SYSTEMS ENGINEERING RESEARCH CENTER

AIRC | ACQUISITION INNOVATION RESEARCH CENTER

TRUST AND TRUSTWORTHINESS IN AI-ENABLED SYSTEMS

Tom McDermott, SERC CTO, Stevens Institute of Technology
Dr. Zoe Szajnfarber, SERC Chief Scientist, George Washington University

FRAMING AI AND SE

Role for Systems Engineers in AI space

AI4SE and **SE4AI**

Focuses on **application of AI in support of systems engineering processes**, enabling enhanced decision-making, optimization, and efficient effort allocation.

Focuses on **leveraging systems engineering principles to develop AIES that are safe,** robust, and efficient AI systems, while extending them in response to the nature of AI enabled systems.

SE4AI applies to AI4SE too, but types of AI tools tend to be different

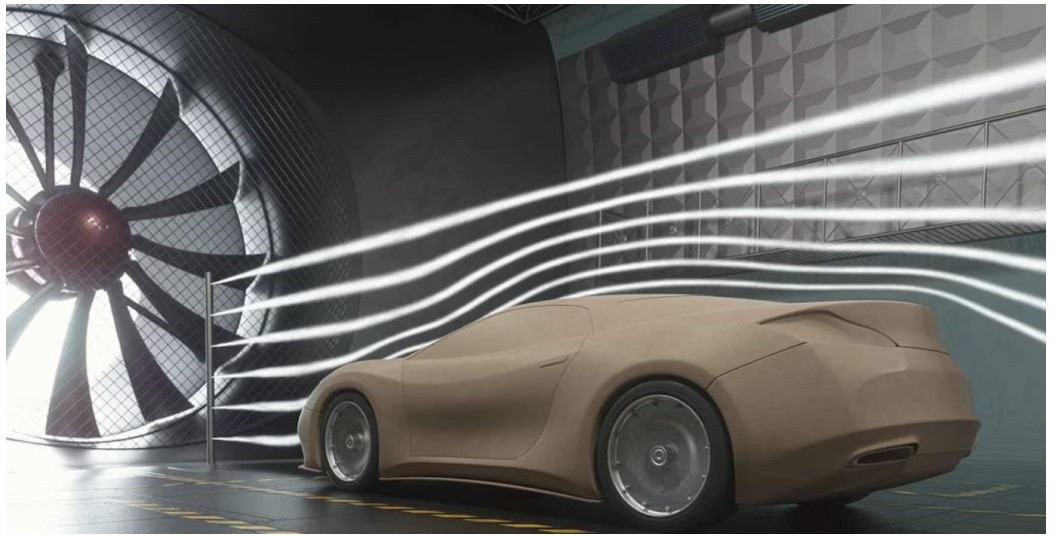
… and AI4SE might change what SEs do too.

Role for Systems Engineers in AI space

Requires trust → **AI4SE** and **SE4AI** → Ensure trustworthy AI

How should AI fit into the system engineer's workflow?

How can SE principles ensure trustworthy AIES?



TRUST AND TRUSTWORTHINESS DEFINITIONS

Do you?

Trust is by the user and is a property of the relationship.

"attitude that an agent (automation or another person) will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability."[1]

Trustworthiness is a property of the artifact.

"ability to meet stakeholders' expectations in a verifiable way; an attribute that can be applied to services, products, technology, data and information as well as to organizations."[2]

Should we?

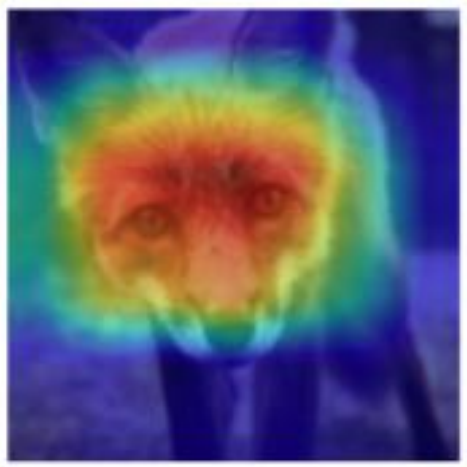
Trustworthy AI combines both concepts

emphasizing properties that generate "AI that can *[should?]* be trusted by humans"[3] Those properties typically include valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.[4]

[1]Cited in NIST RMF Glossary: John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **46**(1):50–80, 2004
[2]Cited in NIST RMF Glossary: ISO/IEC_TS_5723:2022(en)
[3]Cited in NIST RMF Glossary: Mark Coeckelberg (2020) " AI Ethics" MIT Press; [4] NIST RMF

DO YOU TRUST THE "AI" (IN THE SYSTEM)?

Developer



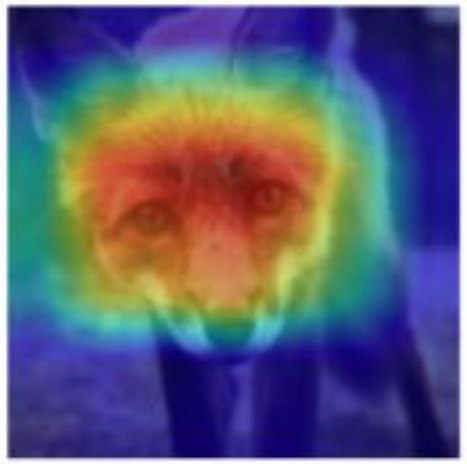
[1]

Accuracy:

If you're a Computer Scientist, you hate this phrasing and want to see the math of this specific algorithm or at least a visualization of the prediction.

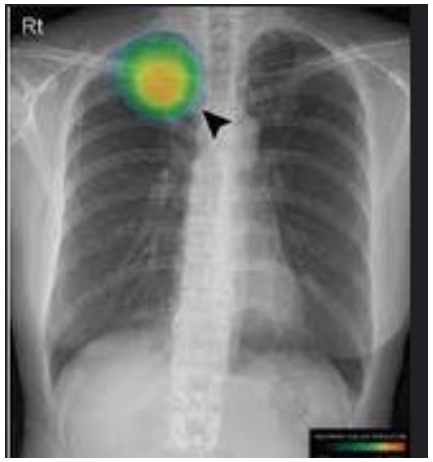
DO YOU TRUST THE "AI" (IN THE SYSTEM)?

