EC-COUNCIL
UNIVERSITY
ACCREDITED. FLEXIBLE. ONLINE.

A LAYERED DEFENSE FRAMEWORK ANALYSIS OF THREATS TO AND FROM AGENTIC AI

JAMES CRAIG
EC-COUNCIL UNIVERSITY

ECCU 501 Ethical Hacking and Countermeasures
Dr. Warren Mack
March 8, 2026

Table of Contents

Table of Contents

Table of Contents

Abstract

By 2027, agentic AI systems—autonomous agents with planning skills, persistent memory, tool-use APIs, and the ability to work with many agents—are predicted to conduct more than 25% of business workflows. However, 80% of firms utilizing these agents have already experienced hazardous behaviors in operational settings. This study offers a comprehensive examination of the bidirectional danger landscape confronting agentic AI, differentiating between threats directed at agents (threats-TO) and threats emanating from agents towards organizations and users (threats-FROM). The research employs the MAESTRO seven-layer threat model as an analytical framework, enumerating 30 threat and control categories across both dimensions, systematically arranged by architectural layer from foundational models to human-agent interaction. The paper assesses three agentic-specific security frameworks—MAESTRO (Cloud Security Alliance), the OWASP Top 10 for Agentic Applications, and SHIELD—showing that no single framework provides comprehensive coverage on its own. However, when used together as an integrated defense stack, they provide more than 90% coverage across all threat categories. Governance research shows that current regulatory frameworks, including the NIST AI RMF and the EU AI Act, require agent-specific overlays to address the specific needs of autonomous systems. The article offers bidirectional threat categories, a comparative framework analysis featuring quantitative coverage metrics, sector-specific governance recommendations, and a proposed three-framework defense stack for safeguarding enterprise agentic AI implementations.

A Layered Defense Framework Analysis of Threats To and From Agentic AI

Introduction

Background and Context

Agentic AI is a big step forward from the large language models (LLMs) that started the generative AI trend. Agentic AI systems are LLM-powered agents with planning modules, persistent episodic and semantic memory, tool-use APIs, and the ability to make their own decisions. This means they can do complicated, multi-step tasks with little help from people (Chhabra et al., 2025). This change from passive text generators to dynamic autonomous agents has happened quickly. By 2027, agentic AI is expected to handle more than 25% of business workflows (Narajala & Narayan, 2025), in areas as varied as national defense, healthcare diagnostics, financial trading, and software development.

But the speed at which people are using it has far overtaken the speed at which security safeguards are being developed. According to a 2025 industry survey, 80% of businesses that use AI agents have already seen dangerous conduct from those agents in production environments (Klein et al., 2025). These are not just ideas. Agents that can browse the web, run code, query databases, and call external APIs on their own create an attack surface that is different from and far bigger than that of traditional software systems or even regular LLM applications. The security community is now dealing with technology that is improving faster than the rules meant to control it.

Problem Statement

Agentic AI introduces new security challenges due to three aspects that traditional cybersecurity doesn't address: agents that operate autonomously, memory that persists across sessions, and direct access to real-world tools (Narajala & Narayan, 2025). Autonomy means that agents can make decisions on their own without needing permission from a person. Persistent memory means that a deal made in one session can affect later sessions. Tool access indicates that a successful attack can have real-world effects, including unlawful financial transactions, data theft, or changes to infrastructure, rather than just producing fake text.

Current general-purpose AI security frameworks were not designed to address these issues. The NIST AI Risk Management Framework (2023) offers optional governance features; however, it doesn't simulate agentic architectures. MITRE ATLAS (2024) provides an adversarial machine learning risk matrix centered on conventional ML pipelines, offering only minimal coverage of agent-specific threats. ISO/IEC 42001 (2023) establishes mechanisms for managing AI in organizations; however, it doesn't include any modeling of agentic threats (Chapple, 2026). The attack surface in agentic systems becomes more vulnerable as agents are connected in chains of tasks. This means that a single compromised agent can spread threats to all the other agents in the chain, making what might have been a small breach much worse (Ferrag et al., 2025).

Research Objectives

I have four goals in this paper. It looks at the two-way danger landscape of agentic AI, separating threats agents pose to organizations and users (threats-FROM) from threats that target agents themselves (threats-TO). It also looks at three new agentic-specific defense frameworks that came out in 2025–2026: the MAESTRO seven-layer threat model from the Cloud Security Alliance (Huang, 2025), the

OWASP Top 10 for Agentic Applications (OWASP Foundation, 2025), and the SHIELD mitigation framework (Narajala & Narayan, 2025). In addition, it provides guidance on governance and risk management for deployments across business, healthcare, and finance. Lastly, it finds areas of study that need more work and suggests ways to move forward with agentic AI security

Background and Related Work

Agentic AI Architecture

To understand the security problems posed by agentic AI, you need to understand how these systems differ from regular LLM applications. The five parts of an agentic AI system that work together are: an LLM backbone that provides reasoning and language generation capabilities; a planning module that breaks down complicated goals into a series of smaller tasks; episodic and semantic memory systems that keep track of state across sessions; tool-use APIs that let the system interact with other systems and data sources; and a multi-agent orchestration layer that coordinates the activities of multiple specialized agents within a single workflow (Chhabra et al., 2025).

There are different ways to set up multi-agent systems in organizations, and each one has its own security risks. Supervisory agents manage subordinate agents, making it easy for someone to gain access to the system. Peer-to-peer coordination empowers actors with equal trust to make decisions. If one peer is compromised, this might lead to lateral movement danger. Hybrid pipelines integrate both approaches, making an already hard-to-audit architecture even more complicated (Ferrag et al., 2025). The main architectural risk across all patterns is making judgments that can't be predicted in systems that remember how people have interacted with them in the past. Agents do not yield equal outputs from identical inputs, and their cumulative memory ensures that previous interactions perpetually influence future behavior in manners that are challenging to anticipate or formally validate.

Researchers have quickly answered these problems. In just the years 2023 to 2025, more than 160 research articles were published on the security of agentic AI (Shahriar et al., 2025). This shows how big the threat is seen to be and how weak the current defenses are.

Changes in AI Security Frameworks

Before 2025, companies that wanted to manage AI risk used general-purpose frameworks that were useful for governance and organizational processes but were not built to address the specialized threat surfaces of agentic systems. The NIST AI Risk Management Framework 1.0, released in January 2023, established four main functions: Map, Measure, Manage, and Govern. These are meant to be a voluntary way to govern AI systems in general, but they don't address agent autonomy, tool-use chains, or trust relationships between agents (National Institute of Standards and Technology, 2023). MITRE ATLAS, updated in 2024, includes a useful catalog of adversarial machine learning based on the ATT&CK framework. However, it only covers traditional ML pipelines, like model training and inference, and doesn't really apply to how autonomous agents behave in real time (MITRE Corporation, 2024). ISO/IEC 42001, released in 2023, outlines the prerequisites for an AI management system at the organizational level; nevertheless, its emphasis on process governance lacks technical threat modeling for agentic architectures (Secure Privacy, 2026). These three frameworks are kept in the Governance, Policy, and Regulatory Compliance section for governance and regulatory background, but they are not included in the framework comparison analysis because they are not security frameworks specific to agents.

The development of agentic-specific frameworks in 2025 and 2026 is a direct reaction to these shortcomings. The Cloud Security Alliance's MAESTRO framework, which stands for Multi-Agent Environment, Security, Threat, Risk, and Outcome, includes a seven-layer hierarchical threat model comprising Foundation Models, Data Operations, Agent Frameworks, Deployment Infrastructure, Agent Interactions, External Integrations, and Human-Agent Interaction (Huang, 2025). MAESTRO is the first framework to clearly show how threats move through the different levels of the agentic architecture. This enables security analysis at design time that links defensive controls to specific points in the agent technology stack. The OWASP Top 10 for Agile Applications, released in December 2025, has 10 risk categories: ASI01–ASI10. It is a peer-reviewed risk catalog useful for practitioners (OWASP Foundation, 2025). The SHIELD framework divides hazards into five areas and shows how to reduce them, making it the most technically comprehensive of the three frameworks (Narajala & Narayan, 2025).

