ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Designing Secure Composite AI Systems: Cross-Domain Holistic Threat Model and Mitigation Framework

Aditya K. Sood | VP Security Engineering and AI Strategy, Aryaka Networks, US | ORCID: 0000-0002-7738-2890

Sherali Zeadally | College of Communication and Information, University of Kentucky, Lexington, KY, US | Academy of Computer Science & Software Engineering, University of Johannesburg, South Africa | ORCID: 0000-0002-5982-8190

Abstract

Composite artificial intelligence (AI) systems are increasingly deployed in mission-critical environments, such as defence, aerospace, industrial control systems, and critical infrastructure, where they enable adaptive, autonomous, and real-time decision-making. However, the growing complexity of these systems introduces multilayered security risks that extend far beyond the assumptions of traditional, component-centric security models. In this work, we introduce a structured taxonomy that decomposes composite AI systems into five tightly interconnected layers: core AI and machine-learning (ML) components, integration and orchestration mechanisms, data flows and shared computational resources, cross-layer system interactions and emergent vulnerabilities, and legacy or deterministic software modules that coexist with AI. Leveraging this taxonomy, we propose a holistic, cross-domain threat modelling approach to systematically identify threats, architectural weaknesses, and design-level security flaws across the entire system lifecycle. Finally, we outline mitigation strategies and architectural best practices aimed at building secure, resilient,

Received: 01.02.2026

Accepted: 07.05.2026

Published: 19.06.2026

Cite this article as:
A.K. Sood, S. Zeadally, "Designing secure composite AI systems: Cross-domain holistic threat model and mitigation framework," ACIG, vol. 5, no. 2, 2026, doi: 10.60097/ACIG/221578

Corresponding author:
Sherali Zeadally, College of Communication and Information, University of Kentucky, Lexington, KY, US; E-mail: szeadally@uky.edu
0000-0002-5982-8190

and trustworthy composite AI systems capable of operating safely under adversarial conditions.

Keywords

AI, threat intelligence, cybersecurity, threat modeling

1. Introduction

The pervasive integration of artificial intelligence (AI) into virtually every facet of modern society has propelled its role from a supportive tool to an indispensable component, particularly in 'mission-critical' [1] applications. These applications, spanning domains, such as defence, healthcare, autonomous systems, and critical infrastructure, are characterised by their extreme sensitivity to failure, with disruptions that can cause significant economic damage or large-scale operational paralysis. Traditional, monolithic AI systems often struggle to meet the stringent requirements of these environments, lacking the necessary robustness, adaptability, and explainability for high-stakes decision-making. This challenge has spurred the development and adoption of composite AI systems. Composite AI [2], also known as hybrid or synergistic AI, is a paradigm where multiple, diverse AI techniques are intelligently orchestrated and integrated to form a cohesive system [3]. Unlike standalone AI models, composite AI leverages the strengths of various methodologies, including deep learning for pattern recognition, symbolic AI for knowledge representation and reasoning, natural language processing (NLP) for human–machine interaction, and expert systems for domain-specific insights. This synergistic approach enables the system to handle complex, real-world scenarios that exceed the capabilities of any single AI technique, resulting in enhanced performance, greater accuracy, and improved decision-making capabilities.

However, the very nature of composite AI, with its interconnected and heterogeneous components, introduces unique and significant security challenges for mission-critical applications. The attack surface of such systems is inherently broader and more complex than that of isolated AI models. Each integrated AI component, as well as the interfaces and data flows between them, presents potential vulnerabilities. Adversarial attacks, such as data poisoning in training datasets, model evasion, or model extraction, can compromise the integrity and reliability of individual AI components, and these compromises can cascade throughout the composite system [4]. Furthermore, the complexity of data management across diverse

AI modules, involving sensitive and often real-time information, necessitates robust data provenance, integrity checks, and stringent access controls.

Threat modelling for composite AI systems demands a paradigm shift. It requires an in-depth discovery phase that focuses on data (inputs, outputs, and training), model processes, and data flow, while accounting for biases and unintended consequences. A comprehensive cross-domain threat model must analyse threats that span multiple domains, such as data, infrastructures, and agents, rather than focusing on isolated components to consider both security threats and vulnerabilities across all the critical components of the composite AI system. Such a model should focus on how vulnerabilities and attacks propagate across interconnected systems, enabling holistic risk assessment and mitigation.

2. Related Work

Traditional threat modelling frameworks, such as spoofing, tampering, repudiation, information disclosure, denial of service (DoS), and elevation of privilege (STRIDE) [5] and process for attack simulation and threat analysis (PASTA) [6], while effective for conventional software, struggle to model the dynamic, probabilistic, and often opaque nature of AI components. These frameworks usually lack the granularity and specific threat categories necessary to address AI-centric vulnerabilities, such as adversarial attacks, data poisoning, and model integrity compromises, as well as the unique risks associated with AI supply chains. While adaptations such as methodologies for threat modelling AI and ML systems [7, 8], integrating AI-specific threat interpretations have emerged, they frequently fall short in systematically modelling the complex interdependencies and emergent behaviours that arise when diverse AI models interact in a larger, hybrid architecture [9] of mission-critical systems. The inherent unpredictability and opacity of specific AI components further challenge the deterministic assumptions underpinning traditional threat modelling, leaving critical blind spots [10].

Vicarte et al. [11] introduced an asset-centric approach to threat modelling of AI systems. They addressed the unique security challenges posed by integrated and autonomous agents. By prioritising assets over attack patterns, the methodology scales with evolving threats and supports increasingly complex, distributed agentic AI environments. Additionally, Tete [12] proposed a framework that integrates STRIDE and Damage, Reproducibility, Exploitability, Affected Users, and Discoverability (DREAD) for proactive threat

identification and risk assessment. Through a case study of a custom Large Language Models (LLM)-based application, we demonstrate an end-to-end threat model aligned with Shostack's four question framework [13], adapted to address LLM-specific risks, to strengthen the security, reliability, and integrity of AI-driven systems. On similar standards, other researchers also proposed a STRIDE-based security methodology for ML pipelines [1418] that defines key security properties, identifies threats, and guides the selection and validation of appropriate controls. Adapted for AI systems using a security-property-driven approach, the methodology was demonstrated through an industrial case study.

Shapira et. al. [15] proposed FRAME, a comprehensive and automated framework for assessing Adversarial Machine-Learning (AML) risks across diverse ML systems. It quantifies risk using a multi-dimensional model that covers deployment context, attack characteristics, and empirical insights, augmented by feasibility scoring, LLM-driven customisation, and a structured AML attack dataset which enable accurate, context-aware assessments. Grosse et. al. [16] analysed the gap between academic threat models and real-world AI deployments. Through a study of common attack models and a survey of 271 industry practitioners, the study found that, while the existing threat models targeting attacks such as poisoning, backdoors, evasion, model stealing, membership inference, adversarial examples, including property inference, are applicable, they often assume unrealistic attacker capabilities, such as extensive access to data, and adversarial examples, highlighting the need for more practical, deployment-aligned threat modelling in AI security.

