements—laying the foundation for scalable and certifiable AI deployment in high-stakes missions. Strategic innovation lies in treating governance actions (like fallbacks, alerts, or model disablement) as explainable, lineage-traceable events, enabling human operators to understand what decisions were made and why—an essential capability in complex, AI-infused SoS where emergent behavior is a real risk. Performance and risk thresholds are mission-specific, allowing operators to tailor AI behavior to operational context, increasing flexibility while maintaining control. A key insight is that model performance drift and environmental shifts aren't just post-mission concerns but can be modeled and monitored from Day 1 using platform-specific data collectors and mission-tuned monitors, enabling predictive mitigation rather than reactive patches.

In an environment where policy and accreditation processes lag behind AI capability, this framework creates a systematized evidence trail that can satisfy assurance requirements. The use of standardized templates in MBSE tools like Cameo ensures plug-ins, monitors, and governance strategies are baked into requirements and design, reframing AI assurance as a designable, testable, and improvable system capability, laying the groundwork for scalable, trustable AI deployment in defense and beyond.

Self-Preference Undermines LM Evaluation

Dr. Shi Feng, *The George Washington University*

The presentation highlighted the widespread use of language models (LMs) in place of humans to evaluate LM outputs and the associated potential for systematic biases in these LM "judges". This bias is important to understand as humans are having increasing difficulty supervising these models as they are applied to ever more difficult and larger tasks. The presentation explained different approaches to expose LM biases in evaluating quality of LM outputs including evaluation of self-preference—where performance or quality is overestimated for models from the same family or models distilled from the judge—and whether pressure for optimization can cause the reward model to diverge from human-defined criteria and produce increasingly worse outputs.

The research found that LM judges can distinguish their own words from other LMs and humans and subsequently favor themselves in pairwise comparisons. Further, divergence of a reward model was exacerbated by self-preference behavior. The work emphasized the need to better understand how LMs perform self-recognition, finding that not all obfuscation methods are equally helpful at mitigating self-preference, and that LM judging decisions can be sensitive to perturbations in self-evaluation situations. The presentation explained the need for SE approaches to create abstractions of these complex protocols, to help understand how these biases are amplified, and to lead to standard methods to improve on these biases.*

Enhanced AI Decision Process for A-eVTOL System

Bing Mak, *Stevens Institute of Technology*

The initiative explored autonomous eVTOL flight and underscored a critical systems-level insight: AI decision-making for flight control cannot be trusted in isolation; it must be deeply embedded within a robust, redundant, and dynamically regulated sensor and communication architecture. The Dynamic Subsystems Management Framework (DSMF) proposed moves beyond conventional redundancy by actively coordinating sensor reliability, communication conditions (LOS/NLOS), and AI logic in real-time, allowing the system to adapt intelligently to operational disruptions. A significant takeaway is that flight safety is no longer just a hardware problem but a dynamic AI governance problem, where real-time sensor prioritization, failure prediction, and adaptive routing are essential for mission assurance.

This approach emphasized unified model architectures (UMA) that abstract across modalities (RF, video, optical) and tasks, simplifying AI deployment while retaining adaptability, a design choice that reduces lifecycle complexity and enhances certification pathways. The integration of Q-learning and fuzzy logic demonstrated a pragmatic blending of autonomous learning with explainable heuristics, recognizing the need to optimize AI performance even under degraded sensing conditions. This work reframes A-eVTOL certification as a SoS reliability challenge, where safety hinges not just on component integrity but on the AI's ability to synthesize imperfect data and maintain operational intent with resilience and transparency.

Reinforcement Learning Qualification Process (RLQP): A Framework for Evaluating Safety and Robustness in Reinforcement Learning

Steven Senczyszyn, *Michigan Technological University*

The Reinforcement Learning Qualification Process (RLQP) strategically reframes how reinforcement learning agents should be evaluated for deployment in safety-critical environments, moving beyond performance metrics to emphasize robustness, safety, and trust. Unlike traditional testing that often assumes static environments, RLQP introduces a structured perturbation and statistical evaluation framework that simulates real-world adversities, thus, identifying the operational limits of RL agents before deployment. Safe reinforcement learning should aim to understand how agents fail under stress, and create a loop where those failures inform retraining and formal safety bounds.

