

SERC RESEARCH REVIEW 2024 | NOVEMBER 12, 2024

Trust and Trustworthiness in AI-Enabled Systems



SYSTEMS
ENGINEERING
RESEARCH CENTER

SERC Strategic Talk

Zoe Szajnfarber, SERC Chief Scientist



STEVENS
INSTITUTE OF TECHNOLOGY
THE INNOVATION UNIVERSITY

GW

VT

MASON

P

GT

GW/TAI
TRUSTWORTHY
AI INITIATIVE

Role for Systems Engineers in AI space

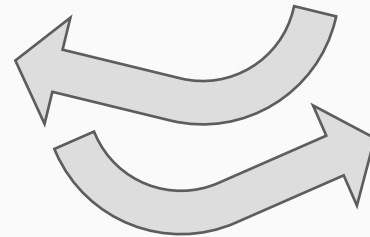
AI4SE

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, trustworthy**, robust, and efficient while extending those tools in response to the nature of AI enabled systems.



SE4AI applies to AI4SE, since we need to trust those tools
... and AI4SE might change what SEs do too.

Role for Systems Engineers in AI space

Requires trust



AI4SE

and

SE4AI

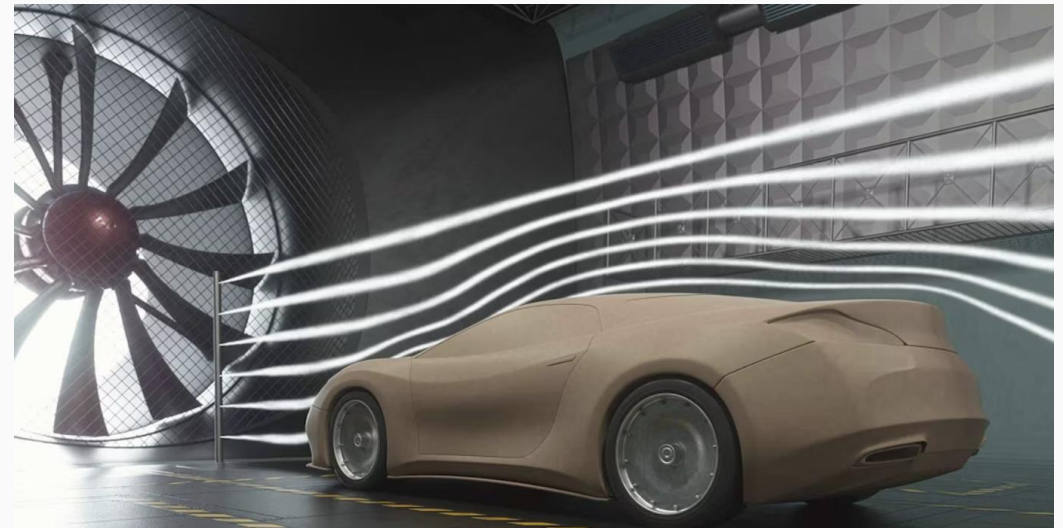
Ensure
trustworthy AI



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

¹Cited 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

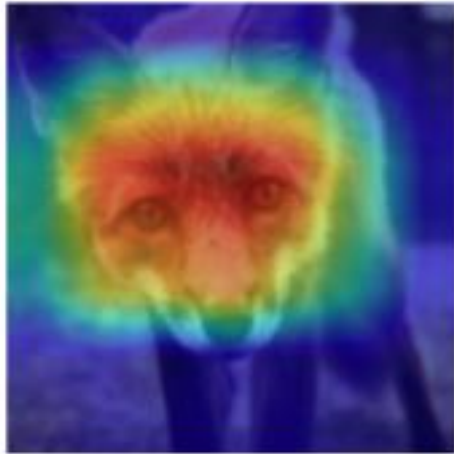
²Cited in NIST RMF Glossary: ISO/IEC_TS_5723:2022(en)

³Cited in NIST RMF Glossary: Mark Coeckelberg (2020) “AI Ethics” MIT Press; ⁴ NIST RMF

Do YOU trust the “AI” (in the system)?

Do YOU trust the “AI” (in the system)?