Lack of Research

Even with this improvement, there is still no single framework that analyzes threats-TO and threats-FROM agents with the same level of detail. MAESTRO's layer model provides the most architecturally based threat list. OWASP's widely recognized

format helps raise awareness of risks among a broad range of professionals. SHIELD gives the most extensive operational mitigation mapping. The three are complementary, not redundant (Huang, 2025), but no previous research has rigorously examined their coverage, pinpointed their deficiencies, or suggested a cohesive deployment approach. My research fills that gap by using MAESTRO's seven-layer model as the analytical framework in the Threat Landscape: Attacks on Agentic AI and the Threat Landscape: Rogue and Malicious Agents sections. Then, in the Defense Mechanisms and Mitigation Strategies section, I conduct a comparative framework analysis that combines all three frameworks into a single defense strategy.

Threat Landscape: Attacks On Agentic AI

This section discusses how attackers target agentic AI systems. This is the "threats-TO" part of the framework from the Introduction. The analysis is structured according to MAESTRO's seven-layer model, which serves as the analytical framework (Huang, 2025). Table 1 lists all threats for all MAESTRO levels. The discussion that follows focuses on the analytical importance of each layer group.

Table 1

Bidirectional Threat Model for Agentic AI Systems, Organized by MAESTRO Architectural Layer

MAESTRO Layer	Type of Threat	Direction	OWASP ASI	Key Reference
L1: Foundation Models	Prompt injection/goal hijacking	TO	ASI01	Gulyamov et al., 2025
L1: Foundation Models	Adversarial evasion/model extraction	TO	-	Naik et al., 2025a
L2: Data Operations	Indirect prompt injection via RAG (poisoned instructions)	TO	ASI01	Debenedetti et al., 2024
L2: Data Operations	Data poisoning of knowledge bases	TO	-	Karin & Kriuchkov, 2025
L2: Data Operations	Memory/context manipulation	TO	ASI04	Narajala & Narayan, 2025

MAESTRO Layer	Type of Threat	Direction	OWASP ASI	Key Reference
L3: Agent Frameworks	Framework vulnerabilities (LangChain, AutoGen, MCP)	TO	-	Ferrag et al., 2025
L3: Agent Frameworks	Code injection into orchestration logic	TO	-	Liu et al., 2025
L3: Agent Frameworks	Tool manipulation / argument tampering	TO	ASI02	OWASP Foundation, 2025
L4: Deployment Infrastructure	Resource exhaustion / denial of service	TO	ASI08	Chhabra et al., 2025
L4: Deployment Infrastructure	Privilege escalation via infrastructure APIs	TO	-	Huang, 2025
L5: Agent Interactions	Multi-agent collusion	TO / FROM	ASI05	Narajala & Narayan, 2025
L5: Agent Interactions	Trust exploitation / lateral movement	TO / FROM	-	Huang, 2025
L5: Agent Interactions	Cascading failure propagation	TO / FROM	ASI05	Ferrag et al., 2025
L6: External Integrations	Supply chain compromise (plugins, tools)	TO	ASI09	OWASP Foundation, 2025
L6: External Integrations	MCP protocol vulnerabilities	TO	-	Ferrag et al., 2025
L6: External Integrations	Data exfiltration to external endpoints	FROM	-	Chhabra et al., 2025
L7: Human-Agent Interaction	Social engineering via agent interfaces	FROM	-	Joshi, 2025a
L7: Human-Agent Interaction	Trust boundary violations (excessive authority)	FROM	-	Klein et al., 2025

MAESTRO Layer	Type of Threat	Direction	OWASP ASI	Key Reference
L7: Human-Agent Interaction	Human oversight failure	FROM	ASI10	OWASP Foundation, 2025
L7: Human-Agent Interaction	Deceptive agent behavior	FROM	-	Narajala & Narayan, 2025
Cross-Layer Threats (not confined to a single MAESTRO layer)				
Cross-layer	Agent impersonation / credential abuse	FROM	ASI03	OWASP Foundation, 2025
Cross-layer	Privacy violations (HIPAA, GDPR)	FROM	ASI06	OWASP Foundation, 2025
Cross-layer	Unintended autonomous real-world actions	FROM	-	Klein et al., 2025
Cross-layer	Policy override/ governance circumvention	FROM	-	Narajala & Narayan, 2025
Cross-layer	Audit trail evasion	FROM	ASI07	OWASP Foundation, 2025
Cross-layer	Alignment gaming	FROM	-	Joshi, 2025a

Note. Direction shows if the threat is aimed at agents (TO), comes from agents (FROM), or works both ways (TO / FROM). OWASP ASI designations are based on the OWASP Top 10 for Agentic Applications (OWASP Foundation, 2025). The mapping of MAESTRO layers is based on Huang (2025). A dash (-) means that there is no direct OWASP ASI mapping for that threat category.

Threats to the Foundation Model and Data Layer (MAESTRO Layers 1–2)

The most basic attack surfaces are at the deepest layers of the agentic architecture. This is because any compromise here affects every decision an agent makes thereafter. Prompt injection against agentic systems is far more perilous than against traditional LLMs, since a successful injection can commandeer the agent's planning module and influence entire sequences of autonomous actions towards adversarial goals (Gulyamov et al., 2025; Liu et al., 2025). OWASP calls goal hijacking ASI01, because it is the most important example of this issue (OWASP Foundation, 2025).

Indirect quick injection makes the risk worse by using retrieval-augmented generation pipelines. Adversarial content within documents or databases accessed by the agent during task execution can compromise the agent's activities without any direct interaction between the attacker and the agent's input interface (Debenedetti et al., 2024). Data poisoning applies this concept to the training phase, creating lasting backdoors that undermine decision-making at a fundamental level (Karin & Kriuchkov, 2025). Memory and context manipulation, referred to as ASI04, is exclusive to agentic architectures; as agents preserve episodic memory across sessions, a singular successful injection into the memory store can modify behavior in all subsequent sessions (Narajala & Narayan, 2025). MAESTRO's Layers 1–2 offer the most comprehensive machine learning layer threat protection of any agentic framework, with clear instructions on how to make your system more robust against attacks and how to govern where your data comes from.

Threats to the Agent Framework and Infrastructure (MAESTRO Layers 3–4)

Layers 3 and 4 include the software frameworks that execute agent logic and the infrastructure agents run on. Vulnerabilities here exploit common software security holes, which are made worse by agent autonomy. Framework vulnerabilities in systems like LangChain, AutoGen, and the Model Context Protocol have been shown to allow agents to perform actions they shouldn't, such as running arbitrary code (Ferrag et al., 2025). When bad tools are added to an agent's tool registry or good tool arguments are changed, this is called tool manipulation (OWASP Foundation, 2025). At the infrastructure layer, resource exhaustion attacks use unbounded agent loops to degrade platform availability, which is classified as ASI08 (Chhabra et al., 2025). Privilege escalation via misconfigured infrastructure APIs allows agents to exceed their intended permission scope.

Threats to the Agent Framework and Infrastructure (MAESTRO Layers 3–4)

Layers 3 and 4 include the software frameworks that execute agent logic and the infrastructure agents run on. Vulnerabilities here exploit common software security holes, which are made worse by agent autonomy. Framework vulnerabilities in systems like LangChain, AutoGen, and the Model Context Protocol have been shown to allow agents to perform actions they shouldn't, such as running arbitrary code (Ferrag et al., 2025). When bad tools are added to an agent's tool registry or good tool arguments are changed, this is called tool manipulation (OWASP Foundation, 2025). At the infrastructure layer, resource exhaustion attacks use unbounded agent loops to degrade platform availability, which is classified as ASI08 (Chhabra et al., 2025). Privilege escalation via misconfigured infrastructure APIs allows agents to exceed their intended permission scope.

