2026 NDIA MICHIGAN CHAPTER
GROUND VEHICLE SYSTEMS ENGINEERING
AND TECHNOLOGY SYMPOSIUM
***DIGITAL ENGINEERING / SYSTEMS ENGINEERING* TECHNICAL SESSION**
AUGUST 11-13, 2026 - NOVI, MICHIGAN

AI-POWERED SAFETY INTELLIGENCE AS A DECISION ENABLER FOR GROUND VEHICLE MBSE: THE DIGITAL SAFETY TWIN APPROACH

Michael Wagner[1], Nelson Santini[1], and Anoop Balakrishnan[1]
[1]Edge Case Research, Inc., Pittsburgh, PA

ABSTRACT

Model-Based Systems Engineering (MBSE) has become a mandated practice for Department of Defense acquisition programs, yet measured benefits remain elusive. The 2024 Defense Science Board found that less than one percent of published literature actually quantified MBSE outcomes, and flagship ground vehicle programs such as the XM30 Infantry Fighting Vehicle have experienced schedule delays attributed directly to insufficient proficiency with model-based approaches. This paper presents the Digital Safety Twin concept: an AI-powered safety intelligence architecture that addresses three of the most labor-intensive and error-prone MBSE workflows. First, the architecture uses hybrid natural language processing and large language model (NLP/LLM) pipelines to auto-formalize unstructured natural language documents into formally structured, traceable requirements. Second, it auto-generates and continuously maintains traceability relationships across requirements, design elements, hazard analyses, and verification artifacts. Third, it provides continuous safety case completeness and confidence assessment through automated Goal Structuring Notation (GSN) synthesis connected to live evidence sources. The approach is grounded in Systems-Theoretic Process Analysis (STPA), the OMG Risk Analysis and Assessment Modeling Language (RAAML), MIL-STD-882E system safety practice, and the UL 4600 safety case framework. We present the methodology, its alignment to the DoD Digital Engineering Strategy, and its applicability to ground vehicle autonomy programs including next-generation infantry fighting vehicles and robotic combat vehicles. We also discuss the limitations, risks, and cultural barriers that must be addressed for AI-augmented safety engineering to achieve acceptance in mission-critical defense applications.

Citation: M. Wagner, N. Santini, A. Balakrishnan, "AI-Powered Safety Intelligence as a Decision Enabler for Ground Vehicle MBSE: The Digital Safety Twin Approach," In *Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium* (GVSETS), NDIA, Novi, MI, Aug. 11-13, 2026.

1. INTRODUCTION

The Department of Defense has invested heavily in digital engineering as a means of accelerating acquisition timelines, improving design quality, and reducing lifecycle costs. The 2018 DoD Digital Engineering Strategy [1] established five foundational goals, and DoDI 5000.97 [2], issued in December 2023, formalized the requirement for program managers to implement digital engineering procedures to the maximum extent possible. Model-Based Systems Engineering sits at the center of this vision, promising a transition from document-centric to model-centric development where an authoritative source of truth enables cross-disciplinary collaboration and automated analysis.

The reality has not matched the ambition. The Defense Science Board's 2024 Task Force on Digital Engineering [3] reported that insufficient and inconsistent architectures, standards, shared digital infrastructure, and intellectual property rights continue to impede realization of department-wide benefits. RAND's 2024 industry survey [4] concluded that implementation of digital engineering has not yet resulted in notable cost savings or schedule reduction, though it has enabled benefits in other areas such as expanded design trade spaces and more comprehensive modeling and simulation. The Carnegie Mellon Software Engineering Institute's MBSynergy Workshop (reported 2025) [5] identified five primary barriers to MBSE implementation—all cultural rather than technical—including mismatches between program management perspective and MBSE expertise, vague definitions of correct implementation, and lack of mission-driven adoption.

Ground vehicle programs provide particularly stark illustrations. The XM30, designated as one of six Army digital engineering pathfinder programs and the first combat vehicle engineered entirely through a model-based approach, experienced schedule delays that GAO's 2025 Weapons Systems Assessment [6] linked to both contractors delivering software and hardware that did not comply with modular open systems requirements. Program officials reported to GAO that they had assumed the contractors possessed proficiency in using a models-based engineering approach, an assumption that proved incorrect. The Robotic Combat Vehicle (RCV) program faced similar frustrations, with an acquisition approach characterized by industry observers as simultaneously overambitious and procedurally constrained.

