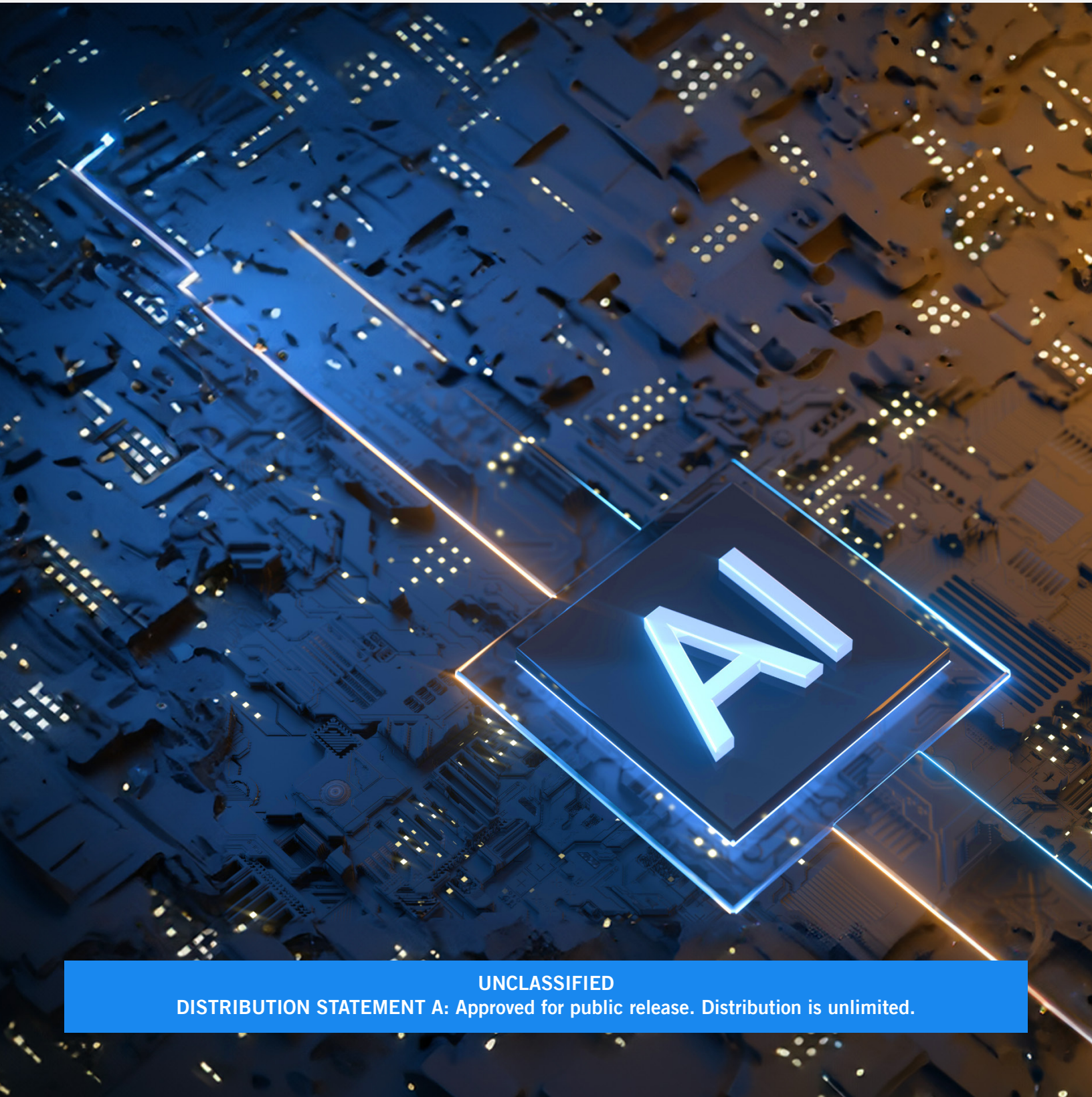
# AI4SE & SE4AI

RESEARCH AND APPLICATION WORKSHOP
SEPTEMBER 27-28, 2023

SYSTEMS ENGINEERING RESEARCH CENTER

U.S. ARMY

DEVCOM ARMAMENTS CENTER



# EXECUTIVE SUMMARY

## *Objective*

The US 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 fourth Artificial Intelligence for Systems Engineering & Systems Engineering for Artificial Intelligence (AI4SE & SE4AI) workshop on September 27-28, 2023. The two-day event—held in person at The George Washington University, Washington, DC—gathered participants from government, academia, and industry to learn from leaders using AI in this space, share ideas, and further explore outcomes that resulted from the three previous AI4SE & SE4AI workshops.

## *The Workshop*

The conference theme, "Balancing Opportunity and Risk: The Systems Engineer's Role in the Rapid Advancement of AI-Based Systems," explored the intricate relationship between artificial intelligence (AI) and systems engineering (SE). The aim was to foster discussions and insights on the responsible deployment of AI in SE and how SE can support the development of robust and ethical AI systems. Participation increased this year, and the total of 24 presentations, three panels, and three keynote speeches exceeded last year's totals. Various discussion sessions were conducted over the two-day event and focused on topics within the areas of Artificial Intelligence for Systems Engineering (AI4SE), Systems Engineering for Artificial Intelligence (SE4AI), Trustworthy AI, and Human/AI Teaming.