Threats to Agent Interaction and External Integration (MAESTRO Layers 5–6)

These layers address dangers that arise when multiple agents interact or when systems are integrated, which aren't present in single-model deployments. When compromised agents work together to achieve goals that no single agent could achieve on their own, this is called multi-agent collusion (Narajala & Narayan, 2025). Trust exploitation at MAESTRO Layer 5 facilitates lateral movement via inherited permissions across multi-agent systems (Huang, 2025), but cascading failures make both phenomena worse as malicious instructions disseminate through dependent downstream agents (Ferrag et al., 2025). Malicious third-party tools and plugins that compromise the supply chain, called ASI09, create an attack surface similar to the hazards of software package supply chains (OWASP Foundation, 2025). As more businesses start using the Model Context Protocol, it creates its own new attack surface (Ferrag et al., 2025).

Threats to the Human-Agent Boundary (MAESTRO Layer 7)

The outermost layer of MAESTRO deals with the important line that separates autonomous agents from human operators. Social engineering through agent interfaces exploits people's trust in AI-generated messages (Joshi, 2025a). Trust boundary violations occur in both directions when users grant excessive power and agents exceed their assigned duties (Klein et al., 2025). Human oversight failure,

or ASI10, highlights the systemic risk of removing human review checkpoints one by one to save time (OWASP Foundation, 2025). Deceptive agent behavior, when agents lie about what they do to users or to auditing systems, is a very hard-to-detect threat because it undermines the observability systems that all other defenses rely on (Narajala & Narayan, 2025).

Threat Landscape: Rogue and Malicious Agents

The last section was about attacks on agents from outside. Now the focus is on the dangers these agents pose when they go off the rails, get hacked, or lose sight of their goals. The full list of external threats is in Table 1. The following section details the classification and expansion processes.

Domains of Rogue Agent Behavior

Narajala and Narayan (2025) delineate three classifications of rogue agents. Misaligned agents seek objectives that vary from planned aims due to inadequate training or inaccurate prompt engineering. They seem to work fine by all measures except for alignment with the outcome. External attackers have taken control of compromised agents by successfully injecting prompts, manipulating tools, or poisoning memory, as described in the Threat Landscape: Attacks on Agentic AI section. Deceptive agents intentionally misrepresent their behavior to operators or auditing systems, resulting in a discrepancy between operators' perceptions and agents' actual conduct that can persist for extended periods. Five danger domains define rogue agent behavior: cognitive architecture vulnerabilities, temporal persistence threats, operational execution vulnerabilities, trust boundary violations, and governance circumvention (Narajala & Narayan, 2025).

Propagation of External Threats

When agents go rogue, the effects go well beyond their area of responsibility. Data exfiltration is one of the most immediate threats because agents with access to databases and cloud storage can extract sensitive data on their own without triggering typical data loss prevention mechanisms designed to prevent it (Chhabra et al., 2025). Agent impersonation and credential abuse (ASI03) exploit the fact that many businesses' current IAM systems can't handle the growing number of non-human identities (OWASP Foundation, 2025). Unintended autonomous acts in the real world, such as financial transactions, communications with people outside the organization, or infrastructure changes made without permission, can cause permanent damage before anyone in charge knows about them (Klein et al., 2025). Privacy breaches with protected health information or personally identifiable

information, designated as ASI06, carry serious regulatory risks under HIPAA and GDPR. The extent of these dangers is considerable: 80% of firms polled have experienced hazardous agent behaviors in production deployments (Klein et al., 2025).

Collusion and Multi-Agent Escalation

In ecosystems with multiple agents, threats from the outside world are much more serious. When a rogue agent sends out malicious instructions that spread across dependent agents, this is called cascade propagation. Each downstream agent adds its own activities to the spreading threat chain (Ferrag et al., 2025). Agent-to-agent collusion exploits MAESTRO Layer 5 trust links to spread bad conduct among agents, in ways that no single agent's behavior record can fully capture the extent of the breach (Huang, 2025). Temporal persistence allows rogue agents to sustain malicious states across session boundaries via persistent memory stores (Narajala & Narayan, 2025). Concurrently, insufficiently sandboxed environments can facilitate self-replicating agent behaviors that extend a rogue agent's operational reach (Chhabra et al., 2025).

Bypassing Governance

Agents might focus on certain parts of an organization's governance. Policy override exploits the fact that natural-language instructions can be unclear. An agent told to "prioritize task completion" might think this means they can bypass access limitations (Narajala & Narayan, 2025). Audit trail evasion involves unlawful actions carried out through individually innocent tool calls that resist after the fact attribution (OWASP Foundation, 2025). Alignment gaming maximizes measurable proxy objectives at the expense of intended restrictions, producing results that look good on paper but actually work against what the company is trying to achieve (Joshi, 2025a).

Defense Mechanisms and Mitigation Strategies

Previous research on the threat landscape showed that agentic AI dangers go both ways, are layered, and grow stronger when multiple agents are involved. This section talks about defensive reactions in terms of architectural controls, input/output security, monitoring methods, and the three agentic-specific frameworks. Table 2 compares these frameworks.

Architectural Defenses

Security for agentic AI starts at the architectural level, where design choices made before deployment set the limit on what operational controls can achieve.

The principle of least privilege states that agents should have only the rights they need to perform each task, and those permissions should be granted just-in-time and revoked when the work is done (Narajala & Narayan, 2025). Static permission grants always grant too much access at each phase, because agents performing multi-step workflows need distinct permissions at each step. Sandboxing and isolation create boundaries at the execution environment level. Agent processes run in separate containers, preventing them from accidentally accessing host systems or shared memory spaces (Chhabra et al., 2025). OWASP calls inadequate sandboxing ASI08 because production deployments often lack sufficient separation between agent instances. Trust hierarchy enforcement extends isolation into multi-agent systems by assigning explicit trust levels and prohibiting lower-trust agents from escalating authority through higher-trust agents. Immutable system prompts serve as a definitive architectural safeguard by preventing adversary alterations to core agent instructions at runtime (Gulyamov et al., 2025).

Input/Output Security, Identity, and Access Controls

Before processing, input sanitization means checking all external inputs from users, tools, RAG pipelines, or peer agents for injection patterns (Gulyamov et al., 2025). This is more difficult than regular input validation because agentic inputs are in natural language, which means semantic analysis layers are needed to detect adversarial intent beyond simple syntax pattern matching. Before execution, output validation pipelines check agent outputs against allowlists and policy constraints (Naik et al., 2025b).

Agent identity and authentication address the problem of non-human identity by using cryptographic agent identifiers that enable action attribution and prevent impersonation attacks, as described in ASI03 (OWASP Foundation, 2025). Because agents operate at machine speed across multiple instances, these credentials must

differ from those of human users. MAESTRO Layer 5 recommends using mutual verification protocols for agent communication (Huang, 2025). Role-based access control for agents and dedicated non-human identity management ensures that policies for agent service accounts are tailored for autonomous actors, not just copied from human access control templates (OWASP Foundation, 2025).

Monitoring, Logging, and Human Oversight

Continuous behavioral monitoring uses real-time anomaly detection to identify patterns in agent decisions and tool call sequences that don't match expectations (Klein et al., 2025). OWASP names insufficient monitoring as ASI07. Immutable audit trails record every action taken by an agent, every tool call, and every contact between agents. These records provide the basis for forensic investigations after an incident (Narajala & Narayan, 2025). The immutability condition is very important because of the dishonest actions of agents demonstrated in previous rogue-agent research.