This paper argues that the missing element is not better MBSE tools in the conventional sense but rather an intelligent integration layer that automates the most labor-intensive aspects of safety-driven systems engineering while serving as a decision enabler for program managers and chief engineers. We introduce the Digital Safety Twin concept: an AI-powered safety intelligence architecture that auto-generates traceability relationships, auto-formalizes unstructured natural language into traceable requirements, and continuously assesses safety case completeness against applicable standards. The contribution is methodological rather than tool-centric: we present a workflow that connects STPA-derived safety properties to MBSE models to MIL-STD-882E compliance evidence through automated reasoning, and we discuss both the promise and the significant limitations of this approach.

2. BACKGROUND AND RELATED WORK

2.1. The MBSE Adoption Gap in Defense

MBSE adoption in defense programs has been characterized by a persistent gap between stated policy goals and measurable outcomes. The DSB [3] cited a literature

review finding that less than one percent of published reports actually measured MBSE benefits versus merely perceiving them. Tool interoperability remains deeply problematic: programs attempting model exchange between tools via XMI have reported that sequence diagrams arrived mostly empty and had to be reconstructed manually. A Department of Defense model migration of 220,000 elements and 300 diagrams between modeling platforms was characterized as a task that would have taken months or years without specialized intervention.

The XM30 experience is instructive. Building the system architecture model generated significantly more growth in data and specifications than program officials anticipated [6], revealing a fundamental challenge: MBSE does not merely digitize existing workflows but forces explicit articulation of relationships and interfaces that were previously implicit or assumed. This is particularly acute for safety-critical systems where traceability between requirements, design decisions, hazard analyses, and verification evidence must be comprehensive and auditable.

The underlying mechanism, rarely named in official assessments, is cognitive overload. MBSE requires practitioners to simultaneously maintain mental models of system architecture, behavioral specifications, interface definitions, traceability relationships, and standards compliance—all within unfamiliar toolchains that impose their own conceptual overhead. The pattern is not unprecedented. When parametric CAD and finite element analysis tools entered widespread use in the early 1990s, adoption followed a similar trajectory: institutional mandates preceded workforce readiness, early implementations were slower than the manual processes they replaced, and broad adoption occurred only when the tools became sufficiently transparent that engineers could focus on engineering problems rather than tool mechanics. The critical difference is that MBSE's scope is broader—it touches every engineering discipline simultaneously—and the cognitive burden scales with system complexity in ways that individual-discipline CAD tools did not. This suggests that sustainable MBSE adoption requires not just training and cultural change but active reduction of the cognitive load imposed by the toolchain itself.

2.2. AI-Augmented Requirements Engineering

The application of artificial intelligence to requirements engineering has progressed through three generations. Early approaches used information retrieval techniques such as TF-IDF and Latent Semantic Indexing for requirements classification and trace link recovery. The second generation applied deep learning, exemplified by Guo et al.'s [23] work using bidirectional recurrent neural networks for semantically enhanced traceability. The current generation leverages large language models for requirements formalization, quality assessment, and automated generation.

Recent work demonstrates both the promise and limitations of LLM-based requirements engineering. An IEEE 2025 paper demonstrated an LLM assistant guided by INCOSE and IREB writing rules that identifies ambiguities, classifies requirement types, and serializes structured requirements into SysML v2 using KerML. However, a Riverside Research investigation of ChatGPT for Internet of Battlefield Things SysML modeling [19] found that AI-generated requirements frequently lack the structural rigor, traceability, and formal syntax essential for direct SysML implementation. Zhang et al.'s 2026 MBSE Co-Pilot research roadmap [20] proposes NLP, ML, and computer vision to augment model development but acknowledges that

hallucination risk remains a fundamental constraint for safety-critical applications.