The importance of workability between humans and AI/machine learning (ML) in varying systems contexts was a consistent thread throughout the two days. The workshop also underscored the importance of a holistic approach to AI development that emphasizes technical prowess, ethical considerations, and human-centric design. In the realm of SE, AI integration requires meticulous consideration of ethical implications, adversarial threats, and counter-autonomy measures, with ongoing research aimed at balancing security needs and standards. Continuous testing, evaluation, and adaptation are vital for cultivating robust and trustworthy AI systems, especially in complex and dynamic environments. Human cognition, decision-making, and the psychological aspects of AI interactions highlight the need for user-friendly and transparent systems. The standardization of AI infrastructure and data formats is crucial for seamless collaboration, and calibrated metrics are essential for measuring trust in AI systems. Throughout, workforce development was emphasized, particularly the need for professionals well-versed in addressing the ethical, technical, and societal challenges posed by AI.

### *Outcomes*

Future research should adopt a multidisciplinary approach, prioritizing ethical considerations, user-centric AI system development, and continuous learning and adaptation in dynamic environments. Collaboration among academia, industry, and government entities remains essential to address the multifaceted challenges associated with AI development and deployment. Recommendations for future research included:

- Seamless integration of AI systems into human workflows, with consideration of user experience, cognitive load, and common standards.
- Robust testing methodologies to identify vulnerabilities in AI systems and ensure the resiliency and adaptability of AI across dynamic and complex environments.
- Development of comprehensive ethical frameworks and guidelines for responsible AI development and deployment, particularly in warfare and manufacturing.
- Enhanced methods for model explainability and interpretability, particularly techniques that capture intuitive judgments and the "gist" of model outputs, to increase transparency and user trust.
- Robust assessment and assurance of the quality of datasets used in AI training, especially in mission-critical applications.
- Standardization of AI infrastructure, development tools, and data formats to enable interoperability and collaboration across diverse sectors.
- Factors such as adversarial attacks, real-world complexities, and the ethical implications of AI in warfare to develop trustworthy autonomous systems for military applications.
- Counter-autonomy measures and the integration of AI in multi-object recognition and targeting.
- Educational programs and workforce training to equip professionals across disciplines.
- Development of AI systems that exhibit dynamic learning and adaptability to evolving environments.
- Measurement of trust in AI systems, including calibrated metrics for trustworthiness in human-machine teams, and understanding how to instill user trust.
- Active learning methods and adaptive systems that can intelligently sample observations to maximize information gain.

Cultivating user trust remains critical for effective design and deployment of AI-enabled systems, particularly as human-AI teaming continues to increase. Key to engendering user trust is model explainability and interpretability, and the 2023 workshop advocated for comprehensible and transparent calibrated metrics to gauge trust in AI systems, encompassing noted factors such as explainability, transparency, and the capacity to adapt to dynamic conditions. Ensuring the quality of training data and emphasizing the importance of evaluating biases, reliability, and representativeness also averts adverse consequences in AI applications that undermine trust.

A human-centered design ethos needs to be adopted, with a focus on human cognition, decision-making processes, and the psychological dynamics of AI interactions. Additionally, ethical considerations remain paramount, especially in sensitive domains such as warfare, requiring the adoption of responsible AI practices, the development of ethical frameworks, adherence to principles, and ongoing evaluation of AI's societal impact.

Continuous learning, real-time adaptation, and adept decision-making in uncertain environments emerged as critical for the success of AI applications. The emphasis on active learning and adaptive systems underscored the importance of intelligently sampling observations to optimize information gain, particularly in contexts characterized by prevalent uncertainty.

****

The AI4SE & SE4AI Research Workshop has grown in attendance and scope in its four years. Each year, the workshop has allowed an exchange on progress, challenges, and goals and has acknowledged the importance of confidence and trust in new technologies and systems. Safety remains a primary concern, with the greatest risk passed on to the warfighter. Workforce development and the idea that people make progress possible have received continued emphasis to ensure individuals understand their role within the larger, interconnected digital ecosystem focused on delivering capabilities to the warfighter at the speed of relevance.

Within the mix of industry, academia, and government represented at the annual workshops is where the answers and solutions can be developed. The workshop organizers and participants look forward to a fifth conference in 2024 and continued guidance on evolving efficiently and effectively into the future.

# TABLE OF CONTENTS

## INTRODUCTION

This workshop was the fourth Artificial Intelligence for Systems Engineering & Systems Engineering for Artificial Intelligence (AI4SE & SE4AI) Research Workshop, jointly sponsored by the US 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 event was held in-person at The George Washington University. The conference theme, "Balancing Opportunity and Risk: The Systems Engineer's Role in the Rapid Advancement of AI-Based Systems," aimed to foster discussions and insights on the responsible deployment of artificial intelligence (AI) in systems engineering (SE) and how SE can support the development of robust and ethical AI systems.