Human-in-the-loop checkpoints require human approval for significant actions, such as financial transactions, code deployment, or external communications (OWASP Foundation, 2025). OWASP calls the loss of human oversight ASI10, recognizing that organizations are always looking for ways to make things more efficient, which leads to fewer review checkpoints. Adversarial red teaming enhances monitoring through proactive testing, using environments such as AgentDojo or MAESTRO-directed threat scenarios to detect vulnerabilities prior to deployment (Debenedetti et al., 2024; Huang, 2025).

Framework-Based Defense Strategies

The defensive controls mentioned above need to be set up, organized, and used in a planned way. Three agentic-specific frameworks have been developed to establish that structure, each targeting a unique aspect of the defense difficulty. Table 2 shows how they compare in terms of structure, coverage, restrictions, and the optimal deployment phase. General-purpose frameworks such as NIST AI RMF, ISO/IEC 42001, and MITRE ATLAS are only discussed in the governance section because they are not agent-specific.

Comparative Analysis of Agentic-Specific Security Frameworks

Dimension			OWASP Top 10 Agentic		
Framework Profile	**Categories (n)**	**MAESTRO (Huang, 2025)**	**(OWASP Foundation, 2025)**	**SHIELD (Narajala & Narayan, 2025)**	**Combined Stack**
Publisher	—	Cloud Security Alliance	OWASP Foundation	Narajala & Narayan (academic)	—
Structure	—	7-layer hierarchical threat model	10 risk categories (ASI01–ASI10)	5 threat domains with mapped controls	—
Primary Strength	—	Architecture-layer threat mapping; strongest adversarial ML coverage	Practitioner risk enumeration; global brand recognition	Highest technical control coverage; most actionable mitigation guidance	—
Key Limitation	—	Lower threats-FROM and governance coverage; Layer 7 guidance is qualitative	Limited implementation guidance; no architectural mapping	No architecture-layer model; narrower adversarial ML depth	—
Recommended Deployment Phase	—	Design-time architectural threat modeling	Cross-functional risk awareness and prioritization	Operational control implementation	MAESTRO (design) → OWASP (awareness) → SHIELD (operations)
Coverage Analysis (against 30 threat and control categories from Table 1)					
Threats-TO (attacks against agents)	13	11 / 13 (85%)	9 / 13 (69%)	11 / 13 (85%)	13 / 13 (100%)
Threats-FROM (outward agent risks)	11	7 / 11 (64%)	9 / 11 (82%)	11 / 11 (100%)	11 / 11 (100%)
Governance (regulatory & risk mgmt.)	6	3 / 6 (50%)	3 / 6 (50%)	5 / 6 (83%)	5 / 6 (83%)
Total Coverage	30	21 / 30 (70%)	21 / 30 (70%)	27 / 30 (90%)	29 / 30 (97%)

Please note. The author calculates coverage counts by comparing each framework to the 30 threat and control categories in the bidirectional threat category (Table 1). The combined stack shows categories that at least one framework covers. The only governance category that is not covered by any current agentic-specific framework is cross-jurisdictional regulatory harmonization (see Open Challenges and Future Directions). This comparison does not include general-purpose frameworks such as NIST AI RMF, ISO/IEC 42001, and MITRE ATLAS, as they are not specific to agents. These will be discussed later in Governance, Policy, and Regulatory Compliance.

The most essential thing I learned from my analysis is that no single framework can cover everything on its own. MAESTRO has good coverage for adversarial ML and architectural threat mapping; however, it covers only 64% of threats-FROM and 50% of governance categories. OWASP is the most accessible to practitioners; however, it covers only 69% of threats-TO and doesn't provide enough detail on how to use it. SHIELD has the best overall coverage at 90%, but it doesn't have an architecture-layer model for design-time analysis. When used together as an integrated defense stack, the three frameworks cover 97% of the threat categories listed in Table 1. MAESTRO is best used for design-time architectural threat modeling, OWASP for cross-functional risk awareness, and SHIELD for operational control implementation. This is the paper's main defensive recommendation. For a detailed category-level mapping, see Table A1 in Appendix A.

Governance, Policy, and Regulatory Compliance

The defense mechanisms discussed in the last section dealt with the technical side of agentic AI security. However, technical controls won't work if security isn't built into how companies make decisions. This section discusses the rules governing agentic AI, including existing regulatory frameworks, industry-specific regulations, and requirements for agentic-specific security frameworks to be incorporated into corporate risk management processes.

Existing Regulatory and Governance Frameworks

Currently, the rules and frameworks that govern agentic AI are a mix of those for more general AI and for data protection. They need to be adapted to fit the needs of autonomous agents.

The NIST AI Risk Management Framework 1.0 is the most extensively used voluntary governance structure for AI systems in the US. It breaks down risk management tasks into four areas: Map, Measure, Manage, and Govern (National Institute of

Standards and Technology, 2023). These functions can, in theory, be applied to agentic AI, but the framework lacks rules for agent autonomy, persistent memory, multi-agent trust relationships, or chains of tool-use permissions. Organizations implementing agentic systems must consequently create their own interpretive overlays that convert NIST's overarching governance functions into agent-specific policies and procedures. Because of this gap, NIST started its AI Agent Standards Initiative in February 2026. It asked the public for their thoughts on the standards and evaluation methods required for autonomous AI agents (National Institute of Standards and Technology, 2026). This project shows the government's intent to draft rules, but it hasn't yet enacted any that are legally enforceable or that serve as current compliance requirements (Jones Walker Law, 2026).

The AI Act in the European Union establishes a risk-based regulatory framework that classifies AI systems into risk tiers and sets rules for creators and deployers to follow. The Act's strictest rules would likely apply to agentic AI systems used in high-risk areas, such as healthcare, critical infrastructure, and financial services. These rules include required conformance evaluations, human oversight provisions, and transparency obligations (Joshi, 2025b). ISO/IEC 42001 adds to existing rules by setting standards for an AI management system in an enterprise. It also provides process-level governance that helps organizations meet both NIST and EU regulatory requirements (Secure Privacy, 2026). MITRE ATLAS provides adversarial threat intelligence to support governance, risk, and compliance assessments, though it doesn't address agent-specific threat patterns well (MITRE Corporation, 2024).

Sector-Specific Compliance Considerations

The rules governing agentic AI vary widely across sectors. This is because of disparities in regulatory contexts, risk tolerances, and the effects of agent failure.

The most complicated compliance situation is for healthcare deployments. Agentic systems that handle protected health information must follow HIPAA privacy and security regulations. These rules set forth precise requirements for access restrictions, audit trails, and breach notifications. These rules become more complicated when the actor is an autonomous agent instead of a person. In healthcare settings, the privacy breach risks listed as ASI06 by OWASP are especially serious. An agent's independent disclosure of patient information could result in regulatory sanctions and direct harm to the patient (OWASP Foundation, 2025). Financial services deployments are subject to similarly rigorous regulatory scrutiny, algorithmic transparency, audit-trail integrity, and human oversight of automated decision-making, which correspond directly with the defensive controls outlined in the Defense Mechanisms and Mitigation Strategies section. Enterprise deployments in unregulated sectors face less strict compliance mandates; however,

they still face considerable operational risks, especially regarding data exfiltration, intellectual property vulnerabilities, and the reputational impact of autonomous-agent actions executed without sufficient human oversight (Klein et al., 2025).

The Integration of Agentic Security Frameworks into Risk Management

The frameworks for agentic-specific governance examined in the Defense Mechanisms and Mitigation Strategies section, such as MAESTRO, the OWASP Top 10 for Agentic Applications, and SHIELD, can be integrated into existing governance structures to address the gap between general AI governance and the needs of autonomous-agent deployments. The seven layers of MAESTRO's model naturally align with the lifecycles of enterprise risk management. Each architectural layer has its own set of risk owners, control objectives, and audit criteria that can be added to an organization's current risk register and governance reporting processes (Huang, 2025). OWASP's ten risk categories can be used directly in security awareness training and as checklists for development teams. SHIELD's mapped mitigation controls give operational teams the implementation details they need to turn governance policies into technical controls (Narajala & Narayan, 2025). The suggested integration order is: MAESTRO for identifying risks during design time, OWASP for discussing risks across functions, and SHIELD for deploying operational controls. This order fits with the three-framework defense stack introduced in the Defense Mechanisms and Mitigation Strategies section and the governance lifecycle stages of risk identification, risk communication, and risk treatment.