Within the GVSETS community specifically, Peterson and Petrotta's 2018 paper [17] proposed augmented intelligence for systems engineering, leveraging SysML's machine-readable properties. Multani et al. [18] at GVSETS 2025 developed a ModelExtractor tool for Cameo that uses AI to overcome barriers to MBSE adoption. At INCOSE 2024, Johns et al. [22] demonstrated GPT-4 Turbo integrated into CATIA Magic/Cameo for requirements and block definition diagram generation, while Sugawara et al. [21] at INCOSE 2025 converted system models to graph structures for natural language querying via LLM. Fratrik [11] presented at the 2023 NDIA Systems Engineering Conference on applying MBSE iteratively to MIL-STD-882E safety programs for autonomous ground vehicles, establishing prior art for the approach described in this paper.

2.3. Automated Traceability

Maintaining traceability across the systems engineering lifecycle is among the most labor-intensive activities in defense acquisition. Requirements must trace to design elements, design elements to hazard analyses, hazard analyses to safety requirements, and all of these to verification and validation artifacts. In practice, these trace links are established manually, maintained inconsistently, and frequently become stale as programs evolve.

Automated traceability research has advanced through information retrieval, deep learning, and now LLM-based approaches. A comprehensive survey identified over fifty publications focused on trace link recovery using NLP techniques. Recent work using LLMs for data augmentation in traceability has outperformed all prior baselines across benchmark datasets. However, the frontier is shifting from retrospective traceability recovery—finding links after the fact—to prospective, continuous traceability maintenance, where links are generated and updated as artifacts evolve. Scalability and reliability at industrial scale remain unproven for LLM-based methods.

2.4. Safety Analysis Integration with MBSE

The integration of system safety analysis with MBSE represents the least mature but potentially most impactful intersection point. Leveson's Systems-Theoretic Accident Model and Processes (STAMP) and its associated hazard analysis method STPA [7] provide a theoretically grounded approach to identifying unsafe control actions in complex systems. The OMG's RAAML specification [8] defines SysML extensions supporting FMEA, FTA, STPA, and GSN safety case notation within system models.

Ahlbrecht, Sprockhoff, and Durak [10] published the most comprehensive MBSE-STPA-GSN integration framework to date, demonstrating automated creation of safety analysis tasks, automated generation of safety cases, and tracking of safety case progress—though they caution that automation may introduce a false sense of confidence. Shea's 2025 AFIT dissertation [9] provides the first formal mapping of STPA-Coordination to MIL-STD-882E system safety process elements within SysML-RAAML, delivering 41 design considerations for Preliminary Design Review and 78 for Critical Design Review across a small unmanned aerial system case history.

For automated safety case generation specifically, Sivakumar et al. [15] evaluated GPT-4's ability to generate GSN-structured safety cases, finding moderately accurate and reasonable results. However, no existing commercial MBSE tool integrates AI-driven hazard analysis, automated safety case generation, and MIL-STD-882E compliance

mapping into a unified workflow. This gap defines the contribution space for the Digital Safety Twin.

3. THE DIGITAL SAFETY TWIN CONCEPT

The Digital Safety Twin is a continuously maintained, model-connected safety representation that operates as an intelligent integration layer across the systems engineering lifecycle. Conventional digital twins for autonomous systems focus on mirroring physical system behavior for development and testing purposes [26]. The Digital Safety Twin, by contrast, mirrors and maintains the safety reasoning about a system: the chain of logic from operational concept through hazard identification, safety requirements derivation, design mitigation, and verification evidence to the conclusion that the system is acceptably safe for its intended use.

The concept is realized through three interconnected AI-powered capabilities, each addressing a specific bottleneck in current defense MBSE practice.

3.1. AI-Driven Requirements Formalization

Defense programs generate vast quantities of unstructured natural language: concept of operations documents, operational requirements, threat assessments, lessons learned reports, standards mandates, and informal engineering notes. Converting this material into formally structured, traceable requirements is traditionally a manual, expert-intensive process that introduces inconsistency and delay.