## WORKSHOP AGENDA STRUCTURE AND AUDIENCE

Ms. Jennifer Swanson (Deputy Assistant Secretary of the Army for Data, Engineering & Software (DASA(DES)), ASA(ALT) served as morning keynote speaker for Day 1 of the two-day event. Dr. Kimberly Sablon (Principal Director, Trusted Artificial Intelligence and Autonomy, OUSD(R&E)) served as afternoon keynote speaker for Day 1. Mr. Michael "Rabbi" Harasimowicz (Director, AI Innovations, Lockheed Martin) served as keynote speaker for Day 2. Each speaker provided relevant perspectives for their designated day's sessions. The workshop was attended by multiple Government agencies, industry and academia affiliates, amounting to almost 200 people present. The workshop agenda was structured into the following four tracks:

- Trustworthy AI, which highlighted the critical aspects of safety, reliability, and ethical considerations in developing and deploying AI systems
- Artificial Intelligence for Systems Engineering (AI4SE), which delved into the application of AI in SE processes, enabling enhanced decision-making, optimization, validation, and verification
- Human/AI Teaming, which examined the collaboration between humans and AI, exploring how to maximize the synergistic potential while addressing ethical and social implications
- Systems Engineering for Artificial Intelligence (SE4AI), which focused on leveraging SE principles and methodologies to develop robust and efficient AI systems

Each track had six presentations on relevant topics, and audience members were able to collaborate and ask questions throughout the briefings. Each session was moderated with an interactive discussion and Q&A at the end.

Presentation materials for the entire workshop are available via the SERC website:
https://sercuarc.org/event/ai4se-se4ai-workshop-2023/

## WELCOME & OPENING REMARKS

**DAY 1 -** Dr. John Lach, Dean, *School of Engineering and Applied Science, The George Washington University*
Dr. Jason Cook, *Senior Scientific Technical Manager (SSTM) for Systems Engineering Research, US Army DEVCOM Armaments Center*

In his introduction, Mr. Lach discussed and identified the approach to engineering at The George Washington University (GWU) as "engineering and…," which highlights that many attend GWU not only for engineering but to explore its connections to other fields (e.g., engineering and healthcare, engineering and…, etc.). It was noted that as SE focuses on the interfaces between different disciplines and their designs, the interdisciplinary approach of GWU engineering resonates with SE professionals.

## WORKSHOP KEYNOTES

**DAY 1 - MORNING KEYNOTE | AI @ Speed & Scale: Evolving with AI and within the Army's Digital Transformation**
Ms. Jennifer Swanson, *Deputy Assistant Secretary of the Army for Data, Engineering & Software (DASA(DES)), ASA(ALT)*

Ms. Swanson discussed the ASA(ALT) mission to modernize the Army as part of the Joint Force through rapid and timely delivery of soldier capabilities that deter adversaries and win our nation's wars. In addition, the DASA(DES) builds pathways for digital transformation so programs can deliver overmatch capabilities. The DASA(DES) has embarked on a digital odyssey, ensuring scaling and maturing of AI and ML. The digital odyssey acknowledges that true transformation requires people, process, and systems and is underpinned by a comprehensive set of initiatives focused on elevating ASA(ALT) employees and transforming solutions and accelerating tools.

Ms. Swanson noted data is foundational for AI, and the data ecosystem at ASA(ALT) is founded on the data mesh concept, explained through the relationship among data product, producer, and consumer: the decision makers drive the demand for data products, which are produced by data domain experts and answer the commander's questions. In discussing testing, she mentioned there is no plan to force vendors into a specific testing pipeline.

**DAY 1 - MORNING | US Army DEVCOM Armaments Center Perspectives**
Myron Hohil

Mr. Hohil gave perspectives from his experience within the Army and DEVCOM AC on AI use and the directives given to the Army by DoD. The Department's AI Strategy directs DoD to accelerate adoption of AI and the creation of a force fit for the current environment. Strategic focus areas and implementation goals include evolving partnerships with industry, academia, allies, and partners, establishing a common foundation that enables decentralized execution and experimentation, and leading in military AI ethics and safety.

SE methods and subject matter experts (SMEs) can provide a robust analytic framework with supporting resources and tools and help to address three critical questions: 1. What SE activities and artifacts are best suited to build trust in AI-enabled systems; 2. What infrastructure (data, models, computational resources, tools, test

beds) are needed to train and validate trusted AI-enabled systems; and 3. What are the key workforce skills and abilities required for an integrated product team (IPT) to be successful in the development and management of AI-enabled systems?

**DAY 1 - AFTERNOON KEYNOTE | Secure, Robust and Scalable AI/Autonomy – A Holistic, System of Systems Approach to Development of Trustworthy AI-enabled Systems**
Dr. Kimberly Sablon, *Principal Director, Trusted Artificial Intelligence and Autonomy, Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E))*

Dr. Sablon's presentation addressed AI from a multi-scale design perspective that extends beyond designing operational requirements and deployment logistics to addressing the distributed architecture of AI systems. Key facets of this extended perspective include: decentralization and a robust data strategy to support decision-making; testing that considers responsible AI strategy, ethical, and legal constraints, as well as interoperability and interpretability; verification that includes the warfighter's perspective and collaboration between mission engineers and R&D; a Center for Calibrated Trust Measurement and Evaluation (CaTE) that emphasizes ethics and security by design, warfighter-in-the-loop experimentation, and evaluation; and the development of calibrated metrics frameworks and a common infrastructure to foster collaboration across DoD Science & Technology (S&T).

