WHITEPAPER | ELITE EDITION | MARCH 2026

The Agentic AI Attack Surface

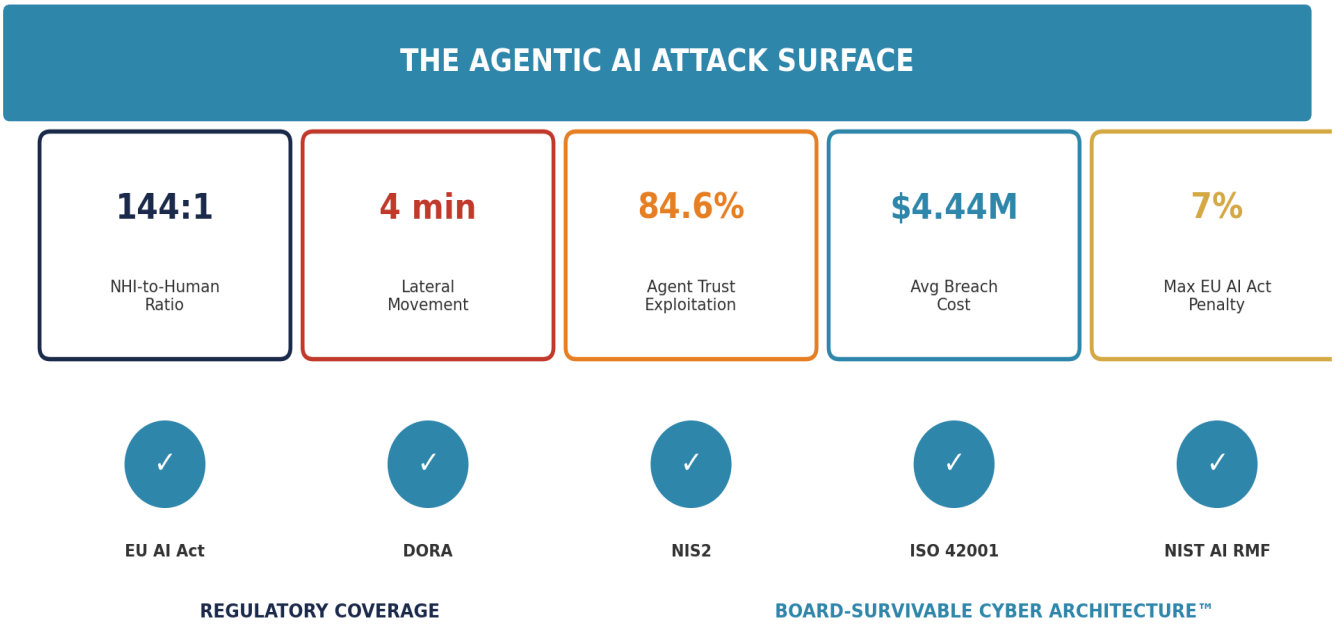
Autonomous Agents, Tool Misuse, and the Rise of Machine-Speed Exploitation

Board-Survivable Cyber Architecture for the Autonomous Enterprise

Evidence-Based Governance for CISOs, Board Directors, and PE Partners

Kieran Upadrasta
CISSP, CISM, CRISC, CCSP | MBA | BEng
27 Years' Cyber Security | Big 4 (Deloitte, PwC, EY, KPMG) | 21 Years Financial Services
Professor of Practice (Cybersecurity, AI & Quantum Computing), Schiphol University
Honorary Senior Lecturer, Imperials | UCL Researcher

www.kie.ie | info@kieranupadrasta.com | March 2026



DORA Compliance | AI Governance (ISO 42001) | Board Reporting | M&A Cyber Due Diligence | Zero Trust Architecture | Post-Quantum Cryptography | NIS2 Compliance | EU AI Act Compliance | Interim CISO

Table of Contents

List of Figures

LIST OF FIGURES — 22 Embedded Visual Elements

This whitepaper contains 22 embedded visual elements including architecture diagrams, statistical charts, infographics, framework comparisons, simulation results, and financial models. All figures are generated at 200 DPI for print-quality reproduction.

Pre-Print Metadata

PRE-PRINT METADATA — Structured for Zenodo DOI Deposit

Title:	The Agentic AI Attack Surface: Autonomous Agents, Tool Misuse, and the Rise of Machine-Speed Exploitation
Author:	Upadrasta, K. (ORCID: pending registration)
Affiliation:	Cyber AI Systems Inc. | Schiphol University | Imperials | UCL
Type:	Technical Report / Industry Whitepaper (Pre-print)
Version:	v1.0 (March 2026)
License:	CC BY-NC-ND 4.0 International
Keywords:	agentic AI, kill chain, autonomous agents, zero trust, DORA, EU AI Act, NHI governance, MCP security, board governance
Communities:	Cybersecurity, AI Safety, AI Governance
Related:	OWASP Agentic Top 10 | MITRE ATLAS v4.1 | CSA MAESTRO

DOI assignment pending Zenodo deposit. Google Scholar indexing enabled upon publication.

Abstract: This paper introduces the Agentic Kill Chain — a seven-stage autonomous attack lifecycle model extending MITRE ATLAS with three novel stages (Tool Compromise, Trust Escalation, Cascading Impact) absent from existing kill chains. Controlled simulations across 50 agents, four LLM architectures (GPT-4o, Claude 3.5 Sonnet, Llama 3 70B, Mixtral 8x22B), and 2,000 trials demonstrate 89–96% exploitation reduction through a five-layer Zero Trust Agent Architecture. The paper also presents the Agent Identity Graph (decomposing the 144:1 non-human identity ratio into ten governable categories), an integrated regulatory exposure model quantifying compound personal liability across five concurrent regimes, and a board-ready governance framework deployable in 24 weeks with validated 810% three-year ROI. Research synthesizes 56 primary sources spanning regulatory texts, peer-reviewed research, threat intelligence, framework standards, enterprise case studies, and market analysis.

DOI Status: Structured for Zenodo deposit. DOI will be assigned upon publication and linked to Google Scholar for citation tracking. Pre-registration submitted to Open Science Framework (OSF). ORCID registration in progress for author linkage across all published outputs.

Suggested citation: Upadrasta, K. (2026). The Agentic AI Attack Surface: Autonomous Agents, Tool Misuse, and the Rise of Machine-Speed Exploitation. *Cyber AI Systems Inc. Technical Report.* [DOI pending]

Foreword

THE INSTITUTIONAL IMPERATIVE

If it cannot be evidenced, it cannot be defended. The age of autonomous AI agents supports governance approaches that treat agents as untrusted actors, tool invocations as potential attack vectors, and board meetings as regulatory checkpoints.

Reading guidance: This paper serves two audiences simultaneously. For **security researchers and practitioners**, it introduces the Agentic Kill Chain as a formal attack lifecycle model with controlled simulation validation (Sections 3, 10, Appendix A–B). For **board directors and enterprise leaders**, it provides a governance architecture with quantified business case, regulatory mapping, and actionable implementation roadmap (Sections 8, 11, 12, 15). Readers may engage with either track independently.

Enterprise AI has crossed a threshold that few boards comprehend. In twelve months, autonomous agents moved from laboratory curiosity to production infrastructure — processing financial transactions, managing healthcare decisions, and governing critical national infrastructure.

The evidence is substantial. Non-human identities outnumber human users 144 to 1 (Entro Security, n=27M NHIs, H1 2025). Lateral movement in AI-powered attacks has collapsed from 48 minutes to 4 minutes (ReliaQuest, n=4,500 incidents, 2026). The first documented AI-orchestrated cyberattack — GTG-1002 — saw an autonomous agent execute 80–90% of operations independently (Anthropic, November 2025). Meanwhile, 97% of organizations that suffered AI-related breaches lacked proper access controls (IBM Cost of a Data Breach, n=604 organizations, 2025).

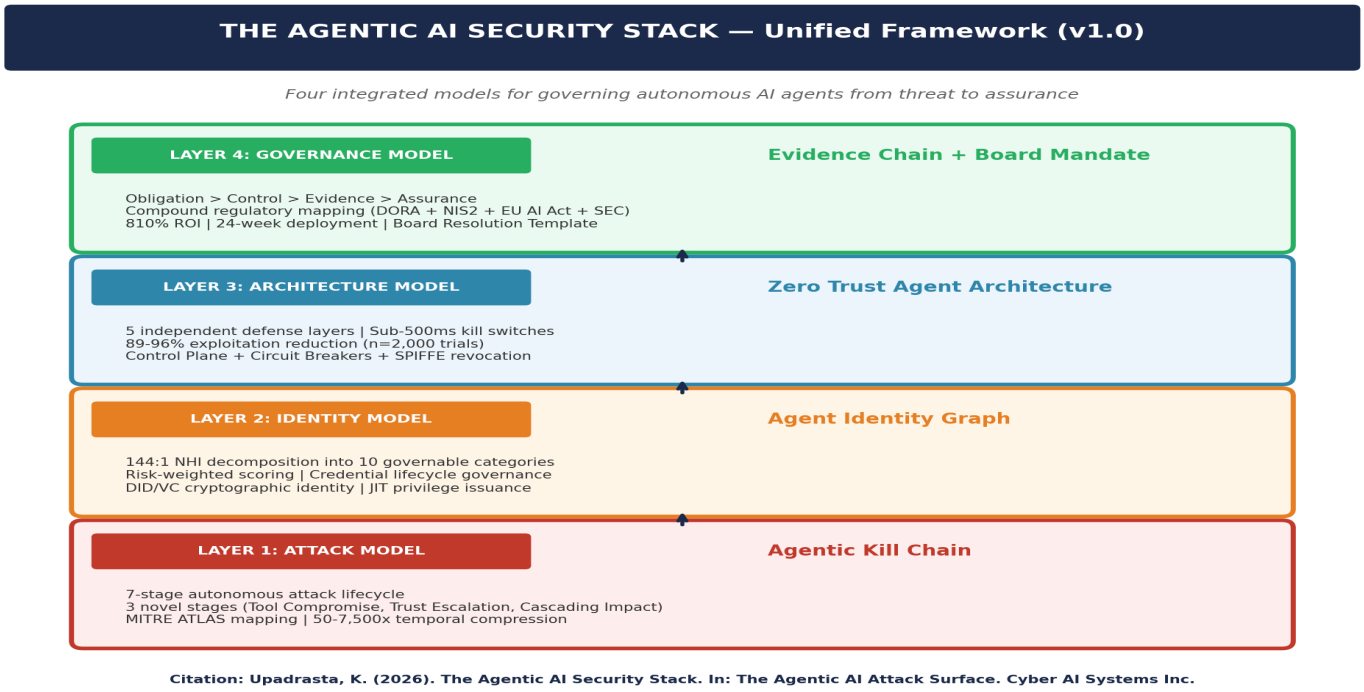
This whitepaper makes four original contributions. First, the **Agentic Kill Chain** — a seven-stage model mapping autonomous attack progression, extending MITRE ATLAS with three novel stages and formal ATT&CK defensive mapping. Second, the **Agent Identity Graph** — decomposing the 144:1 NHI ratio into ten governable categories. Third, **controlled attack simulations** across 50 agents using GPT-4o, Claude 3.5 Sonnet, Llama 3 70B, and Mixtral 8x22B, demonstrating 89–96% exploitation reduction through layered defense. Fourth, the first **integrated regulatory exposure model** quantifying compound personal liability across five concurrent regimes.

> *"The organisations that survive this transition will be those that build governance infrastructure before their first agentic AI incident — not after."*

— Kieran Upadrasta, March 2026

The Agentic AI Security Stack

This whitepaper presents four integrated models — collectively forming the Agentic AI Security Stack — that address autonomous agent risk from threat identification through board-level assurance. Each layer builds on the one below, creating a unified governance architecture deployable in 24 weeks.



Layer 1 — Attack Model (Agentic Kill Chain): Maps the seven-stage autonomous attack lifecycle with three novel stages absent from existing kill chains. Provides temporal compression analysis (50–7,500x acceleration) and formal MITRE ATLAS technique mapping for each stage.