The Digital Safety Twin employs a hybrid NLP/LLM pipeline that processes unstructured text through multiple stages. First, domain-specific named entity recognition identifies system elements, performance parameters, environmental conditions, and safety-relevant terminology. Second, syntactic analysis extracts requirement-like structures and classifies them against INCOSE and EARS (Easy Approach to Requirements Syntax) patterns. Third, an LLM conditioned on the program's existing requirement corpus and applicable standards generates candidate formal requirements in a structured schema. Fourth—and critically—a deterministic verification layer checks each generated requirement against well-formedness rules, identifies ambiguous or untestable language, flags potential conflicts with existing requirements, and assigns confidence scores.

This hybrid architecture directly addresses the hallucination risk that multiple researchers have identified as the primary barrier to LLM use in safety-critical contexts [19][20]. The LLM performs the creative synthesis—transforming unstructured thought into structured form—while deterministic analysis provides the guardrails. No generated requirement enters the authoritative model without human review, but the human reviewer receives a pre-structured, pre-analyzed candidate rather than starting from a blank page.

3.2. Auto-Generated Traceability

The second capability addresses what is arguably the single most time-consuming aspect of safety-driven MBSE: establishing and maintaining traceability relationships. In a typical ground vehicle autonomy program, a single safety requirement may trace upward to an operational need, laterally to multiple design elements and interface specifications, downward to hazard analysis results (STPA unsafe control actions, FMEA failure modes), and forward to verification procedures, test results, and safety case evidence nodes.

The Digital Safety Twin maintains a knowledge graph that represents all program artifacts and their semantic relationships. When new artifacts are created or existing

artifacts are modified, embedding-based similarity analysis identifies candidate trace links. Domain-specific ontologies constrain the link types to those defined by applicable standards: MIL-STD-882E's traceability between hazards, safety requirements, and risk mitigations [14]; RAAML's relationships between STPA control structure elements and GSN argument nodes [8]; and UL 4600's evidence linkages within the safety case [13]. A confidence score accompanies each suggested link, and links below a configurable threshold are flagged for human adjudication rather than auto-accepted.

The key insight is that traceability in safety-critical systems is not merely administrative bookkeeping—it is the mechanism by which an assessor can verify that every identified hazard has been addressed, every safety requirement has been verified, and every piece of evidence connects to a claim in the safety argument. Automating this process does not remove human judgment from the loop; it ensures that the judgment is applied to reviewing and refining a comprehensive set of relationships rather than manually constructing them from scratch.

3.3. Continuous Safety Case Assessment

The third capability synthesizes the first two into a continuously assessed safety case. Following the GSN standard and UL 4600's goal-based safety case structure [13][16], the Digital Safety Twin maintains a living safety argument that connects top-level safety goals through contextual assumptions and argumentation strategies to evidence nodes. Each evidence node links to a specific program artifact—a test result, a review record, an analysis report, a field observation—through the traceability relationships maintained by the second capability.

The assessment operates at two levels. At the structural level, completeness analysis identifies gaps: safety goals without supporting arguments, arguments without evidence, evidence nodes without linked artifacts, and STPA-derived hazards without corresponding safety requirements. At the confidence level, Bayesian reasoning propagates uncertainty from individual evidence nodes through the argument structure to the top-level safety claim, providing a quantitative (though explicitly qualified) measure of safety case maturity.

For MIL-STD-882E compliance specifically, the system maps safety case elements to the standard's required documentation artifacts: System Safety Program Plan, Preliminary Hazard List, System Hazard Analysis, Operating and Support Hazard Analysis, and Safety Assessment Report [14]. Shea's STPA-to-MIL-STD-882E mapping [9] provides the formal correspondence, and the Digital Safety Twin automates the tracking of which standard-mandated artifacts exist, are current, and contain sufficient evidence to support the corresponding safety claims.

4. APPLICATION TO GROUND VEHICLE AUTONOMY PROGRAMS

Ground vehicle autonomy programs present a uniquely challenging environment for safety-driven MBSE. As Koopman and Wagner [12] have argued, autonomous vehicle safety is fundamentally an interdisciplinary challenge that spans perception, planning, control, and human factors. These systems operate in unstructured environments with civilian interaction, incorporate AI/ML perception and planning components whose behavior cannot be fully specified a priori, and must comply with multiple overlapping standards frameworks. The Digital Safety Twin is designed to address several specific pain points documented in current programs.