The warfighter remains at the center, both in terms of technological advancements and international alliances. As the focus shifts to autonomous systems, the roadmap includes cross-echelon, resilient autonomous networks, intelligent and collaborative robotics, human-machine teams, and advancements in navigation and mobility. Counter-autonomy efforts also highlight the need for proactive measures, including machine-generated deception, to stay ahead in the current evolving landscape.

**DAY 2 - KEYNOTE | Lockheed Martin Transformation: You May Not Recognize Us**
Mr. Michael "Rabbi" Harasimowicz, *Director, AI Innovations, Lockheed Martin*

Mr. Harasimowicz examined Lockheed Martin's approach to digital transformation, underscoring the collaborative, innovative, and adaptive approach required for successful integration of AI in SE and military operations. He discussed government partnerships with Microsoft and NVIDIA as indicative of an emphasis on innovation and speed and on an approach to transformation focused on mission-driven transformation, platforms, Systems of Systems (SoS), and agility. He also discussed a principle he described as "Do No Harm," which aims to apply AI to leverage insights from "throwaway" data. Throughout, ensuring trust in advanced systems was noted as essential for use in dynamic environments and modern warfare.

# WORKSHOP PANELS

## *DAY 1 - PANELS*

**Certification of Learning-Based Systems**
Moderator: Dr. Zoe Szajnfarber, *The George Washington University*
Panelists: Dr. Jonathan Barkand, *Program Manager for Learning Systems, DAU*; Dr. Tyler Cody, *Virginia Tech*; MAJ Christine Krueger, *The George Washington University*; Dr. Jason Summers, *ARiA*

The theme of qualification for AI systems and the human operators interacting with them was explored. Highlighted in the discussion were crew resource management, the ability of AI systems to recognize their limitations and seek assistance, and the need for transparency in AI models to build trust. Also addressed were the difficulties in certifying AI systems that continually adapt their code, raising the question of how to perform model selection. The practical challenges of modular design and integration testing were discussed, with a focus on the need for regression testing and user acceptance testing when adding AI modules to existing systems.

Dr. Cody addressed three key takeaways from his research on certification of learning-based systems: 1. properties on the system; 2. the system is more than just the software formatting; and 3. the lifecycle and programs that develop are important. It was noted there are different notions of risks when it comes to adequacy in a learning-based system. There need to be risk mitigation strategies for test inadequacy that vary over the system lifecycle, to ensure that learning can continue and to detect degradation. The general intelligence is already present, not as an auxiliary data analysis but as a concern of system function.

Dr. Summers discussed three examples of actual processes certification and how it could be utilized in learning-based systems: 1. NAVSEA was an example of a 25+ year old software process designed to operate on ad-hoc ML principles; 2. another example illustrated a data and certification process based on the principle “train on all, test on all,” poorly executed with government-defined metrics and without the needed data gathering parameters; and 3. DEVCOM’s SE vision for how to accredit AI that considers tools and artifacts that can help drive and shape the development process.

Dr. Barkand focused on the intersection of AI, federal regulations, and organizational learning systems that requires creative solutions and a nuanced approach. Privacy and data security are paramount concerns as DAU implements AI in federal spaces with varying security levels and a lack of guidelines for AI algorithms, data pipelines, and data transfers. DAU is focused on segregating and training models effectively. Certification for AI systems within the federal space is a challenge as many commercial products cannot be implemented due to certification requirements for federal environments.

MAJ Krueger provided perspective on the difference between certification versus qualification through the lens of his experience as a Blackhawk pilot. Certification is presented as whether or not an aircraft is airworthy, whereas qualification aims to challenge pilots and ensure they are ready. MAJ Krueger addressed AI as a tool, teammate, or a hybrid concept, and noted the need for a hybrid of certification and qualification for learning-based systems for crew resource management.

**Application of LLMs to SE**
Moderator: Dr. Myron Hohil, *US Army DEVCOM Armaments Center*
Panelists: Mr. Amir Abrari, *SPEC Innovations*; Mr. John Crissman, *CNA Corporation*; Dr. Carlo Lipizzi, *Stevens Institute of Technology*

The panel was divided between a broad overview and thoughts on application in SE contexts. The overview described large language models' (LLMs') high level of sophistication and the correlation to intrinsic bias induced by the data. This understanding is important for evaluation and interpretation of results. Regarding application, both Mr. Abrari and Mr. Crissman use LLMs as assistants to perform SE tasks. Mr. Abrari focuses his work on an Innoslate Cloud solution developed by SPEC Innovations to query ChatGPT to produce data relevant to system design. Mr. Crissman uses LLMs and Retrieval Augmented Generation (RAG) to produce functional and non-function requirements and relevant test cases.