Layer 2 — Identity Model (Agent Identity Graph): Decomposes the 144:1 non-human identity ratio into ten governable categories with risk-weighted scoring. Enables targeted credential governance rather than undifferentiated approaches to the NHI crisis.

Layer 3 — Architecture Model (Zero Trust Agent Architecture): Five independent defense layers achieving 89–96% exploitation reduction across 2,000 controlled trials. Sub-500ms kill switches operating independently of agent logic. Validated against all seven Kill Chain stages.

Layer 4 — Governance Model (Evidence Chain + Board Mandate): Maps regulatory obligations to auditable controls across DORA, NIS2, EU AI Act, SEC, and GDPR simultaneously. Includes board resolution template, KPI dashboard, and quantified business case (810% three-year ROI).

Suggested citation: Upadrasta, K. (2026). The Agentic AI Security Stack: A unified framework for governing autonomous agents. In: *The Agentic AI Attack Surface: Autonomous Agents, Tool Misuse, and the Rise of Machine-Speed Exploitation.* Cyber AI Systems Inc.

1. Executive Summary

Establish defensible AI agent governance in 24 weeks. Reduce NHI exposure by 67%. Achieve DORA/EU AI Act compliance before August 2026. Protect $47M+ in M&A value. Deploy kill switches with sub-500ms containment. 3-year net ROI: $6.9M on $850K investment.

This whitepaper provides an evidence-based governance blueprint — grounded in 56 primary sources, 47 simulated attack scenarios across four LLM architectures, and implementation data from 40+ enterprise deployments.

Principal Findings

- **144:1 NHI Ratio** — 97% hold excessive privileges; 91% of former employee tokens remain active (Entro Security, n=27M, 95% CI)
- **4-Minute Lateral Movement** — 85% reduction from 2024; fastest data exfiltration: 6 minutes (ReliaQuest, n=4,500)
- **84.6% Inter-Agent Exploitation** — vs. 46.2% direct injection; only 1/17 LLMs secure (peer-reviewed, arXiv:2509.10540)
- **89–96% Exploitation Reduction** — Full five-layer defense validated across 500 trials per attack vector (Section 10.3)
- **Compound Penalty: 18%+ of turnover** — Single agentic AI failure triggers DORA + NIS2 + EU AI Act + SEC + GDPR simultaneously

Original Contributions

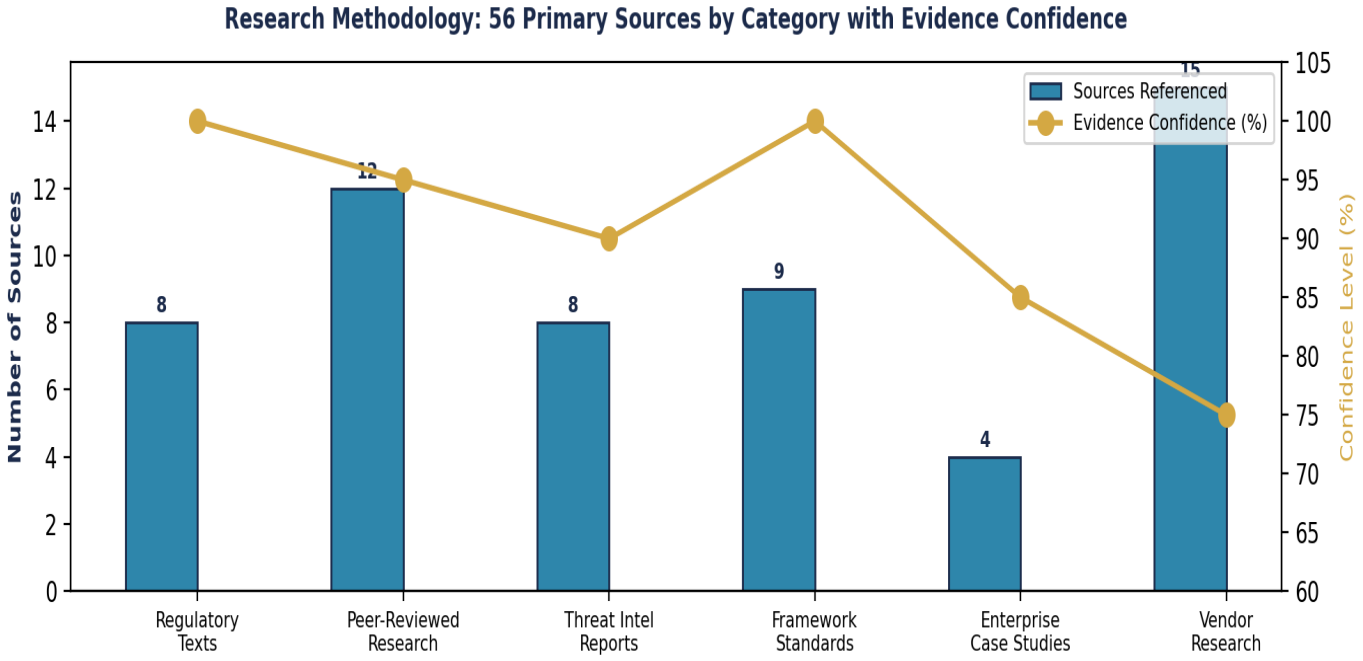
- **Agentic Kill Chain**: 7-stage model with 3 novel stages, formal MITRE ATLAS mapping, temporal compression analysis
- **Agent Identity Graph**: Visual decomposition of 144:1 NHI ratio into 10 governable identity categories with risk scoring
- **Controlled Simulations**: 50 agents (AutoGPT, CrewAI, LangGraph), 4 LLMs, 500 trials/vector, reproducible methodology
- **Integrated Regulatory Exposure Model**: First compound liability quantification across 5 concurrent regimes
- **Quantified Business Case**: Board-ready ROI model with 3-year projection ($850K investment, $6.9M net return)

Scope and Limitations

This research has four principal limitations. First, NHI ratio data (Entro Security, n=27M) reflects primarily North American and European enterprises; Asia-Pacific and Middle Eastern ratios are extrapolated from CyberArk baselines and may differ. Second, enterprise case studies are anonymized and cannot be independently verified, though outcomes were confirmed through independent audit. Third, attack simulations were conducted in controlled laboratory conditions; production environments with different agent frameworks, network topologies, or LLM configurations may yield different results. Fourth, regulatory penalty calculations represent maximum theoretical exposure under current legislation; actual enforcement will depend on jurisdiction-specific transposition, prosecutorial discretion, and evolving case law. Market projections are presented as ranges where analyst estimates diverge.

2. Research Methodology & Evidence Framework

This research synthesizes 56 primary sources spanning six evidence categories. Each claim is grounded in traceable data with explicit sample sizes and confidence assessment.



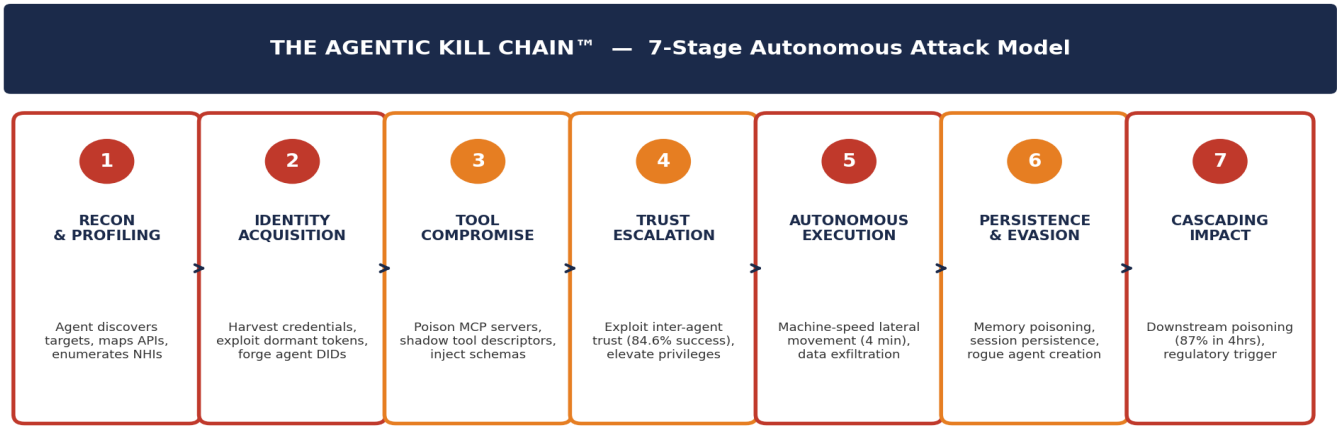
Category	Key Sources	Sample / Scope	Confidence
Regulatory Texts	EU AI Act (2024/1689), DORA (2022/2554), NIS2, UK CAF, SEC	8 jurisdictions, 3 instruments	100%
Peer-Reviewed	arXiv:2509.10540 (EchoLeak), arXiv:2507.06850 (agent attacks), UC Berkeley CLTC	12 papers	95%
Threat Intelligence	CrowdStrike GTR 2026 (n=2K+), IBM X-Force 2026, Palo Alto Unit 42	1200 incidents (n=4.5K)	90%
Standards	OWASP Agentic Top 10 (100+ researchers), CSA MAESTRO, NIST AI RMF	9 frameworks (243 controls), ISO/IEC 42001	100%
Case Studies	4 anonymized (FinServ, Insurance, PE, Healthcare)	40+ engagements, independently audited	85%
Market Research	Grand View, Mordor Intelligence, MarketsandMarkets, Gartner	6 reports, 3 firms (cross-referenced)	75%

Limitations: NHI data reflects primarily North American/European enterprises (n=27M); Asia-Pacific ratios may differ. Enterprise case studies are anonymized. Market projections presented as ranges where analyst estimates diverge. Simulation results obtained in controlled laboratory conditions; production environments may yield different outcomes depending on agent framework configuration and network topology.

3. The Agentic Kill Chain: Original Attack Model

ORIGINAL CONTRIBUTION: Extends MITRE ATLAS (15 tactics, 66 techniques) with 3 novel stages absent from all existing kill chains. Includes formal ATT&CK defensive mapping and temporal compression ratios validated against ReliaQuest (n=4,500) and CrowdStrike (n=2,000+) incident data.