4.1. Managing Multi-Standard Compliance

A typical autonomous ground vehicle program must demonstrate compliance with MIL-STD-882E for system safety, DoDD 3000.09 [27] for autonomous weapon system safety, potentially ISO 26262 [25] for functional safety of automotive-derived subsystems, ISO 21448 [24] for safety of the intended functionality, and UL 4600 [13] for overall autonomous product safety evaluation. These standards are not contradictory but are also not harmonized: each defines its own terminology, process steps, artifact requirements, and evidence expectations.

The Digital Safety Twin maintains cross-standard mappings that allow a single body of safety analysis work to simultaneously address requirements from multiple frameworks. An STPA-identified unsafe control action, for example, maps to a MIL-STD-882E hazard (with severity and probability classification), an ISO 26262 safety goal (with ASIL assignment), an ISO 21448 triggering condition or functional insufficiency, and a UL 4600 safety case sub-goal. The auto-generated traceability ensures that when analysis is updated in one context, the corresponding elements in other frameworks are flagged for review.

4.2. Decision Enablement for Milestone Reviews

The Digital Safety Twin's primary decision-enablement function is providing program managers and chief engineers with an evidence-based answer to the question: how complete and how confident is our safety case, right now?

At Preliminary Design Review, the system can report which STPA unsafe control actions have been identified, which have been allocated to safety requirements, which requirements have been allocated to design elements, and which design elements have been verified. At Critical Design Review, the system identifies residual gaps between the committed safety case and the available evidence. At Test Readiness Review, it assesses whether the planned test program covers the identified hazard scenarios with sufficient depth. At each milestone, the output is not a pass/fail determination—that remains a human judgment—but a structured, quantified input to that judgment that makes the current state of safety knowledge explicit and auditable.

4.3. Addressing Cognitive Overload and the Cultural Challenge

The SEI MBSynergy Workshop [5] identified cultural barriers as the primary impediment to MBSE adoption, but as discussed in Section 2.1, the underlying mechanism is cognitive overload: practitioners are asked to operate across multiple complex toolchains while simultaneously learning new modeling paradigms, maintaining traceability across hundreds of artifact types, and satisfying compliance demands from multiple overlapping standards. The Digital Safety Twin directly attacks this overload by absorbing the mechanical complexity that currently falls on human engineers.

Consider the cognitive burden of maintaining traceability in a program subject to MIL-STD-882E, ISO 26262, and UL 4600 simultaneously. An engineer modifying a single safety requirement must mentally trace the implications to the system hazard analysis, the relevant design elements, the verification plan, the safety case argument node, and the compliance artifacts for each applicable standard. Today this requires navigating between three or more disconnected tools and manually updating relationships in each. The Digital Safety Twin reduces this to a single action: modify the requirement, review the auto-generated impact analysis across all linked artifacts, and

accept or adjust the proposed updates. The engineer's cognitive effort shifts from bookkeeping to judgment—the same transition that made parametric CAD successful when it finally matured past the steep learning curve of its early adoption.

This human-in-the-loop philosophy is not merely a design choice but a technical necessity. Safety arguments in defense systems must be defensible under independent assessment, and the chain of reasoning must be attributable to qualified engineers who can explain and justify every claim. AI acceleration of the mechanical aspects of safety analysis—the assembly, linking, completeness checking, and confidence assessment—preserves the engineer's intellectual authority over the substantive safety judgments while freeing them from the administrative burden that currently consumes the majority of their time.

5. DISCUSSION, LIMITATIONS, AND RISK MITIGATIONS

Intellectual honesty requires acknowledging significant limitations and open challenges in the approach presented here. For each identified risk, we describe the specific mitigation strategies incorporated into the Digital Safety Twin architecture. These mitigations do not eliminate risk—they bound it and make residual risk visible to decision-makers.

5.1. Hallucination and Reliability

LLMs generate plausible but sometimes incorrect output. In requirements formalization, a hallucinated requirement that enters the authoritative model could propagate errors downstream into design and verification, with potentially severe consequences for safety-critical systems. The failure mode is insidious: a well-formed but factually incorrect requirement may pass syntactic quality checks while introducing an undetected specification error.