### *DAY 2 - KEYNOTE PANEL*

**Comparing Ethical and Safety Frameworks for Promoting Trust in AI Systems**
Moderator: Mr. Andy Lacher, *Chief Technologist for Future Airspace Operations, NASA Langley Research Center*
Panelists: Dr. David Broniatowski, *Associate Professor of Engineering Management and Systems Engineering, The George Washington University*; Mr. Chuck Howell, *Senior AI Advisor, MITRE*; Dr. Kim Wasson, *Autonomy Certification Lead, Joby Aviation*

Dr. Szajnfarber filled in for a missing panelist. In her discussion of trust, she made the point that "we only care about 'trust' when we can't 'know.'" It was noted that focus needs to be given to how risk assessments need to change with the use of AI and the potential consequences. In discussing ethics, it was noted that ethics aims to align different parties on values, which can then be applied in specific contexts and applications.

# WORKSHOP PRESENTATIONS – DAY 1

## *TRACK 1: Trustworthy AI*

**Establishing a Measurement Science for Joint Work with Models**
*National Institute of Standards and Technology*

The presentation explored the challenge of engaging persons with minimal background in model formulation. The concept of declarative understanding—focusing on what is wanted rather than how to achieve it—was emphasized. The role of machine agents and the use of domain-specific language (DSL) were discussed, emphasizing the importance of human knowledge in problem domains and human willingness to learn DSL during conversations with machine agents. Illustrative examples of machine agents were presented, and the decision-making process of machine agents was explored.

Highlighted throughout the presentation were the aim of achieving best practices in joint formulation, the importance of communication, and refinement in model formulation, noting its role in enhancing understanding of causality and computational methods. The presentation concluded by acknowledging model formulation as a foundational element in the modern multidisciplinary world.

**Application of Systems Theoretic Process Analysis (STPA) to the Mission Assurance of AI-Enabled Systems**
*The Aerospace Corporation*

The application of System-Theoretic Process Analysis (STPA) can enhance the mission assurance of AI-enabled systems, specifically within the domain of SE for AI. STPA identifies various types of potential issues within a system, enhances system requirements, and mitigates the occurrence of system failures. Practical and illustrative examples were presented, including the application of STPA to Natural Language Processing (NLP) and computer vision image processing systems, which involved defining enterprise risks. An overall control structure for the NLP system was outlined, involving tools that define control actions and provide feedback for model tuning and training.

Mission Assurance Guidance and the Trusted AI Framework were explored, including trusted sources, AI framework threads, and fault and redundancy management within established guidelines. An illustrative example involving a conceptual image processing system for space domain awareness was presented. Early conclusions of the work presented suggest using Model-Based Systems Engineering (MBSE) tools to maintain the STPA database during concept development, thereby informing system architecture. Next steps involve completing internal reviews and engaging with internal customers in the coming fall. An STPA Handbook is available through the MIT website.

**Application of the Traditional Materiel Release Process to the DoD Ethical Principles of Artificial Intelligence and Roadmap**
*US Army DEVCOM Armaments Center*

The presentation focused on materiel release, specifically the release of equipment, and how this framework encompasses the journey to trusted and assured AI/ML systems. The roadmap for AI/ML development was

delineated, spanning conceptualization, design, optimization, and verification stages, as well as a data sets category that emphasizes data assurance. The significance of data science in the context of AI/ML emphasizes the need to identify and evaluate datasets for risk and readiness. Human Systems Integration (HSI) was identified as pivotal in developing the appropriate mental models and optimizing interfaces to convey the right information needed for trusted and assured systems.

The concept of "relatability" was discussed, focusing on identifying potential failure modes and ensuring that AI-enabled systems are integrated safely into larger systems. Risk reduction from various perspectives was addressed, and the budget implications of these initiatives were acknowledged. The assurance case for the DoD Ethical Principles of AI was discussed, acknowledging the evolving nature of ethical principles and advocating for responsible AI. The Department's ethical principles—Responsible, Equitable, Traceable, Reliable, Governable—were highlighted, with a call to apply existing processes to demonstrate adherence.

Next steps and future work involve continuing execution, socializing the roadmap with the community, encouraging continuous feedback from stakeholders, and assessing alignment with Responsible AI Strategy elements. The importance of reliability in design to mitigate failures and identifying off-nominal characteristics was emphasized.

**Using Artificial Intelligence to Accelerate Deployment of Learning-Based Systems**
*US DEVCOM Army Research Laboratory, University of Southern California*

The presentation outlined a vision to expedite testing and deployment processes, using and leveraging AI to augment decision-making capabilities and framing experimentation within the context of information economics. More informative testing aligns with the overarching goal of building trust and explainability as experiential processes, as well as with the broader initiative of the Army Autonomous Enterprise, where experimentation is deemed essential. Sequential decision-making in experimentation leverages human strengths such as prioritization, mission consistency, and context-dependent reasoning. However, human shortcomings, including uncertainty estimation, big-data reasoning, and unbiased decision-making, have prompted efforts to introduce a Decision Support System (DSS) that uses observations to inform and support human decision-makers across different stages.

The iterative cycle of encountering challenges, learning from them, and adapting configurations is essential for continuous improvement of learning-based systems, and there is recognition that a DSS would serve as a valuable database. This approach ensures that the knowledge gained from failures and successes alike is systematically captured, enabling informed decisions, refined configurations, and sustained excellence. Overall, a greater understanding of the system will contribute to heightened trustworthiness and explainability, both crucial elements for successful deployment of learning-based ground systems in complex and context-rich environments.

