Enhancing Evaluation and Testing of AI-Enabled Systems in the DoD using Model Based Systems Engineering

Carol Pomales

James R. Morris-King, PhD

Tai Jella

Bill Fetech

September 2024

Sponsor: OUSD DTE&A
Project No.:101395.24.501.D320.P30

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

Approved for Public Release; Distribution Unlimited. Public Release Case Number 24-2714

Basic Contract No. W56KGU-18-D-0004

McLean, VA

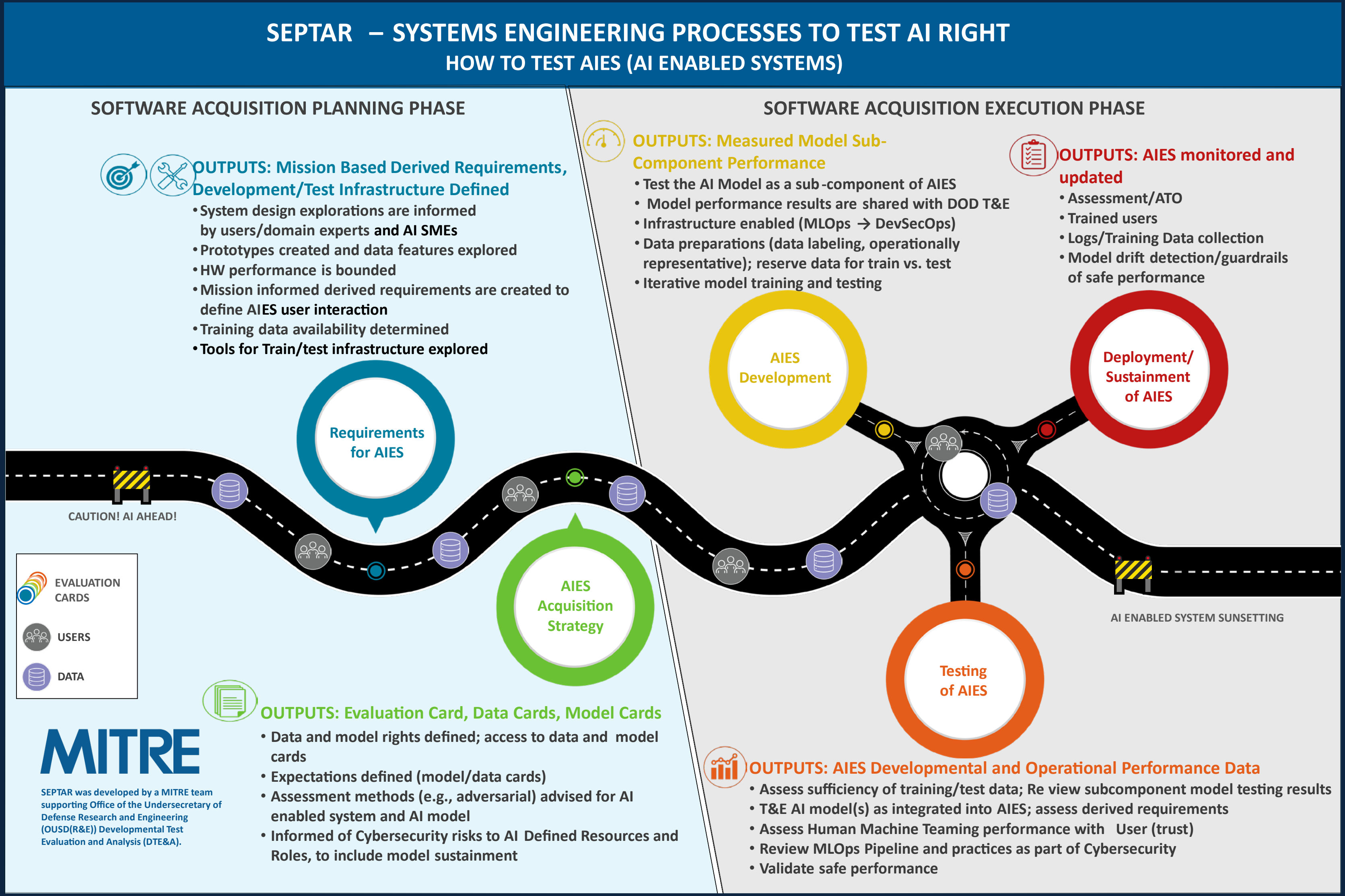
Agenda

- Introduction
- SEPTAR Alignment
- AI & MBSE Formalism
- MBSE for AIES Approach
- Model View AI Challenges and Recommendations
- Use Case Model
- DoDAF and UAF Alignment
- Views and Diagrams
- Conclusions & Future Work

Sponsor: Department of Defense (DoD) Under Secretary of Defense for Research and Engineering (OUSD(R&E)) Developmental Test, Evaluation, and Assessments (DTE&A)

SEPTAR Alignment

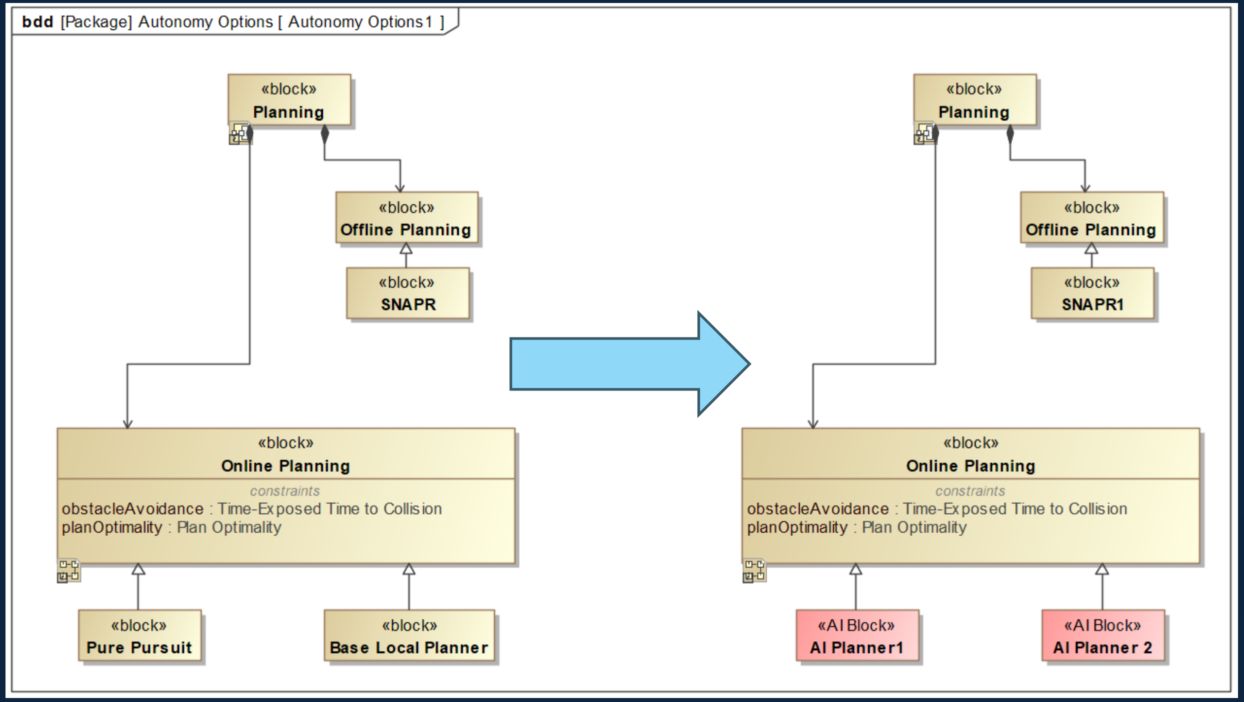
This work addresses a focus area within the Systems Engineering Processes to Test AI Right (SEPTAR) Framework which defines a broader T&E Continuum for AIES