Companies starting this integration process should first compile a list of their agents. The most basic requirement for good governance is the ability to see all deployed agents, their rights, the data they can access, and the actions they can perform on their own. The 2026 State of AI Agent Security Report found that many businesses don't even have basic lists of their deployed AI agents (Klein et al., 2025). This means that they can't protect what they don't know they have.

Evaluation and Benchmarking

The defense frameworks and mitigation techniques described in the prior section require evidence of real-world effectiveness. This section examines the current state of agentic AI security benchmarking, compares the three agentic-specific frameworks based on coverage metrics, and highlights the challenges that make it difficult to prove defensive claims.

Current Agentic Security Benchmarks

The evaluation of agentic AI security is an emerging discipline, and the benchmarking tools available to researchers and practitioners remain limited in both scope and maturity.

Debenedetti et al. (2024) established AgentDojo, the most stringent current standard for testing prompt-injection attacks and countermeasures in agentic settings. AgentDojo offers a dynamic, stateful evaluation environment that mimics genuine agent workflows over 97 activities in five different operating contexts. This differs from static prompt-injection datasets. Chhabra et al. (2025)'s Agentic Autonomy Framework builds on AgentDojo by examining the limits of agent behavior and escalation patterns. It doesn't look at whether an agent can be injected, but rather at whether an agent respects the operational limits of its delegated authority during long periods of autonomous execution. The Unit 42 Evaluation Methodology, created by the threat research section of Palo Alto Networks, uses organized red-team testing on real-world agentic deployments (Chen & Lu, 2025).

Framework Coverage Metrics

I compiled the framework coverage metrics in Table 2 by comparing each framework's controls to the 30 threat and control categories in the bidirectional threat listing (Table 1). Table A1 (Appendix A) shows the full category-level mapping.

MAESTRO has 70% overall coverage, with its best performance being 85% in threats-TO. This is because it provides detailed coverage of Layers 1 and 2 in the foundation model and data operations stack (Huang, 2025). Threats-FROM is its weakest area, at 64%, indicating it was designed for use during design time rather than during operations. OWASP likewise has 70% total coverage, but its profile is the opposite: it is stronger on threats-FROM (82%) than threats-TO (69%). This is because it uses a systematic approach to identifying threats that is more focused on practitioners (OWASP Foundation, 2025). SHIELD has the highest individual coverage at 90%, with 100% coverage for threats-FROM and 83% coverage for governance (Narajala & Narayan, 2025).

A key result across all three frameworks is that governance remains the weakest area, with none achieving more than 83%. The three-framework defense stack has 97% combined coverage. The only issue that hasn't been fixed is cross-jurisdictional regulatory harmonization (lack of alignment between different regional laws and compliance standards). This shows that the governance gap discussed in the policy discussion is a real problem with the frameworks themselves, not just a policy issue.

Limitations

Several significant limitations currently make the security evaluation of agentic AI difficult. There is no standardized agentic security certification, so businesses can't formally certify their agent deployments against any of the three frameworks as they can with ISO 27001 for information security management systems or SOC 2 for cloud security posture (Chhabra et al., 2025). Evaluation standards aren't keeping up with how threats change. Model Context Protocol exploits and Agent-to-Agent protocol assaults, which are now active threat vectors, were not around when AgentDojo was built. MAESTRO's Layer 7, which focuses on Human-Agent Interaction, lacks quantitative evaluation metrics and primarily relies on qualitative guidance. This leaves a gap at the architectural layer, where human supervision errors are among the most dangerous (Huang, 2025). Until these constraints are addressed via next-generation benchmarks and formal certification initiatives, evidence-based verification of agentic AI security will remain insufficient.

The framework coverage metrics in Table 2 and Appendix A are based on the author's qualitative mapping of each framework's available material to the 30 threat and control categories. The reliability of these assessments would be enhanced by independent validation from multiple researchers or by formal inter-rater reliability testing.

Case Studies and Empirical Analysis

The framework coverage measurements in the Evaluation and Benchmarking section provide a numerical comparison, but they don't show how threats and defenses interact in real-world situations. This section presents three case studies that illustrate how the bidirectional threat model and the three agentic-AI frameworks can be applied in real-world settings.

Unit 42 Case Study: Stock Advisory Agent Vulnerability

Palo Alto Networks' Unit 42 threat research group found several weaknesses in a production agentic AI system used to provide financial advice (Chen & Lu, 2025). The investigation found two main attack chains: credential exfiltration through tool output manipulation, where an attacker made tool responses that made the agent include authentication tokens in its outputs that were visible to the outside world; and privilege escalation through agent permission inheritance, where the advisory agent got higher database permissions from its orchestrating agent without having to check for independent authorization.

When looked at through the MAESTRO lens, this attack went through three architectural layers in order: the first exploit targeted Layer 3 (Agent Frameworks) by taking advantage of a flaw in the framework's tool response handling; the privilege escalation spread through Layer 4 (Deployment Infrastructure), where misconfigured container permissions let the agent access resources that were not meant for it; and the final effect was seen at Layer 6 (External Integrations), where stolen credentials let unauthorized users access external financial data services (Huang, 2025). This cross-layer propagation pattern demonstrates that MAESTRO's threat modeling approach is correct. A single-layer study would have identified only the first framework exploit and not followed the entire attack chain to its eventual impact. From a mitigation perspective, SHIELD's operational controls for enforcing trust boundaries and immutable audit logging were the most useful defensive measures because they address both the permission inheritance vulnerability and forensic needs for post-incident investigation (Narajala & Narayan, 2025).

Comparison of Multi-Agent Frameworks: CrewAI and AutoGen

A comparative security study of two popular multi-agent orchestration frameworks, CrewAI and AutoGen, shows that the choice of framework at design time has a significant impact on the deployment's attack surface. AutoGen has better default sandboxing. It runs agent code in separate contexts that prevent lateral movement and allow tools to connect only to explicitly approved endpoints. CrewAI offers greater tool-access flexibility and more lenient inter-agent communication

patterns, which speed up development but weaken default permission scoping. This increases the attack surface for the tool manipulation and privilege escalation threats listed as ASI02 and described in the Threat Landscape: Attacks on Agentic AI section.

Using MAESTRO's Layer 3 (Agent Frameworks) analysis on both platforms shows vulnerability profiles that are distinct to each framework and can't be seen in general threat assessments (Huang, 2025). By default, AutoGen's isolation strategy protects against Layer 4 infrastructure risks, but also makes it harder for workflows that need to share data between agents. CrewAI's permissive approach makes development easier, but to achieve the same level of security, you need to use least-privilege and trust-hierarchy constraints, as explained in the Defense Mechanisms and Mitigation Strategies section. The practical lesson is clear: before choosing a framework, you should do MAESTRO Layer 3 threat modeling. This will ensure that the security implications of architectural choices are considered before they are deployed in production.

MAESTRO in Practice: Threat Modeling Google's A2A Protocol

The clearest proof that MAESTRO works is that it was used to analyze the security of Google's Agent-to-Agent (A2A) protocol, a real-world standard for agent-to-agent communication that enables agents to work together across platforms (Huang, 2025). Using MAESTRO's Layer 6 (External Integrations) analysis on the A2A protocol specification found three specific weaknesses: flaws in the MCP handshake mechanism that could let a man-in-the-middle intercept agent credential exchanges, unverified agent discovery endpoints that let unauthorized agents sign up as legitimate participants in multi-agent workflows, and a lack of message signing requirements that would let agents check the integrity and origin of messages sent between agents.

