A Systems Engineering Perspective on AI Test and Evaluation: Explainability and Counterfactuals

Ali K. Raz
George Mason University
Fairfax, VA, USA
araz@gmu.edu

Erik P. Blasch
MOVEJ Analytics
Fairborn, OH, USA
erik.blasch@gmail.com

Kuo-Chu Chang
George Mason University
Fairfax, VA, USA
kchang@gmu.edu

INTRODUCTION[1]

Artificial intelligence and machine learning (AI/ML) are becoming ubiquitous in modern systems, and the information fusion community has seen a recent surge in AI/ML-driven solutions for data fusion challenges. Advancements in AI/ML algorithms and technologies are quickly finding their way into software- and even hardware-based components of complex systems, often enabling unprecedented capabilities in performance and efficiency. Deep neural networks (DNNs) typically serve as the cornerstone for implementing modern AI/ML algorithms, encompassing supervised, unsupervised, and reinforcement learning (RL) paradigms. As DNNs and their underlying AI/ML implementations inevitably become integral components of complex systems, it is imperative to approach the design, development, integration, and testing of AI/ML from a holistic systems perspective. In this article, we advocate for the incorporation of systems engineering (SE) principles into the realm of AI/ML and discuss two emerging approaches—explainable AI (XAI) and counterfactual test and evaluation (cT&E)—to aid toward building a systems perspective of AI/ML implementation and deployment.

FROM AI/ML IMPLEMENTATION TO INTELLIGENT ENGINEERED SYSTEMS

An engineered system (ES) is essentially a collection of components that interact with each other and their operational environment to fulfill an intended purpose that cannot be achieved by the individual components alone. With the integration of AI/ML technology into these system components, the underlying ES transitions to a new class of intelligent engineered systems (IESs) with machine autonomy. The IESs present a unique set of challenges, stemming from the complexities inherent from traditional ES as well as those arising from the incorporation of AI/ML technologies. The AI/ML methods are broadly classified into ruled-based (e.g., expert systems), model-based (e.g., mathematical model), and data-driven, DNN-based methods [2]. The rule-based and model-based methods, which have been established for decades, are considered mature, well understood, and have been extensively studied for design, development, and integration into complex systems; the same cannot be said about the contextual and data-driven behavioral characteristics of DNNs.

The DNN implementation involves creating intricate mapping between inputs and outputs (I/O) through multiple hidden layers, using large datasets in supervised or unsupervised learning and environment/reward models in RL. The curated datasets are typically divided into training and validation sets, with roughly 80% of the data used for training the algorithm and the remaining 20% for validating the learning outcomes. Once trained, the DNN effectively operates as a black (i.e., invisible) box, lacking interpretable information regarding the decision-making processes within the hidden layers and the underlying I/O map (Figure 1 [1]). For example, once a DNN is trained for multimodal data fusion, it is difficult to know which sensor data input contributed more to the decision output; how much data are required to mitigate a future data imbalance situation; what the intended use or who the intended user is [3], and the extent to which information is labeled for data association [4].

Regarding IES, a DNN black box represents a component in a complex system that must effectively interact with other systems' components and the operational environment to meet the system's intended purpose. However, DNN development has primarily focused on algorithmic advancements and computational efficiencies, particularly within specific data and application domains such as image or face recognition; this presents a major limitation in AI/ML development, and challenges associated with integrating DNNs with other system components have often been neglected [5].

CHALLENGES OF AI/ML IMPLEMENTATION IN AN IES

Employing DNNs to integrate AI into IES poses several challenges because of their backbox nature. Relying solely on an 80-20 data split for training and validation, DNNs cannot be guaranteed to be fit for purpose to meet the overarching systems objectives. These objectives emphasize the necessity for systems to not only achieve their intended purposes but also to exhibit resilience to real-world operations, minimize unintended actions, address adverse effects, and acknowledge consequences. In complex systems, the behavior of a system emerges

[1] The ideas presented in this article are adapted from an earlier publication [1].