Developer



[1]

Domain Expert



[2]

Accuracy:

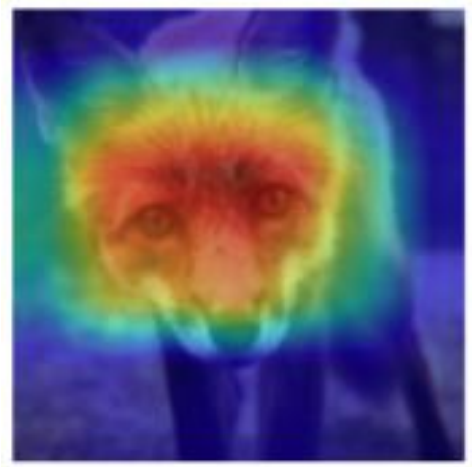
If you're a Computer Scientist, you hate this phrasing and want to see the math of this specific algorithm or at least a visualization of the prediction.

Agrees with me:

If you're a radiologist diagnosing pathology on an image, you might want to see the tool agree with you often enough.

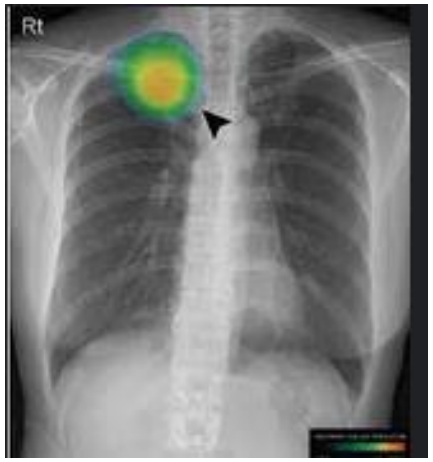
DO YOU TRUST THE "AI" (IN THE SYSTEM)?

Developer



[1]

Domain Expert



[2]

End User

U.S. Department of Transportation

[3]

Accuracy:

If you're a Computer Scientist, you hate this phrasing and want to see the math of this specific algorithm or at least a visualization of the prediction.

Agrees with me:

If you're a radiologist diagnosing pathology on an image, you might want to see the tool agree with you often enough.

Trusted 3rd Party:

If you're an AV passenger, you might want to be told that someone reputable certified it's safety… and not have heard of any fiery crashes lately!

ENGINEERING TRUST IS NOT A NEW CONCEPT

Situational awareness (SA)allows the pilot to evaluate the system behaviors in a larger context, to make a judgement about the final decision action. The SA is developed in training and real-world use.

Trust in the Domain

Trust the Developers

Transparency in the underlying algorithms and behaviors created by engaging the user in the development process and matured in critical reviews.

ROLE OF AI IN COMPLEX SYSTEM MATTERS FOR TRUST FORMATION

Replacing/augmenting existing task

[4]

Developer: Inspect algorithm

Domain Expert: Compare to what I would do

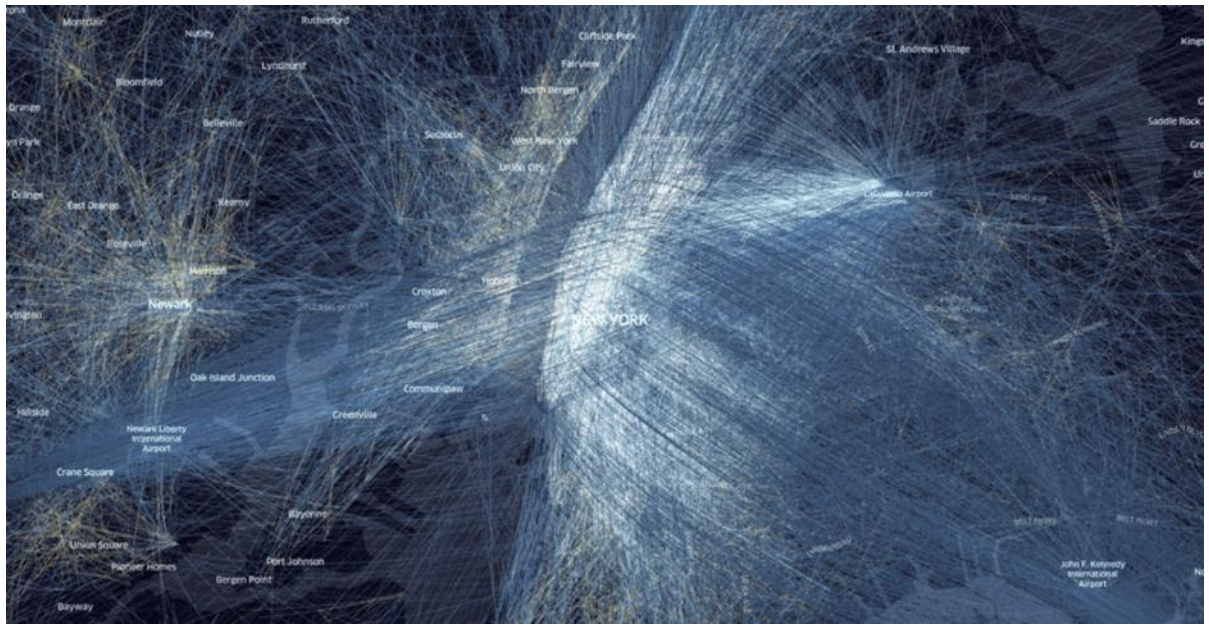
End User: Reputable source (logo/medallion)

ROLE OF AI IN COMPLEX SYSTEM MATTERS FOR TRUST FORMATION

Replacing/augmenting existing task

[4]

Solving new system level problem



What should the answer look like?

KEY TAI QUESTIONS FOR SYSTEMS ENGINEERING

How do the multiple H-AI, AI-AI etc. interactions impact how trust is built in modern complex all-domain systems?

(For AIES) What level of monitoring and re-engineering capacity is required post deployment?

- …and how does this interact with T&E? What does this mean for system resilience?
- What role will the human play in deciding on re-engineering

(For AIES) How will training need to shift left and be considered as part of system co-development?