Developer



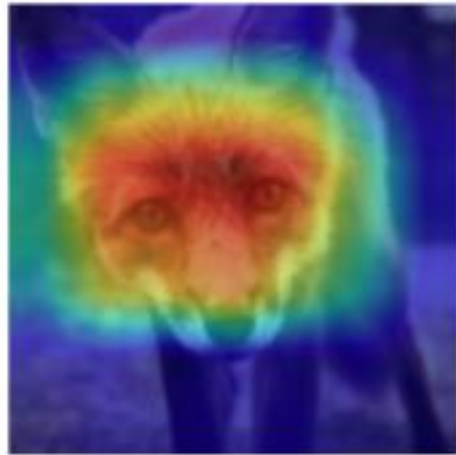
[1]

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

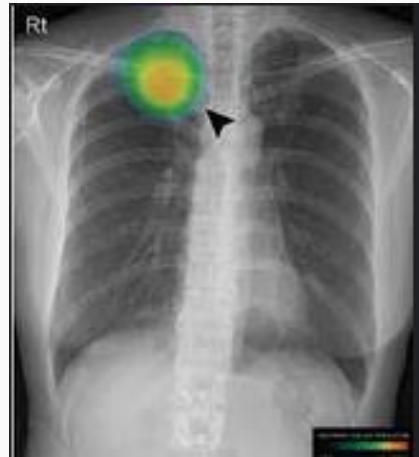


[1]

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.

Domain Expert



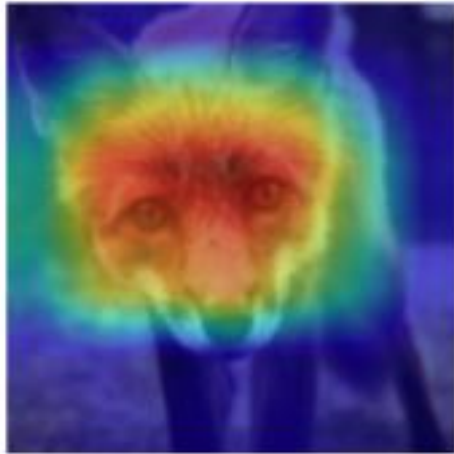
[2]

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

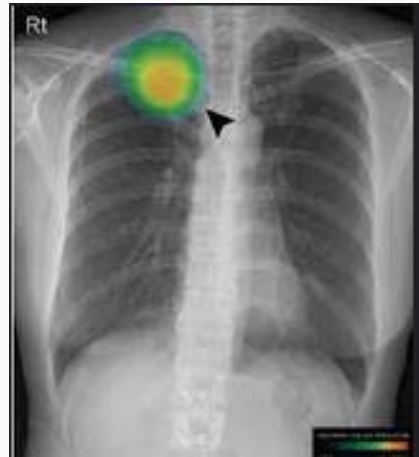


[1]

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.

Domain Expert



[2]

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.

End User



[3]

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!

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?

SERC Role in Defining Key Areas of Inquiry

- In process of updating 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

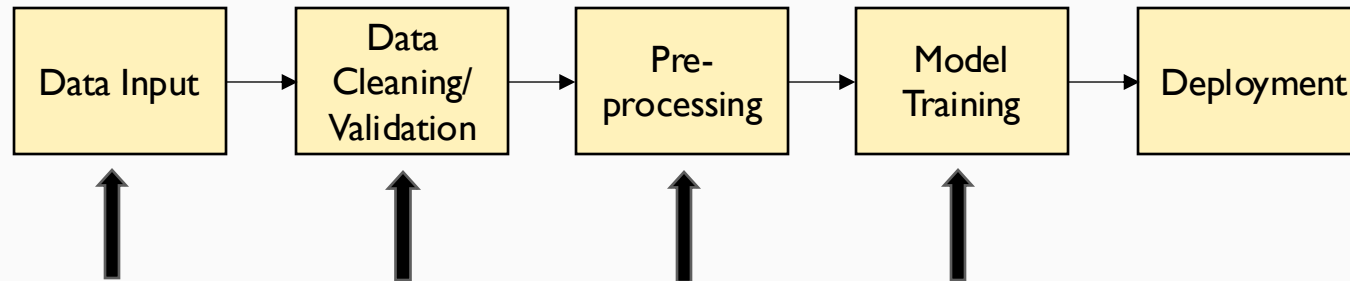
Axel Hahn, DLR-Institute of Systems Engineering for Future Mobility