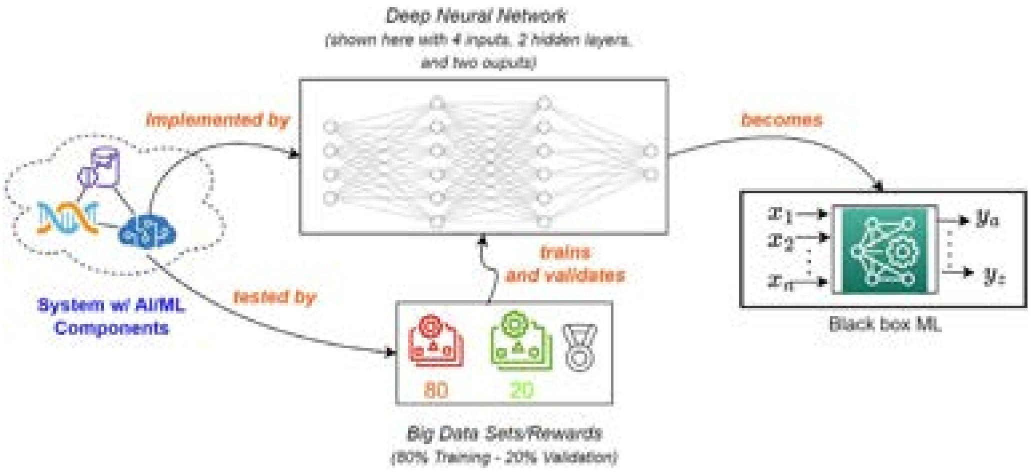

Figure 1
Black box nature of DNNs within an intelligent engineered system [1].

from the interactions between system components and their environment [6]. In this regard, the implementation and integration of opaque AI/ML must include broader IES considerations to ensure compatibility with the complex system dynamics and objectives.

The AI/ML and the SE communities now recognize the limitations for the testing, evaluation, and integration of DNNs into IES. These limitations stem from the lack of robust systems methods, varying and inadequate evaluation methods, and limited approved standards. Considering the wider system operational considerations and their manifestation on DNN training and validating datasets, Barclay Brown's book, *Engineering Intelligent Systems,* postulates this problem as "the green school bus problem" [7]. This hypothetical problem posits that an AI system trained on a dataset primarily comprising military vehicles (typically green in color) may likely classify a green school or commercial bus as a military vehicle unless the dataset includes examples of green school buses (Figure 2). Although green school (and commercial) buses are rare in the United States, their existence is not impossible. A green bus scenario highlights the inherent biases and limitations within AI systems when they are not adequately trained on diverse and representative datasets [7].

Figure 2
Green bus (*from* https://commons.wikimedia.org/wiki/File:The_Green_Bus_school_bus_381_Volvo_Olympian_Northern_Counties_Palatine_II_R381_LGH_in_Birmingham_2_November_2008.jpg).

Judea Pearl, one of the leading researchers in AI/ML, has highlighted the lack of, and need for, cause and effect understanding of AI/ML methods [8], [9]. The absence of structured process models for designing, testing, and integrating AI/ML models has resulted in the lack of reproducibility of AI algorithms [10]. It is vital that these structured system life-cycle development models from SE (e.g., Vee Model [11]) become commonplace in AI/ML to ensure rigorous and replicable AI-development processes. Similarly, the technical debt of AI/ML algorithms, where the full cost and implications are not recognized until the integration stages, is gaining attention. This AI/ML oversight is often attributed to lack of operational considerations during the design and construction phases of DNN algorithms [12]. Furthermore, little consideration has been given for the operations, maintenance, and sustainment of models to evolve with changing situations.

In an IES, AI/ML may handle not only various decisions but also their interactions with other system components with and without AI/ML. These interactions can lead to emergent behaviors—positive and negative—that have significant implications for system performance and safety. The system engineers and designers must strive to establish and validate confidence toward testable, repeatable, and auditable actions, outputs, and decisions made by AI/ML systems. Additionally, it is critical to develop an understanding of failure mechanisms, modes, and consequences, along with effective failure mitigation techniques for certification and assurance. Furthermore, the AI/ML algorithms integrated into IES must function not only as intended within narrowly defined use cases, as dictated by the training and validation data set, but also must effectively operate within the broader operating envelop of the ES. Deployment requires careful consideration of system dynamics, potential interactions, and the robustness of AI/ML algorithms [13].