Underlying theme: unit of analysis is the **socio-technical** system. Need for testing, training and research platforms that capture enough of the key SoAS interactions to represent **behavior.**

SERC ROLE IN DEFINING KEY AREAS OF INQUIRY

Regularly updating our Research Roadmaps

Holding workshops to gather input:

- Research Council Workshop (March 2023 at U of Arizona)
- Archimedes Partner Workshop (June 2024 at GWU)
- AI4SE/SE4AI Workshop (Sept 2024 at GMU)

Identified in humanities and social sciences

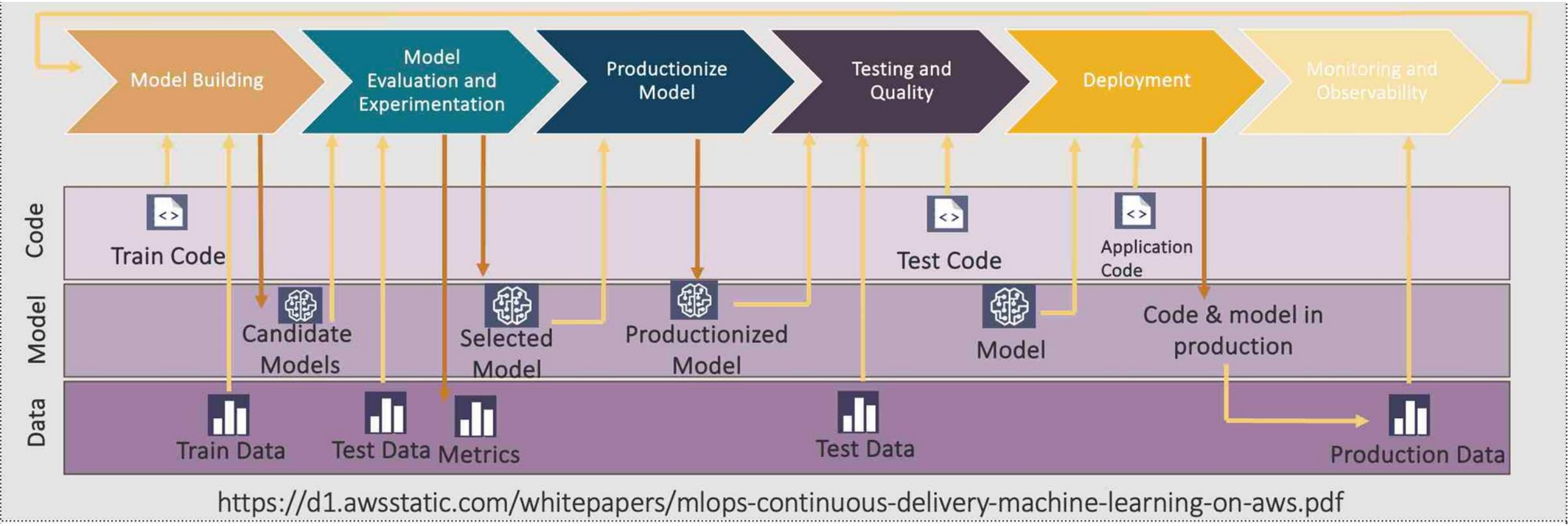
- **Abilities**: skills, competencies and characteristics of the system
  - Open question: How to implement this in systems engineering
- **Benevolence**: "good will" of trustee or believe in trustee that he will do good.
  - Objectifiable characteristics/metrics for "good will" are needed for systems engineering.
- **Integrity**: acting according to norms, standards and principles
  - Systems engineering: technical background on standards;
  - Social Background on standards needed

Axel Hahn, DLR-Institute of Systems Engineering for Future Mobility

AI-generated by GPT-4

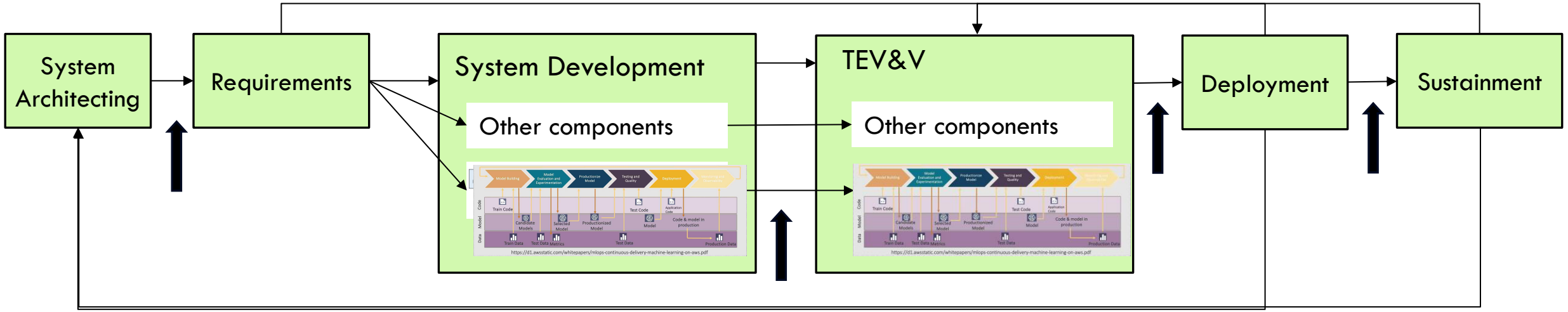
AI SYSTEMS ARCHITECTING: THE AI/ML PIPELINE TODAY

Typical representation of AI/ML pipeline:



Correct and without bias … but this is still focused on the AI model as the system.