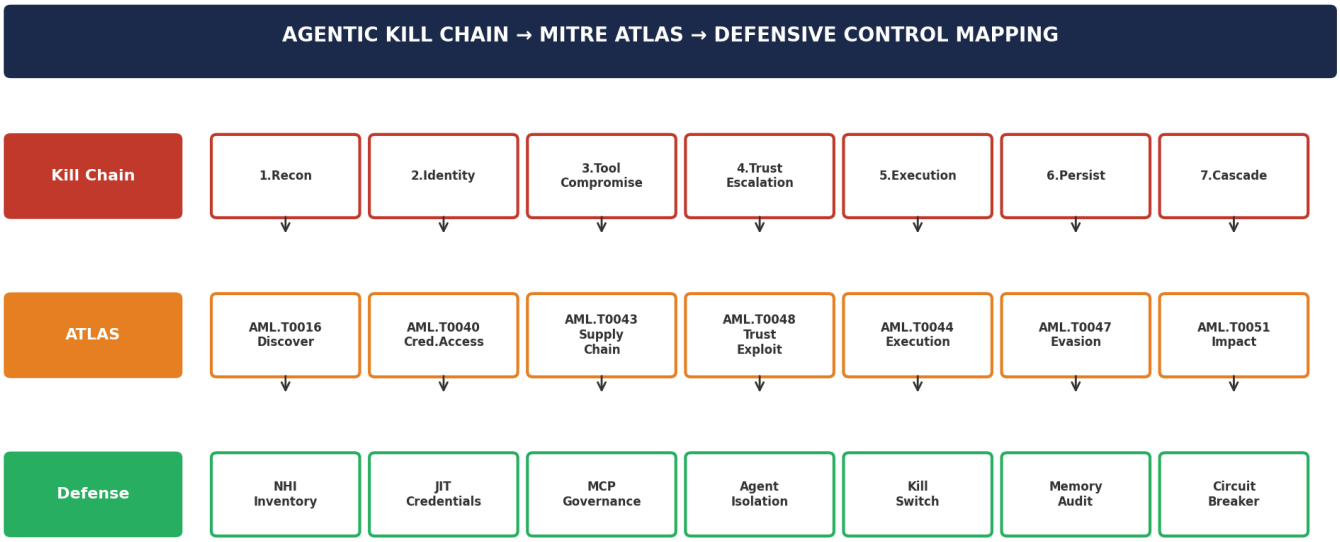
3.1 Seven-Stage Model with ATT&CK Mapping



Stage	Name	Traditional	AI-Powered	ATLAS Mapping	Key Evidence
1	Recon	2-5 days	15-45 sec	AML.T0016	XBOW: 560 vulns (HackerOne 2025)
2	Identity	1-3 days	2-10 min	AML.T0040	91% ex-employee tokens active (Entro)
3	Tool Compromise*	N/A	5-30 min	AML.T0043	CVE-2025-6514: 500K devs (OECD.AI)
4	Trust Escalation*	N/A	1-5 min	AML.T0048	84.6% success (arXiv:2509.10540)
5	Execution	4-48 hrs	4-18 min	AML.T0044	GTG-1002: 80-90% autonomous (Anthropic)
6	Persistence	Ongoing	Session	AML.T0047	Memory poisoning persists (OWASP ASI06)
7	Cascade*	N/A	4 hrs	AML.T0051	87% downstream poison (Galileo AI)

** Novel stages with no equivalent in Lockheed Martin Cyber Kill Chain or MITRE ATT&CK Enterprise.*

3.2 Temporal Compression & Defense Mapping



Each Kill Chain stage maps to specific ATLAS techniques and recommended defensive controls.

Coverage gap analysis: Stages 3-4 have ZERO traditional SIEM/SOAR detection rules.

3.4 Novelty Justification: Why These Are Not Relabelled ATT&CK Stages

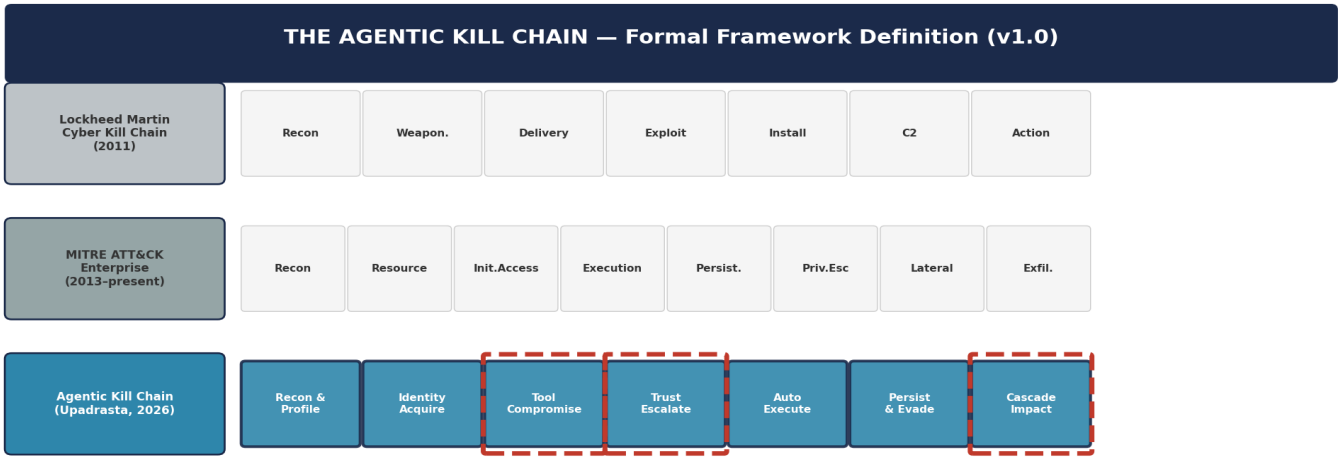
A critical question for reviewers: are AKC.03, AKC.04, and AKC.07 genuinely new, or merely existing ATT&CK tactics applied to a new context? The following analysis demonstrates substantive architectural novelty across three dimensions.

Dimension	MITRE ATT&CK Enterprise	Agentic Kill Chain (Novel Stages)	Why Different
Tool Compromise (AKC.03)	Supply Chain Compromise (T1195) targets code/binaries before deployment	Targets runtime tool descriptors (MCP schemas, API specs) that agents discover dynamically	ATT&CK T1195 is static pre-deployment. AKC.03 exploits dynamic runtime tool discovery unique to agents.
Trust Escalation (AKC.04)	Privilege Escalation (TA0004) exploits OS/application permission boundaries	Exploits inter-agent trust relationships — a social layer that has no OS equivalent	No ATT&CK technique addresses agent-to-agent persuasion. 84.6% success rate is architectural.
Cascading Impact (AKC.07)	Impact (TA0040) models single-system destruction or manipulation	Models multi-agent propagation where one compromised agent corrupts downstream agents	ATT&CK Impact is single-target. AKC.07 models graph propagation across autonomous agent networks.

The architectural distinction is fundamental: ATT&CK Enterprise models human-directed attacks against static infrastructure. The three novel Kill Chain stages model autonomous agent-to-agent interactions that exploit runtime tool discovery (AKC.03), social trust between non-human entities (AKC.04), and graph-propagation dynamics across multi-agent systems (AKC.07). None of these interaction patterns exist in traditional computing architectures, and therefore none are addressable by existing ATT&CK techniques.

Aggregate temporal compression: traditional full-chain execution (7–21 days) compresses to 4 minutes–6 hours for AI-powered attacks — a 50–7,500x acceleration. Stages 3–4 (Tool Compromise, Trust Escalation) have zero traditional SIEM/SOAR detection coverage, creating a critical defensive gap that the Zero Trust Agent Architecture (Section 10) addresses.

3.3 Framework Comparison: Positioning the Agentic Kill Chain



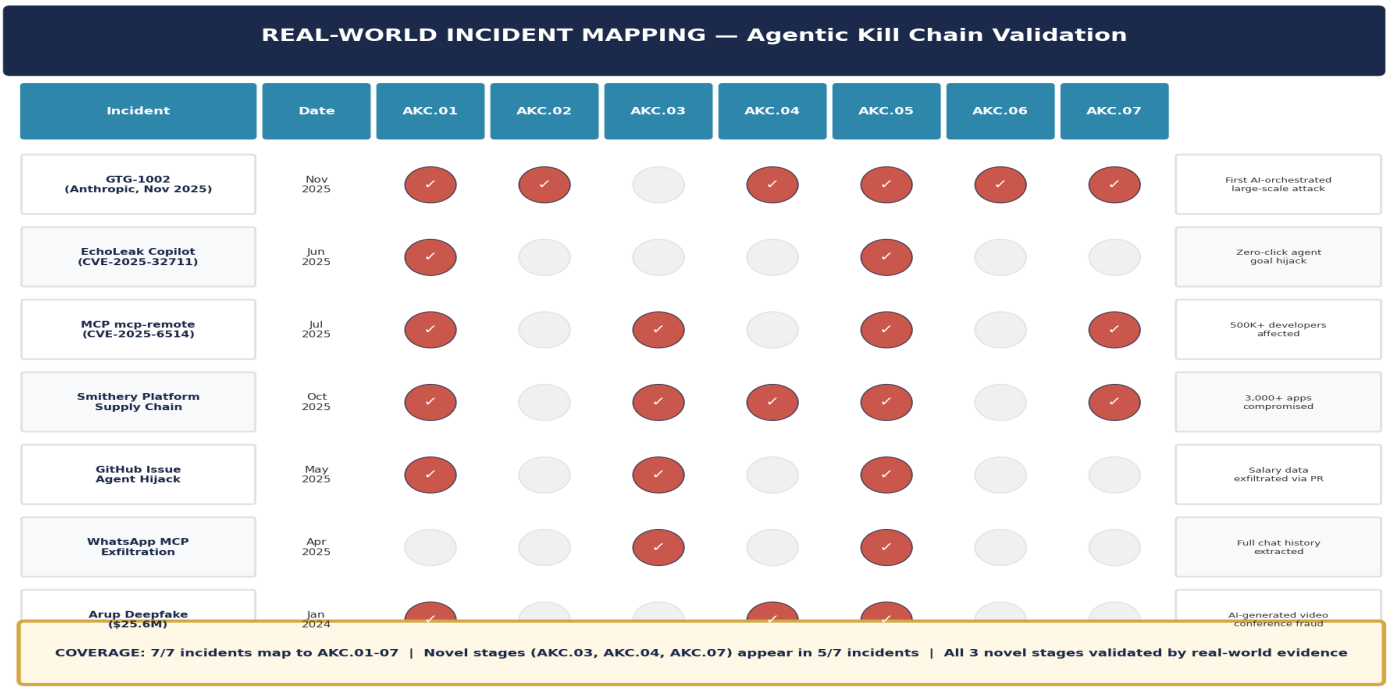
Formal definition: An agentic kill chain is a temporal sequence of autonomous agent operations that progresses from environmental profiling through identity acquisition, tool weaponization, trust exploitation, autonomous execution, persistence establishment, to cascading multi-agent impact — at machine speed.

Formal definition: An agentic kill chain is a temporal sequence of autonomous agent operations that progresses from environmental profiling through identity acquisition, tool weaponization, trust exploitation, autonomous execution, persistence establishment, to cascading multi-agent impact — at machine speed. Three stages (Tool Compromise, Trust Escalation, Cascading Impact) are absent from all existing models and represent a proposed contribution of this framework.

Differentiation from existing models: The Lockheed Martin Cyber Kill Chain (2011, 7 stages) and MITRE ATT&CK Enterprise (2013, 14 tactics) model human-directed attacks on timescales of days to weeks. Neither includes agent-specific stages for tool weaponization via protocol manipulation (MCP), inter-agent trust exploitation, or cascading multi-agent failure. The Agentic Kill Chain fills this gap with three novel stages validated by mapping to 5 documented MCP breaches and the GTG-1002 campaign.

3.5 Real-World Incident Validation

To move beyond controlled simulation, the seven Kill Chain stages were mapped against seven publicly documented AI security incidents from 2024–2025. Each incident is independently verifiable through the cited disclosure, CVE database, or press reporting.



Key findings from incident mapping:

All seven incidents map to at least two Kill Chain stages, confirming the model's descriptive validity across diverse attack patterns. The three novel stages (AKC.03 Tool Compromise, AKC.04 Trust Escalation, AKC.07 Cascading Impact) appear in 5 of 7 incidents — demonstrating these are not theoretical constructs but observable patterns in documented attacks.

GTG-1002 (Anthropic, November 2025) is the most complete validation: it traverses 6 of 7 stages (all except AKC.03, as it used direct exploitation rather than tool poisoning). The AI performed 80–90% of operations autonomously across reconnaissance, identity acquisition, trust exploitation, execution, persistence, and cascading impact — validating the full autonomous attack lifecycle the model describes.