**Conformal Prediction for Testing and Evaluation of Intelligent/ML Systems**
*Georgia Tech Research Institute*

The application of conformal prediction within operations or supply chain teams involves strategic planning to move beyond a binary decision and introduces the concept of probability. Traditional evaluation metrics focus

on characterizing performance unconditionally, raising the need for a nuanced approach that considers regions where performance is robust versus areas of uncertainty.

Conformal prediction hinges on a pre-specified parameter alpha, offering a methodology with minimal assumptions about the dataset. The two method branches, full and split conformal, present different trade-offs. The full conformal uses every observation and can be computationally intensive. In contrast, split conformal involves retraining the model on a calibration set, making it more aligned with the test-and-evaluate model paradigm. Overall, conformal prediction provides a guaranteed form of uncertainty quantification that applies to marginal coverage, making it a valuable tool for validating models, especially in complex and dynamic scenarios. This method is suitable for an algorithm that may be more hypercomplex and not necessarily used for the specific model being reviewed.

**Evaluating the Explainability and Interpretability of AI Systems**
*The George Washington University, Cornell University*

Explainability and interpretability are distinct yet interconnected measures that must be meticulously assessed to foster trust in AI. The psychometric evaluation of explainability and interpretability plays a pivotal role in influencing human judgment, as encapsulated in the formalization and application of Fuzzy Trace Theory (FTT). This leading theoretical framework delineates between gist and verbatim representations, illustrating how humans, in their decision-making processes, often rely on the gist, or the core essence of information. Model output metrics are crucial in considering interoperability and observations of human decision-making following the receipt of model output underscore the significance of gist endorsement.

## *TRACK 2: AI4SE*

**AI-Assisted Generation of SysML Models to Accelerate Agile Digital Engineering**
*Leidos*

Work conducted at Leidos aims to compare models produced by human modelers to models produced by ChatGPT. A small case study was presented that looks at associates with less SE experience and SE Subject Matter Experts (SMEs) and each group's varying level of access to ChatGPT: no access, a set of documents created with ChatGPT to use as a starting point, and open access to ChatGPT with proposed prompts to get started. The case study found that open access to ChatGPT provided the highest qualitative and quantitative scores for both associates and SMEs. GPT used on its own scored last but still produced 40% of the requested content in a fraction of the time.

**Requirement Discovery Using Embedded Knowledge Graph with ChatGPT**
*NASA Langley Research Center, Collins Aerospace*

NASA and Collins Aerospace presented two different approaches to using LLMs for requirement discovery. The work at Collins explores the use of LLMs with Retrieval-Augmented Generation (RAG) to produce results while NASA uses LLMs in conjunction with a graph database to produce results. A noted use of the LLMs/graph database approach uses ChatGPT to produce syntactically valid queries to the Neo4J graph database. The pure LLM approach performed better on open queries but did not do well with counting-based queries. It was also

found that the LLM/graph database approach reduced the number of hallucinations created by the LLMs by connecting these to a ground truth graph database.

**LLM-Based SysML Virtual Assistant**
*Texas A&M University*

Presented was research related to using LLMs within a Cameo plugin to produce SE artifacts, specifically a custom parser that used carefully prompted LLM output to create Systems Modeling Language (SysML) artifacts within Cameo. The effort demonstrated the capacity to use LLMs directly within a SysML authoring tool and to use prompt engineering to create predictable outputs that can be interpreted by a tool.

**Melding Minds and Machines: Harnessing AI Capabilities to Further Model-Based Guided Engineering**
*Booz Allen Hamilton*

Presented was Booz Allen Hamilton's Model-Based Systems Engineering (MBSE) AI Plugin Prototype (MAPPy), which uses LLMs to produce requirements and then validate those requirements based on INCOSE standards. The plugin, produced in Cameo, used ChatGPT to develop model components. Additionally, the team used MBSE to guide development of the plugin and saw significant advantages in using a digital model for the development process, including a centralized model for reviewing technical baseline information such as use cases and requirements and the ability to track burn down of feature implementation across the team.

**A Constructed System of Analysis to Enable Automation and Reasoning in Multi-Model Analysis**
*Stevens Institute of Technology*

Presented was the use of symbolic logic-based AI to provide model verification of a Digital Thread. The Digital Engineering Framework for Integration and Interoperability was used to align analysis threads with ontology-aligned data. This form of knowledge representation showed how Description Logic reasoning, closed-world constraint analysis, and more general graph-based algorithms can be aggregated to provide rich verification capabilities.

**Large Language Model Enabled Generation of Systems Engineering Artifacts**
*Virginia Tech*

Presented was research that used LLMs to generate acquisition document elements, specifically the Bulldog dataset was used as ground truth and compared multiple LLMs (GPT 4, GPT 3.5, and Claude). The MAUVE framework was used for comparing generated results to ground truth using an iterative prompt development approach. It was found that more targeted prompts produced better results, but the LLMs struggled with some tasks, including defining Development Thresholds, Objectives, and numeric operations. It was concluded that LLMs can be useful when used with an expert in the loop.