The Digital Safety Twin employs a layered mitigation strategy. First, the hybrid NLP/LLM architecture separates generative synthesis from deterministic verification. The LLM proposes candidate formalizations; a rule-based verification engine independently checks each candidate against well-formedness constraints, domain ontology consistency, and contradiction detection against the existing requirement corpus. Candidates that fail deterministic checks are rejected before any human ever sees them, reducing the review burden to outputs that are at least structurally sound. Second, every AI-generated artifact carries provenance metadata that records the source input, the model version, the confidence score, and the specific verification checks passed or failed. This provenance chain is immutable and auditable, enabling post-hoc analysis if an error is later discovered. Third, the system implements staged promotion: AI-generated requirements enter a "candidate" state distinct from the authoritative baseline and can only be promoted through an explicit human acceptance workflow that includes peer review. No LLM-generated content enters a safety-critical baseline without qualified human review. The system is designed to make such review efficient and well-informed, not to bypass it.

5.2. Explainability and Auditability

Defense safety assessments require that every claim be traceable to a rationale. When an AI system suggests a traceability link or flags a safety case gap, the basis for that suggestion must be explainable in terms that a safety assessor can evaluate. Current LLM architectures provide limited native explainability, and this remains an active research area with no near-term general solution.

The Digital Safety Twin mitigates explainability risk through three architectural choices. First, traceability link suggestions

are grounded in retrievable evidence: the system presents the specific text spans, model elements, or artifact excerpts that contributed to a suggested link, allowing the reviewer to evaluate the basis rather than trusting an opaque score. This retrieval-augmented approach ensures that even when the underlying similarity computation is not fully interpretable, the inputs to the decision are transparent. Second, the system distinguishes between high-confidence suggestions (where multiple independent signals converge) and low-confidence suggestions (where the basis is thin or ambiguous), routing each to different review workflows. High-confidence links may be batch-reviewed; low-confidence links receive individual scrutiny. Third, all AI-assisted analysis is logged in a format compatible with the safety case audit trail required by MIL-STD-882E [14] and UL 4600 [13]. The audit record captures not only what was suggested and what was accepted but also what was rejected and why, building an institutional knowledge base that improves both the AI models and the human review process over time.

5.3. Validation at Scale

The methodology presented here is conceptual. We have not yet conducted controlled experiments comparing Digital Safety Twin-assisted workflows to traditional approaches on matched programs at defense-relevant scale. The quantified benefits reported for AI-assisted requirements tools in other contexts—such as the Royal Canadian Air Force's reported 50–75% reduction in requirements review time using QVscribe—are suggestive but not directly transferable to the safety analysis domain, where the consequences of error are categorically different from those in requirements quality checking.

Our mitigation approach is incremental deployment with embedded measurement. Rather than attempting full-scope deployment on a program of record, the validation strategy proceeds through three phases. Phase 1 uses synthetic benchmarks: a controlled test suite of known inputs and expected outputs across the three core capabilities, exercised against fictional but representative program data. This enables regression testing and capability measurement without exposing real program data. Phase 2 applies the Digital Safety Twin in parallel (shadow mode) alongside the traditional workflow on an active program, with defined metrics comparing effort, completeness, accuracy, and time-to-milestone between the AI-assisted and manual paths. Phase 3, contingent on Phase 2 results, integrates the Digital Safety Twin into the primary workflow with continuous metric collection. Each phase has explicit go/no-go criteria, and results from earlier phases inform calibration of confidence thresholds and review workflows in later phases. We identify rigorous multi-program validation as the highest-priority future work.

5.4. Scope of Automation and False Confidence

There is a risk that automating safety analysis mechanics could create what Ahlbrecht et al. [10] describe as a false sense of confidence. If program managers equate a high safety case completeness score with actual safety, the tool could become counterproductive—worse than no tool at all, because it would provide unwarranted assurance.