SE PERSPECTIVE FOR AI/ML CHALLENGES

The SE body of knowledge includes several system concepts and principles that facilitate stakeholder analysis, conceptual design, T&E, and verification and validation of ES. For illustration and discussion purposes, the underlying SE philosophy can be simplified by the Vee Model, which guides the development and evolution of a system from its inception of necessity to throughout its entire system lifecycle (Figure 3 [14]–[16]). The outer yellow boxes highlight modification to process model to include AI/ML components which consist for iterative refinement of SE artifacts with design and experimentation.

The SE approach has been identified as instrumental in expediting the integration of AI into practical systems [16]. Moreover, SE offers a framework for addressing notable challenges encountered by the AI community. Hendrycks et al. have recently delineated four pertinent unsolved challenges in AI/ML, mostly relating to the safety of ML algorithms [17]. These challenges include robustness, monitoring, alignment, and external safety, collectively indicating a lack of SE practices for AI/ML. Table 1 summarizes the AI/ML challenges and aligns them with common SE principles that directly address similar issues encountered in the development of a complex system [1], [11], [17], [18].

The major problem areas for AI/ML safety can typically be found in most introductory texts on SE as issues for most complex systems. Nevertheless, the current challenge for SE is to discover how to perform the SE activities and evaluation thereof (e.g., failure mode analysis, sensitivity analysis) with a lack of requirements and no knowledge of the AI/ML conceptual design or decision-making constructs. The SE goal is to provide testable, repeatable, and auditable actions, outputs, and decisions for AI/ML integration into IES.

Of particular interest here, these challenges compound when AI/ML is employed in the information fusion systems. The AI/ML challenges intersect with fusion challenges that involve inherent uncertainty and a multitude of heterogeneous sources, along with multitiered and interacting fusion processes in both low- and high-level fusion contexts.

Integration of opaque AI/ML must include broader IES considerations to ensure compatibility with the complex system dynamics and objectives.

In the following, two emerging approaches are discussed that, in the authors' views, should become part of standard SE practice for test, evaluation, and integration of AI/ML into IES. These approaches are particularly valuable for information fusion systems because they serve to concurrently address both AI/ML challenges and fusion system challenges from a systems perspective.

EXPLAINABLE AI

XAI transforms the opacity of AI/ML components and the underlying DNN to transparent models that are understandable and interpretable by humans [19]. XAI is an emerging area of research with many proposed approaches offering various forms of explanations; it can be used for understanding and interpreting