AI-generated by GPT-4

Defining trust in the systems context

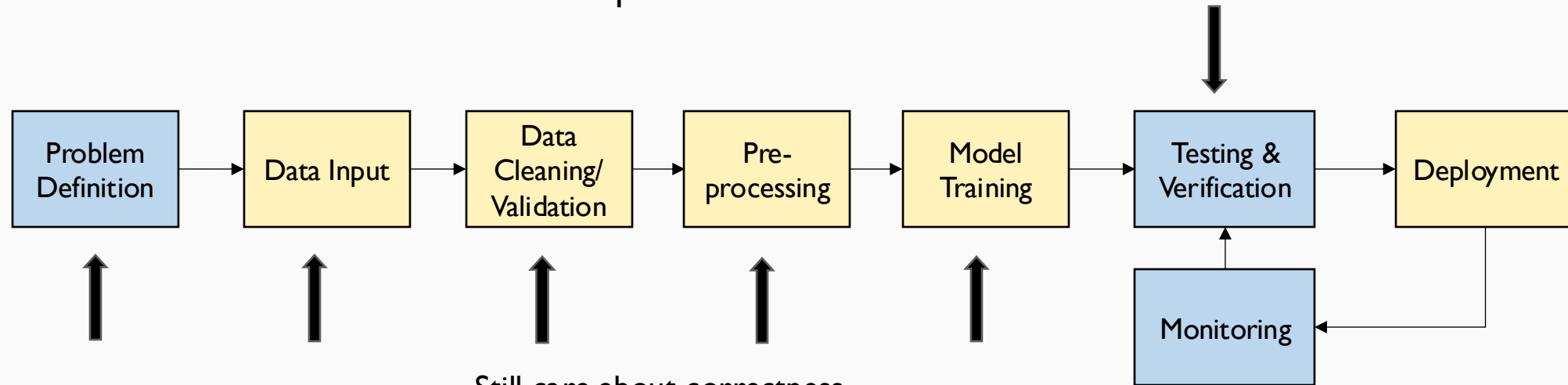
Typical representation of AI/ML pipeline:



Focus of Trust is on doing these steps correctly, and without bias

Defining trust in the systems context

Real-world focused AI researchers and practitioners often add:

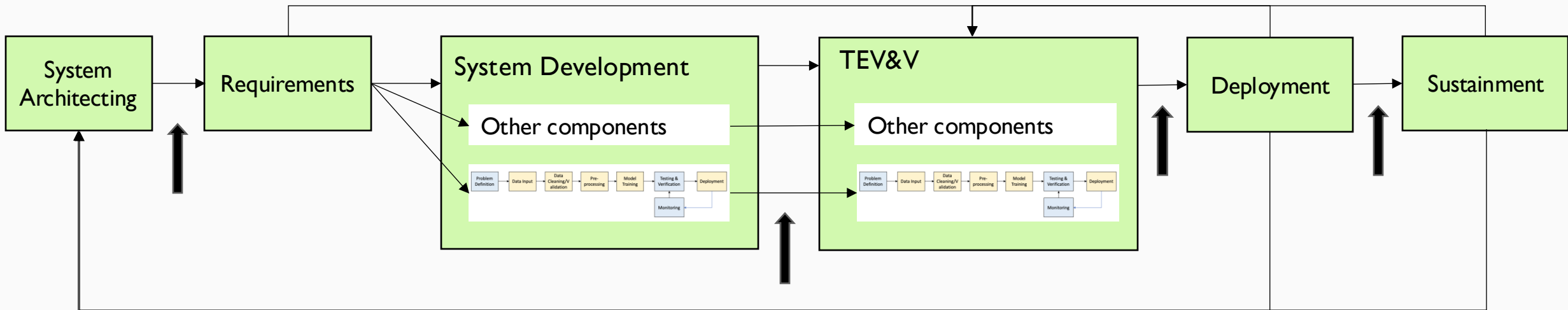


Still care about correctness
and also, if you solved the right problem and model works as intended

... but this is still focused on the model as the system.