Furthermore, the academic community has invested heavily in the verification, validation, and assurance of AI systems, particularly for safety-critical applications [17, 18]. The emergent properties resulting from the interplay of multiple AI and non-AI components, combined with the inherent difficulty of achieving exhaustive test coverage and security assessments for adaptive systems, make the operational lifecycle more difficult. Efforts include applying formal methods to provide mathematical guarantees for certain AI behaviours, developing runtime monitoring techniques for anomaly detection, and creating new testing methodologies to address the non-deterministic nature of learning algorithms [19, 20].

3. Problem Statement

Securing composite AI systems in mission-critical applications faces several profound challenges. While individual aspects

of AI security, threat modelling, and verification/validation have received considerable research attention, a holistic and component-based taxonomy, including a cross-domain threat model for composite AI systems, remains relatively underdeveloped. This, coupled with the inadequacy of the existing threat modelling frameworks to handle their hybrid nature and the persistent challenges in comprehensive testing and validation, underscores the critical need for further academic investigation into securing these complex, interdependent, and mission-critical AI-driven architectures.

4. Research Gaps

We identified several research gaps and addressed them in this research work.

- *First, a critical gap is the lack of an explicit component-based taxonomy.* Without a standardised way to define the individual building blocks of these systems, it is exceedingly difficult to systematically identify and manage security risks. This lack of a common framework prevents the implementation of granular, targeted security controls. It hampers practical risk assessment and compliance for each specific AI model, integration layer, or legacy component.
- *Second, existing threat modelling frameworks are not adequately equipped for the hybrid nature of composite AI systems*. These systems deeply integrate AI into traditional software designs, creating novel and interdependent attack surfaces that conventional, siloed threat models fail to capture. The complex interactions and emergent behaviours among these integrated components demand a unified, hybrid threat-modelling approach to address unique attack vectors.
- *Finally, a significant hurdle is the absence of standardised, comprehensive testing and validation methodologies*. Artificial intelligence components introduce non-determinism and emergent behaviours, making traditional testing paradigms insufficient. This complicates the creation of exhaustive test suites and the reliable assessment of the system's robustness against novel or adversarial inputs. Ensuring stringent safety and reliability in mission-critical contexts thus remains a continuous challenge.

5. Contributions of This Work

Our research contributions are as follows:

- We analysed the use of composite AI systems in mission-critical applications to derive a well-structured taxonomy of the most critical components.

- We proposed a cross-domain, holistic threat model for analysing threats in composite AI systems that utilises the component-based taxonomy.
- We conducted a threat modelling exercise for a mission-critical application, using our proposed threat model to demonstrate threat identification within a specific component of the AI composite system.
- We discussed the substantial impact of the cross-domain holistic threat model and provided supporting reasons.
- Last, we proposed mitigation solutions to secure and defend composite AI systems against threats and cyberattacks.

6. Understanding the Implications of Compromised Composite AI Systems

The architectural complexity of composite AI systems introduces a multidimensional threat landscape. When such systems are compromised, the consequences extend well beyond localised software failures, often resulting in amplified, systemic, and mission-disruptive impacts. We present the implications of compromised composite AI systems below.

- *Operational degradation and mission failure:* Composite AI systems play pivotal roles in real-time decision-making, navigation, target recognition, resource allocation, and threat response. A compromise in any layer, such as a misclassification by a perception module, a logic flaw in orchestration, or data poisoning of shared resources, can degrade decision quality or delay critical responses. In time-sensitive domains, such as autonomous aerial combat or satellite coordination, this degradation can directly lead to mission failure, loss of strategic assets, or even endangerment of human lives.
- *Systemic propagation of faults:* Due to the tightly coupled nature of composite AI architectures, localised attacks can propagate across the system. For example, a manipulated input to an ML component may trigger faulty actuation commands, while a breach in orchestration logic may cause incorrect instructions to cascade across subsystems. These cross-layer fault propagations can lead to emergent behaviours not foreseen during testing, making the system behave unpredictably under adversarial influence.
- *Loss of trust and situational awareness:* Mission-critical systems rely heavily on human operators' ability to trust and interpret AI-generated insights. An adversary's manipulation of AI components, through techniques such as model evasion, spoofing,

or data injection, can erode this trust. In high-stakes scenarios, this may force operators to override automated systems or disregard AI inputs altogether, reducing the overall system effectiveness and increasing cognitive burden during already complex missions.

- *Security breach escalation and privilege abuse:* Composite AI systems often include multiple subsystems with varying privilege levels and access to different types of data. A compromise in one AI module or its associated software interface can allow lateral movement across the system. These attack paths reflect a broader threat vector than traditional software systems due to the dynamic behaviour and opaque decision-making mechanisms of AI components.
- *Strategic and geopolitical consequences:* In defence, aerospace, and infrastructure contexts, the failure of AI systems due to cyber compromise can have significant strategic and geopolitical implications. A disrupted satellite constellation, a disabled swarm of unmanned aerial vehicles (UAVs), or a shutdown of critical infrastructure due to AI misbehaviour may be interpreted as a hostile act or failure of national defence capabilities. This could escalate tensions or result in reputational, financial, and diplomatic costs.
- *Challenge of detection and attribution:* The opaque decision-making of composite AI systems hinders the distinction between benign errors and intentional compromise. This ambiguity complicates rapid recovery, forensic attribution, and effective legal or tactical responses, particularly in adversarial environments that demand real-time decision-making and accountability.

The risks we discussed above underscore the urgent need for comprehensive, system-level threat models, continuous assurance mechanisms, and resilient architectures.

7. Taxonomy of Components of Composite AI Systems

This study helps us understand the systematic operation of composite AI systems, thereby facilitating the design of a cross-domain, holistic threat model. Based on the results, we designed a comprehensive taxonomy of components in composite AI systems (as shown in Figure 1), which forms the basis for developing a cross-domain, holistic threat model.