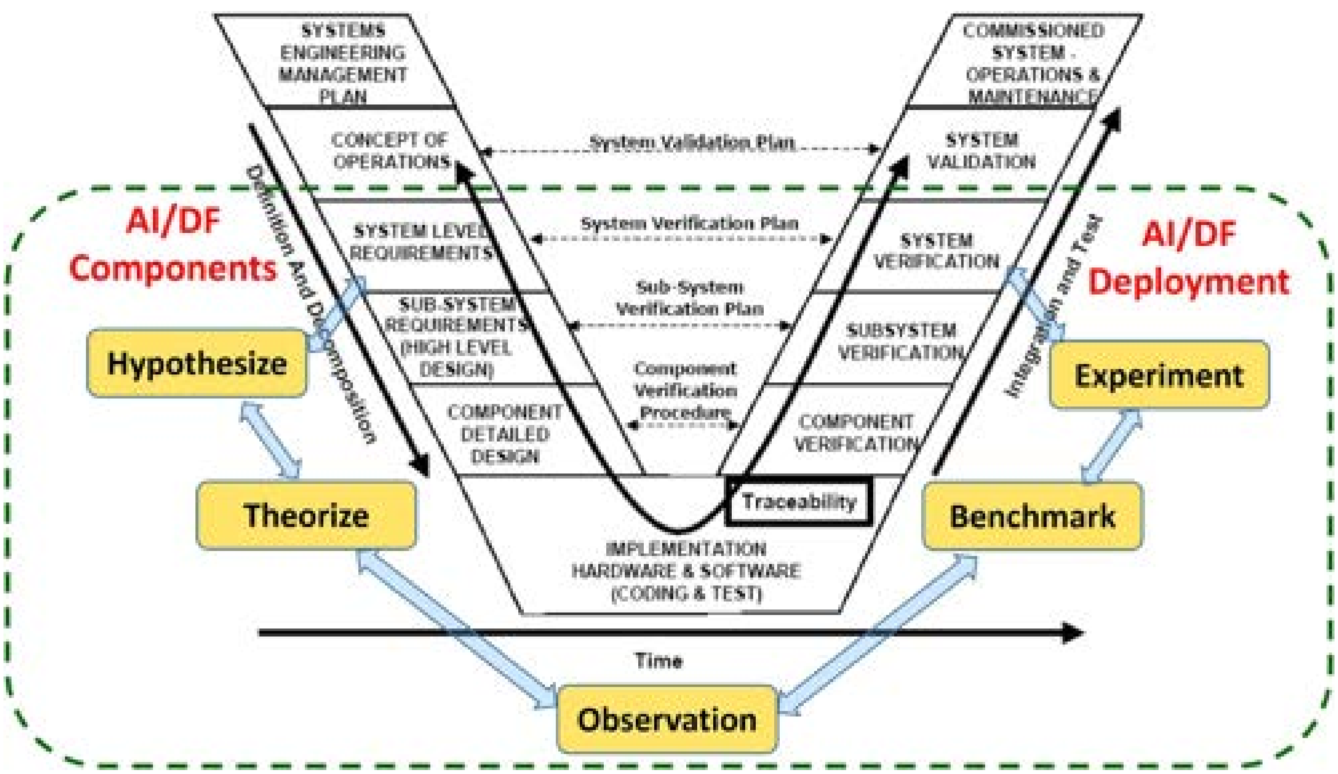

Figure 3
Systems engineering classical Vee Model in gray (adapted from references [14]–[16]).

Table 1

Unsolved Problems in ML and Proposed SE Solutions [1]

ML Unsolved Problem [17]	Systems Engineering Principles and Artifacts as Potential Solutions [11], [18]
Robustness: resilience, black swan events, unusual events	System verification and validation, failure modes and effects analysis (right side of Vee Model)
Monitoring: unexpected model functionality, malicious use	Emergent behavior and interaction analysis, functional analysis (left side of Vee Model) Trade space exploration, design of experiments (both sides of Vee Model)
Alignment: "optimize with difficult to specify human values"	Stakeholder analysis, use case analysis, concept of operations (left side of Vee Model)
External safety: address risks, cyber attacks	External systems diagram, external interfaces, and context analysis (left side of Vee Model)

decision-making constructs of components embedded with AI/ML algorithms. These XAI approaches span from visualizing high-dimensional I/O data spaces and simplifying DNNs model with causality to identifying most the relevant features in the input space that influence the output of DNNs at any given time. Barredo et al. have provided a comprehensive overview of the XAI, while also highlighting its value and applications in the context of information fusion [19].

The absence of structured process models for designing, testing, and integrating AI/ML models has resulted in the lack of reproducibility of AI algorithms.

Employing XAI with the SE standard practice is key for addressing the opacity of the DNNs, establishing component behavior expectations, and assessing how a component with DNNs will interact with other components of the system with and without DNNs. Insights gained from XAI can help developers and testers examine AI system design and implementation issues; it can also help with sharing DNN outcomes and decision-making constructs with stakeholders and subject matter experts (SMEs), as well as support transparency and interpretability.

COUNTERFACTUAL TEST AND EVALUATION

cT&E is used for understanding DNN limitations and testing conformance to expected outcomes under hypothetical scenarios not typically included in training. It investigates the hypothetical "what-if" scenarios to find conditions in inputs, which provokes internal system faults and latent interactions, that could produce an imaginative desired or undesired result [20]. cT&E includes creatively designing metrics for evaluation, developing counterintuitive—and perhaps unlikely—use cases, and systematic design of experiments.