## WORKSHOP PRESENTATIONS – DAY 2

### *TRACK 3: Humans/AI Teaming*

**Is the Machine a Partner or a Tool? A Major Issue of Human-AI Teaming**
*Paris Saclay University (CentraleSupélec) & ESTIA Institute of Technology*

The FlexTech initiative presented encapsulates a holistic methodology that seeks to balance the complexities of technological integration with the imperative for human-centric and adaptable design. It is distinguished from other methodologies by its focus on simplifying complicated processes, thereby rendering complex problems understandable and practically applicable. Technology is introduced into engineering at later lifecycle stages, in contrast to the preferred approach of human-centered design, where system knowledge increases rapidly, design flexibility expands, and resource commitments follow an inverted curve. Within this framework, readiness levels are assessed, ranging from Technology Readiness Levels (TRL) to Human Readiness Levels (HRL) to Operational Readiness Levels (ORL). The challenge lies in allocating the necessary time to implement this approach effectively.

**What Is Human in the Loop, Really?**
*The George Washington University*

Focusing on Human-in-the-Loop (HITL) architectures, the presentation underscored the synergy between humans and autonomous systems and the conviction that this collaborative approach yields superior results and fosters trust. The journey through data, training, construction, and operations prompts a critical exploration of how humans interact with autonomous systems. The challenges inherent in providing feedback were addressed through an example drawn from aviation: pilots were presented with AI solutions that provided options and the flexibility to not adhere to automated suggestions, with the goal of reducing stress. The example demonstrated that effective feedback mechanisms, especially in safety-critical systems, demand a nuanced approach, as poorly delivered feedback may compromise the integrity of complex systems.

**Managing the Continuity of Human Biases in AI SE Applications**
*Colorado State University, MIT Lincoln Laboratory*

The presentation emphasized benchmarking against performance metrics in assessing AI systems, steering away from reliance on the availability heuristic. The advice for creators of LLMs and similar systems is to prioritize specifying individual characteristics and values, as incorporating these aspects at the forefront allows creators to effectively guide the system's responses.

One highlighted perspective explored the triad of models within a data-driven context, an approach that involves situational alignment and in which the desired outcome prompts the system to locate the most fitting model. It illustrated the evolving landscape where users are not merely searching for information but actively seeking models tailored to their specific needs.

**Opportunities and Risks of Incorporating LLMs in the Systems Engineering and Design Workflow: A Case Study of Robotic System Design Process**

*The George Washington University*

The presentation focused on the collaborative dynamic between Human-AI Interaction (HAI) teams and individual engineers and the design decisions inherent in the problem-solving process that necessitate a breakdown of steps, subsystems, and interactions. The challenges and potentials of incorporating AI into the design realm prompt considerations about historical data maintenance and the need for a balance between change and continuity in the creative process.

Comprehensive observation, involving screen recording and transcription, is needed to understand the intricacy of interactions between user specifications and AI-generated solutions. ChatGPT's capacity to independently generate tailored solutions highlights both its strengths and limitations. Opportunities for generating alternatives at different design levels were identified, along with concerns about tracking hard constraints and maintaining a coherent design sequence.

**The Need to Update Design and Certification Performance Standards for Operator Intervention of Autonomous Systems**
*George Mason University*

The presentation emphasized that near-autonomous systems rely on human supervision, particularly in instances of functional delegation where human operators monitor and intervene when necessary. The concept of a trigger window for takeover by a human operator was introduced. The window represents the time available for a human operator to intervene and steer the system back to a safe operating state. The work presented further explores the Time on Procedure (ToP), a metric encompassing detection time, reaction time, and machine response time. It was noted that these processes are not instantaneous and require a comprehensive evaluation.

The presentation mentioned regulatory standards, pointing out that CFR 14 and CFR 25, while addressing various aspects of aviation safety, do not explicitly specify human response times. The NPRM process, a four-year mechanism through which the FAA proposes and implements new processes, reflects the regulatory intricacies that govern changing and updating of standards.

**Mission Engineering in Healthcare and What We Can Apply to the Military**
*Missouri University of Science and Technology*

Parallels between healthcare and the military were drawn to highlight meta-architecture as a crucial framework to structure relationships within complex systems. Meta-architecture strikes a balance between the autonomy of subsystems and the integration of the entire complex system, fostering purposeful design, self-organization, and addressing tensions between stability and change through emergence. An Organ Procurement Organization (OPO) was presented as an illustrative example. The OPO managed a diverse array of features, including blood types and creatinine levels. The emphasis on increasing utility while not disregarding equity aligned with the broader goal of providing equal access to life-saving measures for everyone. Tools such as AnyLogic, pipeline, and Python contributed to the performance evaluation process, including discard rates. The analogy extended to soldiers equipped with sensors on the battlefield, where a meta-architecture was not immediately available but held promise for future development.

Recycling LLMs and transfer learning were discussed, emphasizing the need to scrutinize the feedback cycle in

this context. This holistic view underscores the interconnectedness of systems within the healthcare and military domains and the importance of meta-architectures in navigating complexity.