- *AI components:* These are the core intelligent modules of the composite system, each typically specialised for a particular task or type of reasoning. They can leverage various AI paradigms, such

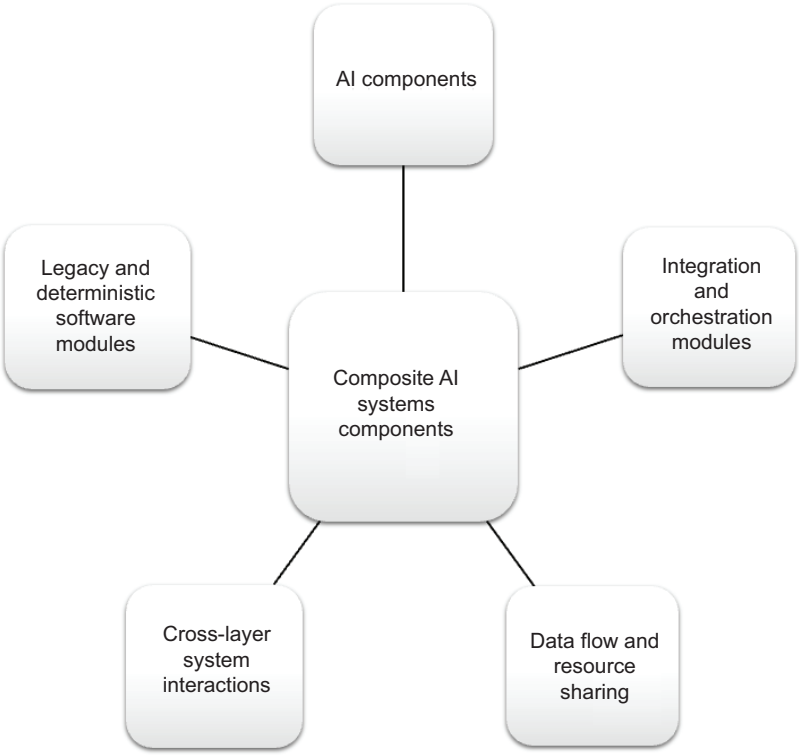

Figure 1. A taxonomy of components of composite AI systems.

as LLMs, ML models (e.g. neural networks for pattern recognition, predictive models), symbolic AI (e.g. rule-based systems, expert systems, and knowledge graphs for logical reasoning), NLP models, computer vision systems, planning and scheduling algorithms, or even generative AI models for creating new data or responses, including agentic systems. In mission-critical contexts, these components are often designed for high accuracy, low latency, and robustness to noise or adversarial attacks.

- *Integration and orchestration modules:* These modules serve as the 'nervous system' of the composite AI system, coordinating the various components and ensuring they work together seamlessly and efficiently. Integration modules manage the technical interfaces and data formats necessary for multiple components to communicate effectively. In contrast, orchestration modules define the workflow and sequence of operations, managing dependencies, resource allocation, error handling, and the overall system state. In mission-critical settings, these modules are essential for ensuring real-time performance, fault tolerance, and predictable behaviour, often employing techniques, such as

message queues, Application Programming Interface (API) gateways, and workflow engines.

- *Data flow and resource-sharing:* It encompasses mechanisms and infrastructure that facilitate the efficient and reliable movement of data between components and the sharing and management of computational resources. In mission-critical systems, data flow must be low-latency, high-throughput, and fault-tolerant, involving efficient data serialisation, transport protocols, and potentially distributed data stores. Resource-sharing mechanisms ensure that computational resources (CPU, GPU, and memory) are allocated optimally among various AI components and other system modules to meet real-time processing demands and avoid bottlenecks, often through dynamic resource allocation, load balancing, and efficient memory management.
- *Cross-layer system interaction:* This component examines how the different layers of a composite system, including hardware, operating systems, and application-level AI modules, interact and influence one another to ensure the overall system performance and reliability. Mission-critical AI systems often have multiple layers of abstraction, ranging from bare-metal hardware to high-level decision-making AI. Cross-layer interaction involves designing interfaces and protocols that enable these layers to communicate effectively, share status information, and adapt their behaviour accordingly. This includes scenarios where the AI module needs to be aware of hardware limitations or operating system priorities. Conversely, lower layers might need to adapt to AI's computational demands or criticality assessments, often through feedback loops.
- *Legacy and deterministic software modules:* These are established, well-understood software components characterised by predictable and reliable behaviour, often extensively tested and certified. They are integrated with AI components to leverage their proven dependability. In mission-critical applications, there are frequently existing, highly reliable software components that perform crucial functions that do not require AI, or where AI is not yet mature enough to provide the necessary level of determinism. 'Deterministic' implies that given the same input, the module will always produce the same output, making its behaviour entirely predictable and verifiable. Integrating these with AI enables the system to benefit from AI's adaptive capabilities while retaining the robustness of proven, deterministic code for foundational operations, which often involve safety-critical functions, low-level controls, hardware interfaces, and regulatory compliance.

Using the above taxonomy, we discussed the real-world case study of robotics in hazardous environments, as shown in Table 1.

Table 1. Mapping component-based taxonomy to robotics to understand the associated composite AI system.

Component	Details and usage	Role of robotic systems in mission-critical operations
AI component	• AI-powered perception using computer vision (object recognition and damage assessment), lidar/sonar (3D mapping, obstacle detection), and sensor fusion (radiation mapping and gas detection). • Navigation and path planning using reinforcement learning or search algorithms for optimal and safe routes through complex, unstructured terrain, avoiding hazards.	• Enhanced situational awareness by providing a comprehensive understanding of the hazardous environment (e.g. exact location of radiation source, structural integrity). • Ensuring autonomous safe navigation to avoid known and unforeseen dangers, minimising the risk of damage or mission failure.
Integration and orchestration modules	• Combines heterogeneous sensor data (e.g. visual, thermal, chemical, and radiation) into a unified environmental model for the robot. • Manages the sequence of operations (e.g. initial reconnaissance, inspection, and intervention), coordinating different AI modules and robot actions.	• Creates a resilient, redundant picture of the hazard, even if a single sensor fails. • Ensures that complex missions are executed logically and efficiently, minimising time spent in high-risk areas.
Data flow and resource-sharing	• Robust communication links using high-bandwidth, low-latency transmission of sensor data (including video streams and 3D scans) and control commands, often with redundancy. • Critical, real-time AI inference (e.g. immediate obstacle avoidance) performed on board (edge); complex planning, large-scale data analysis, and model updates offloaded to remote, powerful computing infrastructures (cloud).	• Enables immediate responses to dynamic changes in hazardous environments. • Optimises the use of limited onboard resources while leveraging powerful remote processing capabilities for complex AI applications.
Cross-layer system interaction	• The robot's physical state (battery, motor temperature, and joint angles) informs the AI's decision-making software; the AI's planned actions are then checked against physical constraints. • AI analyses mission progress and environmental conditions (e.g. high radiation and low light) to optimise power consumption for longevity.	• Ensures the robot operates within its physical limits to maximise mission duration and prevent damage. • Allows early detection of issues, enabling timely intervention (e.g. sending another robot, retrieving the current one) to prevent mission abandonment or catastrophic failure.
Legacy and deterministic software modules	• Deterministic software that manages the precise movement of individual motors and actuators. • Hard-wired and software-based emergency stop protocols that immediately halt all robot movement in critical situations. • Robust, standard, and highly reliable software for maintaining data links under challenging conditions.	• Provides fundamental, irrefutable control mechanisms that can override AI in emergencies. • Ensures basic robot functions operate predictably and reliably, forming a stable foundation for AI.