The development and training of DNNs does not include all conceivable input combinations; therefore, a full spectrum response of deployed DNN remains unknown. The unknown implies that regions of DNNs (within its hidden layers) may never be invoked in a traditional T&E sense. These unknowns are addressed through comprehensive cT&E by proposing hypothetical scenarios and subsequently analyzing the corresponding system behavior. By imaging and exploring different scenarios and examining DNN outcomes, stakeholders can identify bounds of pragmatic use and ensure that AI systems remain safe and adhere to expected outcomes, thereby mitigating risks and enhancing confidence in their deployment.

INFORMATION FUSION SYSTEM APPLICATION OF EXPLAINABILITY AND COUNTERFACTUALS

To illustrate XAI and cT&E for IES with information fusion, we briefly consider the following two applications.

First, a conceptual, high-level information fusion (HLIF) system designed to provide situational awareness based on inputs from heterogeneous sensors is considered. Recent advances in HLIF explore the integration of DNNs with promising results; however, a significant challenge lies in comprehending the decision-making processes within these underlying DNNs. For example, imagine a HLIF DNN tasked with fusing inputs from three sensors to determine the corresponding situation and produce an output (Figure 4[a]). Although traditional T&E approaches may focus on assessing the timeliness and performance of this HLIF DNN, the fusion engineers often lack insight into how sensor inputs are transformed into outputs.

The sensor fusion aggregation for the situational awareness challenge can be addressed by employing feature relevance explainability techniques, such as Shapley Additive Explanations (SHAP) [21]. When applied to the DNN illustrated in Figure 4(a), SHAP enables the derivation of an analytical expression, directly linking input values to the output (Figure 4[b]). By using this analytical expression, fusion SMEs can gain an understanding of whether the HLIF DNN aligns with the established knowledge base for its application.

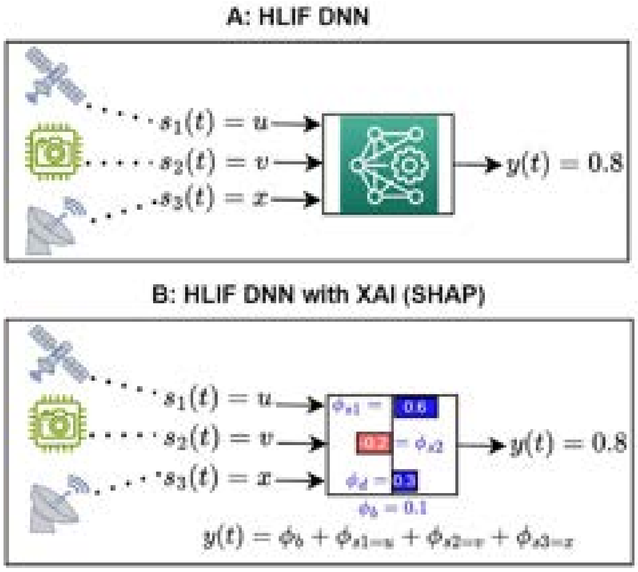

Figure 4
HLIF DNN with multisensory input: (a) black box HLIF DNN; (b) HLIF DNN with explainability.

In the second example, the causal Bayesian Network (BN) method is used to compare and prioritize counterfactual hypotheses for sensor allocation using an RL algorithm in a space situation awareness scenario [22], [23]. RL algorithms, implemented as multilayer DNN, use a model of the environment characterized by a state space, reward structure, and action space to train an RL agent. The agent is tasked with making decisions in dynamic and uncertain environments; however, once trained, the RL agents essentially operate as black boxes, lacking interpretability in their decision-making processes. In the proposed engineered explainable counterfactual evaluation Bayes Net (ExIcEBN) approach, we first construct a baseline "observational" BN model to evaluate the expected reward of an action based on the current environment state and the sensor allocation decision derived from the RL agent. Subsequently, we employ the twin networks model concept to predict the hypothetical world consequent of an event given a potential (counterfactual) antecedent [24].