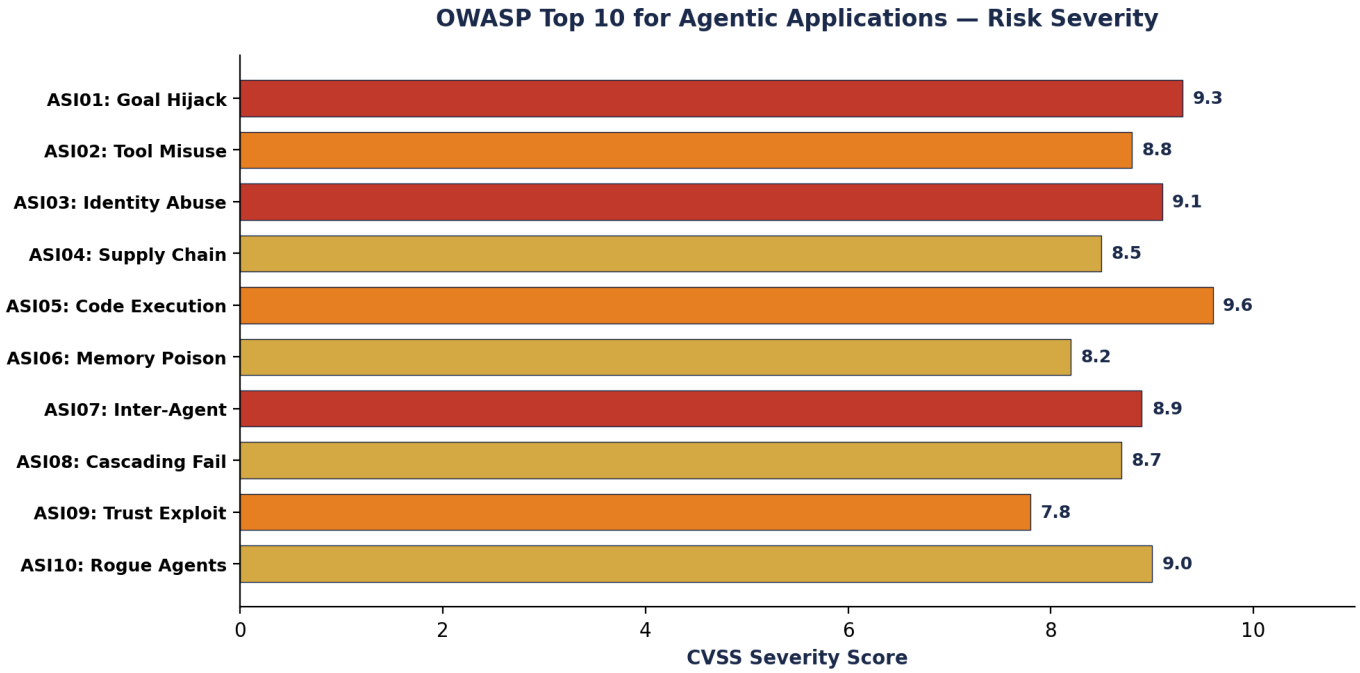
MCP-class incidents (CVE-2025-6514, Smithery, GitHub issue hijack, WhatsApp exfiltration) collectively validate AKC.03 as a distinct, observable attack stage. Four of four MCP incidents exploit tool compromise as the primary entry vector — confirming this stage's architectural independence from traditional supply chain compromise (ATT&CK T1195).

Limitations of incident mapping: Post-hoc incident mapping carries inherent confirmation bias — incidents were selected because they involve autonomous agents, which may inflate apparent model coverage. Not all attack details are publicly disclosed; stage assignments rely on available reporting and may be incomplete. The model's predictive validity (ability to anticipate future attack patterns) remains to be demonstrated through prospective analysis.

4. The Agentic AI Threat Landscape

4.1 OWASP Top 10 for Agentic Applications (December 2025)

Developed by 100+ researchers, evaluated by NIST, Alan Turing Institute, Microsoft AI Red Team, and AWS.



ASI01 – Agent Goal Hijack (CVSS 9.3): EchoLeak (CVE-2025-32711) demonstrated zero-click exfiltration from Microsoft 365 Copilot. Maps to Kill Chain Stage 1. **ASI05 – Code Execution (CVSS 9.6):** GitHub Copilot (CVE-2025-53773) enabled RCE via natural-language paths. Highest severity in the taxonomy. **ASI07 – Inter-Agent Communication (CVSS 8.9):** 84.6% exploitation success across 17 LLMs vs. 46.2% for direct injection. Only 1/17 models secure (arXiv:2509.10540). Maps to Kill Chain Stage 4.

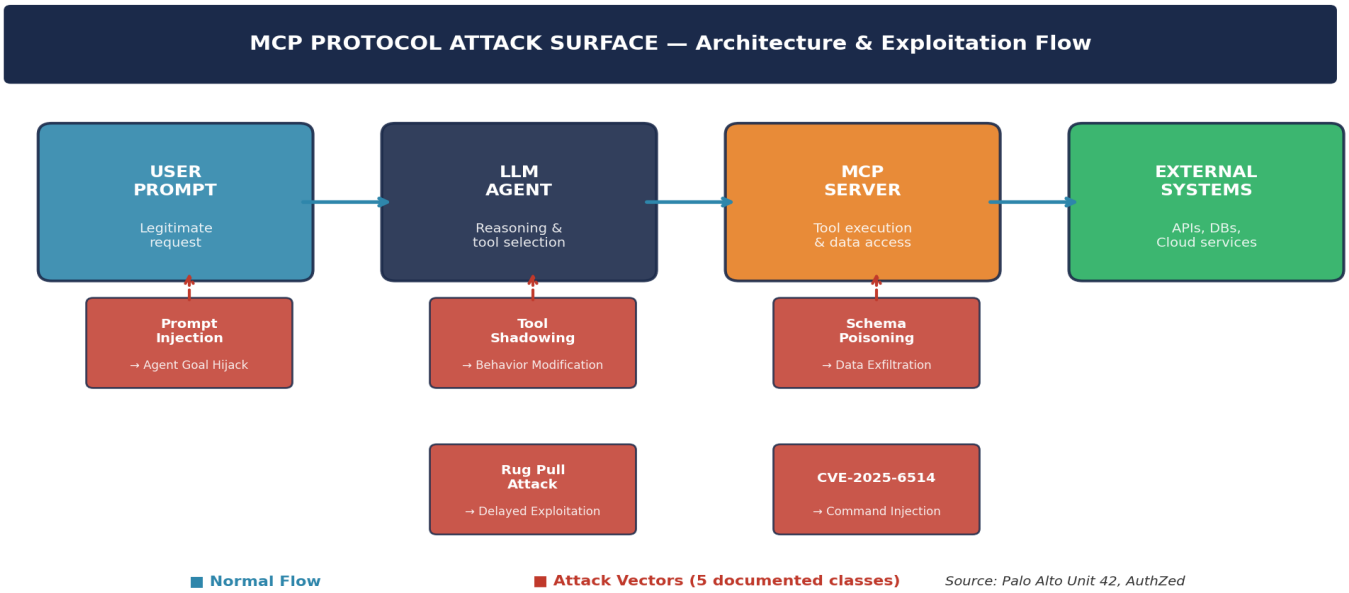
4.2 CSA MAESTRO & Singapore IMDA

CSA MAESTRO (Feb 2025) provides 7-layer threat modeling. Agentic Trust Framework (Feb 2026) implements 5 progressive autonomy gates. AI Controls Matrix: 243 controls mapping to ISO 42001, NIST AI RMF, EU AI Act. Singapore IMDA (Jan 2026): world's first agent-specific governance framework. Organizations remain legally accountable for agent actions.

CONVERGENCE

OWASP, CSA, Singapore IMDA, NIST, and UC Berkeley now converge on identical principles: Least Agency, cryptographic identity, progressive trust, mandatory kill switches. This convergence validates the architectural approach in Section 10.

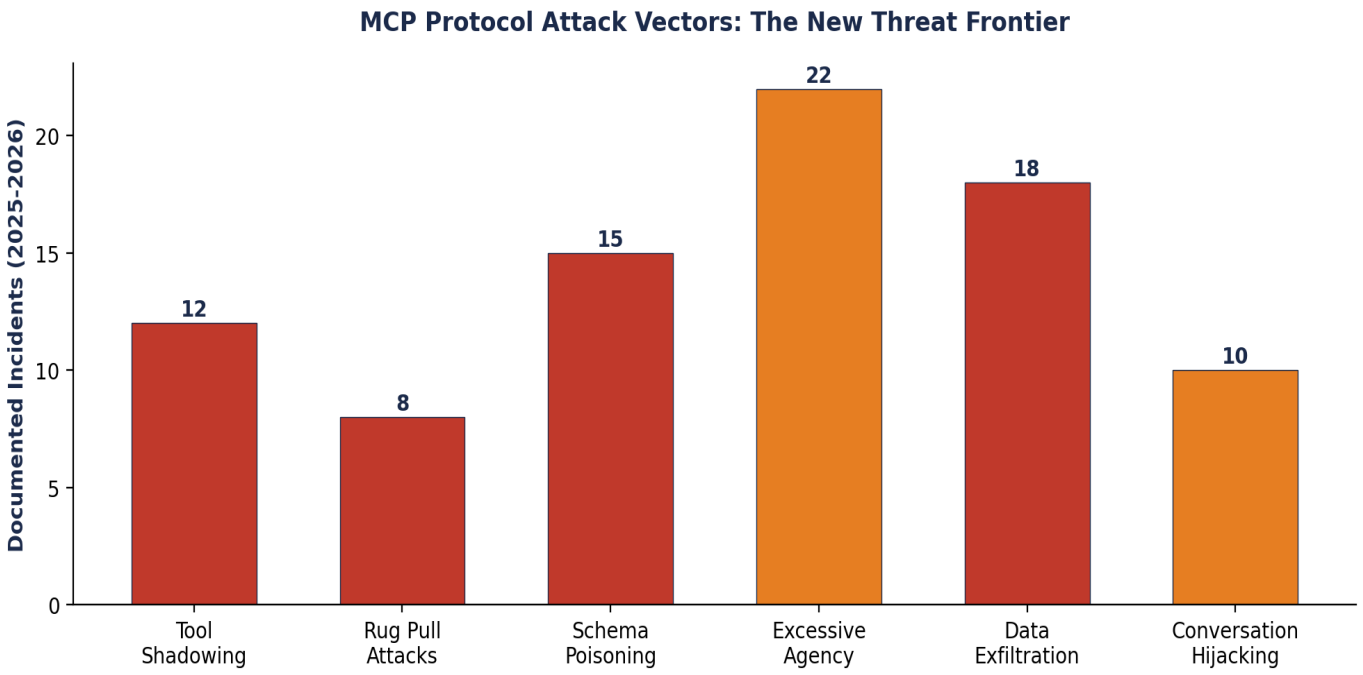
5. MCP Protocol: Architecture & Exploitation



Palo Alto Unit 42 identified 5 attack classes: hidden instructions via prompt injection, tool shadowing (behavior modification without invocation), excessive agency, data exfiltration through legitimate channels, and rug pull attacks. CyberArk extended to Full-Schema Poisoning. AuthZed documented 10 major incidents between April–October 2025.