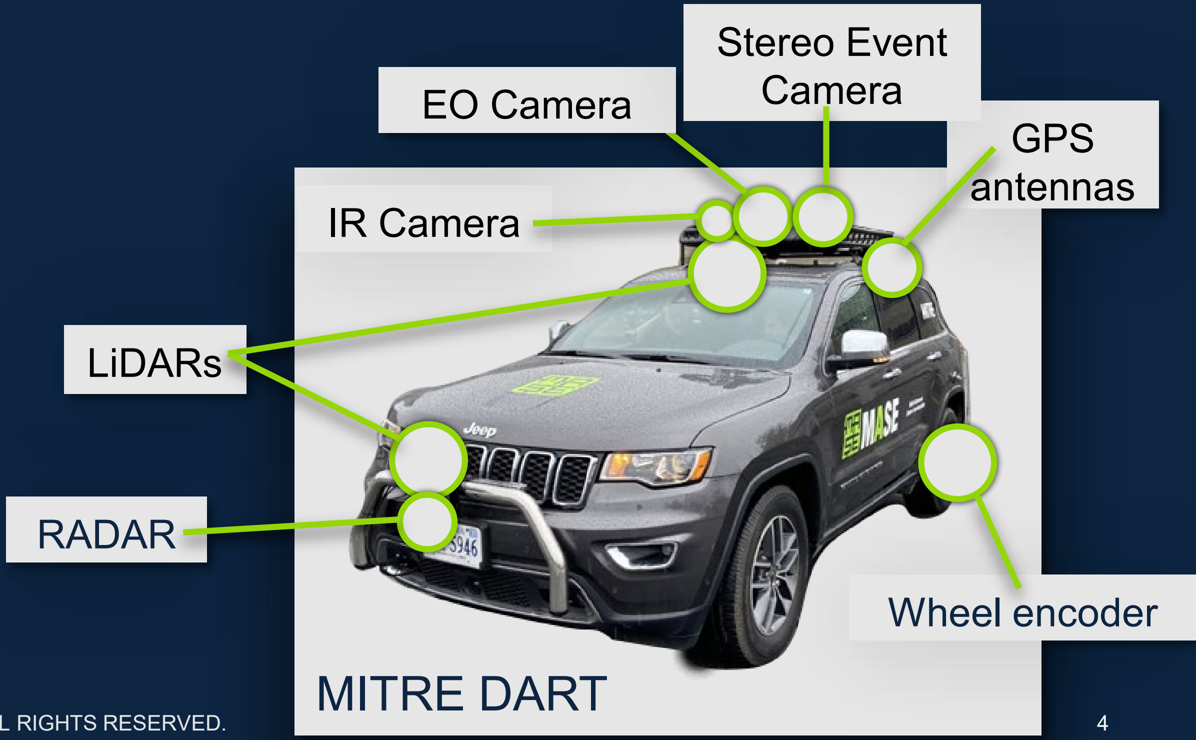


Use Case: Autonomous Ground Vehicle (AGV)

We grounded our work in an MBSE use case of an AI-enabled system (AIES)

- DART-AGV project is a MITRE program in support of Army Research Lab
  - Goal is to evolve an MBSE model for an autonomous vehicle system to include AI components.
- Team modified ARL-provided MBSE model to include an Artificial Intelligence Online Planner
  - Added and modified views to highlight the AI component and its interactions