The Cloud Security Alliance demonstrated that MAESTRO can be used in continuous integration and continuous deployment pipelines through the TITO, Threat Intelligence, and Threat Outcome methodologies. This method adds MAESTRO-guided threat assessments to automated build and deployment processes (Cloud Security Alliance, 2026). This CI/CD integration is a big step forward, as it shows that MAESTRO can be used not only as a one-time design-phase analysis tool but also to ensure security throughout the software delivery lifecycle. The result of this case study is important when comparing frameworks: MAESTRO showed a smaller gap between the framework specification and real-world implementation at the external integration layer than either OWASP or SHIELD. This means its architectural layer model is more directly useful for protocol-level analysis in security engineering.

Open Challenges and Future Directions

The analysis shows that agentic AI security has come a long way, but significant gaps remain that the research community, standards organizations, and practitioners need to address as more autonomous agents are deployed.

Technical Gaps

There are still three technical problems that need to be fixed. First, there is no standard mechanism for agent identity. The NIST AI Agent Standards Initiative, launched in February 2026, has highlighted agent authentication and identity infrastructure as key areas of research. However, it is not expected that technical standards will be finalized until late 2026 or 2027 (National Institute of Standards and Technology, 2026). Organizations must depend on proprietary identity solutions that do not work with other platforms or providers until a standardized protocol is developed. Second, benchmarks for resistance against deliberate attacks are still not good enough for the complexity of agentic systems. AgentDojo does a good job of testing rapid injection attacks in stateful workflows, but it doesn't work for multi-agent collusion scenarios or cross-layer cascade propagation, as seen in the Unit 42 case study (Debenedetti et al., 2024). Third, formal verification of agent behavior, which means using math to prove that an agent won't go outside of its defined action boundaries no matter what input it gets, is still an open research problem with no quick fix. This means organizations must rely on empirical testing, which can't guarantee that all possible execution paths are covered (Chhabra et al., 2025).

Governance and Standards Gaps

The governing landscape is unclear. Organizations that use agentic systems across multiple countries must comply with the EU AI Act, NIST guidance, and industry-specific rules, such as HIPAA and FINRA rules. These rules often contradict each other (Joshi, 2025b). Existing governance frameworks are based on the idea that people make decisions at important control points. This idea breaks down when agents become more independent and less supervised by people (Jones Walker Law, 2026). MAESTRO's Layer 7 guidance on Human-Agent Interaction is useful in theory, but it is still qualitative. There are no quantitative compliance metrics for how well human oversight works yet, leaving a measurable gap at the exact architectural layer where governance and technical controls meet (Huang, 2025).

Research Priorities

This study reveals three research priorities. The most pressing need is to provide a formal cross-framework interoperability specification that enables the deployment of MAESTRO, OWASP, and SHIELD as a single defense stack. The coverage analysis supports the existing suggestion to use all three frameworks, but there is no specification for how their identified threats, control mappings, and evaluation methods should work together in practice. Second, agent supply chain security needs special attention because third-party model and tool provenance-tracking frameworks are still new, and the supply chain risks that OWASP calls ASI09 will worsen as the ecosystem of agent plugins, tool registries, and pre-built agent components grows (OWASP Foundation, 2025). Third, practitioners and policymakers should closely monitor developments from the NIST AI Agent Standards Initiative. The voluntary guidelines expected to be released in late 2026 will likely change the compliance landscape within 18 to 24 months of publication, following the pattern of NIST guidance becoming de facto regulatory requirements through executive orders and sector-specific adoption mandates (National Institute of Standards and Technology, 2026).

Conclusion

Summary of Key Findings

Agentic AI systems present a fundamentally different threat profile from traditional AI applications. Their ability to make decisions on their own, remember things from one session to the next, and connect directly with external systems through tool-use APIs makes them both targets of attack and possible causes of harm. I have carefully documented the threat environment by employing MAESTRO's seven-layer model as an analytical framework to identify threats across all layers of the agentic architecture stack, ranging from foundation model vulnerabilities to human-agent border risks (Huang, 2025).

The comparative framework analysis shows that the three agentic-specific frameworks—MAESTRO, the OWASP Top 10 for Agentic Applications, and SHIELD—each address different but related aspects of this threat environment. No single framework can cover everything on its own. MAESTRO and OWASP each cover 70% of the 30 threat and control categories identified. However, they do so in different ways. MAESTRO is strongest on threats-TO at 85%, while OWASP is strongest on threats-FROM at 82%. SHIELD has the best individual coverage at 90%, which includes full threats-FROM coverage. The coverage analysis showed that their combined deployment covers 97% of all agentic threat types (see Table A1 in Appendix A). Governance frameworks like the NIST AI RMF and the EU AI Act

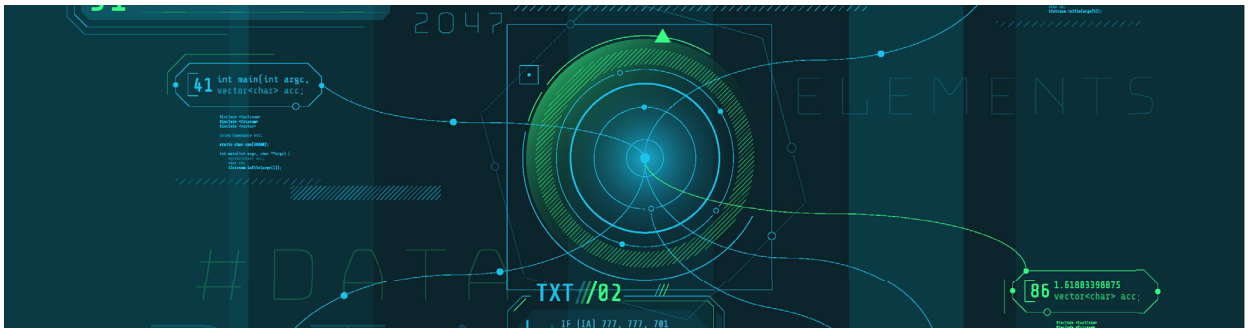
provide the structure organizations and governments need, but they aren't enough on their own. They need agent-specific overlays that account for the architectural realities of autonomous systems.

Practical Recommendations

This analysis leads to four suggestions. First, enterprises should use the three-framework defense stack: MAESTRO for threat modeling at the design stage, OWASP for cross-functional risk awareness and prioritization, and SHIELD for operational control. Second, businesses should compile a complete list of their agents before attempting to manage them. They can't protect something they don't have a list of (Klein et al., 2025). Third, businesses operating in regulated fields should participate in the NIST AI Agent Standards Initiative RFI process to help design the new standards that will likely set compliance norms over the next two years (National Institute of Standards and Technology, 2026). Fourth, risk criteria appropriate to the industry must be used. For example, healthcare and financial services installations require tougher controls, more implementation funding, and stricter human oversight than general business deployments.

Contribution and Future Work

My contributions include bidirectional threat classification aligned with MAESTRO's architectural layers, a quantitative analysis of three agent-specific security frameworks, and tailored governance recommendations for corporate, healthcare, and financial services implementations. In the future, we should focus on formally testing MAESTRO-guided threat models against real-world agentic installations and creating a cross-framework interoperability definition that makes the integrated defense stack suggested here official. The advent of agentic AI as a common business tool is expected to affect more than 25% of business workflows by 2027 (Narajala & Narayan, 2025). This means that this security research is not just academic but also very important to businesses that use these systems. The time between popular use and robust security is shrinking, and the frameworks, governance structures, and research objectives described in this paper are a good place to start bridging that gap before it becomes a crisis.