RLQP emphasizes subtask decomposition within end-to-end learning architectures, allowing safety analysts to extract meaningful control logic from opaque RL behaviors, crucial given that conventional hazard analysis tools often fall short for RL systems. The combination of perturbation testing, curriculum learning, and reproducible evaluation protocols represents a comprehensive effort to build RL systems that recover gracefully from real-world deviations. Ultimately, RLQP lays the groundwork for a qualification standard, closing the gap between cutting-edge research in RL and the engineering rigor demanded in autonomous vehicles, drones, and defense systems.

Operationalizing Tool Selection with an MLOps Tool Evaluation Rubric

Dr. Thomas Serban von Davier, *Software Engineering Institute (SEI), Carnegie Mellon University*

Carnegie Mellon University's SEI has developed a tool evaluation system for teams that need to systematically assess Machine Learning Operations (MLOps) tools needed for AI development. Presented was an evaluation rubric used by the SEI team that provides a structured and customizable capability-driven characterization, enabling practitioners to identify tools that best match their development stack, compliance requirements, and system goals. For example, the rubric assesses whether a tool can operate in air-gapped environments or support high-side deployment scenarios, enabling teams to eliminate tools that don't meet core needs.

SEI's rubric is organized around core MLOps functions: data management, model management, development and testing, and deployment, which are then further decomposed into detailed subcategories. In ongoing work, SEI is building a modular, GUI-driven version of the rubric to improve accessibility and project-specific tailoring for Government stakeholders. Additionally, SEI is starting to pair rubric-based evaluations with hands-on expert testing in controlled environments to provide more accurate assessments on what a tool can reliably deliver in practice.*

Opportunities and Challenges in Integrating System Safety Models into SysML: SysML-based Fault Tree Analysis

Dr. Lance Sherry, *George Mason University*

This presentation explored how SysML-based modeling, particularly through the Risk Analysis and Assessment Modeling Language (RAAML), can integrate safety analysis into the digital engineering workflow, allowing safety insights and design decisions to inform each other in real time. As systems increasingly incorporate AI and migrate toward digital engineering, traditional system safety analysis (SSA) methods—performed separately from system design—struggle to keep pace. Certification processes require safety artifacts such as hazard and fault tree analyses (FTA), often developed in isolation and leading to outdated assumptions and limited feedback between design and safety teams.

The research focused on embedding FTA within SysML models to quantify and manage uncertainty in safety-critical systems. A key challenge addressed was how to estimate missing or uncertain probabilities in fault tree "leaf nodes," often caused by limited data or rare-event behavior. This effort demonstrated methods using Bayesian and interval analysis to represent uncertainty, improving confidence in system-level safety assessments. Linking safety models, design artifacts, and probabilistic analysis within a unified SysML framework supports dynamic, data-informed certification processes and strengthens safety assurance for AI-enabled systems.

A Data-Driven Framework for Trustworthy Situational Awareness Application in DAA Systems

Lydia Zeleke, *North Carolina Agricultural and Technical State University*

The researchers focused on the challenges associated with Detect-and-Avoid (DAA) systems for Unmanned Aircraft Systems (UAS), which are relied upon to replicate a human's "see-and-avoid" capabilities and ensure collision-free flight. DAA systems generate alerts and guidance based on sensor inputs; however, the DAA evaluation datasets consist of many short, multiple multivariate time series representing distinct aircraft encounter scenarios. The dynamic nature of DAA data and its structure makes it difficult for traditional data integration and analysis methods to manage with solid performance.

The research team developed a dual-path approach that brought together near real-time anomaly identification and interpretation with ingestion of DAA data to monitor performance and trace deviations back to the anomaly diagnosis model. The approach employed first-order model-agnostic meta learning for task-adaptive model parameter initialization based on gradient steps and then fine-tuned the meta learning to generalize the model to help it adapt to unforeseen encounter scenarios. The presentation highlighted the advantages of parallelizing training, demonstrating that the approach with the meta-learning capability outperformed other benchmarks. Future work will focus on evaluating DAA performance metrics to quantify how specific sensor anomalies degrade system behavior and on approaches to incorporate model explainability.*

SE4AI: Evolving Role of Digital Engineering

An Adaptive Guidance Tool for Enhanced Education in Systems Engineering

Bryant Baldwin, *University of South Alabama*

The research focused on developing an Intelligent Tutoring System (ITS) to replicate mentorship in SE. Inspired by successes such as Project Sherlock and DARPA's Digital ITS, which rapidly advanced novices to expert-level proficiency, this system adapts to individual learning styles, identifying competency gaps and providing targeted guidance. Built within the GIFT Virtual Open Campus, the platform walks users through SE processes and tailors curricula to build core competencies like systems thinking and stakeholder engagement. Future iterations will integrate virtual conversational agents to simulate stakeholder interactions, allowing learners to practice real-world communication and requirements elicitation. The long-term vision is an AI mentor that evolves with use, democratizing expert knowledge and guiding users through complex design processes autonomously and adaptively.