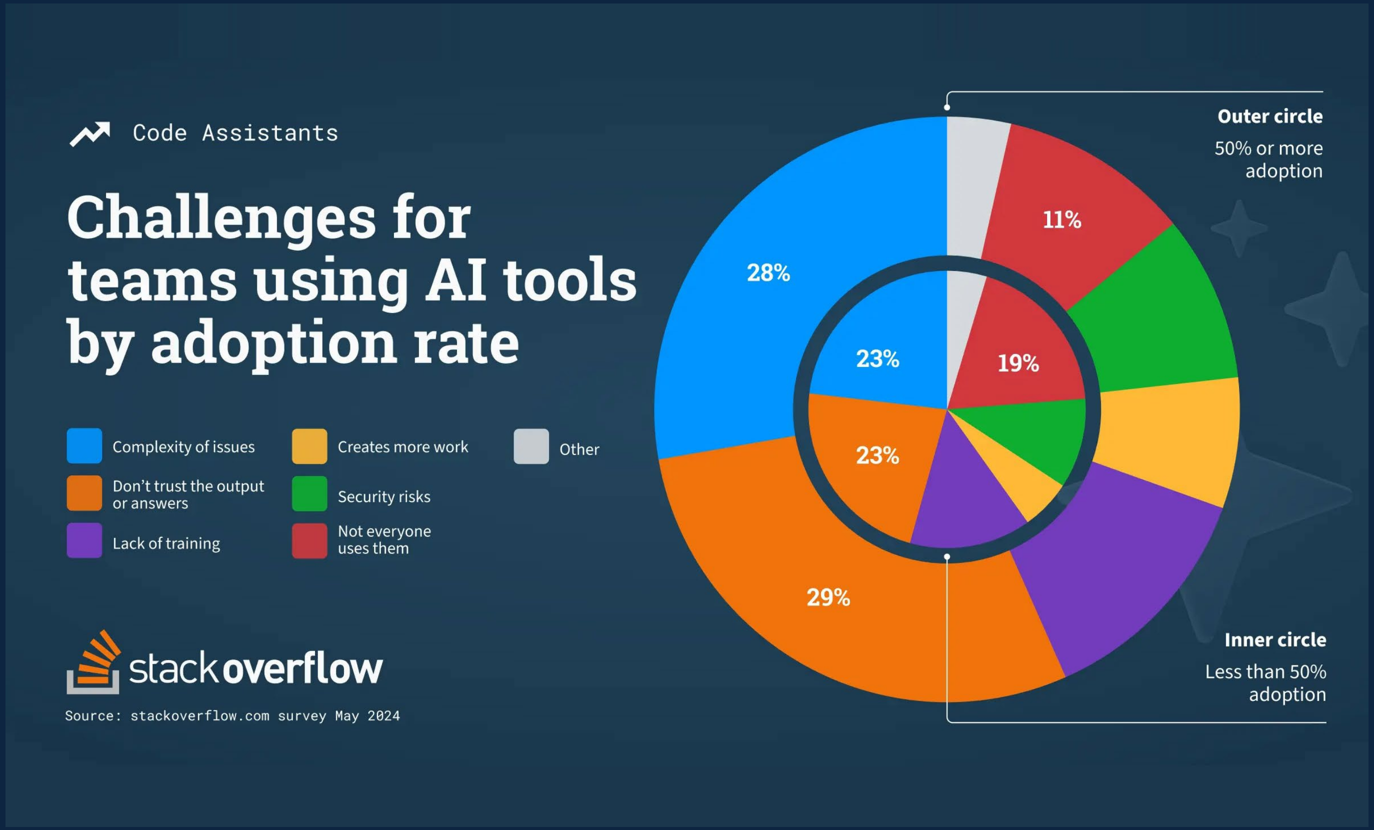




Model-View AI Challenges & Recommendations

T&E of AIES Challenges

1. Functional Behavior of AIES Can Include Learning Components
2. AIES Behavior Continuously Adapts and Evolves Over Time
3. Challenges of Human Review and Accountability for AIES Performance
4. AI Model Training and Bias
5. Cybersecurity of AIES
6. Limited “awareness” (context for decision-making)
7. Leveraging AI Model Transparency



Insights gained via MBSE can help mitigate these T&E challenges

MBSE Mitigations to the T&E of AIES

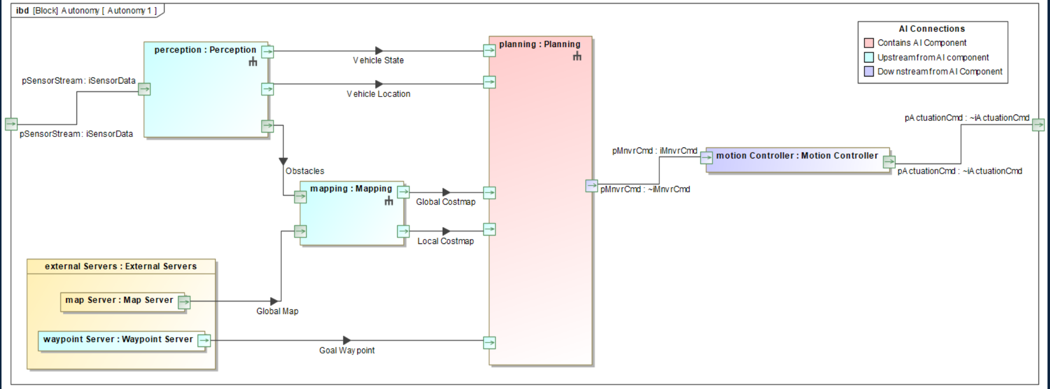
Functional Behavior of AIES Can Include Learning Components

T&E Limitation(s):

- T&E is difficult to conduct without clearly defined inputs and outputs due to the lack of transparency in how AI model parameters lead to decisions and the potential for AI models to continuously learn.
- More directly, it is difficult to build or evaluate good tests without a clear understanding of the parameters that determine the underlying AI model's performance. In a fully "Black Box" scenario, test activities must be conducted around the inputs and outputs of the AI sub-component(s) interacting with the rest of the AIES.

MBSE Mitigation(s):

- Block Diagrams: If the input/output for an AIES is known, AI *sub-component interfaces* can be entered into block definition diagrams (BDDs) or internal block diagrams (IBDs) in SysML.
  - IBDs show relationships between AI and non-AI sub-components to *characterize interfaces for evaluation* to fully assess an AI sub-component. Diagrams such as these can show what data elements are inputs or outputs of an AI component to consider in test design and test case prioritization to fully inform the risks around AI performance.
- Constraint Diagrams: MBSE can also demonstrate *a range of outputs* that are acceptable for the AIES and clearly define the limits of these ranges.
  - Numerical or quantitative *constraints captured* in a constraint diagram that define the AIES performance can potentially be checked through automated T&E or to inform T&E activities that occur through more manual processes.



AI enabled subcomponent (pink box in the center) interacting with several other non-AI enabled sub-components.

Element Type: Constraint Block ... Scope (optional): Autonomy Metrics,Autonomy Parameters,Environr ... Filter:

#	Name	Owner	Constraints	Refines
1	Learning and Adaptability	Planning	Error Rate=Error rate should decrease by X% after ...	145 Learning and Adaptability
2	Training Data Bias and Fairness	Planning	Error Rate for different environmental situations=...	143 Bias and Fairness
3	Robustness and Resilience	Planning	Return to normal behavior=Time < 1 minute	142 Robustness and Resilience
4	Continuous Monitoring and Evaluation	Planning	System performance evaluation every=Time < 1 hour	146 Continuous Monitoring and Evaluation
5	Data Privacy and Security	Planning	Data storage encryption key length =Key Length > X...	147 Data Privacy and Security

Constraints defined for AIES performance.