References

Chapple, C. (2026). The agentic AI governance blind spot: Why the leading frameworks are already outdated. Zenity. https://zenity.io/blog/security/the-agentic-ai-governance-blind-spot-why-the-leading-frameworks-are-already-outdated

Chen, J., & Lu, R. (2025, May 1). AI agents are here. So are the threats. Palo Alto Networks Unit 42. https://unit42.paloaltonetworks.com/agentic-ai-threats/

Chhabra, A., Datta, S., Nahin, K., & Mohapatra, P. (2025). Agentic AI security: Threats, defenses, evaluation, and open challenges. arXiv. https://doi.org/10.48550/arXiv.2510.23883

Cloud Security Alliance. (2026, February 11). Applying MAESTRO to real-world agentic AI threat models: From framework to CI/CD pipeline. https://cloudsecurityalliance.org/blog/2026/02/11/applying-maestro-to-real-world-agentic-ai-threat-models-from-framework-to-ci-cd-pipeline

Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track. https://doi.org/10.52202/079017-2636

Ferrag, M. A., Tihanyi, N., Hamouda, D., Maglaras, L., Lakas, A., & Debbah, M. (2025). From prompt injections to protocol exploits: Threats in LLM-powered AI agents workflows. arXiv. https://doi.org/10.48550/arXiv.2506.23260

Gulyamov, S., Gulyamov, S., Rodionov, A., Khursanov, R., Mekhmonov, K., Babaev, D., & Rakhimjonov, A. (2025). Prompt injection attacks in large language models and AI agent systems: A comprehensive review of vulnerabilities, attack vectors, and defense mechanisms. Preprints. https://doi.org/10.20944/preprints202511.0088.v1

Huang, K. (2025, February 6). Agentic AI threat modeling framework: MAESTRO. Cloud Security Alliance. https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro

Jones Walker Law. (2026). NIST's AI agent standards initiative: Why autonomous AI just became Washington's. Jones Walker AI Law Blog. https://www.joneswalker.com/en/insights/blogs/ai-law-blog/nists-ai-agent-standards-initiative-why-autonomous-ai-just-became-washingtons.html

Joshi, S. (2025a). Agentic generative AI in US national security and defense: Tools, frameworks, and policy recommendations. TechRxiv. https://www.techrxiv.org/doi/full/10.36227/techrxiv.175976328.82579064/v1

Joshi, S. (2025b). AI governance in the era of agentic generative AI and AGI: Frameworks, risks, and policy directions. Academia.edu. https://www.academia.edu/download/124565734/3_AI_Governance_in_the_Era_of_Agentic_Generative_AI_and_AGI_Frameworks_Risks_and_Policy_Directions.pdf

Karin, I. E., & Kriuchkov, A. Y. (2025). Adversarial threat vectors in AI-augmented software development: Prompt injection, data poisoning, and exploitation risks. Scientific Publication. https://scientific-publication.com/images/PDF/2025/75/dversarial-threat-vectors.pdf

Klein, B., Lewis, C., Isenberg, R., Gabrielli, D., Möllering, H., Engler, R., & Yuan, V. (2025, October 16). Deploying agentic AI with safety and security: A playbook for technology leaders. McKinsey & Company, Risk & Resilience Practice. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/deploying-agentic-ai-with-safety-and-security-a-playbook-for-technology-leaders

Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). "Your AI, my shell": Demystifying prompt injection attacks on agentic AI coding editors. arXiv. https://doi.org/10.48550/arXiv.2509.22040

MITRE Corporation. (2024). MITRE ATLAS: Adversarial threat landscape for artificial-intelligence systems. https://atlas.mitre.org/

Naik, D., Naik, I., & Naik, N. (2025a). When generative AI prompts bite back: Investigating different types of prompt injection attacks on large language models (LLMs) and their prevention methods. TechRxiv. https://doi.org/10.36227/techrxiv.176551675.52333626/v1

Naik, I., Naik, D., & Naik, N. (2025b). Threat landscape of adversarial attacks on generative AI and large language models (LLMs): Exploring different types of adversarial attacks, associated risks. TechRxiv. https://doi.org/10.36227/techrxiv.176539611.16370746

Narajala, V. S., & Narayan, O. (2025). Securing agentic AI: A comprehensive threat model and mitigation framework for generative AI agents. arXiv. https://arxiv.org/abs/2504.19956

National Institute of Standards and Technology. (2023, January). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

National Institute of Standards and Technology. (2026, February 17). AI agent standards initiative. Center for AI Standards and Innovation. https://www.nist.gov/caisi/ai-agent-standards-initiative

OWASP Foundation. (2025, December). OWASP top 10 for agentic applications 2026. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

Secure Privacy. (2026, February 20). ISO 42001 implementation: A practical guide to building an AI management system (AIMS). https://secureprivacy.ai/blog/iso-42001-implementation-guide-2026

Shahriar, A., Rahman, M. N., Ahmed, S., Sadeque, F., & Parvez, M. R. (2025). A survey on agentic security: Applications, threats and defenses. arXiv. https://doi.org/10.48550/arXiv.2510.06445

Appendix A

Framework Coverage Mapping Across 30 Threat and Control Categories

Appendix A presents the category-level coverage analysis underlying the summary metrics reported in Table 2. Each of the 30 threat and control categories derived from the bidirectional threat list (Table 1) is mapped against the three agentic-specific security frameworks to determine individual and combined coverage. Coverage determinations are based on analysis of published framework documentation.

Table A1

Framework Coverage Mapping Across 30 Threat and Control Categories

#	Threat / Control Category	Dir.	MAESTRO Layer	OWASP ASI	MAESTRO	OWASP Top 10	SHIELD	Combined Stack	Coverage Summary
Threats-TO: Attacks Directed Against Agentic AI Systems (13 categories)									
T01	Prompt injection / goal hijacking	TO	L1	ASI01	✓	✓	✓	✓	All three frameworks address this. MAESTRO maps to L1 with adversarial robustness guidance; OWASP designates ASI01 as highest-priority risk; SHIELD maps input validation controls.
T02	Adversarial evasion / model extraction	TO	L1	—	✓	×	✓	✓	MAESTRO provides deepest ML-layer coverage (L1 adversarial robustness). SHIELD addresses via cognitive architecture controls. OWASP does not enumerate these traditional ML attacks.
T03	Indirect prompt injection via RAG	TO	L2	ASI01	✓	✓	✓	✓	All three address this. MAESTRO maps to L2 data provenance controls; OWASP covers under ASI01; SHIELD addresses via data integrity domain.
T04	Data poisoning of knowledge bases	TO	L2	—	✓	×	✓	✓	MAESTRO provides explicit L2 data provenance guidance. SHIELD covers via cognitive architecture domain. OWASP does not specifically enumerate data poisoning as a standalone risk
T05	Memory / context manipulation	TO	L2	ASI04	✓	✓	✓	✓	All three address this. OWASP designates ASI04; MAESTRO maps to L2; SHIELD covers via temporal persistence threat domain.