Design Choices for Human/AI Teaming on Complex and Long-running Endeavors

Dr. Peter Denno, *National Institute of Standards and Technology*

The presentation focused on a framework for agent-led interviewing systems that help small and medium-sized manufacturers interface with complex engineering tools through conversational AI. The concept of "tool stewards" envisions AI agents that act as knowledgeable collaborators, eliciting design requirements, guiding users through domain-specific languages (DSLs), and ensuring traceability and understanding. Using the Model Context Protocol (MCP), the system structures dialogue around four key design dimensions—processes, data, resources, and optimization goals—while maintaining transparency and iterative refinement. The agents ask and interpret questions, build structured design representations, and mentor users to ensure comprehension, avoiding black-box automation. This approach blends knowledge elicitation, AI orchestration, and learning-by-doing, supporting ISO-standardized digital manufacturing processes. Ongoing work with ISO/IEC SC 42 aims to formalize best practices for AI-assisted design interviews, ensuring that future digital tools teach as they build.

Ship of Theseus Methodology for Creating Living Engineered Systems

Laura Otero Hernandez, *The George Washington University*

The presentation proposed the Ship of Theseus Methodology, a framework for engineering self-adaptive, "living" systems that evolve while retaining their core identity—much like the mythological ship whose parts were replaced over time. The research explored how digital twins, failure mode analysis, and state-machine modeling can enable systems to recognize when they no longer meet mission requirements, autonomously adjust those requirements, and evolve in operation. Using examples from adaptable technologies—such as SpaceX's reusable boosters and the long-lived B-52 Stratofortress—the presentation illustrated how continuous adaptability can extend system lifecycles while maintaining capability. The methodology seeks to bridge the gap between design and operations, embedding recursive verification and validation directly into the system. Early work aimed to validate this concept through digital twin testing, ultimately defining what qualifies as a living engineered system capable of self-assessment, self-modification, and long-term mission relevance.

SE4AI: Test and Evaluation

T&E of AI Enabled System Case Study – Intelligence Analysis with RAG-LLM

Elena Charnetzki, *MITRE*

The research supported OSD's DTE&A on applying hazard analysis and model-based systems engineering (MBSE) to improve testing of RAG-LLM (Retrieval-Augmented Generation Large Language Model) systems in defense contexts. An analysis case demonstrated how combining Systems-Theoretic Process Analysis (STPA) with MBSE can help identify and prioritize critical risks, especially under worst-case conditions. The research team developed a generalized hazard reference model tracing failures through the full human–AI workflow, allowing programs to reason about where hazards originate and how guardrails such as input vetting, explainability, and user training can mitigate them.

The framework presented grounds analysis in concrete user stories and activity diagrams that capture real operational contexts. Several archetypal hazard scenarios were presented, emphasizing that evaluation must account for both model performance and human expertise in detecting errors. The approach enables earlier, more systematic "tabletop" analysis of potential failures, informing design choices, guardrail development, and test strategies before deployment to support safer and more trustworthy AI integration across the Department.

Discourse Analysis as a Diagnostic Lens into Dialogue Systems: LLM Evaluation Considerations Across Four Functionality Dimensions

Dr. Samantha Finkelstein, *Software Engineering Institute (SEI), Carnegie Mellon University*

The presentation aimed to clarify how inconsistent terminology and vague operationalization complicate evaluation of LLMs. Many so-called paradoxes in AI assessment stem from poorly defined terms, particularly when concepts such as *trustworthiness*, *reliability*, or *safety* are used interchangeably across different system levels. Drawing from SE, distinctions were made between System 1 (model-level performance, e.g., accuracy or latency) and System 2 (application-level outcomes, e.g., human safety or mission impact). It was proposed that LLMs serve three primary functions – conversation, generation, and analysis – each requiring distinct evaluation methods. For conversational systems, measuring *discourse quality* was emphasized, noting that meaning is co-constructed across dialogue turns and depends on conversational context, participant roles, and epistemic alignment.