Defining trust in the systems context

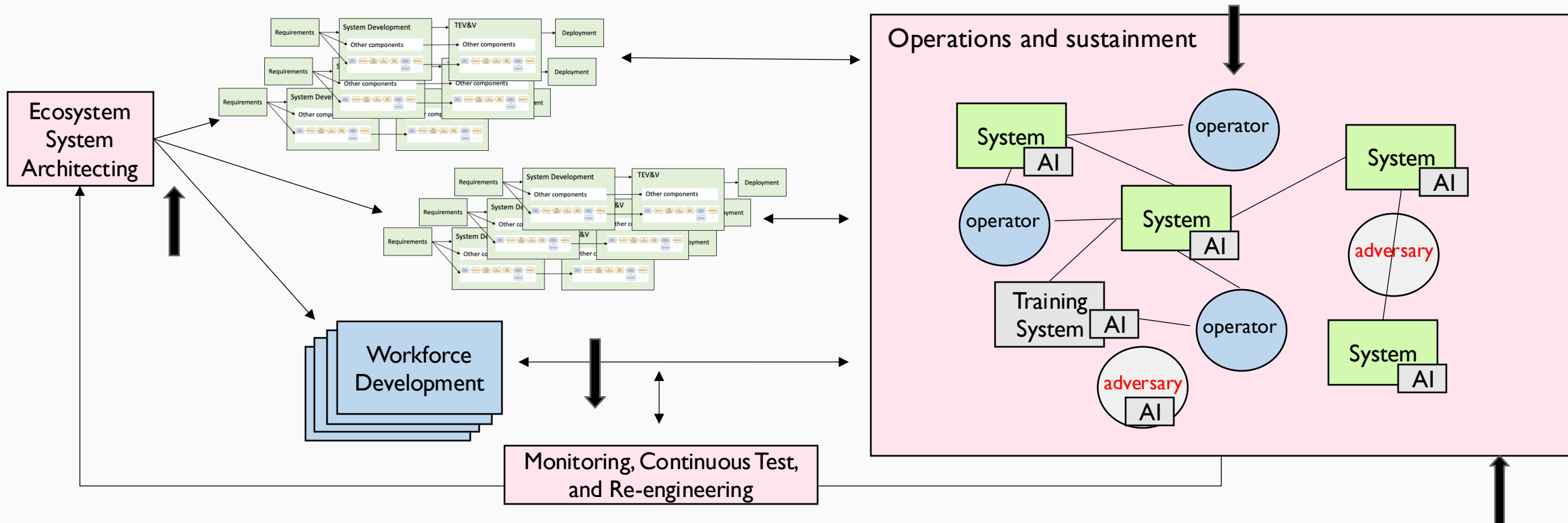
For Systems Engineers, AI is part of a "system"



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

Defining trust in the systems context

DoD operates in a socio-technical systems environment, involving complex interactions among humans and systems that were not always intended to work together in a constantly changing environment.

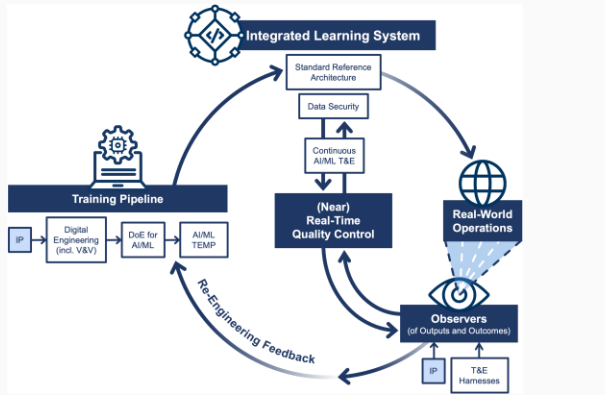


Everything on the previous slides... and extent to which operators use and trust new technology, how risks and functions are allocated to different parts of the overall systems, how changing environment is monitored, and network is updated accordingly

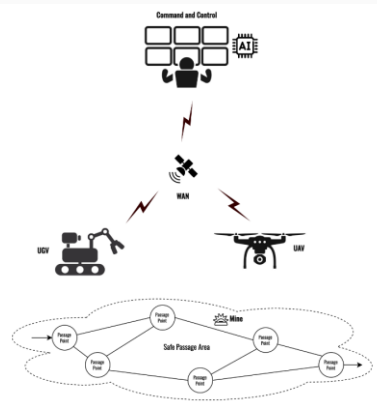
Key TAI questions for Systems Engineering

- 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**.