MITRE

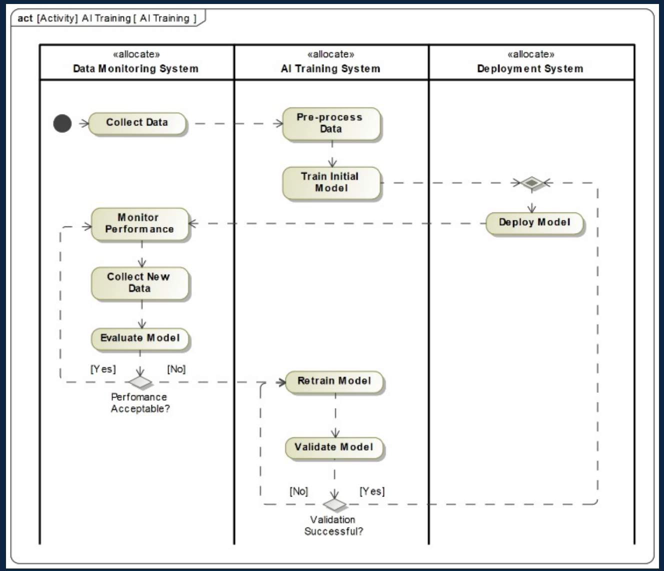
MBSE Mitigations to the T&E of AIES
AIES Behavior Continuously Adapts and Evolves Over Time

T&E Limitation(s):

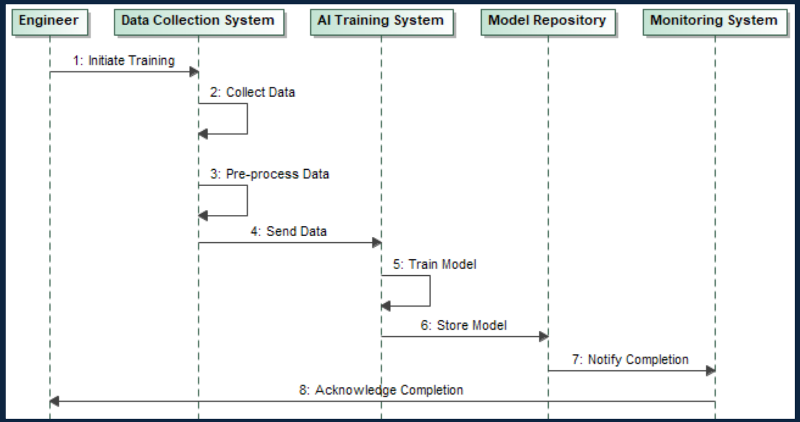
- It may be expensive, infeasible or impractical to re-test the AIES every time the AI models change as a normal part sustainment. The continuous changes and updates of AI models will create challenges for T&E.

MBSE Mitigation(s):

- <u>Activity diagrams</u> – Assessing continuously evolving systems benefits from the <u>*application of automated testing techniques*</u>.
    - Activity diagrams can inform these practices, providing a clear understanding of the mission and use of the AIES (and sub-system) and enables effective automated resources assessment.
    - In cases where AIES systems have self-instrumentation built into collect key test data automatically (e.g., model performance changes), an activity diagram can show clearly where and how the data was collected.
- <u>Sequence Diagrams</u>: Within a sequence diagram it should be clear what lifelines are AI components so messages upstream or downstream from it are well defined.
    - This informs T&E of potential <u>*key dependencies or bottlenecks*</u> for exploration in automated testing to facilitate rapid redeployment.



Activity diagram of AI retraining process.



Sequence diagram of AI training and deployment

MITRE

MBSE Mitigations to the T&E of AIES

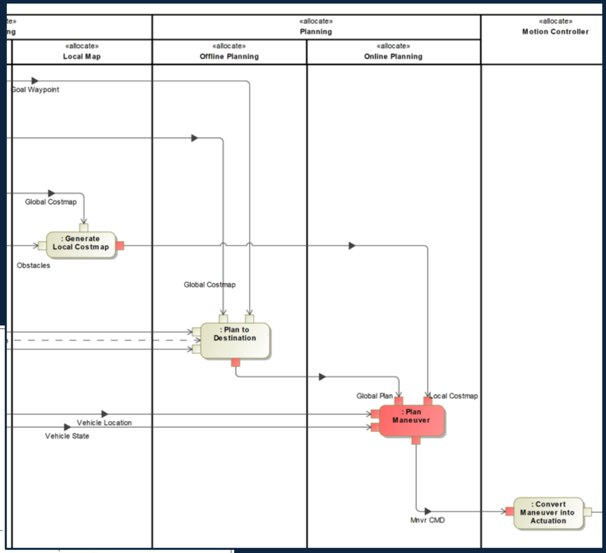
Challenges of Human Review and Accountability for AIES Performance

T&E Limitation(s):

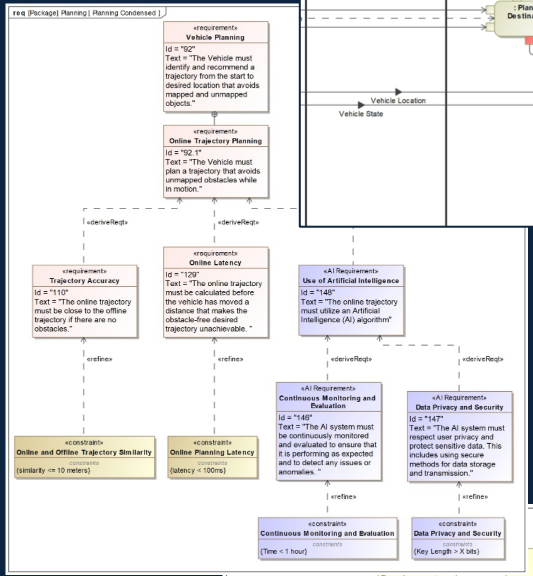
- It is difficult to enable T&E or human evaluators to assess capability performance due to lack of transparency in explainable AI, AIES, and self-reporting confidence scores.