FOR SYSTEMS ENGINEERS, AI IS PART OF A "SYSTEM"

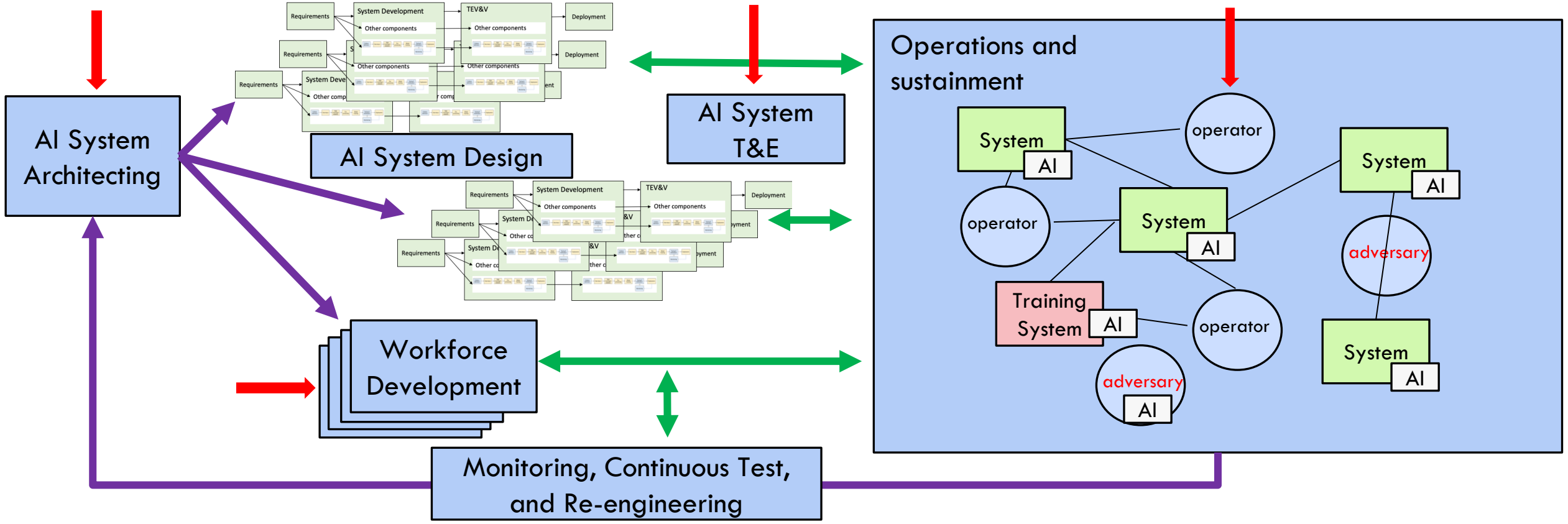


Emphasizes tradeoffs in performance and risk.

Recognizes that system might need to work in unplanned ways over its lifecycle and that behavior (and failures) must be acceptable.

THE REAL WORLD OPERATES IN A SOCIO-TECHNICAL SYSTEMS ARCHITECTURE

Involving complex interactions among humans and systems that were not always intended to work together in a constantly changing environment.



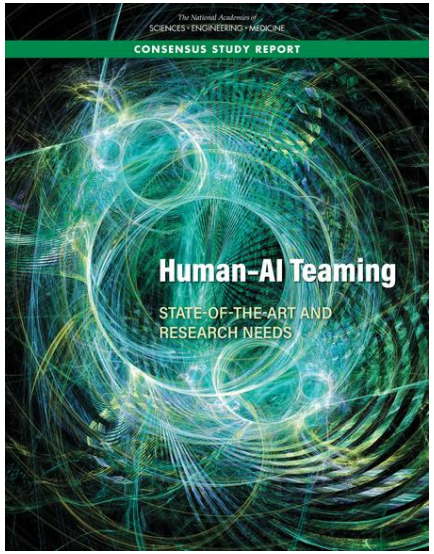
AI in system context; Building user trust; Architecting for long-term trust; T&E as a continuum

SYSTEMS ENGINEERING RESEARCH CENTER

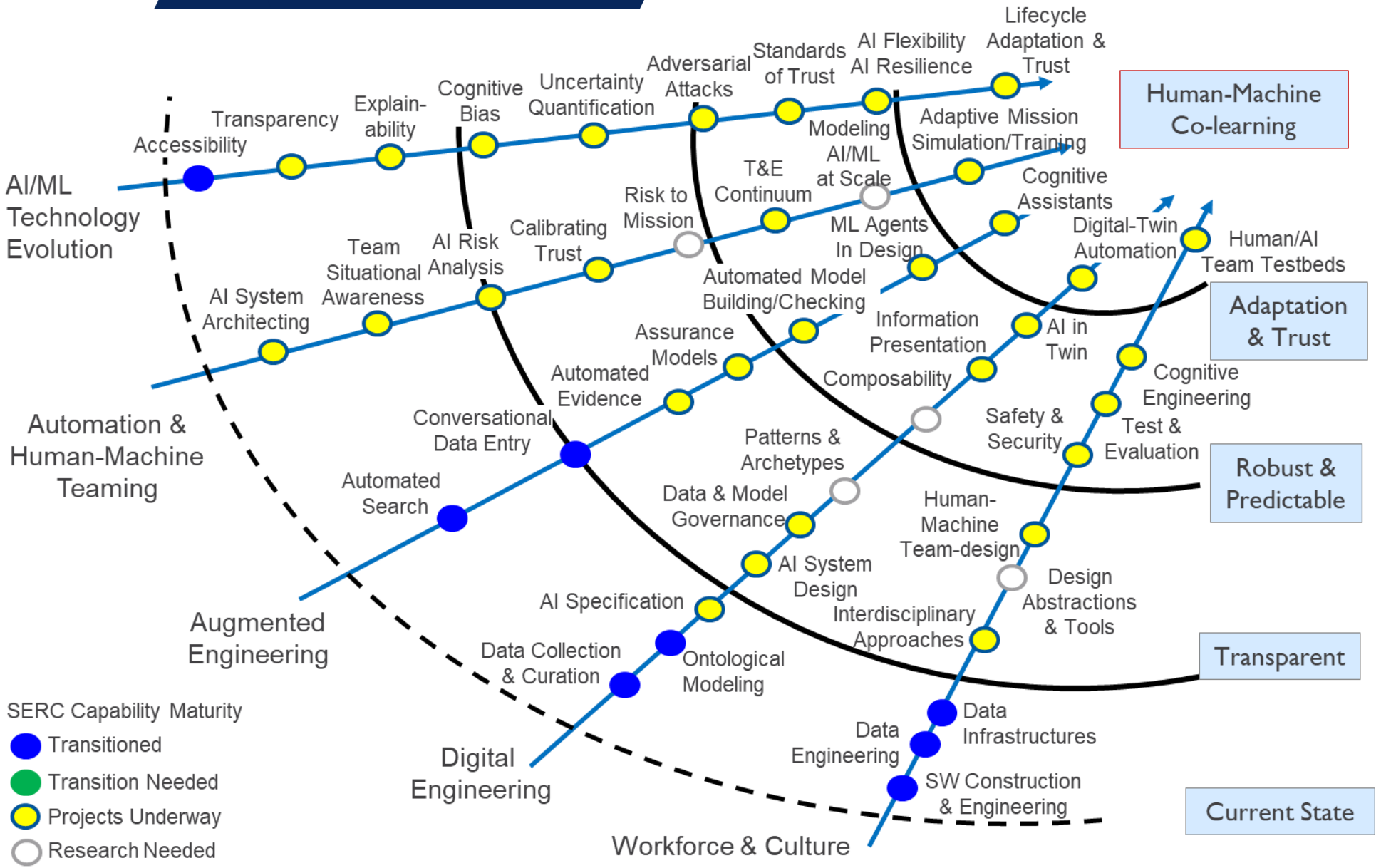
AIRC ACQUISITION INNOVATION RESEARCH CENTER