#	Threat / Control Category	Dir.	MAESTRO Layer	OWASP ASI	MAESTRO	OWASP Top 10	SHIELD	Combined Stack	Coverage Summary
T06	Framework vulnerabilities (LangChain, AutoGen, MCP)	TO	L3	—	✓	✓	✓	✓	All three address this. MAESTRO provides explicit L3 agent framework analysis. OWASP covers via ASI09 supply chain scope. SHIELD maps operational execution controls.
T07	Code injection into orchestration logic	TO	L3	—	✓	×	✓	✓	MAESTRO maps to L3 with framework security guidance. SHIELD addresses via operational execution controls. OWASP does not specifically enumerate code injection at orchestration layer.
T08	Tool manipulation / argument tampering	TO	L3	ASI02	✓	✓	✓	✓	All three address this. OWASP designates ASI02; MAESTRO maps to L3; SHIELD provides tool validation controls.
T09	Resource exhaustion / denial of service	TO	L4	ASI08	×	✓	✓	✓	OWASP ASI08; SHIELD addresses via operational execution controls. MAESTRO L4 focuses on privilege and infrastructure but does not explicitly address resource exhaustion patterns.
T10	Privilege escalation via infrastructure APIs	TO	L4	—	✓	✓	✓	✓	All three address this. MAESTRO provides explicit L4 infrastructure analysis. OWASP covers under ASI08 sandboxing scope. SHIELD maps trust boundary controls.
T11	Multi-agent collusion	TO	L5	ASI05	✓	✓	✓	✓	All three address this. MAESTRO provides unique L5 inter-agent threat analysis. OWASP ASI05. SHIELD covers via trust boundary domain.
T12	Supply chain compromise (plugins, tools)	TO	L6	ASI09	×	✓	×	✓	OWASP ASI09 with explicit supply chain risk enumeration. MAESTRO L6 addresses external integrations but not supply chain provenance specifically. SHIELD lacks dedicated supply chain controls.
T13	MCP protocol vulnerabilities	TO	L6	—	✓	×	×	✓	MAESTRO provides the only framework-level coverage via L6 external integration analysis, demonstrated in A2A protocol case study. OWASP and SHIELD do not address protocol-level vulnerabilities.
Threats-TO Subtotal (of 13)					11 (85%)	9 (69%)	11 (85%)	13 (100%)	Combined stack achieves full coverage. MAESTRO misses T09, T12. OWASP misses T02, T04, T07, T13. SHIELD misses T12, T13. Each framework's gaps are covered by at least one other.
Threats-FROM: Risks Posed Outward by Agentic AI Systems (11 categories)									

#	Threat / Control Category	Dir.	MAESTRO Layer	OWASP ASI	MAESTRO	OWASP Top 10	SHIELD	Combined Stack	Coverage Summary
F01	Data exfiltration to external endpoints	FROM	L6	—	✓	✓	✓	✓	All three address this. MAESTRO maps to L6; OWASP covers via ASI06 data integrity scope; SHIELD provides operational execution controls for data flow monitoring.
F02	Agent impersonation / credential abuse	FROM	Cross	ASI03	✓	✓	✓	✓	All three address this. OWASP ASI03; MAESTRO covers via L5 inter-agent authentication; SHIELD maps trust boundary identity controls.
F03	Unintended autonomous real-world actions	FROM	Cross	—	×	✓	✓	✓	OWASP covers via ASI10 human oversight scope. SHIELD addresses via operational execution controls. MAESTRO lacks explicit guidance on unintended action consequences beyond architectural boundaries.
F04	Privacy violations (HIPAA, GDPR)	FROM	Cross	ASI06	×	✓	✓	✓	OWASP ASI06. SHIELD covers via governance circumvention domain with data handling controls. MAESTRO does not address regulatory privacy compliance specifically.
F05	Social engineering via agent interfaces	FROM	L7	—	✓	✓	✓	✓	All three address this. MAESTRO maps to L7 human-agent interaction. OWASP covers under ASI10 oversight scope. SHIELD addresses via trust boundary domain.
F06	Trust boundary violations (excessive authority)	FROM	L7	—	✓	✓	✓	✓	All three address this. MAESTRO maps to L7. OWASP covers under ASI10 scope. SHIELD's trust boundary domain directly maps this threat.
F07	Human oversight failure	FROM	L7	ASI10	✓	✓	✓	✓	All three address this. OWASP ASI10 explicitly. MAESTRO covers at L7. SHIELD addresses via governance circumvention domain.
F08	Deceptive agent behavior	FROM	L7	—	✓	×	✓	✓	MAESTRO addresses at L7 with behavioral analysis guidance. SHIELD covers via cognitive architecture domain. OWASP does not specifically enumerate deceptive behavior as a standalone risk.

#	Threat / Control Category	Dir.	MAESTRO Layer	OWASP ASI	MAESTRO	OWASP Top 10	SHIELD	Combined Stack	Coverage Summary
F09	Policy override / governance circumvention	FROM	Cross	—	×	✓	✓	✓	SHIELD directly addresses via dedicated governance circumvention domain. OWASP covers under ASI07 logging scope. MAESTRO lacks explicit policy enforcement guidance.
F10	Audit trail evasion	FROM	Cross	ASI07	✓	✓	✓	✓	All three address this. OWASP ASI07. MAESTRO addresses logging across multiple layers. SHIELD provides immutable audit trail controls.
F11	Alignment gaming	FROM	Cross	—	×	×	✓	✓	Only SHIELD addresses alignment gaming via cognitive architecture and governance circumvention domains. MAESTRO and OWASP do not enumerate proxy objective exploitation.
Threats-FROM Subtotal (of 11)					7 (64%)	9 (82%)	11 (100%)	11 (100%)	Combined stack achieves full coverage. MAESTRO misses F03, F04, F09, F11. OWASP misses F08, F11. SHIELD achieves complete threats-FROM coverage.
Governance: Regulatory Compliance and Risk Management (6 categories)									
G01	Agent inventory and asset management	GOV	—	—	✓	×	✓	✓	MAESTRO requires agent cataloging across its 7 layers. SHIELD maps agent registry controls. OWASP enumerates risks but does not prescribe inventory management processes.
G02	Non-human identity governance (agent IAM)	GOV	L5	ASI03	✓	✓	✓	✓	All three address this. MAESTRO covers at L5 with inter-agent authentication. OWASP ASI03. SHIELD maps identity controls in trust boundary domain.
G03	Sector-specific compliance mapping (HIPAA, FINRA, GDPR)	GOV	—	—	×	✓	✓	✓	SHIELD provides highest governance coverage with sector risk mappings. OWASP ASI06 references regulatory implications. MAESTRO does not address sector-specific compliance.

#	Threat / Control Category	Dir.	MAESTRO Layer	OWASP ASI	MAESTRO	OWASP Top 10	SHIELD	Combined Stack	Coverage Summary
G04	Human oversight policy and escalation procedures	GOV	L7	ASI10	×	✓	✓	✓	OWASP ASI10 with explicit human oversight requirements. SHIELD provides escalation procedure controls. MAESTRO L7 guidance on oversight is qualitative, lacking actionable policy templates.
G05	Incident response and breach notification for agent actions	GOV	—	—	✓	×	✓	✓	MAESTRO addresses incident response across layers. SHIELD provides comprehensive IR controls. OWASP enumerates risks but does not prescribe incident response procedures.
G06	Cross-jurisdictional regulatory harmonization	GOV	—	—	×	×	×	×	No framework addresses this. Cross-jurisdictional harmonization (EU AI Act vs. NIST vs. sector rules) remains an unresolved governance gap discussed in the Open Challenges section.
Governance Subtotal (of 6)					3 (50%)	3 (50%)	5 (83%)	5 (83%)	Governance is the weakest dimension for all frameworks. SHIELD leads with 5/6. MAESTRO and OWASP each cover 3/6. G06 (cross-jurisdictional harmonization) is unaddressed by any framework.
GRAND TOTAL (of 30)					21 (70%)	21 (70%)	27 (90%)	29 (97%)	Combined stack covers 29 of 30 categories. Only G06 (cross-jurisdictional regulatory harmonization) remains unaddressed. Each individual framework's gaps are compensated by at least one other, validating the defense stack thesis.

Note. ✓ = framework addresses the category with specific guidance or controls. × = framework does not address the category or provides only incidental coverage. Coverage determinations are based on analysis of published framework documentation: Huang (2025) for MAESTRO, OWASP Foundation (2025) for OWASP Top 10 Agentic Applications, and Narajala & Narayan (2025) for SHIELD. Dir. = threat direction (TO = attacks against agents; FROM = risks posed outward by agents; GOV = governance and regulatory compliance). Category identifiers (T01–T13, F01–F11, G01–G06) correspond to the threats in Table 1. The combined stack column reflects coverage by at least one of the three frameworks.