Consider a robotic deployment in a warfare scenario compounded by an attack on critical systems, such as a cyberattack disrupting power grids amidst conflict and natural disaster, the role of autonomous robots in stabilising critical infrastructure becomes paramount.

8. Proposed Cross-Domain Holistic Threat Model

In this section, we present a detailed threat model (Table 2) that covers various threat classes and maps them to the components of composite AI systems used in mission-critical applications. Defining different classes of threats for components helps to implement granular security controls and enables efficient threat understanding, enabling a robust security posture. We also apply the threat model derived from the component-based taxonomy to autonomous vehicles, dissecting the potential threats.

8.1. Threat Prioritisation Based on Multi-Factor Risk Scoring

We need to integrate risk scoring and prioritisation into the threat model by assigning each identified threat a quantitative or semi-quantitative score based on key dimensions, such as likelihood, impact, and contextual risk factors. Likelihood captures the probability of a threat (e.g. ease of exploitation, attacker capabilities) while impact reflects the potential damage (e.g. data loss, system compromise, and operational disruption). Contextual factors refine the score further by incorporating asset criticality, exposure level, identity privilege, and environmental conditions.

Once scores are computed, threats can be ranked and grouped into priority tiers (e.g. critical, high, medium, and low), enabling teams to first focus on mitigation efforts on the most significant risks. This prioritisation should be dynamic, continuously updated based on telemetry, threat intelligence, and system behaviour (e.g. active-exploitation signals or anomalous activity). Additionally, integrating risk scores with response workflows allows for automated or semi-automated actions, such as escalating alerts, enforcing stricter controls, or triggering containment for high-risk scenarios. This approach transforms the threat model from a static artifact into a living, decision-driven framework that adapts to evolving risks.

Next, we expand this concept to the threat model presented for the composite AI system.

Composite risk function: It defines the overall risk as a combination of multiple contributing factors, each normalised and weighted. It provides a flexible mathematical foundation to aggregate diverse risk signals into a single score. We define risk as a weighted aggregation of multiple normalised factors:

- Risk score (RS) = $\Sigma(w_i \times f_i)$, where f_i is the normalised factor score, and w_i is the weight.

Table 2. Threat model of composite AI systems using the proposed taxonomy and application of the threat model to autonomous vehicle systems.

Composite AI system components	Mapped threats	Threat model application to autonomous vehicle systems

AI components	• *Data poisoning:* Injecting malicious data into training data, leading to biased, inaccurate, or vulnerable models (e.g. misclassifying a hostile drone as friendly).
	  • In applications, attackers can trigger *prompt injections* to inject malicious data into supporting AI models, thereby exploiting functionality and bypassing guardrails.
	  • An attacker can conduct *reinforcement learning (RL) poisoning* by injecting false data *to* corrupt the agent's learning, causing it to develop unsafe, inefficient, or exploitable policies that could lead to catastrophic failures.
	  • An attacker can also execute *label flipping (LF) or label poisoning* to flip the labels of a subset of training data points, often randomly or targeted at specific classes, thereby causing the model to learn incorrect decision boundaries.
	• *Adversarial evasion:* Crafting subtly perturbed inputs that cause misclassification or misbehaviour at inference time (e.g. a self-driving car misinterpreting a stop sign).
	• *Model inversion:* Unauthorised reconstruction of the model's architecture or inference of sensitive training data through repeated queries (intellectual property theft and privacy violation).
	• *Model backdoors:* Hidden vulnerabilities or malicious code embedded [21] during training or the import of third-party malicious models, activated by specific runtime triggers, leading to malicious behaviour (e.g. a diagnostic AI generating a false positive for a particular patient profile).
	• *Integrity compromise:* Tampering with model weights or code directly.
	• *Bias exploitation:* Adversaries leverage inherent model biases (derived from training data) to target or mislead specific groups or scenarios.
		• *Adversarial evasion:* Subtle alterations to visual data (e.g. stickers on road signs) causing misclassification (e.g. stop sign detected as speed limit).
		• *Data poisoning*: Injecting false Global Positioning System (GPS) signals or LiDAR returns to disrupt localisation or object detection, resulting in an incorrect understanding of the environment.
		• *Model integrity attacks:* Tampering with learned models (e.g. planning weights) to introduce unsafe driving behaviours or backdoors.
		• *RL poisoning:* Manipulating rewards or environment feedback during training to make the vehicle learn dangerous driving policies.

Integration and orchestration modules	• *Interface manipulation:* Intercepting commands (data) exchanged between components, leading to misinterpretation or malicious execution (e.g. corrupting sensor fusion output before it reaches the planning module). • *Orchestration logic compromise:* Tampering with the central control flow, decision rules, or routing logic, allowing an adversary to dictate system behaviour or create cascading errors (e.g. forcing a smart grid optimisation AI to cause blackouts). • *Conflicting outputs exploitation:* Forcing diverse AI models to produce contradictory outputs, causing system confusion, paralysis, or incorrect aggregated decisions. • *Timing attacks:* Disrupting the precise synchronisation among functional modules in a composite AI system can lead to highly unpredictable responses. • *Resource exhaustion*: Overloading key orchestration components with excessive requests or data, leading to system-wide unresponsiveness or failure, which results in a denial of service.	• *Interface manipulation:* Maliciously modifying sensor fusion outputs before they reach the planning module, leading to incorrect decisions. • *Orchestration logic compromise:* Gaining control over the central decision-making system to issue arbitrary commands (e.g. turn off braking, change destination). • *Timing attacks:* Disrupting the synchronisation between perception, planning, and control loops, leading to unsafe responses.