UPDATED 2023 ROADMAP

34th Annual INCOSE international symposium
hybrid event
Dublin, Ireland
July 2 - 6, 2024



SYSTEMS ENGINEERING RESEARCH CENTER



HUMAN-MACHINE CO-LEARNING

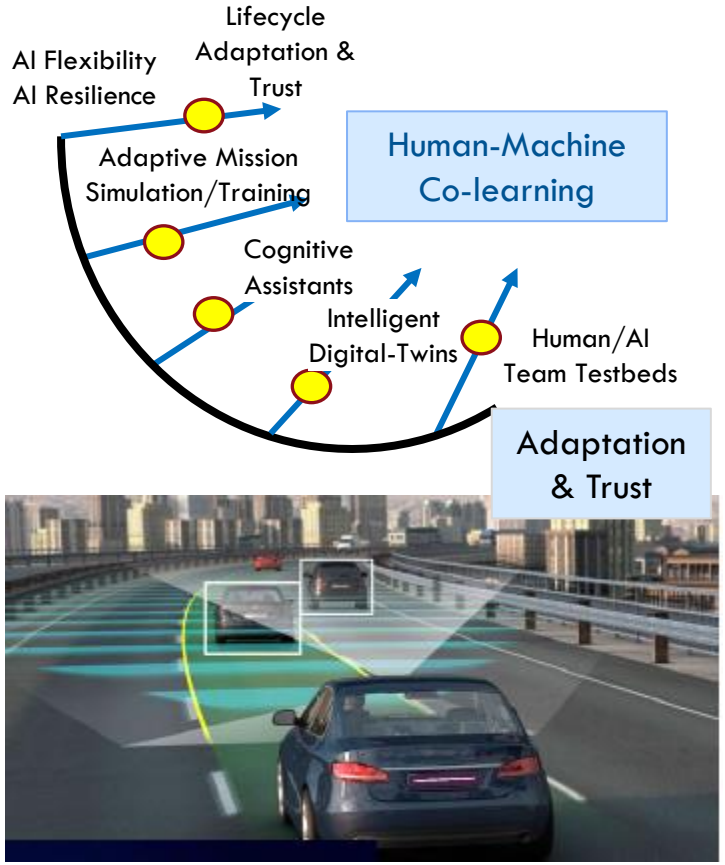
Adaptive Cyber-Physical-Human Systems – intelligent digital twins: modeling of cyber-physical systems as influenced by humans, in testbeds…

Adaptive Mission Simulation/Training –
Simulation and training that supports non-static objectives (pick-up games)

AI Flexibility & Resilience –
AI systems that self-adapt to changing operational boundaries while maintaining rigorous safety and security and policy constraints

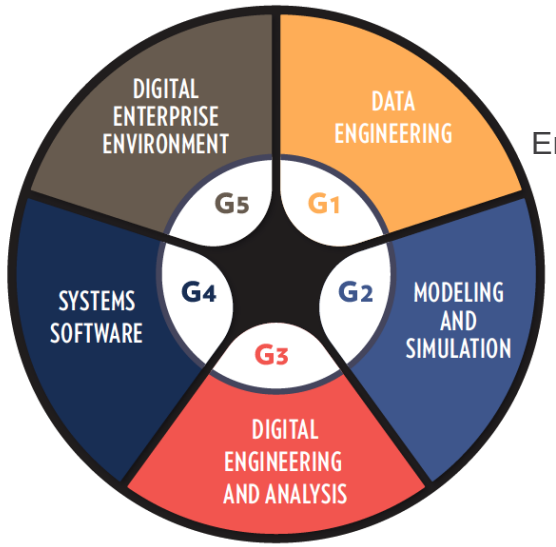
Lifecycle Adaptation and Trust –
Evolution of AI performance and related aspects of human trust over the lifecycle of a system as the system changes/evolves

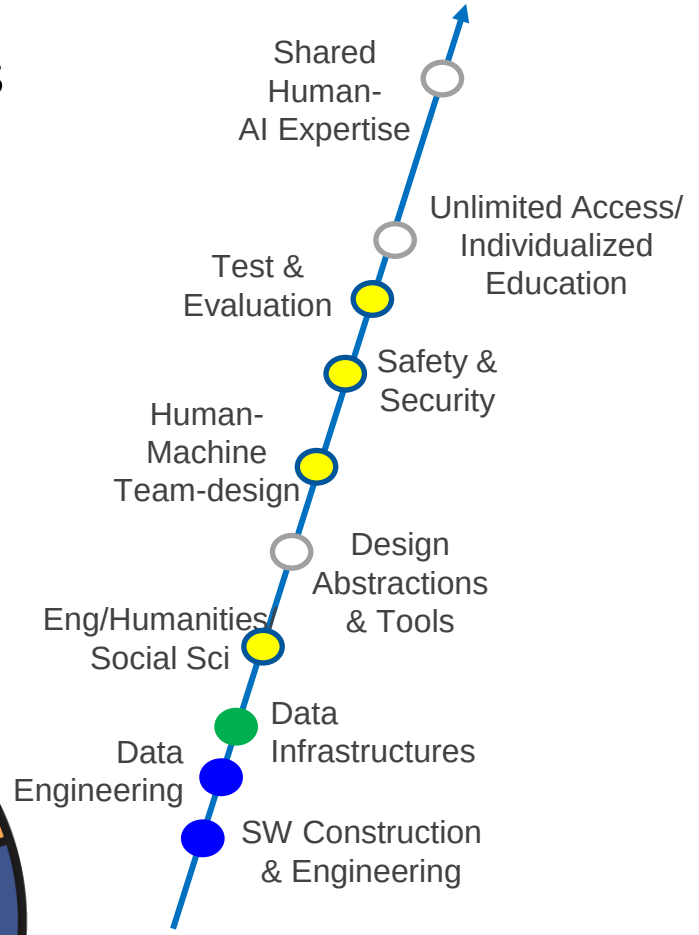
Human/AI Team Testbeds –
Flexible testbeds for evaluating human-AI teams for research, evaluation, design optimization, and adaptation