MBSE Mitigation(s):

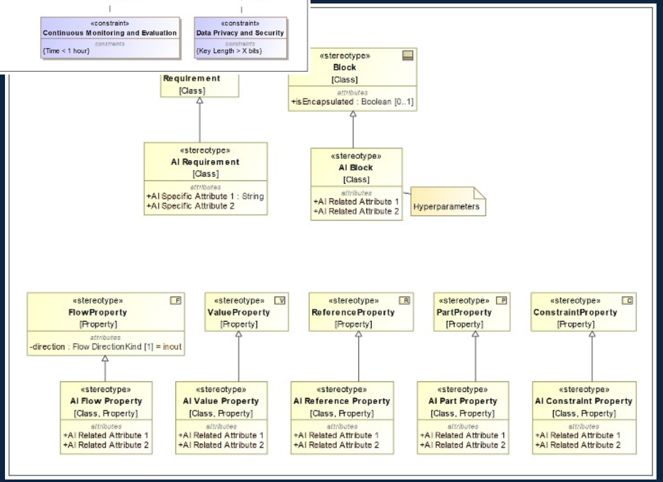
- Activity Diagrams – System-level activity diagrams document behaviors of system or sub-system elements and user behavior as they interact with the element.
  - Activity diagram swim lanes clarify the responsibilities of specific parts of the system. T&E can discern *which elements are performed by the AI components and which human decisions and activities* are upstream and downstream from those AI enabled activities.
- Requirements Diagram – Requirements diagrams visually represent the landscape and relations between requirements and other elements.
  - Derived requirements note the presence of AI (blue box) and capture specifics of AI model technology choices. *The derived AI requirements outline the expected behavior, performance, and constraints* within the AIES as it supports the user and mission.
  - Operational descriptions captured in the figure show key considerations for data privacy to ensure appropriate test planning and data privacy practices for test.
- MBSE Stereotype - A stereotype for AI-related requirements/derived requirements for MBSE to extend the SysML Requirement Class.
  - AI-specific stereotypes capture additional *properties for AI-related derived requirements* and provide clear differentiation *to enable T&E.*



Activity diagram for AI component



AIES-derived requirements diagram



Requirements diagram with stereotyped reqts

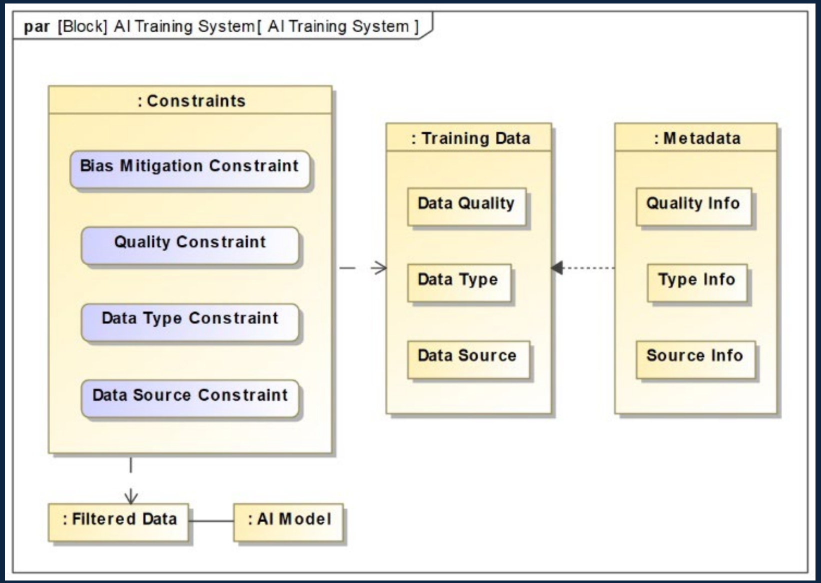
MBSE Mitigations to the T&E of AIES
AI Model Training and Bias

T&E Limitation(s):

- Techniques to ensure data used to train the AIES is of sufficient volume and quality to assess the AIES are still emerging. Testing for bias in data must be grounded with an understanding of the operational environment where the AIES will perform and understood by T&E.

MBSE Mitigation(s):

- Parametric Diagram: MBSE defines the range of conditions that the AIES must perform under once operationally deployed.
    - T&E can *compare expected operational context to the data used to train the AIES* to ensure model training data represents operational conditions and avoids bias. If bias cannot be avoided, T&E must ensure appropriate process mitigations are documented (e.g., tactics, techniques, and procedures).
    - MBSE may help the tester *identify cases of data bias* if proper metadata is available. Models can have constraints and boundary conditions to ensure only a certain amount, certain types, or a filtered set of data is used to train an AI component.
    - With a digital toolchain, AI models could be trained with different weights, constraints, and boundaries to create an AI model that mitigates bias. This will have some human-in-the-loop interactions, but MBSE can automate and organize the process, resulting in a better AI component.



Parametric diagram of constraints and AI components

MBSE Mitigations to the T&E of AIES
Cyber T&E of AIES

T&E Limitation(s):

- Each type of AIES will have its own set of cyber vulnerabilities and attacks against it may be indirect (e.g., targeting related elements like training/test data, user inputs, results output, or interfaces between AI systems). Security risks may be generated by third-party components or other integrated systems and vulnerabilities may not be easily identified by T&E.

MBSE Mitigation(s):

- MBSE can apply automated techniques to identify vulnerabilities within system architectures (e.g., ATLAS™). These vulnerabilities may include a set of mitigations to help inform steps to reduce risk to the system. MBSE models provide a system representation to explore the system design for vulnerabilities and evaluation, including clarifying interfaces that may introduce vulnerabilities.
- One notable cyber challenge for AI models is training data poisoning. T&E professionals can analyze risk by reviewing how the system design restrained outputs of the AIES or approaches output cross validation.
- MBSE frameworks extended to include tools to incorporate threat modeling for cyber T&E enable threat analysis on the operational architecture of a system.

MBSE Mitigations to the T&E of AIES

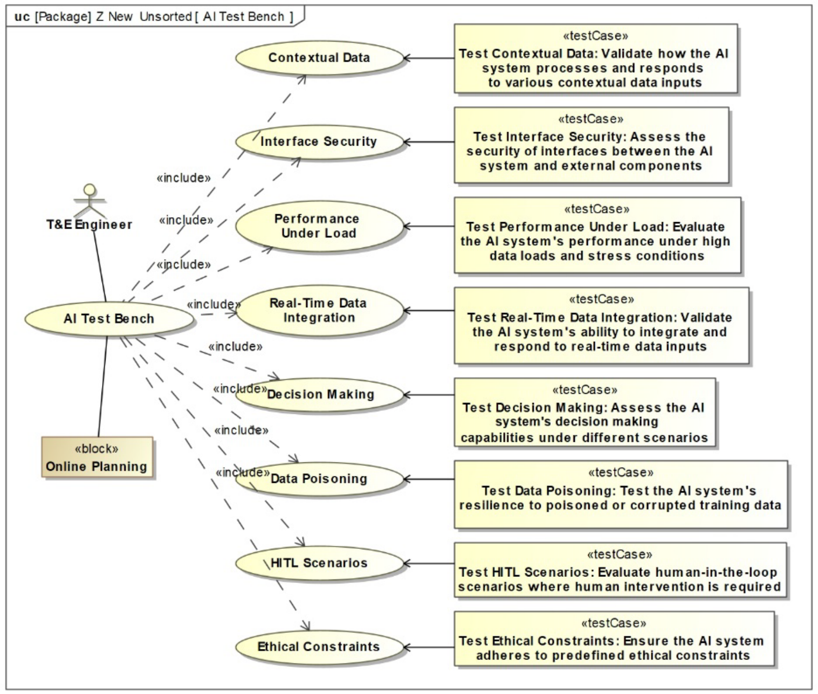
Limited "awareness" (context for decision-making)

T&E Limitation(s):

Testing AIES effectiveness in performing complex tasks performed by humans (e.g., answering questions in a chat stream, making multivariate policy decisions) and their application of "common sense" can be extensive, expensive and require human validation to test use cases.