This risk is addressed through deliberate design of the system's output semantics. The Digital Safety Twin never produces a binary safe/unsafe determination. Instead, it reports three distinct dimensions: structural completeness (what percentage of expected argument nodes have supporting evidence), evidence currency (how recently each evidence artifact was updated relative to the design baseline it references), and confidence

distribution (a Bayesian propagation of uncertainty from leaf evidence through the argument structure, presented as a distribution rather than a point estimate). When completeness is high but confidence is low—indicating that evidence exists but is weak, stale, or inconsistent—the system explicitly flags this divergence. Furthermore, the system identifies what it calls "irreducible review points": nodes in the safety argument where the claim is inherently a matter of engineering judgment rather than verifiable evidence. Examples include the adequacy of an operational design domain definition, the sufficiency of a test coverage strategy, or the acceptability of residual risk. At these nodes, the system requires documented human rationale and will not auto-populate confidence scores. This forces the safety case to be honest about where automation ends and professional judgment begins—a property we consider essential for the system's integrity and for its acceptance by independent safety assessors.

5.5. Organizational Adoption and Change Management

Beyond the technical risks, the SEI MBSynergy Workshop [5] makes clear that cultural and organizational factors are the primary determinants of success or failure for any MBSE initiative. Introducing AI-powered tooling into an environment already struggling with basic MBSE adoption could compound rather than alleviate resistance if the approach is perceived as adding yet another layer of tool complexity.

The mitigation strategy is to position the Digital Safety Twin not as an additional tool but as a reduction in the number of disconnected tools and manual processes that engineers must manage. The auto-generated traceability replaces manual spreadsheet-based traceability matrices. The requirements formalization pipeline replaces ad hoc document-to-model translation. The continuous safety case assessment replaces periodic, labor-intensive safety review preparation cycles. Each capability is deployable independently, allowing programs to adopt the capabilities that address their most acute pain points first and expand incrementally. Critically, the system must demonstrate value within the first iteration cycle—if engineers do not see tangible effort reduction within weeks rather than months, adoption will stall regardless of long-term potential. The phased validation approach described in Section 5.3 is designed to generate these early demonstrations of value.

6. CONCLUSIONS

The Department of Defense's investment in digital engineering and MBSE for ground vehicle programs has yet to deliver the measured benefits that policy mandates envision. The gap is not primarily technological but rather reflects the enormous manual burden of maintaining traceability, formalizing requirements, and constructing auditable safety arguments across complex, multi-standard environments.

The Digital Safety Twin concept presented in this paper addresses this gap through three AI-powered capabilities: auto-formalization of natural language into traceable requirements, auto-generation and continuous maintenance of traceability relationships, and continuous safety case completeness and confidence assessment. Grounded in STPA, RAAML, MIL-STD-882E, and UL 4600, the approach is designed as a decision enabler that augments rather than replaces qualified safety engineers.

We have been forthcoming about the limitations—hallucination risk, explainability gaps, absence of at-scale validation data, false confidence risk, and organizational adoption challenges—and have described specific architectural and

procedural mitigations for each. These mitigations bound but do not eliminate risk, and their effectiveness must be validated through the phased deployment approach described in Section 5. The defense ground vehicle community should hold AI-augmented safety tools to the same evidentiary standard it applies to the systems those tools are meant to assess.

The engineering problem is clear and urgent. Programs such as the XM30, RCV, and future autonomous logistics platforms will require safety cases of unprecedented scope and complexity. The current approach—manual traceability, document-centric safety analysis, periodic safety reviews disconnected from the live engineering baseline—cannot scale to meet this demand. Whether the specific architecture described here or an alternative approach prevails, the defense ground vehicle community must engage seriously with AI-powered safety intelligence as a necessary complement to MBSE adoption.

For practitioners who wish to begin developing experience with AI-augmented safety analysis rather than waiting for a fully matured platform, existing safety intelligence tooling offers a pragmatic entry point. Edge Case's Guardian platform, for example, implements several of the building blocks discussed here and can help teams establish the traceability discipline, provenance practices, and human review workflows that any AI-augmented approach—the Digital Safety Twin among them—will ultimately depend on. Beginning with such tooling on a bounded, well-instrumented problem is consistent with the incremental, measurement-driven adoption strategy advocated in Section 5.

7. REFERENCES

[1] Department of Defense, "DoD Digital Engineering Strategy," June 2018.