It was argued that many LLM "hallucinations" are *interpretive overreaches*, misalignments between user intent and model inference, best addressed through dialogue design rather than retraining. Discourse is explainable and dialogue is designable. Evaluators can adopt established discourse-analytic methods to distinguish between linguistic misunderstandings and genuine capability failures, ultimately enabling more rigorous, human-centered evaluation of conversational AI.

Safe Experimentation with LLM-Controlled UAVs: An Agile Systems Engineering Approach to Requirements Development for Autonomous Systems
Matthew Harris, *SAIF Autonomy*

This work addressed the challenge of developing and assuring safety for non-deterministic autonomous systems. Traditional assurance methods often lag behind the pace of AI-enabled innovation, limiting safe experimentation. The research team introduced a Runtime Assurance (RTA) architecture that enables safe testing of LLM–based UAV controllers by separating safety functions from the autonomy itself. The RTA module monitors for violations (such as leaving geofences or exceeding payload limits) and intervenes independently of the AI controller to maintain safe operation. Live UAV trials demonstrated the RTA can predict and prevent violations in real time, while also capturing system data to refine design and assurance requirements.

The project resulted in two sets of elicited requirements, one for the LLM controller (covering mission constraint awareness, command formalism, and language safety) and one for the RTA system (covering recovery effectiveness, stability, latency, and transparency). Together, they bridge the gap between traditional certification frameworks and AI assurance, showing how agile SE can support adaptive, data-informed requirements for autonomy. Natural language control through LLMs reduced warfighter cognitive burden by allowing more intuitive interaction with UAVs, though further sophistication is needed to realize the full potential of this human–AI teaming.*

Learning Complex Degradation Signal Manifolds for Accurate RUL Prediction via Conditional Diffusion Models
Donghyun Ko, *North Carolina State University*

Traditional Remaining Useful Life (RUL) prediction models rely on simplified assumptions about how systems degrade, often using exponential or Brownian motion error models that cannot capture the nonlinear, multimodal behaviors found in real-world data. This research introduced Conditional Diffusion Models as a generative, data-driven alternative capable of directly learning the manifold of degradation signals without predefined distributional assumptions. Comparing diffusion-based approaches against Bayesian update models and machine learning benchmarks like PCA + Lognormal regression demonstrated that diffusion models outperform conventional methods, particularly when degradation involves irregularities, sudden shifts, or multiple failure modes.

The presentation highlighted diffusion models' consistency and reliability in predicting errors, with lower root mean square and mean absolute percentage errors showing they can generalize effectively to complex, noisy environments. Their strength lies in capturing subtle nonlinear relationships and handling diverse operational conditions, making them a robust framework for real-world prognostics. Diffusion models represent a new paradigm for trustworthy, adaptive system health management, offering the potential to transform predictive maintenance and mission-readiness assessments across industries.*

Tailorable Risk-Informed AI Test and Evaluation Strategy

Dr. Erin Lanus, *Virginia Tech National Security Institute*

The research focused on developing a risk-informed testing and evaluation (T&E) framework for AI-enabled systems (AIES), particularly those integrated into electronic warfare (EW) and cognitive EW (CogEW) applications. While traditional T&E approaches can still apply, AI introduces distinctive risks tied to data quality, model opacity, and non-determinism that complicate assurance and validation. Analysis of existing frameworks and literature identified challenges such as inadequate training data characterization, data or model drift, implicit requirements derived from data, and the difficulty of tracing causal faults in opaque models.

The research team proposed a hierarchical, SoS view of AI testing, ranging from software and component testing to system integration and human–machine teaming, to better capture dependencies and evolving system behavior. A dynamic test scheduling (one-time, periodic, or continuous) would align with how frequently AI models update or learn online, and would require a formalized, risk-based AI T&E strategy that tailors methods to specific AI traits (e.g., learning type, lifecycle phase, and human or environmental interactions). Ultimately, AI testing must move beyond static, point-in-time validation toward continuous, adaptive evaluation across the system's lifecycle to meaningfully reduce risk and ensure operational trustworthiness.