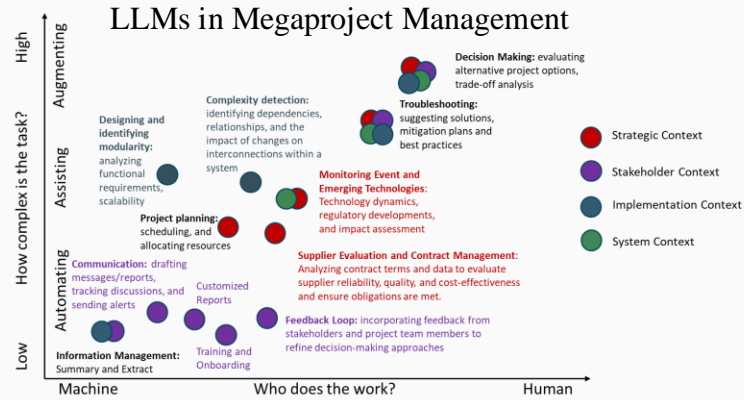
Snapshot of Ongoing SERC/AIRC Research



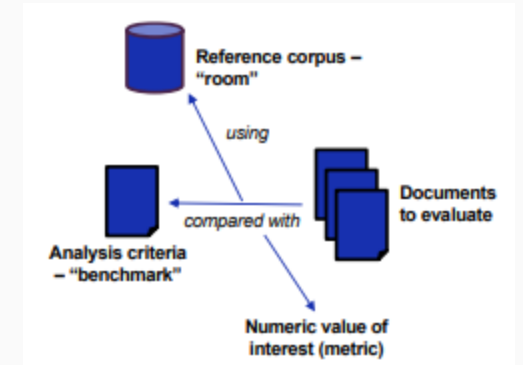
Framework For AI Resilience Through Evaluation Of Systems And Technology (FAIREST)



Trusted AI Systems Engineering Challenge



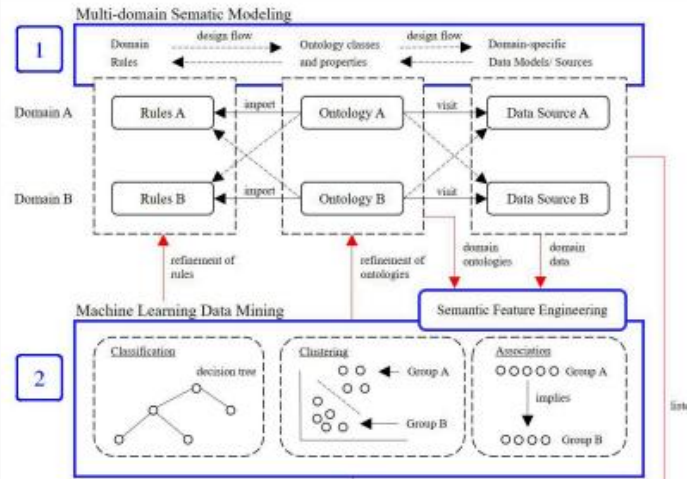
Future of Managing Megaprojects



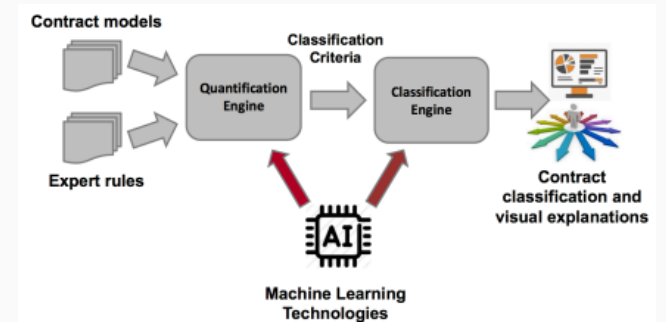
Meshing Capability and Threat-based S&T Resource Allocation