	• *Middleware exploitation:* Exploiting inherent vulnerabilities in middleware components to trigger unauthorised actions. • *Exploiting integration bridges:* Exploiting the interfaces or adaptors connecting modern AI components with older, legacy systems (e.g. a vulnerability in a SCADA gateway exposing the grid control AI).	• *Middleware exploits:* Vulnerabilities in communication protocols (e.g. Robot Operating System [ROS], Data Distribution Service [DDS]) that allow unauthorised access or command injection.
Data flow and resource-sharing	• *Data leakage or exfiltration across components:* Vulnerabilities in data handling or access controls allowing sensitive information processed by one component to be exposed to unauthorised components or external entities (e.g. driver's data from an NLP module leaking to an insecure analytics component). • *Shared knowledge base poisoning:* Corrupting central data stores or knowledge graphs used by multiple AI models, leading to widespread misinformed decisions (e.g. injecting false intelligence into a military Intelligence Surveillance Reconnaissance (ISR) knowledge base. • *Synchronisation attacks:* Disrupting the timing or coordination of data exchange, resulting in outdated information, race conditions, or incorrect sequential operations throughout the system. • *Resource contention:* Maliciously consuming shared computational resources, starving critical components of necessary processing power, leading to degraded performance or DoS.	• *Data leakage or exfiltration*: Stealing sensitive telematics, mapping, or passenger data during transmission or storage. • *Emergent misbehaviour:* Corrupting map data or real-time sensor data streams to cause navigation errors or system crashes and intercepting or spoofing Vehicle-to-Everything (V2X) communications, resulting in false traffic warnings and vehicle-to-vehicle collision messages. • *Resource exhaustion:* Overloading the vehicle's processing units with excessive data or tasks, resulting in system paralysis or performance degradation.

Cross-layer system interaction	• *Cascading failures:* A successful attack on the vulnerability of one component (e.g. data poisoning in a perception model) leads to a ripple effect of failures or misbehaviour in downstream components that rely on its output (e.g. faulty perception results in incorrect path planning).
	• *Emergent misbehaviour:* Unpredictable, undesirable behaviours arising from the complex, non-linear interactions of components, which were not designed or tested in isolation. Different types of attacks include:
	  • Attacks on data integrity could also result in unexpected behaviour impacting the system's availability and security.
	  • Attacks that target communication protocols via spoofing and injecting arbitrary commands or payloads.
	• *Over-reliance on a single source of truth:* If a critical component acting as a 'single source of truth' (e.g. sensor fusion in autonomous vehicles) is compromised, the entire system relies on its flawed output, leading to systemic failure.
	• *Human–AI interface manipulation:* Exploiting the human–machine interface to feed false information to the human operator, misinterpreting human commands, or bypassing human oversight mechanisms (e.g. a tampered alert system suppressing critical warnings).
	• *Cyber-physical attacks*: These attacks target the vital communication and feedback pathways that connect the high-level system or AI logic to the physical world. The attacker can cause a cyber-originated attack to manifest as a direct physical failure.
		• *Cyber-physical attacks:* Combining digital exploits (e.g. software vulnerability) with physical manipulation (e.g. a specific light pattern) to achieve a coordinated, highly effective attack.
		• *Cascading failures:* A compromise in a lower-level perception module results in faulty input to the prediction module, which then generates hazardous plans that override safety features.

(*Continues*)

Table 2. Continued.

Composite AI system components	Mapped threats	Threat model application to autonomous vehicle systems

Legacy and deterministic software modules	• *Insecure APIs and misconfiguration:* Insecure APIs, often due to weak authentication, authorisation flaws, or improper data validation, can expose sensitive information, allow unauthorised command execution, and disrupt precise operations. Similarly, system misconfigurations, such as default credentials, overly permissive access controls, or unpatched services, create exploitable pathways.
	• *Lack of AI-aware security:* Legacy systems often lack built-in security features to address AI-specific threats, making them susceptible to new attack vectors introduced by AI integration (e.g. traditional firewalls may fail to detect adversarial ML attacks).
	• *Outdated security patches and protocols:* Legacy components may have unpatched vulnerabilities or use weaker security protocols, which can serve as entry points for compromising the entire composite system.
	• *Fixed logic bypass:* Adversaries exploit the deterministic, rule-based nature of legacy components to predictably bypass or manipulate their intended functions, thereby affecting AI layers.
	• *Supply chain attacks:* These attacks can compromise the integrity and security of the entire system. These attacks exploit vulnerabilities in the development and distribution of trusted software components. Several attacks include:
	  • *Compromised open-source dependencies:* An attacker injects malicious code into widely used open-source libraries integrated into the deterministic module, creating backdoors or logic bombs.
	  • *Malicious updates or patches:* Attackers compromise a vendor's distribution channel to inject malicious code into a seemingly legitimate software update, altering the module's behaviour.
	  • *Poisoned build environments:* By infiltrating the continuous integration/continuous delivery (CI/CD) pipeline, an attacker injects malware directly into the compiled, signed binary of the deterministic software before deployment.
	  • *Firmware or hardware tampering:* Malicious code is injected into the module's firmware or hardware during manufacturing, creating a persistent and difficult-to-detect vulnerability that can subvert the AI's control signals.
		• *Unpatched vulnerabilities:* Exploiting known security flaws in older electronic control units (ECUs), infotainment systems, or low-level firmware.
		• *Insecure interfaces:* Legacy protocols or APIs that lack modern authentication and encryption, serving as entry points to more critical systems (e.g. Controller Area Network (CAN) bus exploits).
		• *Bypassing safety interlocks:* Manipulating deterministic safety features (e.g. hardcoded speed limits) via a compromised interface.
		• *Configuration errors:* Misconfigurations in legacy modules allow unintended access or functionality.
		• *Supply chain attacks:* Injecting malicious code or hardware components at any stage of vehicle manufacturing or software development. Similar variations of the attack include injecting malicious updates into the vehicle's critical systems.

Core risk factors: They represent the key dimensions influencing risk, such as likelihood, impact, exposure, and privilege. These factors capture both technical and contextual aspects of threats in a structured way. Table 3 presents core risk factors.

Weighted risk model: It assigns relative importance to each risk factor by weighting it to compute a composite risk score. This allows prioritisation to reflect organisational context and security priorities. Table 4 presents an example of weight configuration.

Risk classification: It maps numerical risk scores into categories, such as low, medium, high, and critical. It enables consistent decision-making and guides appropriate response actions based on severity. Table 5 presents an example of a generic score range for mapping to priority levels.

Table 3. Core risk factors.

Factor	Symbol	Description
Likelihood	(L)	Probability of threat occurrence
Impact	(I)	Severity of potential damage
Asset criticality	(A)	Importance of the affected system/data
Exploitability	(E)	Ease of exploitation
Exposure	(X)	Degree of external accessibility
Privilege level	(P)	Level of access/authority involved
Behavioural anomaly	(B)	Degree of deviation from expected behaviour
Threat intelligence	(T)	External evidence of active exploitation

Table 4. A configuration example of weights against core risk factors.