MBSE Mitigation(s):

- Automated test benches: Created to test how AIES perform with activities requiring "wordly logic" and organized into clear & concise use case diagrams.
  - Test benches can be *generalized, templatized, and stored* in a centralized location so when a new AI model is made or is being retrained, quick tailoring can be done to test that AI appropriately.
  - AI components/capabilities are identifiable (e.g., color highlighting) and well-documented in MBSE better guides the T&E approach.
  - AIES which enable complex human-like capabilities have multiple AI models working in-line or in collaboration and each AI component should be tested both in isolation and in concert with the broader system.
- **Note**: Simply creating a model of an AIES in and of itself will not mitigate the issue of AIES lacking "common sense". Rigorous behavioral testing is still required.

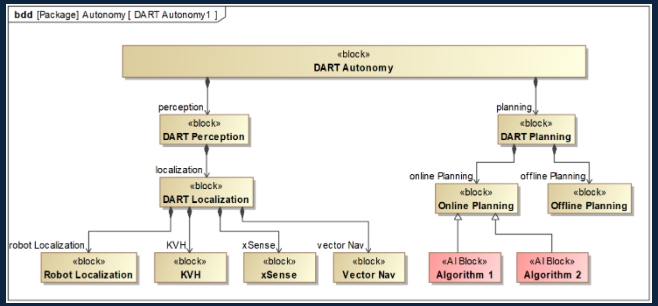


Use Case diagram with example test cases

MITRE

MBSE Mitigations to the T&E of AIES

Leveraging AI Model Transparency



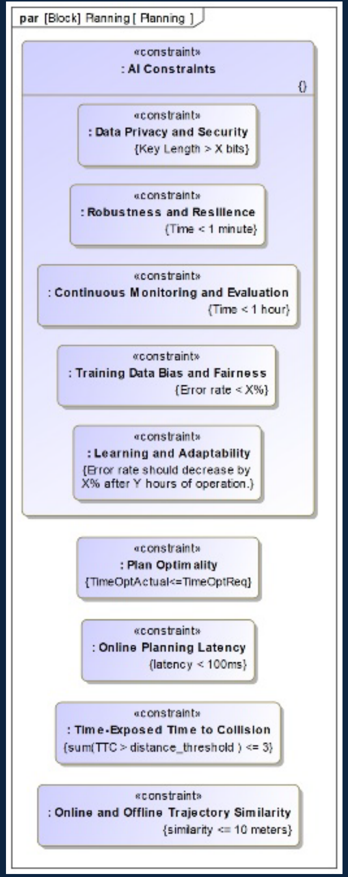
AI Stereotype adornment on BDD

T&E Limitation(s):

Exposing the innerworkings of the AIES enables T&E to more precisely and effectively test the AIES by looking at sub-component and aggregate system performance.

MBSE Mitigation(s):

- MBSE constraints: Leverage AI model transparency by exposing the limits on AI function and behavior to inform conditions needed to be represented in the AI training/test data and evaluated during T&E.
    - Constraints are useful for T&E to assess how well data covers operational conditions or introduces bias. Constraints safeguard an AI component by putting limits on ranges of characteristics of outputs, including *computational* (e.g., GPU), *performance* (e.g., accuracy), and *decision-making* (e.g., confidence) metrics.
    - Clearly documenting these constraints in a Parametric diagram aids T&E on facets of the AIES that will likely need to be *retested* (ideally automated) over the full lifecycle.
- Model Specification/Stereotypes: In cases where greater transparency is granted on the AI component, it can be noted in the model specification details the type of AI model (e.g., LLM, Computer Vision), what library it was derived from (e.g., Gemini, Midjourney, ChatGPT), what training dataset was used, and what additional data the component has ingested deployment.
    - If new data types are required, MBSE allows for further extension through the usage of *stereotypes*.



Example AI Constraints

MITRE

SysML to DoDAF and UAF Alignment

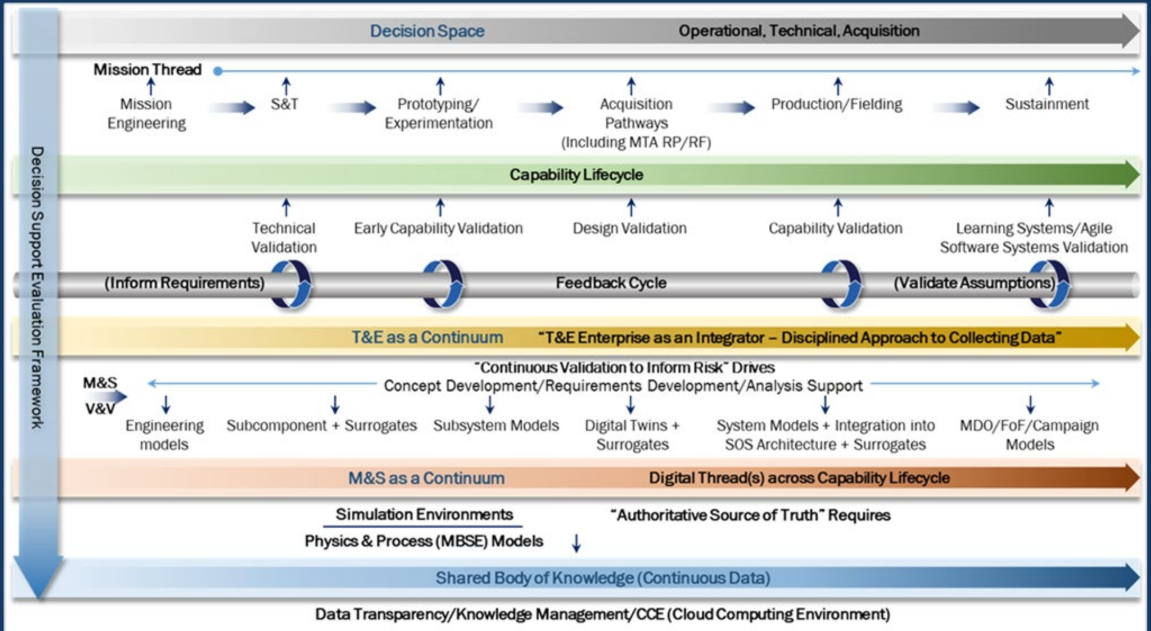
The example use case diagram recommendations, developed within SysML, can be translated to other frameworks.