START WITH WORKFORCE AND CULTURE

- Integrating AI/ML experts with domain experts, all disciplines
- Trust: Engineering/Humanities/Social Sciences
- Evolving tools to align with design and disciplinary abstractions
- Human Systems Engineering: no longer a specialty discipline
- Threat models, safety, security, resilience, and other ‘ilities
- Evolving test and evaluation competency
- Fundamentally changing education





SERC DIGITAL ENGINEERING COMPETENCY FRAMEWORK

DIGITAL ENGINEERING TRANSFORMATION

Data Collection and Curation - data collection, management, curation and governance

Ontological Modeling – schematic representation to semantic representation

AI Specification – what will be allocated to the machine, in both product and process

AI System Design – system design as a mechanism for generalization of AI performance factored into design activities

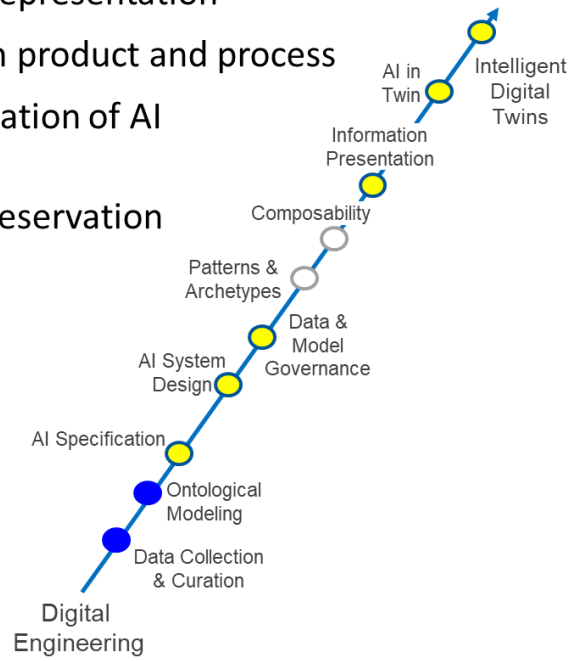
Data & Model Governance – Lifecycle management, control, preservation and enhancement of models and associated data

Patterns and Archetypes – learning from modeling artifacts

Composability – training and evaluating for design in context

Information Presentation – representing the decision space for human understanding and learning

Digital Twin Automation –
AI in Intelligent Digital Twins: real-time continuous learning from real system and front-running simulations



Convergence of Data Science and Systems Engineering Disciplines

Models become central to defining complex systems of systems

Results in Product plus Digital Twins of Product

Human-Machine interfaces and Visualization of complex interrelationships

"AI4SE"

Maybe:

It's life, Jim, but not as we know it

Creating the perfect assistant…

What would you include?

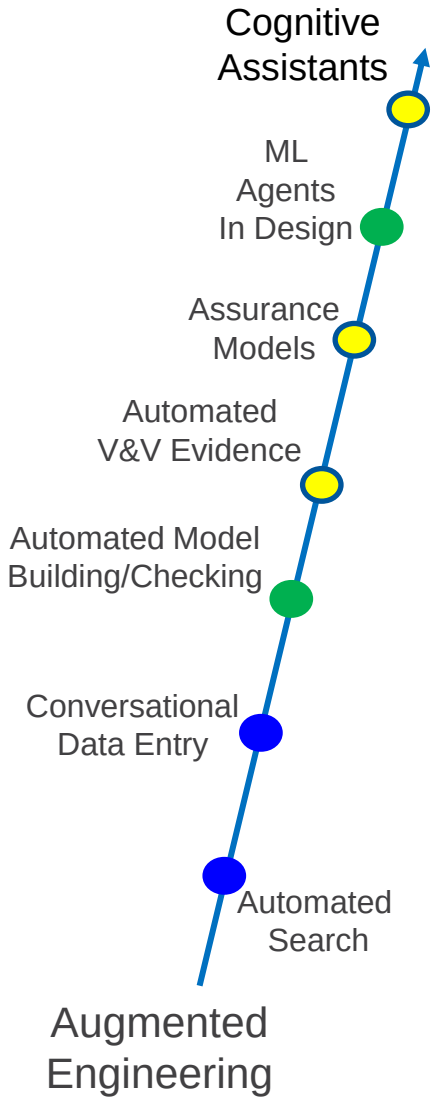
Would you make it sentient
(with rights, goals, desires, independent thoughts)

What should it know?

How should it behave?