Management And Business Knowledge Representation For Decision Making

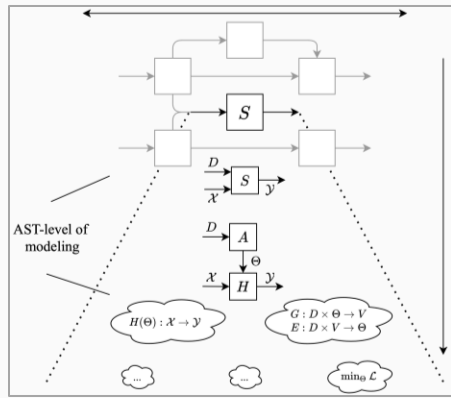


Architecting for Digital Twins with AI/ML

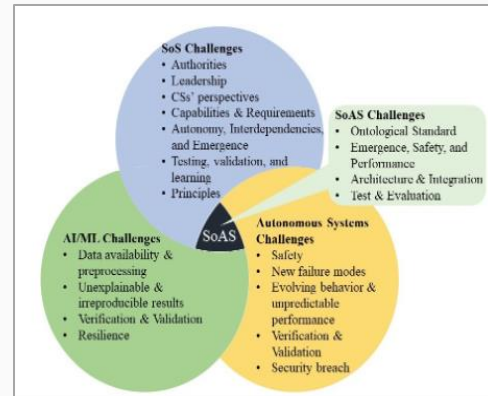


Analyzing and Assessing Contracts for Embedded Risk

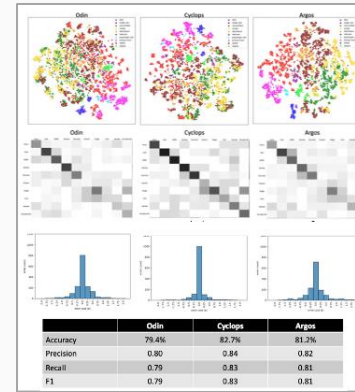
Snapshot of Ongoing SERC/AIRC Research



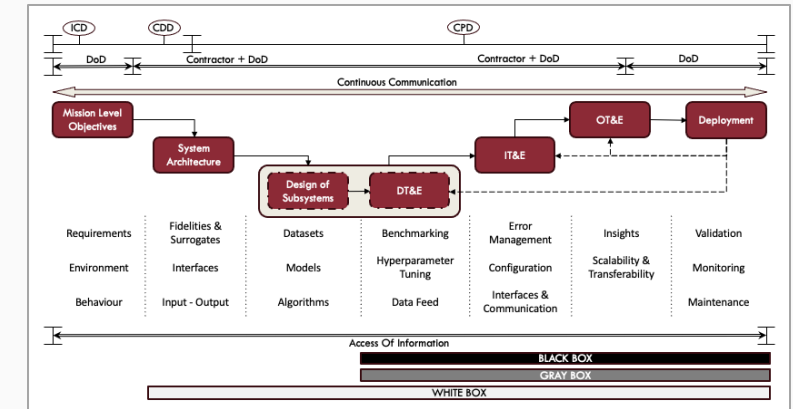
Foundational SE Theory to Model AIES



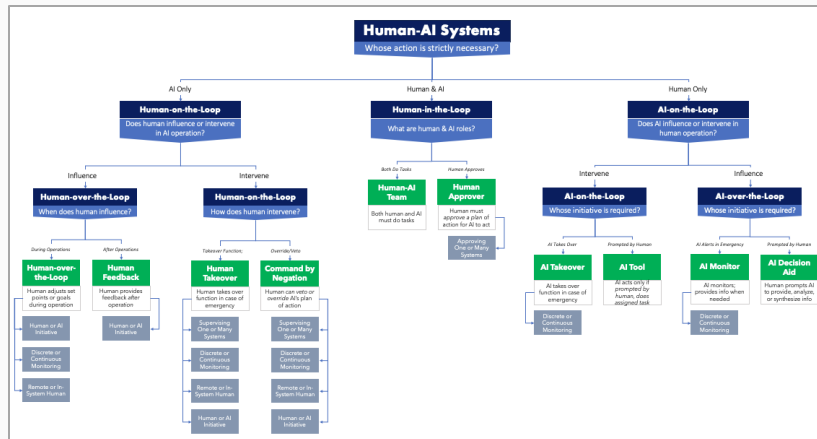
Analytical Methods and Tools to Support SoAS