Factor	Symbol	Weight (w)
Likelihood	(L)	0.15
Impact	(I)	0.20
Asset criticality	(A)	0.15
Exploitability	(E)	0.10
Exposure	(X)	0.10
Privilege level	(P)	0.10
Behavioural anomaly	(B)	0.10
Threat intelligence	(T)	0.10

Table 5. Example of a normalised factor score with priority levels.

Normalised factor score range (0–1)	Priority level
0.75–1.0	Critical
0.50–0.74	High
0.25–0.49	Medium
0.0–0.24	Low

8.2. Data Poisoning Threat: Risk Score Calculation

We apply the multi-score risk scoring above to a potential data poisoning attack targeting AI components. We use the weights presented above for various core risk factors to calculate the risk score of data poisoning attacks.

$$\text{Risk score (data poisoning)} = (wL \times L + wI \times I + wA \times A + wE \times E + wX \times X + wP \times P + wB \times B + wT \times T).$$

We assume normalised values.

$$L = 0.8,\ I = 1.0,\ A = 0.9,\ E = 0.7,\ X = 0.8,\ P = 0.9,\ B = 0.6,\ T = 0.7$$

The final risk score is:

$$\begin{aligned}\text{Risk score} &= (0.15 \times 0.8) + (0.20 \times 1.0) + (0.15 \times 0.9) + (0.10 \times 0.7) \\ &\quad + (0.10 \times 0.8) + (0.10 \times 0.9) + (0.10 \times 0.6) + (0.10 \times 0.7) \\ &= 0.12 + 0.20 + 0.135 + 0.07 + 0.08 + 0.09 + 0.06 + 0.07 \\ &= 0.825.\end{aligned}$$

The risk severity of the data poisoning threat is critical because the risk score falls within the range 0.75–1.0. Based on this score, the system prioritises immediate action, such as restricting or revoking access to risky entity for AI component, enforcing stricter authentication, deploying tamper-detection controls, or triggering alerts and initiating an investigation. This example shows how risk scoring helps to move from identifying threats to prioritising and acting on them effectively.

This scoring matrix transforms a cross-domain holistic threat model into a quantitative, decision-driven framework, enabling consistent prioritisation, faster response, and alignment with real-world risk dynamics.

8.3. Evaluating the Effectiveness of the Proposed Cross-Domain Threat Model

The proposed threat model extends beyond theoretical vulnerabilities to provide practical insights into the unique security challenges associated with composite AI systems. We discuss the effectiveness of the threat model below.

- *Comprehensive delineation of composite risk surface:* Traditional threat modelling often focuses on individual software components or standalone AI models, rather than the broader composite risk surface. However, a cross-domain holistic approach is crucial because it accounts for the amplified attack surface inherent in composite AI systems. By examining the interplay among diverse AI paradigms, classical software modules, orchestration layers, and data pipelines, it identifies emergent vulnerabilities arising specifically from their integration. This includes, for instance, cascading failures, subtle interface manipulations, and sophisticated attacks that exploit the sequential or parallel processing across multiple components.
- *Driving a proactive security paradigm:* This threat model helps security practitioners to implement a proactive 'security-by-design' methodology. By thoroughly articulating potential attack vectors across the entire composite AI system lifecycle, this threat model empowers developers and security architects to embed resilience and robust defences from the outset. This pre-emptive approach significantly reduces the likelihood of costly and catastrophic security incidents in mission-critical applications where real-time patching may be infeasible or too late.
- *Enabling targeted mitigation strategy development:* Understanding the precise nature of threats that exploit the interconnectedness and heterogeneity of composite AI systems enables the formulation of highly tailored mitigation strategies. Instead of applying generic security controls, security practitioners can use the threat model to obtain the granular insights necessary to design specific defences against, for example data poisoning affecting a particular model, adversarial examples exploiting sensor fusion, or a compromise of the orchestration layer. This precision enhances the effectiveness of security measures and optimises their deployment.
- *Optimising security resource allocation:* Resource constraints are a common challenge in security initiatives. By providing a clear, prioritised view of threats unique to composite AI architectures, this threat model enables organisations to make more informed investment decisions. It highlights where security efforts will have the greatest impact, enabling strategic allocation of

computational resources, specialised talent, and defensive technologies to protect the most critical components and interfaces within the integrated system.

- *Informing robust human–AI teaming protocols:* In mission-critical contexts, humans often remain 'in the loop' or 'on the loop'. By understanding how the composite AI systems could be deceived, malfunction, or present misleading information, this threat model helps define clear protocols for human monitoring, interpretation, and intervention. This fosters appropriate trust in the AI system, guiding human operators on when to rely on AI outputs and when to exercise caution or perform manual override.
- *Enhancing system resilience and trust:* This threat model facilitates the design of resilient and trustworthy composite AI systems, thereby improving the overall system resilience and confidence. By providing a deep understanding of potential failure modes and attack vectors, it paves the way for designing composite AI systems that can gracefully degrade, self-recover, or maintain critical functionality even under duress. This increased resilience, coupled with a clearer understanding of security risks, is crucial for fostering public and stakeholder trust in AI technology.

8.4. Proposed Mitigation Solutions for Securing Composite AI Systems

Securing composite AI systems requires a holistic, defence-in-depth approach spanning the entire lifecycle and all architectural layers. Table 6 presents several mitigation solutions for various categories. We have mapped the enforcement of proposed mitigations to showcase impact on the security of autonomous vehicle system.

By adopting a holistic, cross-domain threat modelling and security approach throughout the lifecycle of composite AI systems, the unique security risks in mission-critical applications can be effectively addressed, resulting in more resilient and trustworthy autonomous capabilities.

9. Limitations of the Proposed Threat Model Framework

While the proposed threat model framework offers a structured lens to understand emerging patterns in composite AI systems, it is important to recognise several inherent limitations. It is primarily derived from conceptual analysis, real-world AI security risks, and observed behaviours in limited environments. Consequently, the proposed relationships, categories,

Table 6. Mitigation solutions to secure and prevent threats targeting composite AI systems.