A new kind of being:

- Neither human nor sentient
- No soul; no goals
- Vast knowledge and information
- Creativity of a sort
- Little (or no) judgment
- Great speed
- "Eager" to please
- Happy to re-do things many times

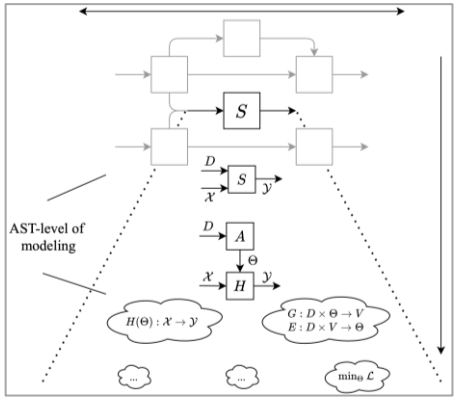


Barclay Brown, Living in a Generative World

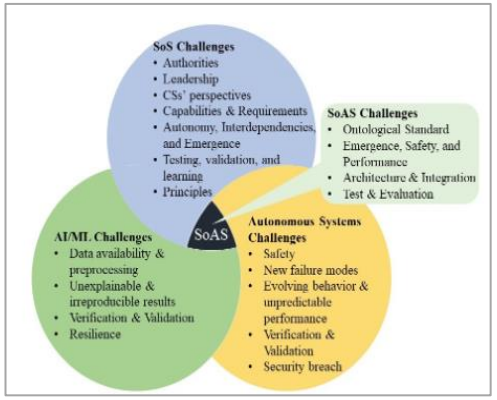
SYSTEMS ENGINEERING RESEARCH CENTER

AIRC ACQUISITION INNOVATION RESEARCH CENTER

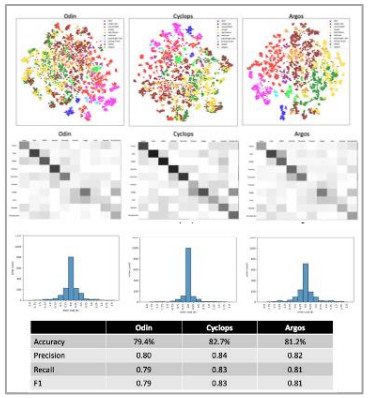
SNAPSHOT OF ONGOING SERC/AIRC RESEARCH



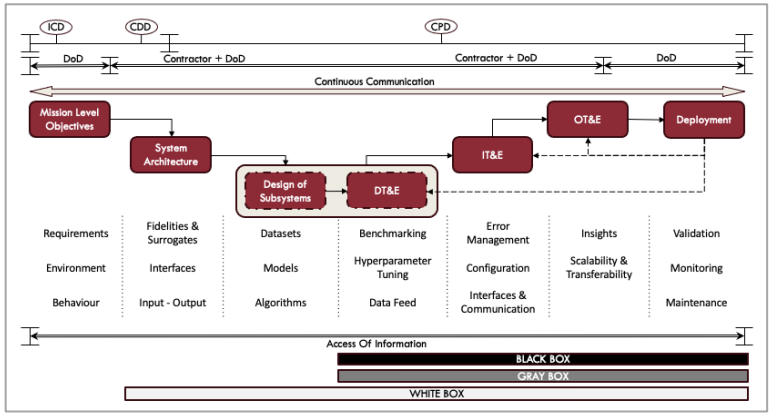
Foundational SE Theory to Model AIES



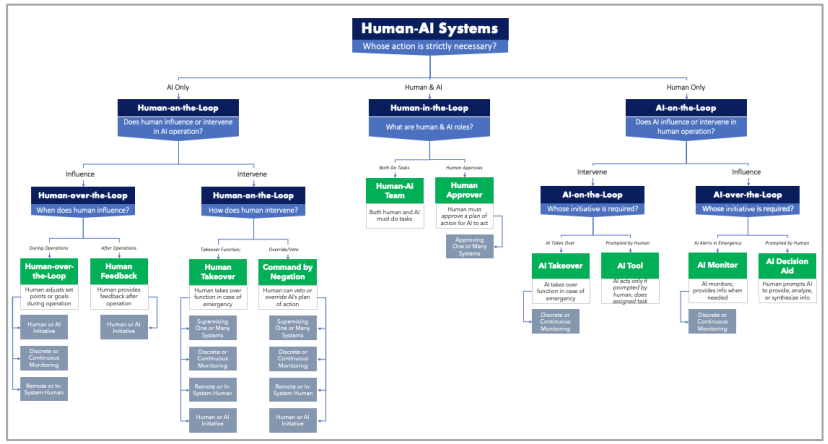
Analytical Methods and Tools to Support SoAS



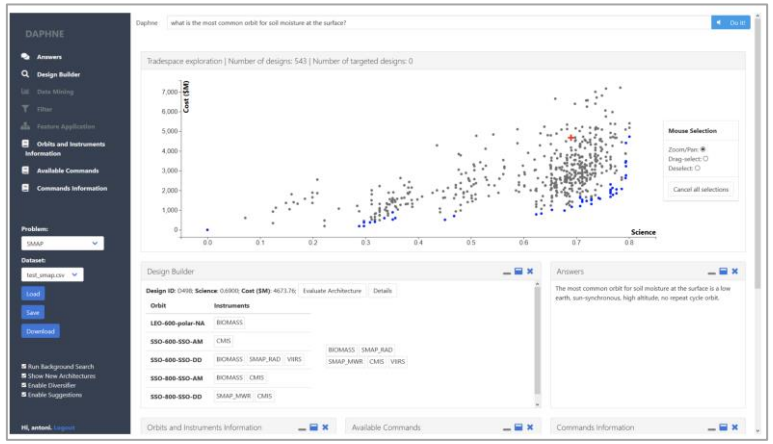
Explainability and Interpretability Techniques



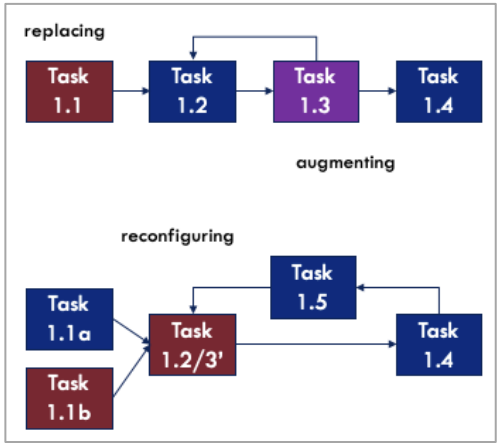
T&E of AIES



Multiple "architectures" of human AI Collaboration

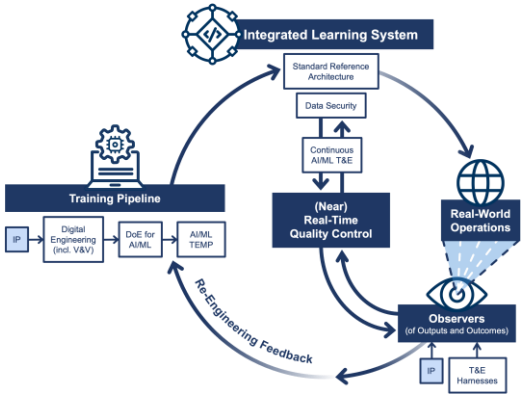


Cognitive Assistants for SE/Acquisition tasks form Cost Estimation to Model Generation