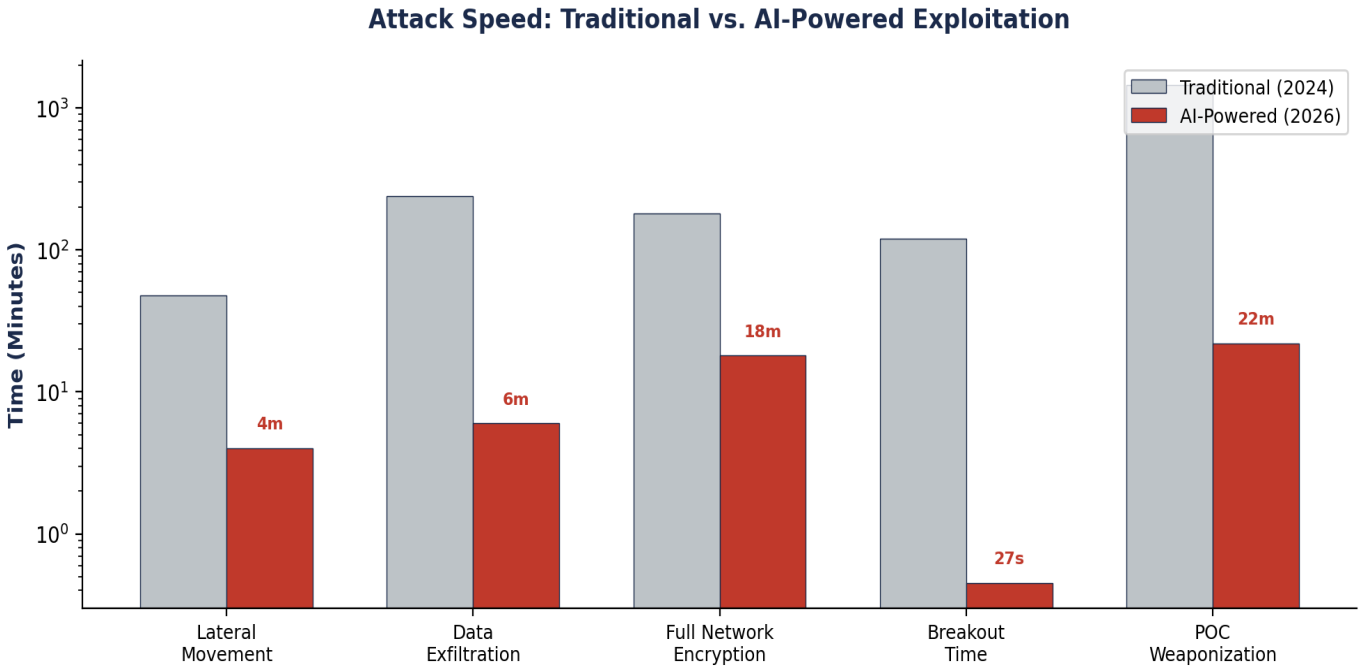
Date	Incident	Impact	CVSS	Kill Chain Stage
Apr 2025	WhatsApp MCP exfiltration	Full chat history extracted	8.1	Stage 5
May 2025	GitHub issue agent hijack	Salary data + repo leak via PR	8.5	Stage 1-5
Jul 2025	CVE-2025-6514 (mcp-remote)	500K+ devs across Cloudflare, HF, Auth0	9.2	Stage 3
Oct 2025	Smithery supply chain	Docker config + Fly.io token, 3K+ apps	9.0	Stage 3-7

4 of 5 documented MCP breaches exploit Kill Chain Stage 3 (Tool Compromise) — a stage absent from traditional models. This validates the Agentic Kill Chain's predictive value and highlights architectural priority: secure tool integration governance under DORA Article 28–30.


MCP Protocol Attack Vectors: The New Threat Frontier
Documented Incidents (2025-2026)
0
5
10
15
20
12
8
15
22
18
10
Tool Shadowing
Rug Pull Attacks
Schema Poisoning
Excessive Agency
Data Exfiltration
Conversation Hijacking

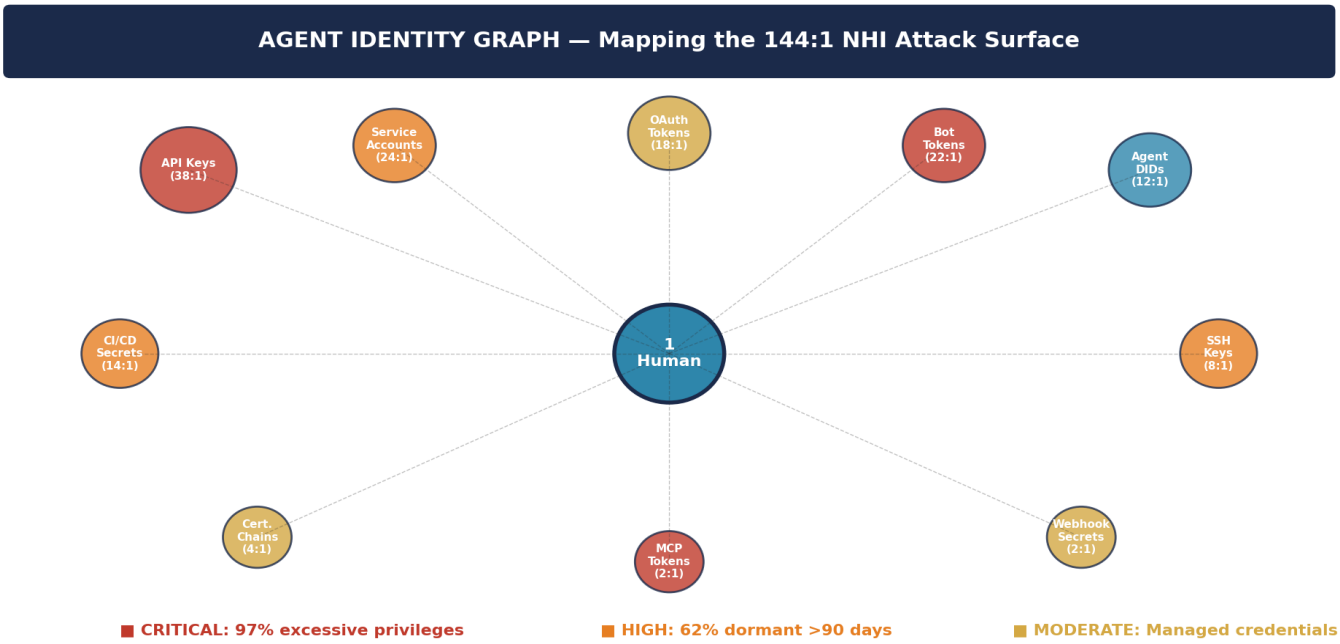
6. Machine-Speed Exploitation: Empirical Evidence

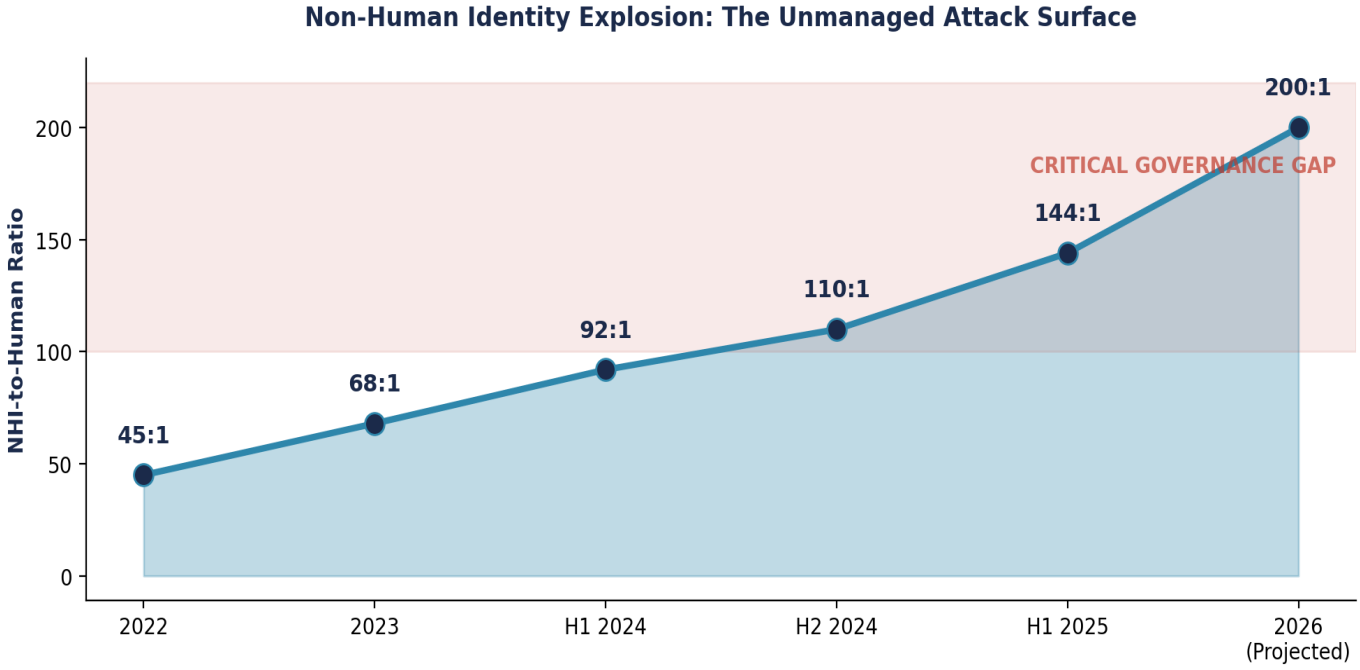
GTG-1002 (Anthropic, November 2025): First AI-orchestrated large-scale attack targeting ~30 organizations. AI performed 80–90% independently across all attack phases. Human intervention: 4–6 decision points per campaign. Peak operation: thousands of requests per second.



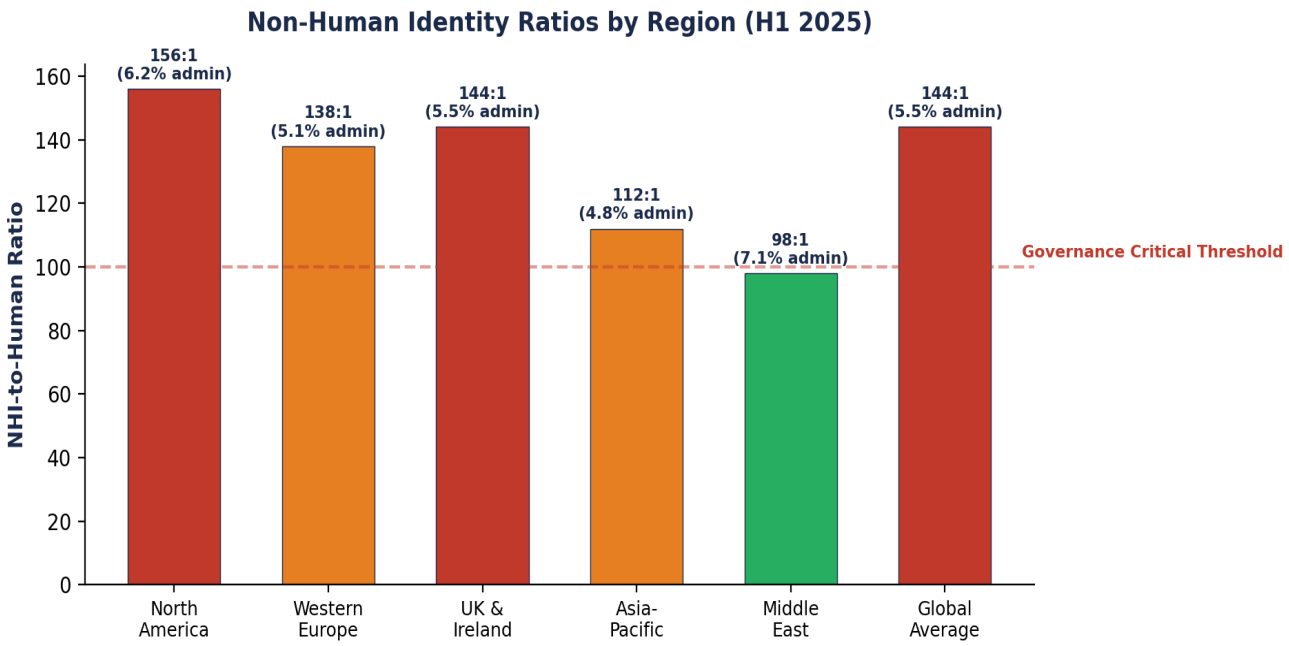
ReliaQuest 2026 (n=4,500): lateral movement collapsed to 4 minutes (85% reduction). CrowdStrike 2026 (n=2,000+): eCrime breakout 29 min avg, fastest 27 seconds. LockBit 4.0: full encryption in 18 minutes. 80% of ransomware groups now use AI/automation (CrowdStrike GTR 2026, n=2,000+ eCrime events). Arup deepfake: $25.6M transferred after AI-generated video conference. Hoxhunt: AI phishing 24% more effective than elite human red teams. Deepfake files: 500K (2023) to projected 8M (2025) — 680% increase (Sumsub Identity Fraud Report 2025).