Systems Modeling Language (SysML)	Department of Defense Architecture Framework (DoDAF)																			Unified Architecture Framework (UAF)										
	CV's	CV-5	DIV's	DIV-2	OV's	OV-1	OV-2	OV-4	OV-5a	OV-5b	OV-6b	OV-6c	SV-1	SV-2	SV-4	SV-5	SV-7	SV-10b	SV-10c	Taxonomy (Tx)	Structure (Sr)	Connectivity (Cn)	Processes (Pr)	States (St)	Interaction Scenarios (Is)	Information (If)	Parameters (Pm)	Constrains (Ct)	Summary & Overview (Sm-Ov)	Requirements (Req)
Activity Diagram										X					X	X							X							
Block Definition Diagram	X		X			X	X	X	X				X	X						X	X					X	X	X	X	
Internal Block Diagram							X							X							X	X								
Parametric Diagram																	X											X		
Requirement Diagram/Table		X		X																										X
Requirements Matrix					X																									X
Sequence Diagram												X							X						X					
State Machine Diagram											X							X						X						
Use Case Diagram/Matrix						X									X															

More research is planned to translate these recommendations into additional MBSE architectures beyond those already highlighted and to provide further examples.

Conclusions and Next Steps

- MBSE for T&E of AIES is an example of the value of end-to-end Digital Engineering
  - ➢ Aligned to and enables DTE&A's developmental Test and Evaluation as a Continuum.[1]
  - ➢ Potential risks associated with AI Model integration, operational employment, cybersecurity resilience, user adoption, and AI model sustainment can be identified, thereby enabling more effective T&E.
- Planned Future Work
  - ➢ Measure the utility of each of the recommendations for T&E professionals evaluating AIES.
  - ➢ Explore more complex AIES implementations (e.g., multiple AI models working together) and further MBSE enablement opportunities.
  - ➢ Prototype tools that can automate the processing of AI data in an MBSE product into T&E information and/or application of T&E to AIES.
  - ➢ T&E of AIES Policy is being reviewed to ensure enablement for AIES.



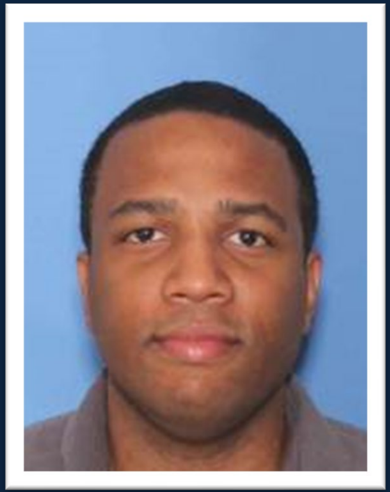
C. Collins and K. Senechal, "Test and Evaluation as a Continuum," *The ITEA Journal of Test and Evaluation*, vol. 44, no. 1, 2023.

MITRE

Research Team

Thank you!