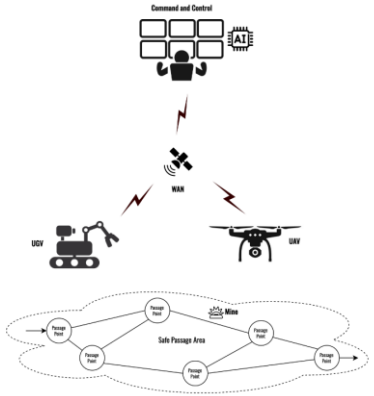


New AI-supported work processes to support e.g., contracting work

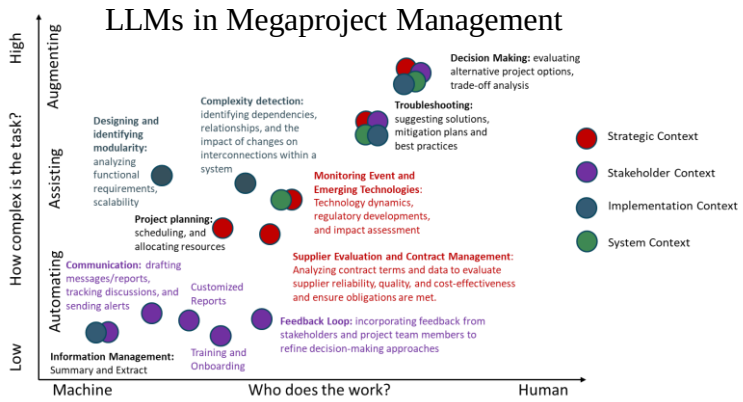
SNAPSHOT OF ONGOING SERC/AIRC RESEARCH



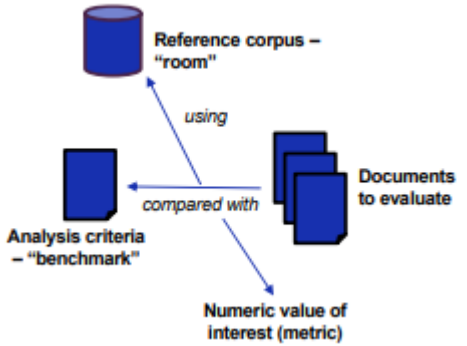
Framework For AI Resilience Through Evaluation Of Systems And Technology (FAIREST)



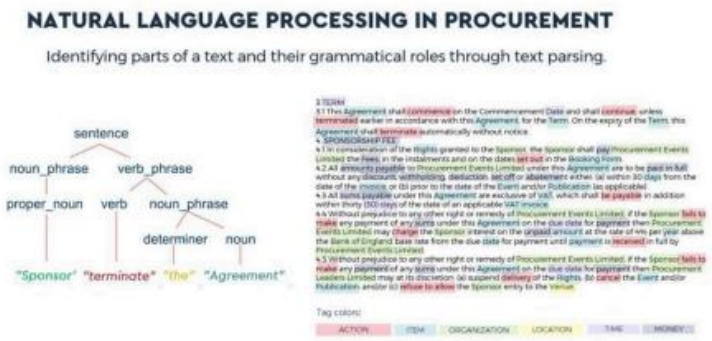
Trusted AI Systems Engineering Challenge



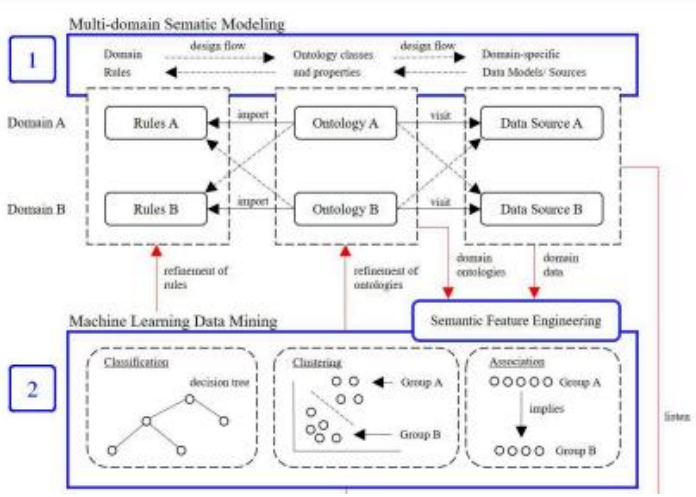
Future of Managing Megaprojects



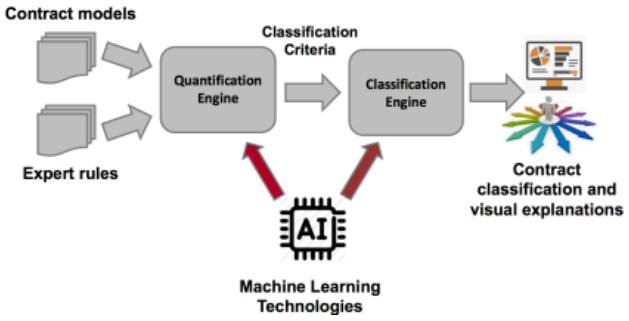
Meshing Capability and Threat-based S&T Resource Allocation



Management And Business Knowledge Representation For Decision Making



Architecting for Digital Twins with AI/ML



Analyzing and Assessing Contracts for Embedded Risk

SERC 5TH ANNUAL AI4SE & SE4AI WORKSHOP

2023 SUMMARY REPORT

The conference theme, "Safer AI-Enabled Complex Systems: Responsible Deployment of AI through Systems Engineering," aims to foster discussions and insights on how systems engineering can support the development of robust and ethical AI systems, and how AI tools can in turn transform the practice of systems engineering.

https://sercuarc.org/event/ai4se-se4ai-workshop-2024/#dates