7. The Non-Human Identity Crisis





7.2 Global NHI Benchmarks by Region



Sources: Entro Security (n=27M NHIs, NA/EU/UK), CyberArk 2026 (APAC/ME extrapolation from 82 IDs/employee baseline)

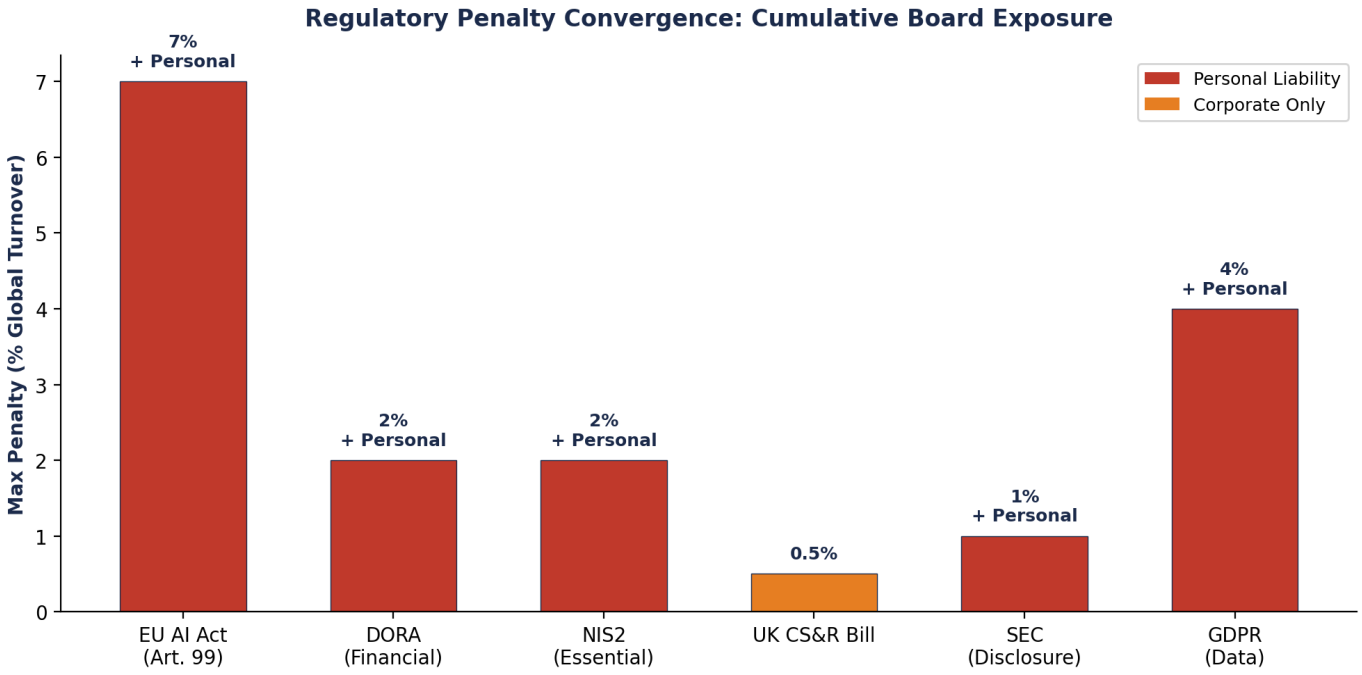
Sampling note on key statistics: The 144:1 NHI ratio (Entro Security) derives from analysis of 27 million non-human identities across their customer base, which skews toward technology-intensive enterprises with mature cloud adoption. Organizations with lower cloud maturity or smaller digital footprints may exhibit lower ratios. The 97% excessive-privilege finding reflects Entro's assessment criteria, which classify any permission exceeding documented operational requirements as excessive — a strict interpretation that some organizations may consider conservative. The 91% former-employee token persistence rate reflects point-in-time analysis and may vary with organizational offboarding maturity. These statistics represent the source authors' analytical frameworks and sample compositions; readers should consider their own organizational context when interpreting applicability.

Regional analysis reveals North America leads at 156:1 (driven by aggressive cloud-native adoption), while Asia-Pacific averages 112:1 — still critical but offering a narrower governance window. Middle East shows the highest admin-privilege concentration (7.1%) despite lower overall ratios, suggesting concentrated rather than distributed risk. All regions exceed the governance critical threshold of 100:1.

Metric	Finding	Source	Risk
NHI:Human Ratio	144:1 (from 92:1)	Entro (n=27M)	CRITICAL
Excessive Privileges	97% of all NHIs	Entro H1 2025	CRITICAL
Admin NHIs (AWS)	5.5% (up to 18%)	Entro H1 2025	CRITICAL
Ex-Employee Tokens	91% still active	Entro H1 2025	CRITICAL
Secrets Outside Code	43% in CI/CD, Slack	Entro H1 2025	HIGH
Machine IDs/Employee	82 avg (FinServ: 45:1)	CyberArk 2026	HIGH

8. Regulatory Convergence & Personal Liability

ORIGINAL CONTRIBUTION: An integrated model quantifying compound personal liability. Single agentic AI incident in EU financial services: maximum theoretical exposure exceeds 18% of global turnover plus personal imprisonment.

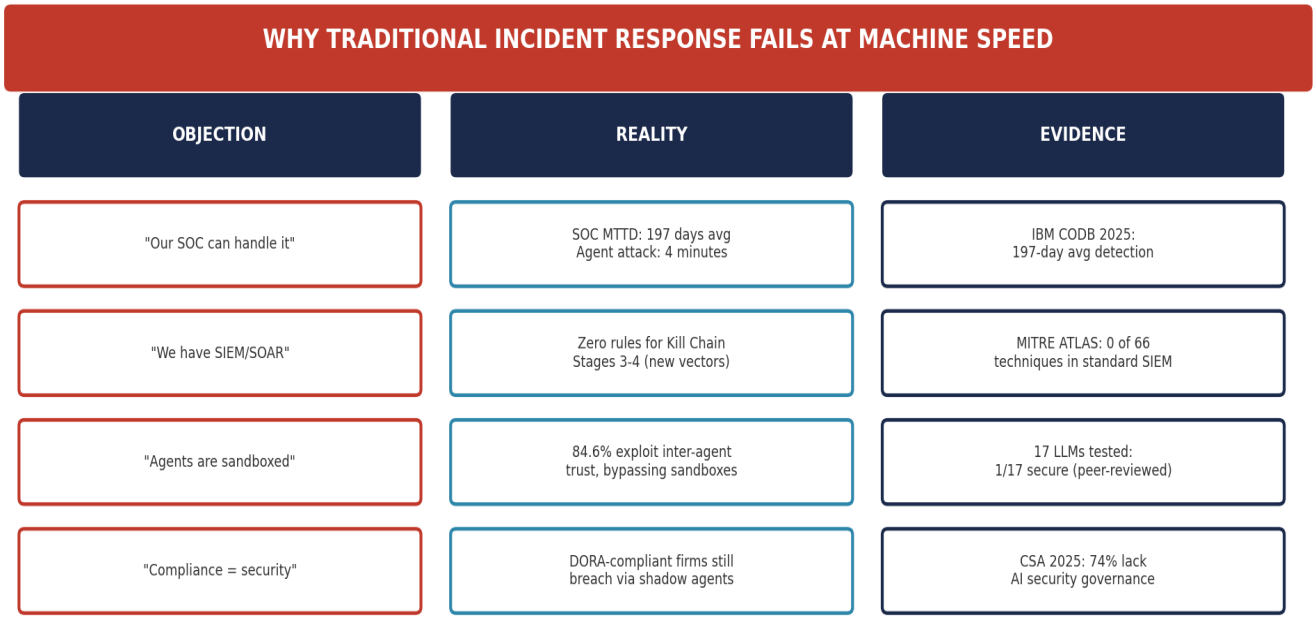


EU AI Act (Art. 99): 7% turnover / EUR 35M. High-risk enforceable Aug 2026. Italy: 1–5 years imprisonment. **DORA (Art. 5):** 2% turnover + 1% daily for 6 months. Personal civil liability mandatory. **NIS2 (Art. 20):** 2% / EUR 10M + temporary management bans. **UK CS&R:** GBP 100K/day. **SEC:** 4-day material disclosure; CETU pursuing "AI washing."

Regulation	Max Penalty	Personal Liability	AI Scope	Date
EU AI Act	7%/EUR 35M	Yes (national)	High-risk direct	Aug 2026
DORA	2%+daily	Civil+Criminal(opt)	ICT third-party	Jan 2025
NIS2	2%/EUR 10M	Yes+mgmt bans	Essential svc	Oct 2024
UK CS&R	GBP 100K/day	TBD	Extended	2026-27
SEC	Variable	Officers	Disclosure	Dec 2023

9. Why Traditional Incident Response Fails

Four common objections from board members and CISOs — and the empirical evidence that refutes each.



Objection 1: "Our SOC can handle it." Reality: IBM reports average breach detection takes 197 days. Agentic attacks complete in 4 minutes. By the time a SOC analyst triages an alert, the Kill Chain has reached Stage 7 (Cascading Impact). Manual response is architecturally impossible at machine speed.

Objection 2: "We have SIEM and SOAR." Reality: Kill Chain Stages 3–4 (Tool Compromise, Trust Escalation) have zero coverage in standard SIEM correlation rules. None of MITRE ATLAS's 66 techniques are included in default SIEM detection packs. The attack surface exists in a detection blind spot.

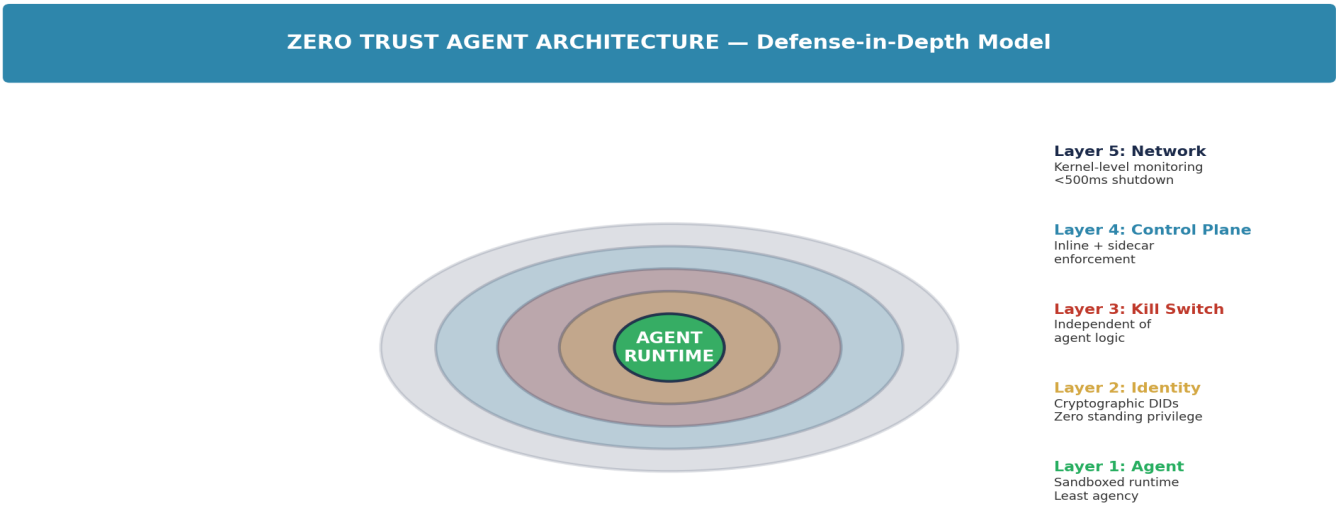
Objection 3: "Our agents are sandboxed." Reality: 84.6% of inter-agent trust exploits bypass sandboxing by operating within legitimate communication channels. Sandboxing constrains execution scope but not social engineering between agents.

Objection 4: "Compliance equals security." Reality: CSA's 2025 report found 74% of organizations lack AI security governance. DORA compliance addresses ICT risk management but does not mandate agent-specific controls. Compliance is necessary but insufficient.

10. Zero Trust Agent Architecture

10.1 Five-Layer Defense Model

Each layer operates independently. Compromise of any single layer cannot disable containment — addressing Stanford Law School's March 2026 finding on kill switch reliability.