## *TRACK 4: SE4AI*

**Hiring Trained Animals: Generative AI Patterns and Practices for Systems Engineering**
*Collins Aerospace*

Presented were best practices and foundational information on how to approach LLMs in the SE setting. These included basic LLM operation, different LLMs available, prompt engineering, and formatted output. Discussed was use of LLMs as components in systems or applications and an example was provided of using ChatGPT to prepare panelists for a panel on sustainability in systems.

**Systems' Perspectives for AI Acceleration in Large Enterprises and Government Agencies**
*MITRE*

Presented was the use of the SE lifecycle to accelerate AI. Considered were business needs and adoption hurdles and how SE approaches can be used to address these. A Continuous Integration/Continuous Deployment (CI/CD) pipeline for addressing testing of AI-based systems in an Agile environment was presented and showed how a plugin developed for Cameo can be used to produce executable definitions of CI/CD pipelines using SysML Activity Diagrams.

**T&E of Complex AI Use Cases**
*MITRE*

Research was presented on Testing and Evaluation (T&E) of complex AI using a forward area language converter that uses natural language processing (NLP) and translation as a sample use case. It was noted that errors in inputs and/or outputs from AI components can cascade and that higher performance of single AI components may not lead to higher system performance. It was also noted that evaluation needs to be performed at multiple levels of abstraction (e.g., component, subsystem, system). Emphasized was the complex process of T&E that may require large data sets and error tolerances from subsystems interacting with AI components. The importance of evaluation from both operational and lifecycle perspectives was noted.

**Systems Engineering Processes to Test AI Right (SEPTAR)**
*MITRE*

Presented was MITRE work on the Systems Engineering Processes to Test AI Right (SEPTAR). Proposed is the standardization of AI deliverables and the concept of the "Data Card," an evaluation card to maintain continuity when personnel change. Discussion of training data when engaging in contracting was encouraged. It was emphasized that AI is ongoing, so there is a need to maintain a pipeline for T&E that can be returned to in the future. It was noted that T&E needs to go deeper when the AI used is opaquer; involving users in the AI development process allows T&E to be more tailored because there is greater understanding of the underlying technologies and approaches.

**A Systems Engineering Perspective on Safety of AI-based Systems**
*University of Alabama in Huntsville*

An extended example of a car crash involving Tesla's Autopilot driving functions was presented to explore safety in AI-based systems. Several risks associated with incorporating AI into systems were noted, as well as existing SE principles that address risk. It was proposed that new research is needed to address identified gaps between unique risks associated with AI and existing SE principles. The R3 concept was presented to focus on robustness, reliability, and resilience in relation to safety.

**Simulating the Tradespace with Synthetic Data**
*Clemson University, US Army DEVCOM Ground Vehicle Systems Center*

Research and work were presented on the creation and use of synthetic data to perform tradespace exploration. In working with sponsors, the research team found a limitation in the data related to tradespace exploration. A solution was developed to overcome this limitation by creating synthetic data within the tradespace and breaking down objectives into the underlying variables (which could include constants, binary variables, discrete values, and continuous variables). The third generation of the synthetic data will be open sourced and is expected to be available in 2024 for use in agent-based simulation of decision makers to provide assistance in viewing various trades.

# 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 yet again an opportunity to bring the community together to advance SE and AI.

**WORKSHOP ORGANIZERS**

**Executive Hosts:**

Dr. Dinesh Verma, *SERC Executive Director, Stevens Institute of Technology*
Dr. Jason Cook, *SSTM for SE Research, US Army DEVCOM Armaments Center*

**Workshop Leads:**

Mr. Tom McDermott, *SERC – Stevens Institute of Technology*
Ms. Kara Pepe, *SERC – Stevens Institute of Technology*
Mr. Albert Stanbury, *US Army DEVCOM Armaments Center Systems Engineering Directorate*

**Moderators:**

Dr. Zoe Szajnfarber, *The George Washington University*
Dr. Myron Hohil, *US Army DEVCOM Armaments Center*
Mr. Andy Lacher, *NASA Langley Research Center*

**ACRONYM LIST**

AI/ML – Artificial Intelligence/Machine Learning
CaTE – Calibrated Trust Measurement and Evaluation
CI/CD – Continuous Integration/Continuous Deployment
DoD – Department of Defense
DAU – Defense Acquisition University
DE – Digital Engineering
DEVCOM – Combat Capabilities Development Command
DNN – Deep Neural Network
DSL – domain-specific language
DSS – decision support system
FTT – Fuzzy Trace Theory
HAI – Human-AI Interaction
HITL – Human-in-the-Loop
HIS – Human Systems Integration
IPT – integrated product team
LLM – Large Language Model
MBSE – Model-Based Systems Engineering

ML – machine learning
NLP – Natural Language Processing
OUSD(R&E) – Office of the Under Secretary of Defense for Research and Engineering
RAG – Retrieval Augmented Generation
SE – Systems Engineering
SED – Systems Engineering Directorate
SEPTAR – Systems Engineering Processes to Test AI Right
SERC – Systems Engineering Research Center
SME – subject matter expert
SoS – System of Systems
SSTM – Senior Scientific Technical Manager
S&T – Science and Technology
STPA – Systems Theoretic Process Analysis
SysML – Systems Modeling Language
T&E – Testing and Evaluation
ToP – Time on Procedure
UARC – University Affiliated Research Center
V&V – Verification and Validation