Category	Mitigation solutions	Impact of security controls enforcement on autonomous vehicle system
Component-level security	• Integrate security from the initial design phase, considering threat models for each component and their interactions. • Apply robust security measures to each AI model and software module (e.g. adversarial robustness, data integrity checks, secure training environments, and model integrity verification).	• Prevent unsafe interactions among perception, planning, and control modules to prevent system-level exploits and cascading failures. • Detect and resist adversarial manipulations of cameras, LiDAR, radar, or stop signs to prevent misperception. • Validate sensor data to prevent GPS spoofing, corrupted telemetry, or false localisation/navigation. • Prevent poisoned data or backdoor models from introducing malicious behaviours in autonomous vehicle AI models. • Ensure only verified, authentic models are deployed, blocking unauthorised model updates.
Secure integration and orchestration	• Implement authentication, authorisation, input validation, and rate limiting for all inter-component APIs. • Secure microservices by applying best practices for securing distributed systems. • Harden the orchestration logic by rigorous testing, formal verification, and configuring minimal privileges. • Secure communication channels by encrypting all data in transit between components.	• Restrict access to trusted subsystems to prevent unauthorised command injection via APIs. • Block malformed or excessive requests to mitigate DoS and system overload attacks. • Isolate services to prevent lateral movement attacks across autonomous vehicle subsystems. • Ensure safe command sequencing and prevent unsafe execution paths. • Protect vehicle-to-cloud and vehicle-to-vehicle communications from Man-in-the-Middle (MITM) attacks and data tampering.
Data provenance and integrity	• Deploy an immutable data logging capability to log all data inputs, outputs, and transformations. • Implement strict data validation at every ingress and egress point for each component to ensure data integrity. • Design-federated learning or privacy-preserving AI to reduce reliance on centralised sensitive data.	• Provide tamper-proof audit trails to detect log tampering or forensic evasion. • Filter invalid or malicious sensor data or actuator commands before they affect autonomous vehicle decisions. • Reduce risk of centralised data leakage during learning or model updates.
Continuous monitoring and anomaly detection	• Establish baselines for normal operational and user behaviour for each component and the system as a whole. • Cross-component anomaly detection: monitor for unusual data flows, processing loads, or conflicting outputs across the system to detect cross-component anomalies. • Use eXplainable AI (XAI) techniques with anomaly mapping to understand component decisions and detect unexpected reasoning.	• Detect deviations in driving behaviour or autonomous vehicle decision drift to identify stealth attacks. • Flag inconsistencies between perception, planning, and actuation layers for multi-stage attacks. • Reveal abnormal AI reasoning to detect hidden manipulations in autonomous vehicle decisions.

(*Continues*)

Table 6. Continued.

Category	Mitigation solutions	Impact of security controls enforcement on autonomous vehicle system
Human–AI teaming and override	• Define clear points of human intervention and escalation procedures within human-in-the-loop and human-on-the-loop protocols. • Ensure that human operators can safely and effectively take control or halt operations, even under duress, by implementing robust override mechanisms. • Follow the procedure for transparent decision-making by providing human operators with sufficient context about AI decisions.	• Enable human operators to validate or halt unsafe autonomous decisions. • Allow immediate manual control or emergency stop in case of actuator hijacking or runaway autonomy. • Improve operator situational awareness for timely corrective actions.
Supply chain security	• Implement a component vetting process to thoroughly vet all third-party AI models, libraries, and frameworks. • Maintain Software Bill of Materials (SBOM), a detailed inventory of all software components and their versions. • Verify the model and code for integrity using cryptographic signatures.	• Prevent malicious third-party sensors, models, or software from entering the autonomous vehicle system. • Track software dependencies and ensure vulnerable components are patched. • Verify the authenticity and integrity of autonomous vehicle software and AI models.
Resilience and redundancy	• Design the system to gracefully degrade or recover from component failures or compromises as part of its fault-tolerance capability. • Use diverse AI architectures or models for critical functions to reduce common mode failures. • Isolate components to limit the blast radius of a successful attack.	• Maintain safe operation or controlled stop under partial system failures (e.g. sensor/actuator malfunction). • Reduce risk from adversarial attacks on a single AI model via cross-validation. • Limit the blast radius of compromised AI components to contain attacks. • Cross-validate multiple sensors to detect spoofing or failure, ensuring safe navigation.

and boundaries may fall short of fully reflecting the complexities encountered in real-world deployments. Large-scale benchmarking, longitudinal studies, and controlled experiments are needed to validate the taxonomy's robustness and completeness.

Second, there are notable concerns about generalisability. The proposed threat model framework is influenced by specific verticals, such as composite AI systems, robotics, and autonomous systems, which may not directly translate to other contexts, such as financial, healthcare, or decentralised ecosystems. Different domains introduce unique constraints, interaction models, and risk profiles that may require adaptations or extensions to the taxonomy. Consequently, its applicability to heterogeneous systems and industries may be limited without further domain-specific refinement.

Third, the proposed threat model framework is built on a set of simplifying assumptions. These include assumptions about relatively stable system behaviour, clear identity boundaries, observable interactions, and predictable tool usage patterns. In practice, composite AI systems can exhibit highly dynamic, non-deterministic behaviour, influenced by evolving context, external inputs, or adversarial manipulation. Additionally, visibility into decision-making and interactions may be incomplete, especially in closed or proprietary systems, leading to potential gaps in classification and interpretation.

Furthermore, the current proposed threat model framework does not fully consider the collective impact of adversarial conditions and adversaries' system manipulation techniques, tactics, and procedures (TTPs) when multiple correlated threats are active, which will affect the behaviour of composite AI systems in subtle and hard-to-detect ways. There is also a risk of oversimplification, where complex, multi-layered behaviours are reduced to discrete categories, potentially overlooking nuanced interactions across layers of the system.

Finally, the rapidly evolving nature of composite AI systems introduces temporal limitations. New architectures, protocols, and attack patterns are emerging quickly, which may render parts of the taxonomy outdated or incomplete over time. We need continuous iteration, community validation, and integration of real-world feedback to keep the taxonomy relevant and accurate.

In summary, while the proposed threat model framework provides a useful foundation for understanding and structuring threats

specific to composite AI systems, it should be viewed as an evolving framework that requires ongoing validation, cross-domain adaptation, and refinement to address the dynamic, complex nature of these systems. It is also worth mentioning that the development of solutions to address these limitations is beyond the scope of this paper but we will address them as part of our future work.

10. Future Challenges for Securing Composite AI Systems

By explicitly detailing interdependencies and emergent attack surfaces, our proposed cross-domain threat model would provide a foundational security blueprint crucial for enhancing rigorous verification, validation, and assurance (VVA) processes, enabling more targeted and effective testing strategies. In the future, we will focus on several key areas to expand this threat model, each of which is critical.

- *Integrating threat model with security tools:* First, we aim to integrate the proposed threat model into automated security analysis tools. This would enable programmatic identification of attack surfaces and potential paths for vulnerability propagation across the hybrid architecture.
- *Threat intelligence integration for advanced insights:* Second, we will explore integrating dynamic threat intelligence directly into the model. This would enable real-time adaptation to the evolving threat landscape, allowing the system to dynamically prioritise risks and suggest countermeasures as new adversarial techniques or system changes emerge.
- *Designing tools for risk assessment:* Finally, we will develop specialised tooling and methodologies for quantitative risk assessment unique to composite AI systems. This would move beyond qualitative threat identification to provide measurable impacts of vulnerabilities, supporting more informed and data-driven security investment decisions in mission-critical environments.