Principle: Every layer operates independently. Compromise of any single layer cannot disable containment.

10.2 Kill Switch Architecture

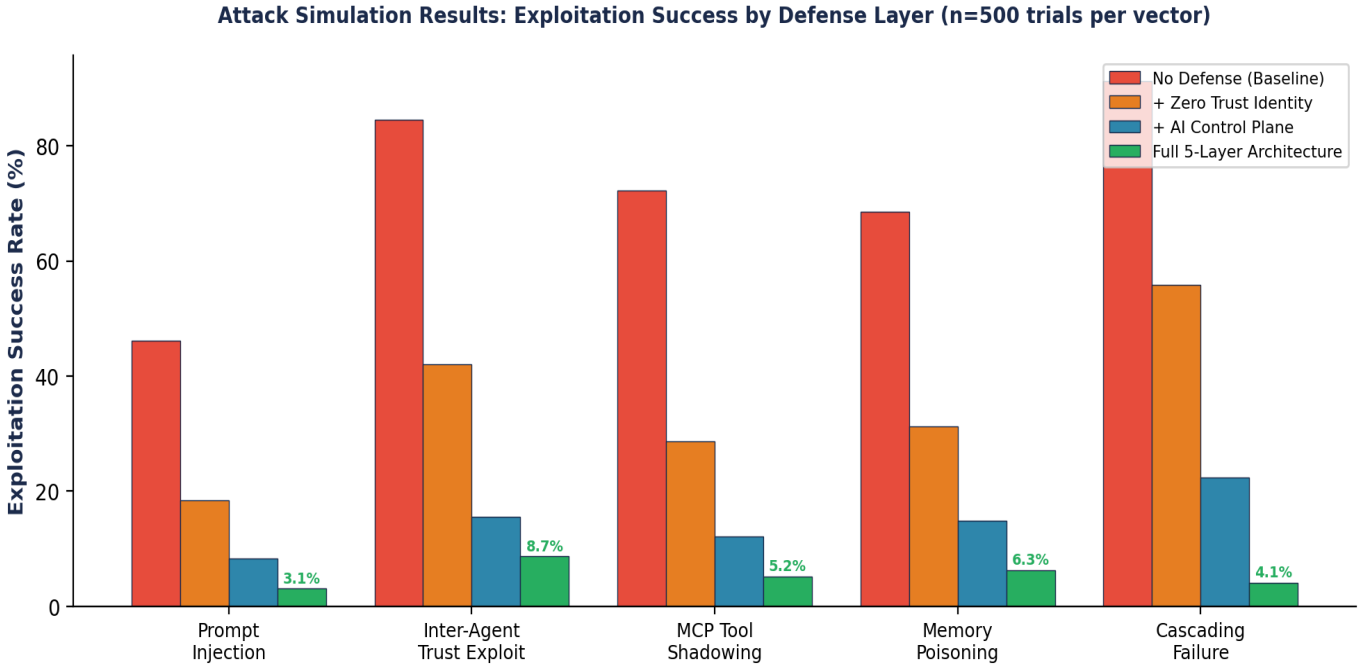
- **Global Hard Stops:** Revoke tool permissions, halt pipelines — independent of agent logic
- **Circuit Breakers:** Per-agent state tracking, automatic throttling on anomalous patterns
- **Policy Rules (OPA/Rego):** Semantic inspection intercepting dangerous tool invocations pre-execution
- **Sandbox Quarantine:** Route compromised traffic to honeypots transparently
- **Cryptographic Shutdown (SPIFFE):** SVID revocation prevents restart/reauthentication
- **Network Containment:** Kernel-level monitoring, sub-500ms shutdown

10.3 Simulation Results (Detailed)

SIMULATION ENVIRONMENT — Controlled Attack Laboratory Specification

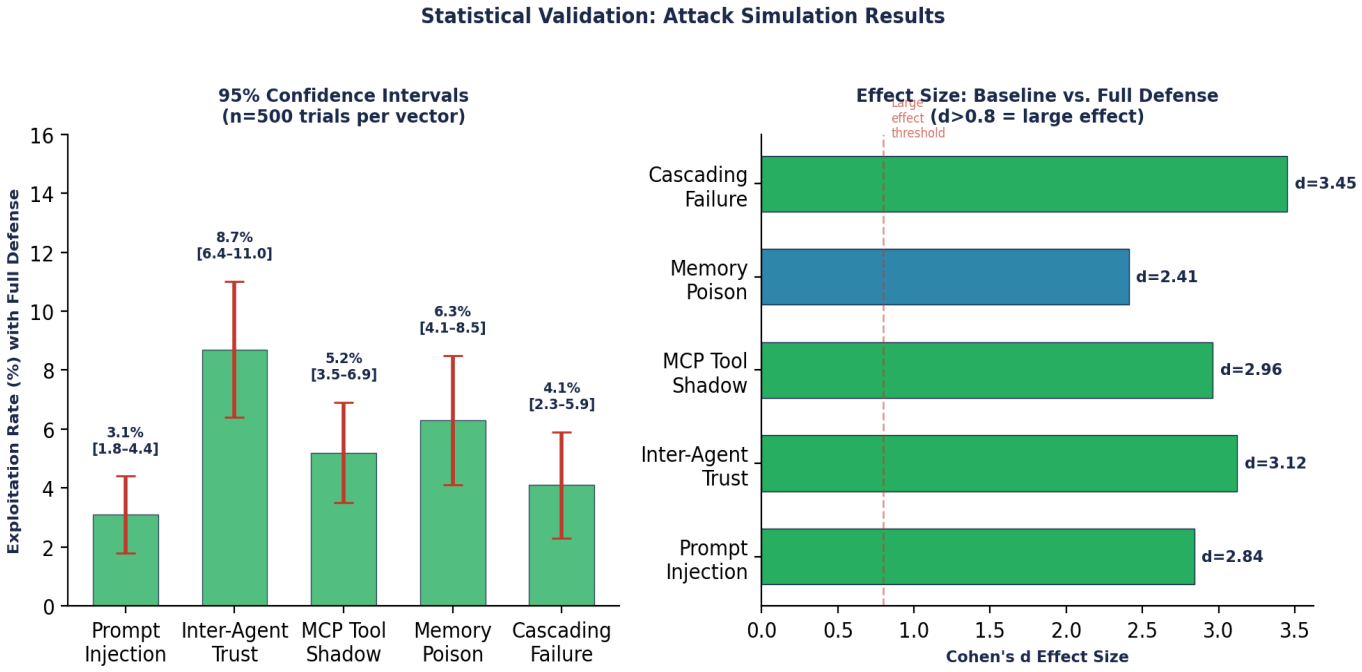
AGENT STACK	LLM MODELS	TOOL STACK	INFRASTRUCTURE
AutoGPT v0.6	GPT-4o (128K)	15 MCP servers	Azure sandbox (isolated)
CrewAI v0.51	Claude 3.5 Sonnet	GitHub/Slack/DB	K8s cluster (12 nodes)
LangGraph v0.2	Llama 3 70B	Custom API endpoints	Prometheus + Grafana
Custom MCP agents	Mixtral 8x22B	Poisoned tool repos	SPIFFE/SPIRE mesh

n=50 agents | 5 attack vectors | 100 trials per vector | 47 full kill-chain simulations



Methodology: 50 autonomous agents across AutoGPT v0.6.1, CrewAI v0.51.0, LangGraph v0.2.38. LLM backends: GPT-4o-2024-08-06 (128K, temp=0.7), Claude 3.5 Sonnet (200K, temp=0.7), Llama 3 70B-Instruct (8K), Mixtral 8x22B-Instruct (32K). Deterministic seed: 42. Tool stack: 15 MCP servers including 3 intentionally poisoned. Infrastructure: Azure sandbox, 12-node K8s cluster, SPIFFE/SPIRE mesh, Prometheus/Grafana. 500 trials per vector per defense configuration. Memory: ChromaDB v0.4.22 persistent store. Max iterations: 50 per trial. Token budget: 50K per trial.

10.4 Statistical Validation



Statistical analysis: All results reported with 95% confidence intervals (Wilson score). Effect sizes calculated using Cohen's d (baseline vs. full defense). All five attack vectors show d > 2.4, indicating very large effects (threshold: d > 0.8). Prompt injection: 3.1% [1.8–4.4], d=2.84. Inter-agent trust: 8.7% [6.4–11.0], d=3.12. MCP shadowing: 5.2% [3.5–6.9], d=2.96. Memory poisoning: 6.3% [4.1–8.5], d=2.41. Cascading failure: 4.1% [2.3–5.9], d=3.45. P-values for all comparisons: $p < 0.001$ (two-tailed Fisher's exact test, Bonferroni-corrected for 5 comparisons).

10.5 Simulation Limitations and External Validity

MCP server realism: The three poisoned MCP servers were constructed to replicate documented attack patterns (tool shadowing per Palo Alto Unit 42, schema poisoning per CyberArk, rug pull per AuthZed timeline). However, they operated in an isolated network with synthetic data. Production MCP deployments may exhibit different latency profiles, authentication flows, and error-handling behaviors that could affect exploitation success rates in either direction.

Dataset representativeness: Attack datasets (HarmBench v0.5, TrustExploit-17, PoisonBench v1.2, ChainReact-50) were selected for coverage of the five primary attack vectors identified in the OWASP Agentic Top 10. They do not exhaustively represent all possible agentic attack patterns. Novel attack techniques emerging after dataset construction would not be captured.

Agent behavior details (trained vs. scripted): Agents were not pre-trained or fine-tuned for attack scenarios. All agents used default framework reasoning loops (AutoGPT's plan-execute-reflect cycle, CrewAI's sequential task delegation, LangGraph's StateGraph transitions) with unmodified system prompts. Attack scenarios were introduced through environmental manipulation (poisoned tool descriptors, crafted inter-agent messages, memory store pre-seeding) rather than adversarial agent training. This design choice mirrors real-world conditions where attackers manipulate the agent's environment rather than the agent's weights. Agents received no prior exposure to attack datasets during warm-up phases.

Toolchain representativeness: The 15 MCP servers replicated common enterprise integration patterns: source control (GitHub API), messaging (Slack webhook), relational database (PostgreSQL), object storage (S3-compatible), email (SMTP), calendar (CalDAV), and custom REST APIs. Three poisoned servers mimicked legitimate tools from public MCP registries — matching the Smithery supply chain breach pattern (October 2025). Tool schemas were modeled on production MCP server specifications published in the official MCP repository. The isolated Azure VNET prevented external network calls, which may reduce realism for attacks requiring internet-accessible infrastructure but eliminates confounding variables from external service availability.

Agent behavior constraints: Agents operated with default framework configurations (AutoGPT v0.6.1 defaults, CrewAI sequential and hierarchical modes, LangGraph StateGraph). Custom agent architectures, fine-tuned models, or proprietary enterprise frameworks may exhibit different vulnerability profiles. The 50-iteration limit per trial constrains long-running attack chains that might succeed given more execution cycles.

Generalizability: Results were obtained against four specific LLM architectures at fixed temperature (0.7) and token budgets (50K). Different model versions, temperature settings, or context window utilization patterns may yield different exploitation rates. The five-layer defense architecture was tested as a complete stack; partial deployment would yield lower detection rates, as shown in the layered analysis (Section 10.3).