XAI transforms the opacity of AI/ML components and the underlying DNN to transparent models that are understandable and interpretable by humans.

In twin networks model, a "factual world" twin represents what actually occurred in the event, whereas a "counterfactual world" indicates what could have happened to the "factual world" twin had the antecedent been different. The two worlds share a common set of domain-specific conditions. The evidence from the factual world is used to update past information on the shared contextual variables. The updated information is then used to predict the hypothetical outcome in the counterfactual world, considering both the updated contextual variables and the newly established antecedent [24], [25].

For example, consider a system event where a specific action (A = A1) was taken, and a high level of risk (R = high) was observed based on the estimated target uncertainty and activity. To assess the system under different potential circumstances, a counterfactual hypothesis might be proposed: "If the action had been different, the risk level could have been lower". The low-risk hypothesis aims to address the question, "Could the risk level have been lower if a different action had been taken?" (Figure 5).

To predict the hypothetical world outcome based on the updated information on the contextual variables and the newly established antecedent, we apply *do*-operator on the "control" variable [25]. The *do*-operator facilitates an intervention in the counterfactual world by enforcing a different value for the antecedent variable from the ones observed in the factual world and removes all incoming edges to the variable (Figure 5).

The counterfactual query process enables decision makers to compare the actual occurrences in the real world with what would have happened under a different scenario in a hypothetical world. The process can also assess whether the situation

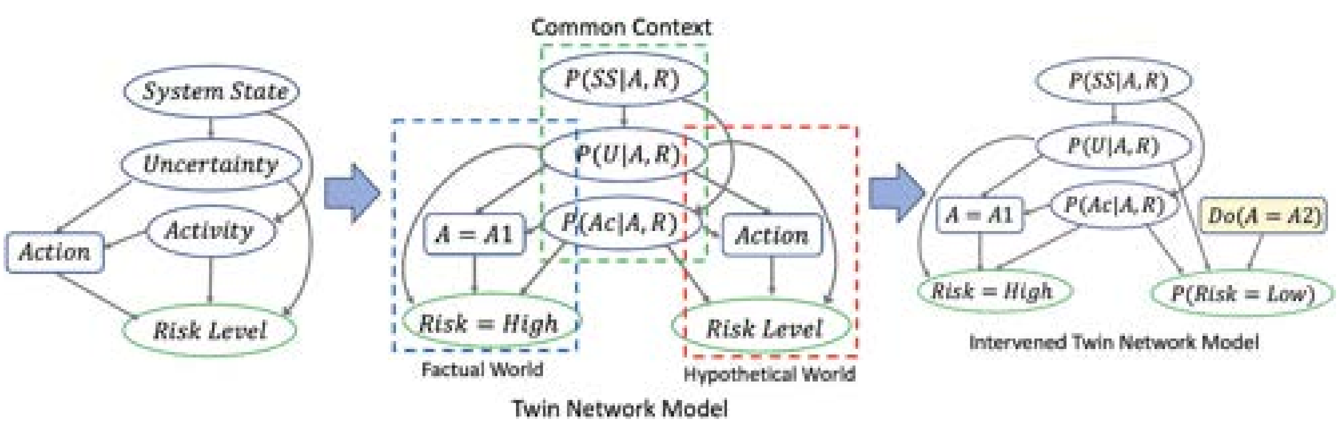

Figure 5
A twin network model example.

could have been managed more effectively if a different action had been taken or if additional information (e.g., target activity) had been known during the original assessment phase of the decision-making process.

CONCLUSIONS

Adopting AI/ML methods for IES and information fusion systems requires integrating these technologies into the system lifecycle from its inception to deployment. To appropriately understand and mitigate unintended outcomes and to ensure the development of safe and reliable systems, rigorous and thorough T&E and verification and validation processes are indispensable. This article highlights the need to employ explainable methods and counterfactual exemplars to better manage expectations, contextual use, and bound performance. Moreover, leveraging data fusion to reduce uncertainty in engineering information systems underscores the imperative to expand the SE Vee Model. The Vee Model expansion should include benchmarks for standardization, comparison, and certification, thereby providing a structured framework for evaluating the effectiveness and reliability of AI/ML integration within information fusion systems and future IES.