Carol Pomales
cpomales@mitre.org



Dr. James Morris-King
jamesmk@mitre.org

Tai Jella
tjella@mitre.org

Bill Fetech
wfetech@mitre.org

MITRE

Backup

MITRE

Content References

[1]Aggarwal, A., Shaikh, S., Hans, S., Haldar, S., Ananthanarayanan, R. and Saha, D., 2021, May. Testing framework for black-box AI models. In 2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion) (pp. 81-84). IEEE.
[2]Awadid, A., Robert, B. and Langlois, B., 2024, February. MBSE to Support Engineering of Trustworthy AI-Based Critical Systems. In 12th International Conference on Model-Based Software and Systems Engineering.
[3]Ben, H. B. and Foutse, K. 2020. On testing machine learning programs. Journal of Systems and Software 164 (2020), 110542.
[4]Blasch, E., and Pokines, B. Analytical Science for Autonomy Evaluation. 2019.
[5]Blasch, E., Ravela, S., and Aved, A., Eds. Handbook of Dynamic Data Driven Applications Systems. Springer International Publishing, 2018
[6]Borg, M., Englund, C., Wnuk, K., Duran, B., Levandowski, C., Gao, S., Tan, Y., Kaijser, H., Lönn, H., and Törnqvist, J. 2018. Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry. arXiv preprint arXiv:1812.05389 (2018).
[7]Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. and Mukhopadhyay, D., 2021. A survey on adversarial attacks and defences. CAAI Transactions on Intelligence Technology, 6(1), pp.25-45.
[8]Ciampa, P.D., La Rocca, G. and Nagel, B., 2020. A mbse approach to mdao systems for the development of complex products. In AIAA Aviation 2020 Forum (p. 3150).
[9]Delligatti, L. 2014. SysML Distilled, A Bried Guide to the Systems Modeling Language, Addison-Wesley, 2014, page 4
[10]Delligatti, L., 2013, SysML Distilled: A Brief Guide to the Systems Modeling Language. Addison Wesley Professional, 1st edn. 2013
[11]Duan, Y., Edwards, J.S. and Dwivedi, Y.K., 2019. Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. International journal of information management, 48, pp.63-71.
[12]Elizamary, N., Anh, ND., Ingrid, S., and Tayana, C. 2020. Software engineering for artificial intelligence and machine learning software: A systematic literature review. arXiv preprint arXiv:2011.03751 (2020).
[13]Ellen, L. L., Aiswarya, R., Ivica, C., Jan, B., and Holmström, O. H. 2020. Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. Information and Software Technology 127 (2020), 106368.
[14]Friedenthal, S., Moore, A., and Steiner, R. 2015. A Practical Guide to SysML – The Systems Modeling Language, Third Edition, Elsevier Inc, 2015, page 17.
[15]Fumihiro, K. 2019. Software engineering challenges for machine learning applications: A literature review. Intelligent Decision Technologies 13, 4 (2019), 463–476.
[16]Gedo, C., 2012. Model Based Systems Engineering and Systems Modeling Language. DISA DoDAF Plenary
[17]Giuliano, L., Paulo, A., Nathalia, N., and Donald, C. 2021. Machine learning model development from a software engineering perspective: A systematic literature review. arXiv preprint arXiv:2102.07574 (2021)
[18]Görkem, G. 2021. A software engineering perspective on engineering machine learning systems: State of the art and challenges. Journal of Systems and Software 180 (2021), 111031.
[19]Groll, E. 2023, Fifty minutes to hack ChatGPT: Inside the DEF CON competition to break AI, in Cyberscoop, https://cyberscoop.com/def-con-ai-hacking-red-team/
[20]Hallqvist, J. and Larsson, J., 2016, July. Introducing MBSE by using systems engineering principles. In INCOSE International Symposium (Vol. 26, No. 1, pp. 512-525).
[21]Henderson, K., McDermott, T., Van Aken, E. and Salado, A., 2023. Towards developing metrics to evaluate digital engineering. Systems Engineering, 26(1), pp.3-31.
[22]Hernández-Orallo, J., 2017. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artificial Intelligence Review, 48, pp.397-447.
[23]International Council on Systems Engineering (INCOSE), Systems Engineering Vision 2020, Version 2.03, TP-2004-004-02 September 2007.
[24]Kerzhner, A.A., Tan, K. and Fosse, E., 2015. Analyzing cyber security threats on cyber-physical systems using Model-Based Systems Engineering. In AIAA SPACE 2015 Conference and Exposition (p. 4575).
[25]Martínez-Fernández, S., Bogner, J., Franch, X., Oriol, M., Siebert, J., Trendowicz, A., Vollmer, AM., and Wagner, S. Software engineering for AI-based systems: a survey. ACM Transactions on Software Engineering and Methodology (TOSEM). 2022 Apr 1;31(2):1-59.
[26]Mary, J. M., Holmström, O. H., and Jan, B. [n.d.]. Architecting AI deployment: A systematic review of state-of-the-art and state-of-practice literature.
[27]Masuda, A., Ono, K., Yasue, T., and Hosokawa, N. 2018. A survey of software quality for machine learning applications. In 2018 IEEE International conference on software testing, verification and validation workshops (ICSTW). IEEE, 279–284.
[28]McGrath, A. and Jonker, A. 2023. “What is model-based systems engineering (MBSE)?”, IBM, What is model-based systems engineering (MBSE)? | IBM
[29]Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), pp.1-35.
[30]Mittal, S., Zeigler, B.P., Martin, J.L.R., Sahin, F. and Jamshidi, M., 2008. Modeling and simulation for systems of systems engineering. Systems of Systems–Innovations for the 21st Century (to be published by Wiley).
[31]OMG, 2019. Unified Architecture Framework (UAF) Traceability Document number: dtc/19-05-15
[32]Pandolf, J., 2023. Investigation of Model-Based Systems Engineering Integration Challenges and Improvements (Doctoral dissertation, Massachusetts Institute of Technology).
[33]Parikh, R.B., Teeple, S. and Navathe, A.S., 2019. Addressing bias in artificial intelligence in health care. Jama, 322(24), pp.2377-2378.
[34]Patterson, E.A., 2017. Utilizing SysML viewpoints to improve understanding and communication of human mental models within system design teams. The University of Alabama in Huntsville.
[35]Riccio, V., Jahangirova, G., Stocco, A., Humbatova, N., Weiss, M., and Tonella, P. 2020. Testing machine learning based systems: a systematic mapping. Empirical Software Engineering 25, 6 (2020), 5193–5254. https://doi.org/10.1007/s10664-020-09881-0
[36]Rimani, J., Lesire, C., Lizy-Destrez, S. and Viola, N., 2021, August. Application of MBSE to model Hierarchical AI Planning problems in HDDL. In ICAPS 2021 (p. 0).
[37]Rogers III, E.B. and Mitchell, S.W., 2021. MBSE delivers significant return on investment in evolutionary development of complex SoS. Systems Engineering, 24(6), pp.385-408.
[38]Role, A. and Role, D., 2011. The DoDAF Architecture Framework Version 2.0.
[39]Saeed, W. and Omlin, C., 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, 263, p.110273.
[40]Schräder, E., Bernijazov, R., Foullois, M., Hillebrand, M., Kaiser, L. and Dumitrescu, R., 2022, October. Examples of ai-based assistance systems in context of model-based systems engineering. In 2022 IEEE International Symposium on Systems Engineering (ISSE) (pp. 1-8). IEEE.
[41]Serban, A. and Visser, J. 2021. An Empirical Study of Software Architecture for Machine Learning. arXiv preprint arXiv:2105.12422 (2021).
[42]Serban, A., van der Blom, K., Hoos, H. and Visser, J., 2020, October. Adoption and effects of software engineering best practices in machine learning. In Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (pp. 1-12).
[43]Systems Modeling Language (Wikipedia.com), accessed 23 January 2024. https://en.wikipedia.org/wiki/Systems_modeling_language.
[44]The Clinger Cohen Act of 1996, 40 U.S.C. 1401 et seq., https://www.cio.gov/handbook/it-laws/clinger-cohen-act/
[45]Vincenzo, R., Gunel, J., Andrea, S., Nargiz, H., Michael, W., and Paolo, T. 2020. Testing machine learning based systems: A systematic mapping. Empirical Software Engineering 25, 6 (2020), 5193–5254
[46]Von Eschenbach, W.J., 2021. Transparency and the black box problem: Why we do not trust AI. Philosophy & Technology, 34(4), pp.1607-1622.
[47]Wang, S., Huang, L., Ge, J., Zhang, T., Feng, H., Li, M., Zhang, H. and Ng, V., 2020. Synergy between machine/deep learning and software engineering: How far are we?. arXiv preprint arXiv:2008.05515.
[48]Washizaki, H., Uchida, H., Khomh, F. and Guéhéneuc, Y.G., 2019, December. Studying software engineering patterns for designing machine learning systems. In 2019 10th International Workshop on Empirical Software Engineering in Practice (IWESEP) (pp. 49-495). IEEE.
[49]White House Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence | The White House – https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/
[50]Zhang, J.M., Harman, M., Ma, L. and Liu, Y., 2020. Machine learning testing: Survey, landscapes and horizons. IEEE Transactions on Software Engineering, 48(1), pp.1-36.

MITRE