[2] Department of Defense, "DoDI 5000.97: Digital Engineering," December 2023.

[3] Defense Science Board, "Task Force on Digital Engineering: Final Report," May 2024.

[4] RAND Corporation, "The Impact of Digital Engineering on Defense Acquisition and the Supply Chain: Insights from an Industry Survey," 2024.

[5] W. Hayes, J. Hugues, P. Capell, and N. Shevchenko, "Report on the First MBSynergy Workshop," CMU/SEI-2025-TR-004, Carnegie Mellon Software Engineering Institute, July 2025.

[6] Government Accountability Office, "Weapons Systems Annual Assessment," GAO-25-107569, June 2025.

[7] N. Leveson, "Engineering a Safer World: Systems Thinking Applied to Safety," MIT Press, 2011.

[8] OMG, "Risk Analysis and Assessment Modeling Language (RAAML), v1.1," 2023.

[9] D. Shea, "Model-Based Approach to Support Safety Driven Design and Satisfy MIL-STD-882E Requirements with STPA Coordination," AFIT Dissertation, 2025.

[10] S. Ahlbrecht, M. Sprockhoff, and U. Durak, "A System-Theoretic Assurance Framework for Safety-Driven Systems Engineering," Software and Systems Modeling, 2024.

[11] F. Fratrik, "Model-Based Systems Engineering for Iterative MIL-STD-882E Autonomous Ground Vehicle Safety Programs," NDIA Systems Engineering Conference, 2023.

[12] P. Koopman and M. Wagner, "Autonomous Vehicle Safety: An Interdisciplinary Challenge," IEEE Intelligent Transportation Systems Magazine, 2017.

[13] UL Solutions, "UL 4600: Standard for Safety for the Evaluation of Autonomous Products," 3rd Edition, 2023.

[14] MIL-STD-882E, "Department of Defense Standard Practice: System Safety," Change 1, September 2023.

[15] S. Sivakumar et al., "Prompting GPT-4 to Support Automatic Safety Case Generation," Expert Systems with Applications, 2024.

[16] M. Wagner and P. Koopman, "An Overview of Draft UL 4600," Edge Case Research, 2020.

[17] A. Peterson and G. Petrotta, "Implementing Augmented Intelligence in Systems Engineering," GVSETS, 2018.

[18] V. Multani et al., "Utilizing AI to Overcome Barriers to MBSE Adoption," GVSETS, 2025.

[19] J. Brooks, "Leveraging AI-Driven Requirements for SysML Modeling of the IoBT," Riverside Research, 2024.

[20] Zhang et al., "MBSE Co-Pilot: A Research Roadmap," Systems Engineering, 2026.

[21] Sugawara et al., "Extracting Information from System Model as Graph Structure by Large Language Model in MBSE," INCOSE International Symposium, 2025.

[22] Johns et al., "AI Systems Modeling Enhancer (AI-SME): Initial Investigations into a ChatGPT-enabled MBSE Modeling Assistant," INCOSE International Symposium, 2024.

[23] J. Guo et al., "Semantically Enhanced Software Traceability Using Deep Learning Techniques," ICSE, 2017.

[24] ISO 21448:2022, "Road Vehicles — Safety of the Intended Functionality (SOTIF)."

[25] ISO 26262:2018, "Road Vehicles — Functional Safety."

[26] ITEA Journal, "Tailoring the Digital Twin for Autonomous Systems Development and Testing," Vol. 44, No. 4, 2023.

[27] Department of Defense, "DoDD 3000.09: Autonomy in Weapon Systems," January 2023.

8. CONTACT INFORMATION

Michael Wagner
Edge Case Research, Inc.
Pittsburgh, PA

9. ACKNOWLEDGMENT

The authors thank their colleagues at Edge Case Research for discussions that informed the development of the Digital Safety Twin concept.

10. ACRONYMS

GSN	Goal Structuring Notation
LLM	Large Language Model
MBSE	Model-Based Systems Engineering
NLP	Natural Language Processing
RAAML	Risk Analysis and Assessment Modeling Language
RCV	Robotic Combat Vehicle
STPA	Systems-Theoretic Process Analysis
UCA	Unsafe Control Action