REFERENCES

1. Raz, A. K., Miller, W., Chang, K.-C., Lin, Y., and Blasch, E. Explainable AI and counterfactuals for test and evaluation of intelligent engineered systems. *INCOSE International Symposium*, Vol. 33, 1, (2023), 1524–1539, doi:10.1002/iis2.13096
2. Majumder, U. K., Blasch, E. P., and Garren, D. A. *Deep Learning for Radar and Communications Automatic Target Recognition*. Norwood, MA: Artech House, 2020.
3. Blasch, E., Cruise, R., Aved, A., Majumder, U., and Rovito, T. Methods of AI for multimodal sensing and action for complex situations. *AI Magazine*, Vol. 40, 4 (2019), 50–65, doi:10.1609/aimag.v40i4.4813
4. Blasch, E., et al. Machine learning/artificial intelligence for sensor data fusion–opportunities and challenges," *IEEE Aerospace and Electronic Systems Magazine*, Vol. 36, 7 (July 2021), 80–93, doi:10.1109/MAES.2020.3049030
5. Freeman, L. Test and evaluation for artificial intelligence. *INSIGHT*, Vol. 23, 1 (2020), 27–30, doi:10.1002/inst.12281
6. Raz, A. K., Llinas, J., Mittu, R., and Lawless, W. F. Engineering for emergence in information fusion systems: a review of some challenges. In *Human-Machine Shared Contexts*, W. F. Lawless, R. Mittu, and D. A. Sofge, Eds. Cambridge, MA: Academic Press, 2020, pp. 241–255. doi:10.1016/B978-0-12-820543-3.00012-2
7. Brown, B. *Engineering Intelligent Systems*. Hoboken, NJ: Wiley, 2022, https://www.oreilly.com/library/view/engineering-intelligent-systems/9781119665595/, last access Dec. 13, 2022 [Online].
8. Pearl, J. Reasoning with cause and effect. *AI Magazine*, Vol. 23, 1 (Mar. 2002), doi:10.1609/aimag.v23i1.1612
9. Pearl, J., and Mackenzie, D. I. AI can't reason why. *Wall Street Journal*, May 18, 2018. https://www.wsj.com/articles/ai-cant-reason-why-1526657442, last access Apr. 27, 2020 [Online].
10. S. Higginbotham, Hey, data scientists: show your machine-learning work. *IEEE Spectrum*, 2020. https://spectrum.ieee.org/computing/software/hey-data-scientists-show-your-machinelearning-work, last access Apr. 27, 2020 [Online].
11. INCOSE. *INCOSE Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities* (4th ed.). Hoboken, NJ: Wiley, 2015.
12. D. Sculley et al., Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, 2503–2511, last access Apr. 29, 2020 [Online]. http://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf
13. Blasch, E., Bin, J., and Liu, Z. Certifiable artificial intelligence through data fusion. arXiv, last access Nov. 2, 2021. doi:10.48550/arXiv.2111.02001
14. Kossiakoff, A., Biemer, S., Seymour, S., and Flanigan, D. *Systems Engineering Principles and Practice* (3rd ed.). Hoboken, NJ: Wiley, 2020.
15. Blasch, E., Raz, A. K., Sabatini, R., and Insaurralde, C. Information fusion as an autonomy enabler for UAV traffic management. In *AIAA Scitech 2021 Forum*, American Institute of Aeronautics and Astronautics, 2021. doi:10.2514/6.2021-0658
16. Raz, A. K., Blasch, E. P., Guariniello, C., and Mian, Z. T. an overview of systems engineering challenges for designing AI-enabled aerospace systems. In *AIAA Scitech 2021 Forum*, American Institute of Aeronautics and Astronautics, 2021. doi:10.2514/6.2021-0564
17. Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ML safety. *ArXiv210913916 Cs*, Apr. 2022. Last access May 7, 2022 [Online]. http://arxiv.org/abs/2109.13916
18. Buede, D. M., and Miller, W. *The Engineering Design of Systems Models and Methods* (3rd ed.). Hoboken, NJ: John Wiley & Sons, 2016.
19. Barredo Arrieta, A., et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, Vol. 58 (Jun. 2020), 82–115. doi:10.1016/j.inffus.2019.12.012
20. Roese, N. J. Counterfactual thinking. *Psychological Bulletin.* Vol. 121, 1 (1997), 133–148. doi:10.1037/0033-2909.121.1.133
21. Lundberg, S. M., and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Last access Dec. 13, 2022 [Online]. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
22. Fan, Z., Chang, K.-C., Raz, A. K., Harvey, A., and Chen, G. Sensor tasking for space situation awareness: Combining reinforcement learning and causality. In *2023 IEEE Aerospace Conference*, Big Sky, MT, IEEE, 2023, 1–9.
23. Fan, Z., Chang, K.-C., Chen, G., Khan, S., and Franco, M. Explainable deep reinforcement learning for space situational awareness: counterfactual explanation approach. In *2024 IEEE Aerospace Conference*, Big Sky, MT, IEEE, Mar. 2024.
24. Pearl, J. *CAUSALITY* (2nd ed.). Los Angeles, CA: Cambridge University Press, 2009. Last access Dec. 13, 2022 [Online]. http://bayes.cs.ucla.edu/BOOK-2K/
25. Pearl, J., and Mcckenzie, D. *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books, 2018.