Through these enhancements, the threat model can evolve into a dynamic, actionable framework for securing the next generation of complex AI systems.

11. Conclusion

As AI capabilities advance and become tightly integrated into mission-critical operational workflows, the need to secure these technologies becomes increasingly urgent. In this work,

we propose a timely cross-domain threat model for composite AI systems that effectively captures their complexity and interdependence. Through layered analysis, the model reveals how vulnerabilities can spread across software logic, AI components, data integration paths, and orchestration mechanisms. The implementation of this cross-domain holistic threat model enables practitioners to transition from *ad hoc* mitigation to a structured, system-level assurance. As autonomy and AI capabilities continue to scale, this research promotes the shift from isolated robustness to integrated assurance, ensuring that AI systems not only operate intelligently but also function securely, predictably, and safely under mission-critical conditions.

Acknowledgement

We thank anonymous reviewers for their valuable comments which helped us improve the organisation, content, and presentation of this work.

References

[1] S.T. Spantideas, A.E. Giannopoulos, P. Trakadas, "Smart Mission Critical Service Management: Architecture, Deployment Options, and Experimental Results," *IEEE Transactions on Network and Service Management*, vol. 22, no. 2, pp. 1108–1128, April 2025, doi: 10.1109/TNSM.2024.3498348.

[2] A.P. Sheth, K. Roy, R. Venkataramanan, V. Nadimuthu, "C^{3}AN: Custom, compact and composite AI systems – A neuroSymbolic approach: 4th-generation evolution of intelligent systems," *Scholar Commons*, Artificial Intelligence Institute at Scholar Commons, 2025. [Online] Available: https://scholarcommons.sc.edu/cgi/viewcontent.cgi?article=1646&context=aii_fac_pub [Accessed: July 20, 2025].

[3] Rapid Innovation (n.d). *Composite AI: The future of hybrid AI – Use cases*. [Online]. Available: https://www.rapidinnovation.io/post/composite-ai-concepts-benefits-industry-applications-implementation-strategies-challenges-future-outlook. [Accessed: June 20, 2025].

[4] Joint Cybersecurity Information, CISA, NSA, FBI. (May 22, 2025). *AI data security. Best practices for securing data used to train & operate AI systems*. [Online]. Available: https://www.ic3.gov/CSA/2025/250522.pdf. [Accessed: Jun. 25, 2025].

[5] A. Shostack. *Threat modeling: Designing for security*. Hoboken, NJ: Wiley, 2014.

[6] N. Pape, C. Mansour, "PASTA Threat Modeling for Vehicular Networks Security," in *7th International Conference on Information and Computer Technologies* (ICICT), Honolulu, HI, USA, 2024, pp. 474–478, doi: 10.1109/ICICT62343.2024.00083.

[7] National Institute of Standards and Technology (NIST). (2022). *Adversarial machine learning: A taxonomy and terminology of attacks and defenses*, NIST AI 100-2. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf. [Accessed: Jun. 10, 2025].

[8] C. Koball, Y. Wang, B.P. Rimal, V. Vaidyan, "Machine learning security: Threat model, attacks, and challenges," *Computer*, vol. 57, no. 10, pp. 26–35, 2024, doi: 10.1109/MC.2024.3396357.

[9] S.T. Spantideas, A.E. Giannopoulos, P. Trakadas, "Smart mission critical service management: Architecture, deployment options, and experimental results," *IEEE Transactions on Network and Service Management*, vol. 22, no. 2, pp. 1108–1128, 2025, doi: 10.1109/TNSM.2024.3498348.

[10] M. Andreoni, W.T. Lunardi, G. Lawton, S. Thakkar, "Enhancing autonomous system security and resilience with generative AI: A comprehensive survey," *IEEE Access*, vol. 12, pp. 109470–109493, 2024, doi: 10.1109/ACCESS.2024.3439363.

[11] J.S. Vicarte, M. Spoczynski, M. Elsaid, "Threat modeling for AI: The case for an asset-centric approach," 2025, arXiv:2505.06315.

[12] S.B. Tete, "Threat modeling and risk analysis for large language model (LLM)-powered applications," 2024, arXiv:2406.11007.

[13] J. Beyer, "Adam Shostack on threat modeling," *IEEE Software*, vol. 37, no. 6, pp. 110–112, 2020, doi: 10.1109/MS.2020.3017406.

[14] L. Mauri, E. Damiani, "STRIDE-AI: An approach to identifying vulnerabilities of machine learning assets," in *IEEE International Conference on Cyber Security and Resilience (CSR)*, Rhodes, Greece, 2021, pp. 147–154, doi: 10.1109/CSR51186.2021.9527917.

[15] A. Shapira, S. Shigol, A. Shabtai, "FRAME: Comprehensive risk assessment framework for adversarial machine learning threats," 2025, arXiv:2508.17405.

[16] K. Grosse, L. Bieringer, T.R. Besold, A. Alahi, "Towards more practical threat models in artificial intelligence security," in *Proceedings of the 33rd USENIX conference on security symposium* (SEC '24), Berkeley, CA: USENIX Association, 2024, Art. no. 274, pp. 4891–4908.

[17] L. Myllyaho, M. Raatikainen, T. Männistö, T. Mikkonen, J.K. Nurminen, "Systematic literature review of validation methods for AI systems," *Journal of Systems and Software*, vol. 181, Art. no. 111050, 2021, doi: 10.1016/j.jss.2021.111050.

[18] J. Renkhoff, K. Feng, M. Meier-Doernberg, A. Velasquez, H.H. Song, "A survey on verification and validation, testing and evaluations of neurosymbolic artificial intelligence," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 8, pp. 3765–3779, 2024, doi: 10.1109/TAI.2024.3351798.

[19] C. Kyrkou, A. Papachristodoulou, A. Kloukiniotis, A. Papandreou, A. Lalos, K. Moustakas, "Towards artificial-intelligence-based cybersecurity for robustifying automated driving systems against camera sensor attacks," in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, Limassol, Cyprus, 2020, pp. 476–481, doi: 10.1109/ISVLSI49217.2020.00-11.

[20] M. Hataba, A. Sherif, M. Mahmoud, M. Abdallah, W. Alasmary, "Security and privacy issues in autonomous vehicles: A layer-based survey," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 811–829, 2022, doi: 10.1109/OJCOMS.2022.3169500.

[21] A.K. Sood, S. Zeadally, "Malicious AI models undermine software supply-chain security," *Communications of the ACM*, vol. 68, no. 6, pp. 62–71, 2025, doi: 10.1145/3704724.