Appropriate Levels of Human Judgement for Autonomy

Dr. Elizabeth Mezzacappa, *U.S. Army DEVCOM Armaments Center*

This research explored how to design and test autonomous systems that preserve appropriate levels of human judgment—a requirement codified in DoD Directive 3000.09, which mandates that autonomous weapons be designed to allow commanders and operators to exercise meaningful control. Through a literature review conducted for the Office of the Secretary of Defense (OSD), the work identified human factors engineering (HFE) and human systems integration (HSI) as the core scientific frameworks for ensuring this standard. The challenge lies in translating policy language about "appropriate judgment" into measurable, testable requirements for system design, usability, and interaction. Effective testing must begin before system fielding and include structured feedback loops with users, subject matter experts, and warfighters.

The presentation emphasized that testing for human judgment is essentially testing for optimal human–system integration. Metrics should capture performance outcomes (mission success) and effectiveness measures (situation awareness, fatigue, workload, and trust). Progress can be seen in ongoing work by IDA and MITRE and recent DoD progress, including the adoption of the Human Readiness Level standard. Ultimately, ensuring appropriate levels of human judgment requires embedding usability, feedback, and trust calibration throughout design and evaluation.*

Measuring AIES Trustworthiness

Carol Pomales, *MITRE*

The presented work was conducted in support of OSD's Developmental Test, Evaluation, and Assessments (DTE&A) on advancing methods for testing and evaluating AI-enabled defense systems. AI testing must be treated as a continuous, lifecycle process grounded in SE principles. Building on MITRE's *Systems Engineering Processes to Test AI Right (SEPTAR)* framework, the research team proposed a multidimensional trustworthiness metric to help program and test teams identify, evaluate, and mitigate risk across development phases. The metric integrates technical, procedural, and contractual factors (spanning R&D, prototyping, cybersecurity, data governance, and requirements definition) to generate actionable recommendations rather than a single score.

The research distinguished trustworthiness, a property of the system (safety, security, explainability, fairness, reliability), from trust, a human perception of system dependability. These constructs were linked to NIST definitions and the DoD's shift toward continuous test and evaluation (CTE), which reuses evidence throughout the lifecycle to reduce late-stage risk and cost. The approach aims to standardize evaluation practices, automate evidence generation through tools such as data and model cards, and promote repeatable, mission-aligned testing of AI systems. An invitation to collaborate to refine and apply the framework was extended to attendees.

Mis-classification Testing in Open Source Supervised Learning Projects

Dr. Akond Rahman, *Auburn University*

This research examined a blind spot in how open-source software projects that use supervised learning handle testing: Among 278 Python-based projects analyzed, 76% had no test cases at all, and none explicitly tested for mis-classifications, a vulnerability that can arise from adversarial "poisoning" attacks where clean training data is subtly corrupted to produce incorrect predictions. Such mis-classifications can lead to serious downstream consequences in domains like healthcare and finance. The study aimed to help developers detect these vulnerabilities by automatically generating test cases that can identify accuracy degradation caused by poisoned data.

The researchers developed an automated approach using a loss-based label perturbation method that proved more effective and faster than traditional probability-based approaches for simulating and detecting mis-classifications. Introducing these automated unit tests into supervised learning pipelines allows developers to proactively assess model robustness and improve security against data poisoning. The work underscored the growing need to extend software testing practices to the ML lifecycle treating mis-classification detection not as a research curiosity, but as a fundamental element of responsible machine learning engineering.*

A New Test & Evaluation Regime for Human-AI Systems

Aditya Singh, *The George Washington University*

The presentation focused on the development of a new test and evaluation (T&E) regime for human–AI systems that accounts for how integration architectures shape performance, risk, and mission outcomes. The research addressed a gap in current AI T&E efforts: while much focus is placed on data, models, and algorithms, less attention is paid to how humans and AI actually interact in operational workflows. Human–AI system architecture was defined as the allocation of functions, authority, and decision control between human operators and AI agents, illustrating its significance through examples ranging from driver-assist features to autonomous driving systems.