Ali K. Raz, Ph.D. CSEP, is an assistant professor at George Mason University Systems Engineering and Operations Research department and an assistant director of Intelligent Systems and Integration at the C5I Center. His research and teaching interests are understanding collaborative autonomy and developing systems engineering methodologies for integrating autonomous systems with machine learning and information fusion. He holds a BSc. and MSc. in electrical engineering from Iowa State University and a Ph.D. in aeronautics and astronautics from Purdue University. He is a co-chair of INCOSE AI Working Group and a senior member of AIAA and IEEE.

Erik P. Blasch, Ph.D., is a program officer with the Air Force Research Laboratory. He received his B.S. in mechanical engineering from the Massachusetts Institute of Technology and his Ph.D. in electrical engineering from Wright State University in addition to Master's degrees in mechanical engineering, industrial engineering, electrical engineering, medicine, military studies, economics, and business. Additionally, his assignments include USAF Officer (retired)., adjunct associate professor, and president of professional societies. He is an experienced engineer with a demonstrated history of working in government administration, military operations, academic settings, and industrial innovation. His research includes information fusion, space-aware tracking, industrial avionics, and human factors designs, compiling 11 books, 62 patents, 170+ journal articles, and 500+ papers. He is a Fellow of AIAA (astronautics), IEEE (electrical), MSS (sensing), RAeS (aerospace), and SPIE (optical) societies.

Kuo-Chu Chang, Ph.D., is a professor of systems engineering and operations research department at George Mason University. He received his M.S. and Ph.D. degrees in electrical engineering from the University of Connecticut. His research interests include estimation theory, data fusion, Bayesian inference, machine learning, and financial engineering. He has more than 40 years of industrial and academic experience and has published more than 300 papers in the areas of multitarget tracking, distributed sensor fusion, Bayesian Networks technologies, and quantitative finance. He is an IEEE Fellow.

2023 marks the 25th Anniversary of the International Society of information Fusion. As part of this celebration, we would like to honor and remember not only the technical achievements in our field, but also the people, places, events, and more. ISIF is collecting videos, photos, and short stories (250 words max.) from its members. Please find the form to make your contribution on: https://isif.org/isif-25th-anniversary-celebration-0

Disclaimer: This content may be used on our social media, website site, marketing materials, and more. Content must be owned by the submitter and cannot contain copyrighted material. By submitting photos, videos, and written content, you grant ISIF the right to use and distribute the submitted materials.