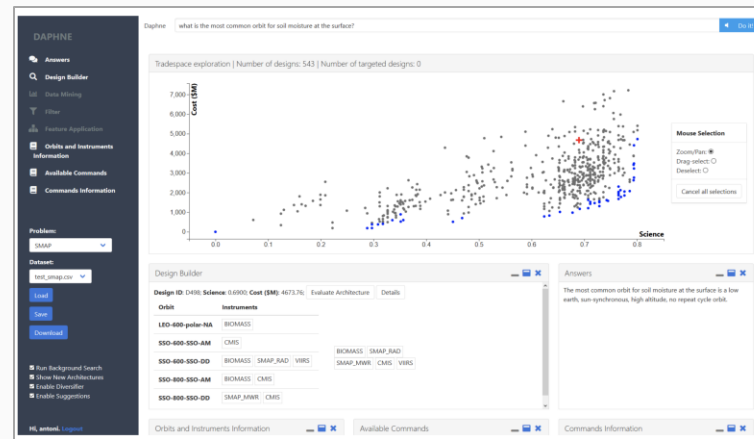
Explainability and Interpretability Techniques



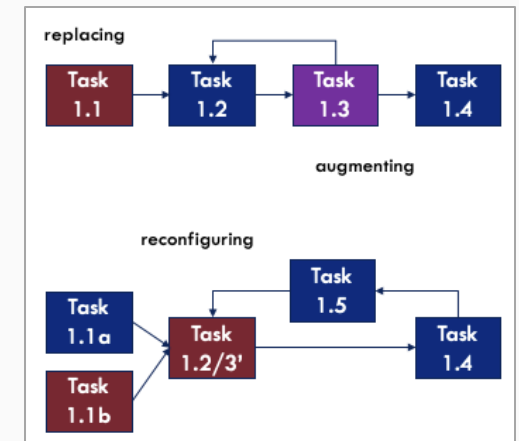
T&E of AIES



Multiple "architectures" of human AI Collaboration



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

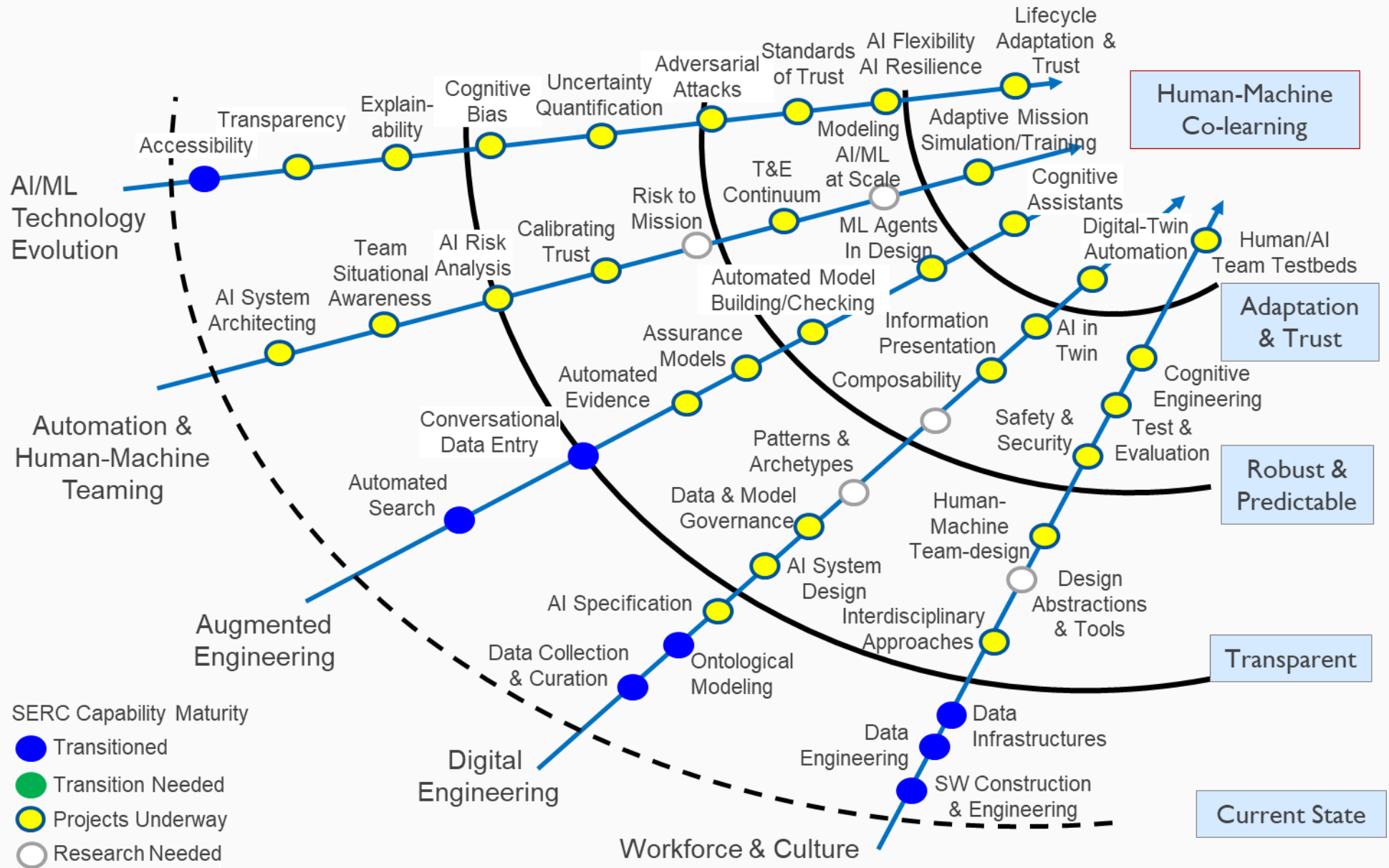
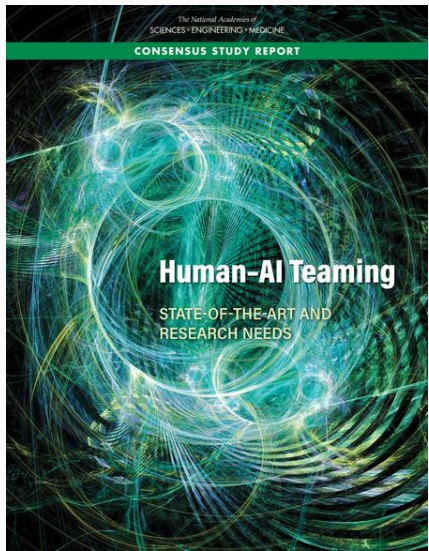


New AI-supported work processes to support e.g., contracting work

Snapshot of Ongoing SERC/AIRC Research

	AI4SE	SE4AI	Workforce
Prototype tools	Cognitive assistants, early flags	T&E Dashboards, visualizations	DCTC, LLMs as personalized support
Theory and methods	HAI architectures, SoAS methods	System theoretic approaches, AI. “-ilities”	Competency models, user studies
Towards testbeds	Pilot deployments + user studies	SE TAI Challenge, test harness	Pilot deployments + user studies

Updated 2023 SERC AI & Autonomy Roadmap






SYSTEMS
ENGINEERING
RESEARCH CENTER

Thank you!

Stay connected with SERC Online:



Email the presenter: Zoe Szajnfarter

 zsajnf@gwu.edu

Email the research team:

 [Peter Beling, Ali Raz, Tom McDermott, Val Sitterle, Jitesh Panchal, Doug Buettner](#)