Using a simulated Army challenge scenario, an unmanned vehicle navigating a minefield, multiple architectures were compared, from fully human-operated to fully autonomous and hybrid configurations. Findings showed that system outcomes (e.g., traversal time, number of IEDs struck) did not follow a simple gradient between human-only and AI-only performance. Instead, mixed human–AI architectures often exhibited nonlinear and unpredictable trade-offs, with outcomes heavily influenced by factors such as environmental complexity and mutual calibration of human and AI trust. It was concluded that human–AI T&E must expand the system boundary to include the human as an integral system component, emphasizing that only through such architectures and simulation-based experimentation can evaluators uncover emergent behaviors, resilience limits, and hidden dependencies that shape real-world mission effectiveness.

Stress Testing Safety-Critical Learning Enabled Systems with Optimization and Adaptive Sampling

Jon Vigil, *OptTek Systems*

This work was conducted under an Army-sponsored Small Business Technology Transfer (STTR) effort focused on verification, validation, and assurance of machine learning systems for safety-critical applications. Conventional model validation techniques, such as accuracy metrics and unbiased test datasets, often fail to expose hazardous or edge-case behaviors that could compromise safety in operational environments. To address this, the research team developed stress-testing methods that deliberately search for model failures using optimization-based falsification and adaptive sampling. Treating models as black boxes and iteratively biasing tests toward unsafe behaviors demonstrated how these approaches can uncover vulnerabilities and unsafe decision boundaries overlooked by traditional testing.

Examples such as a reinforcement learning driving simulator, collision-avoidance systems, and adaptive control models showed how optimization algorithms and Gaussian process–based adaptive sampling can map “safety contours” and identify failure thresholds even in complex or uncertain environments. These methods support certification, training data refinement, and runtime monitoring when exhaustive testing is infeasible. Ongoing work explores scaling the approach to high-dimensional and image-based models using generative AI to create realistic perturbations. Stress testing was illustrated as a necessary evolution in AI assurance, shifting from proving model performance to disproving safety assumptions, and offering a structured path toward integrating model-level evidence into system-level assurance frameworks.

ACKNOWLEDGEMENTS

The organizers would like to express thanks to the presenters in this workshop who generously shared their knowledge, expertise, and experience. Thank you to DEVCOM AC Systems Engineering Directorate and SERC for planning and facilitating, and to all the attendees for the open discussion, ideas, and information exchange. It was again an opportunity to gather the community together to advance SE and AI.

WORKSHOP ORGANIZERS

Executive Hosts:
Dr. Dinesh Verma, *SERC Executive Director, Stevens Institute of Technology*
Mr. Edward W. Bauer, *Director of the Systems Engineering Directorate (SED), U.S. Army DEVCOM Armaments Center (AC)*

Technical Committee Leads:
Dr. Zoe Szajnfarber, *GWU/SERC*
Mr. Tom McDermott, *Stevens Institute of Technology/SERC*
Mr. Al Stanbury, *U.S. Army DEVCOM AC*
Dr. Val Sitterle, *Stevens Institute of Technology/SERC*

Moderators:
Ms. Syeda Anjum, *Stevens Institute of Technology/SERC*
Dr. Peter Beling, *University of Virginia*
Dr. Myron Hohil, *U.S. Army DEVCOM AC*
Dr. Bryan Mesmer, *University of Alabama in Huntsville*
Dr. Jitesh Panchal, *Purdue University*
Mr. Benjamin Schumeg, *U.S. Army DEVCOM AC*
Mr. Al Stanbury, *U.S. Army DEVCOM AC*
Dr. Ralph Tillinghast, *U.S. Army DEVCOM AC*
Mr. Benjamin Werner, *U.S. Army DEVCOM AC*

ACRONYM LIST

AI/ML – artificial intelligence/machine learning
AV – autonomous vehicle
DE – digital engineering
DEVCOM – Development Command
DoD – Department of Defense
DSL – domain-specific language
GenAI – generative AI
HSI – human systems integration
LLM – large language model
MBSE – model-based systems engineering
ML – machine learning
RAG – retrieval-augmented generation
RLQP – Reinforcement Learning Qualification Process
SE – systems engineering
SED – Systems Engineering Directorate
SEPTAR – Systems Engineering Processes to Test AI Right
SERC – Systems Engineering Research Center
SoS – system of systems
STPA – systems theoretic process analysis
SysML – systems modeling language
T&E – testing and evaluation
UARC – University Affiliated Research Center
UAS – unmanned autonomous system
UAV – unmanned autonomous vehicle
VLM – visual language model
V&V – verification and validation