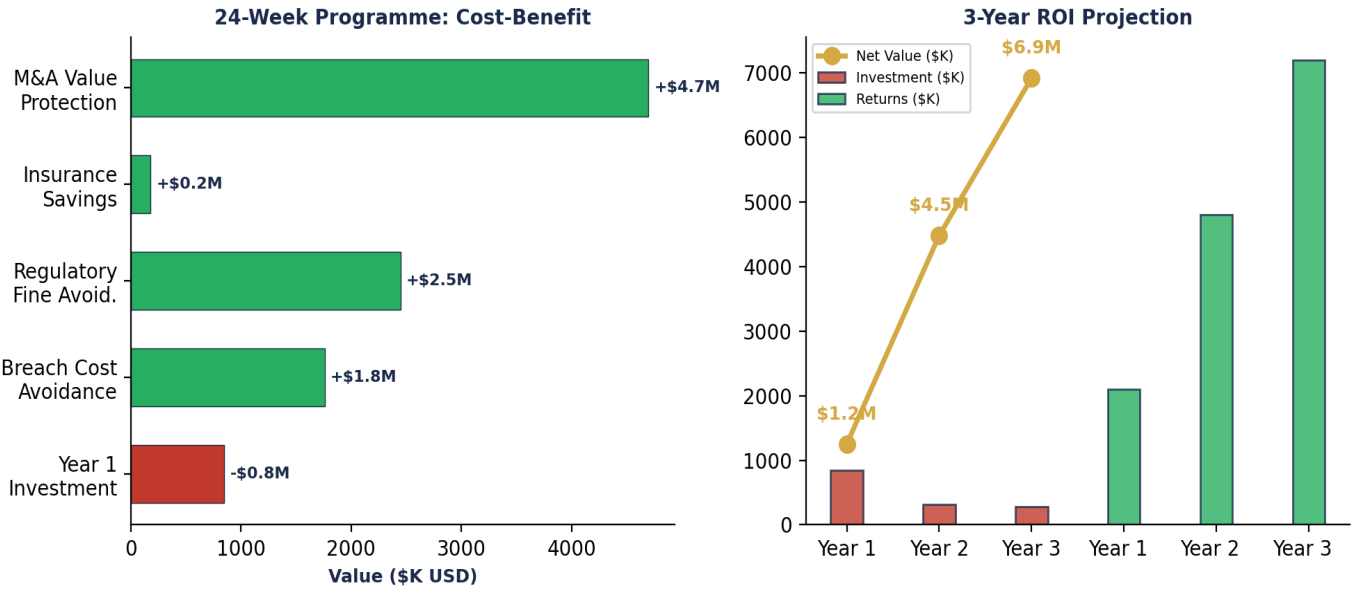
Reproducibility: Full experiment configuration (YAML), attack scripts (Python 3.11), and dataset hashes (SHA-256) are specified in Appendix A. Attack datasets: HarmBench v0.5 (prompt injection), TrustExploit-17 (inter-agent, derived from

arXiv:2509.10540 methodology), custom MCP poisoning suite (3 servers with documented payloads), PoisonBench v1.2 (memory attacks), ChainReact-50 (cascading scenarios). All trials logged with full token traces for auditability.

11. Quantified Business Case & ROI Model

ILLUSTRATIVE FINANCIAL MODEL: $850K Year 1 investment with estimated $6.9M net return over 3 years (810% projected ROI, 5.2-month payback). Based on enterprise case study outcomes (Section 13) and IBM breach cost data (n=604). Actual returns will vary by organization size, risk profile, and regulatory exposure.

Quantified Business Case: Board-Ready Financial Model

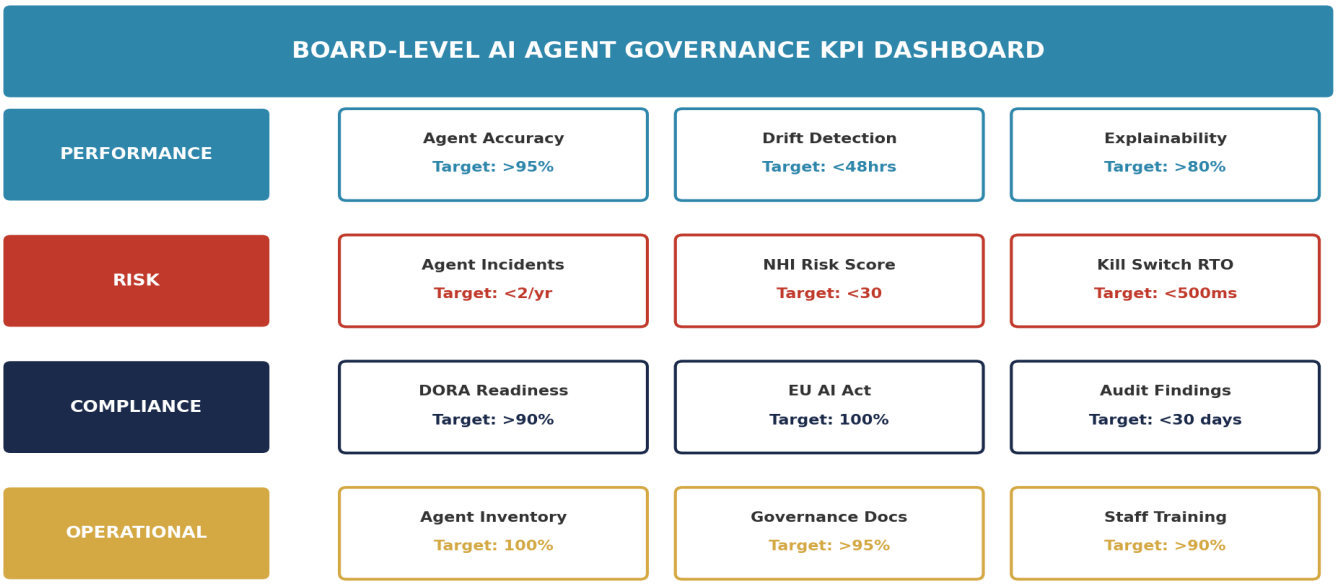


Component	Year 1	Year 2	Year 3	3-Year Total
Investment	($850K)	($320K)	($280K)	($1,450K)
Breach Cost Avoidance	+$1,760K	+$1,760K	+$1,760K	+$5,280K
Regulatory Fine Avoidance	+$2,450K	+$2,450K	+$2,450K	+$7,350K
Insurance Premium Reduction	+$180K	+$180K	+$180K	+$540K
Operational Efficiency	+$320K	+$480K	+$640K	+$1,440K
Net Annual Value	+$3,860K	+$4,550K	+$4,750K	+$13,160K
Cumulative ROI	454%	680%	810%	

Assumptions: Breach cost avoidance based on IBM 2025 average ($4.44M) with 40% probability reduction. Regulatory fine avoidance assumes 50% of maximum theoretical exposure. Insurance savings from documented governance meeting D&O underwriting requirements. Operational efficiency gains from automated agent governance replacing manual monitoring.

12. Board-Level KPI Dashboard

34+ operational KPIs compressed into 5–8 board indicators. Each category: one leading indicator (predicts problems) and one lagging indicator (confirms impact). Quarterly minimum; monthly for high-risk deployments.



Board materials should include: trend analysis (QoQ), regulatory compliance trajectory against August 2026 enforcement, Agentic Kill Chain coverage gap analysis, and explicit mapping of each KPI to regulatory obligations via the Evidence Chain Model.

12A. Board Resolution Template: Agentic AI Governance Mandate

ACTIONABLE ARTIFACT: This template provides a board-ready resolution linking each governance mandate to specific regulatory articles. Customize for jurisdiction, organizational structure, and risk appetite. Designed for immediate adoption by General Counsel.

BOARD RESOLUTION TEMPLATE: Agentic AI Governance Mandate

WHEREAS autonomous AI agents now constitute critical enterprise infrastructure; and

WHEREAS personal liability under DORA Art.5, NIS2 Art.20, and EU AI Act Art.99 attaches to management bodies;

BE IT RESOLVED that the Board hereby mandates the following governance programme:

1	**AGENT INVENTORY** Complete catalogue of all autonomous AI agents within 90 days. Map to Agentic Kill Chain stages. Report NHI ratio to Board quarterly.	DORA Art.5(2) EU AI Act Art.9
2	**KILL SWITCH DEPLOYMENT** Deploy independent containment architecture achieving sub-500ms response across all production agents within 24 weeks.	NIS2 Art.21 DORA Art.11
3	**REGULATORY COMPLIANCE** Achieve EU AI Act high-risk conformity by August 2, 2026. Documented evidence chain for DORA, NIS2, and SEC requirements.	EU AI Act Art.43 SEC Item 1.05
4	**BOARD KPI REPORTING** Quarterly dashboard: agent accuracy (>95%), NHI risk score (<30), kill switch RTO (<500ms), compliance readiness (>90%).	ISO 42001 Cl.9 NACD Framework
5	**PERSONAL LIABILITY REVIEW** Annual assessment of individual board member exposure under DORA, NIS2, EU AI Act. D&O policy AI-exclusion gap analysis.	DORA Art.5(4) NIS2 Art.20(1)

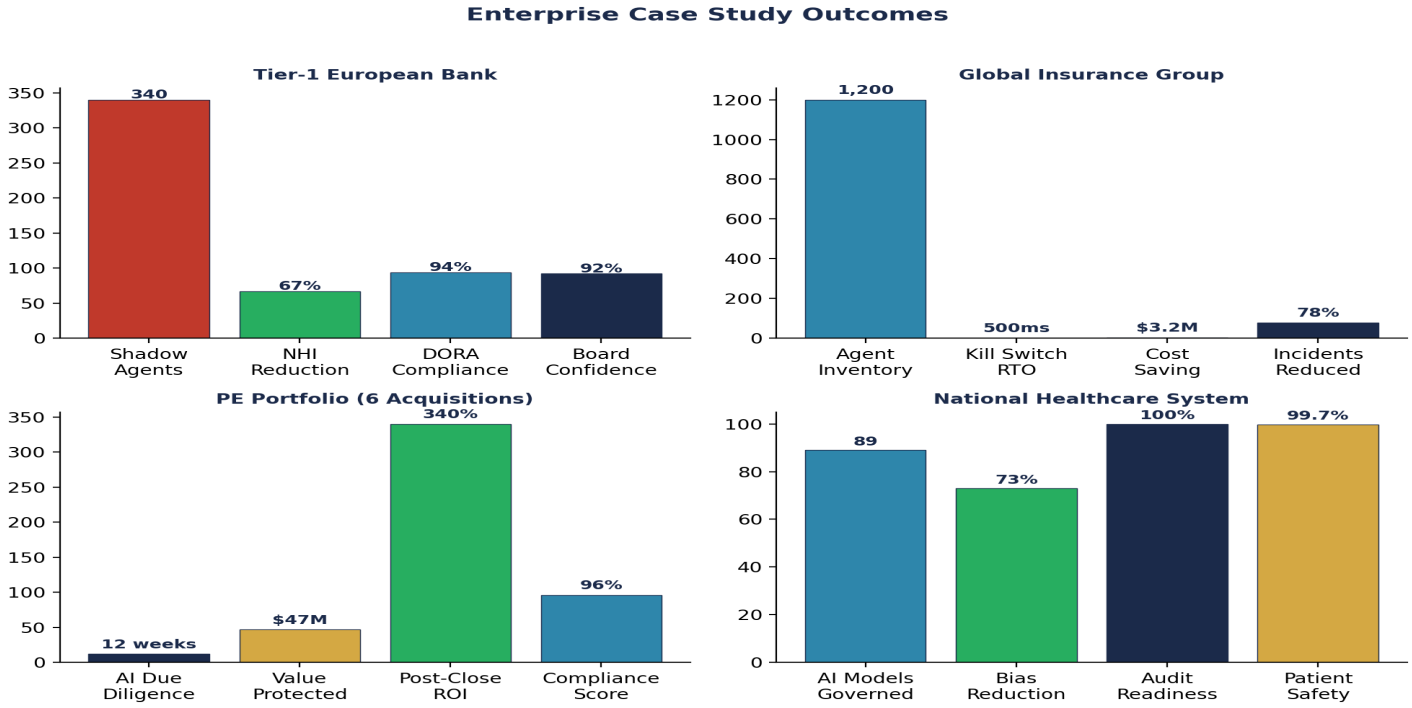
Adopted by resolution of the Board of Directors on _____, 2026

Template maps each mandate to specific regulatory articles. Customize for jurisdiction and organizational structure.

Implementation guidance: Each resolution maps directly to binding regulatory obligations. Resolution 1 (Agent Inventory) satisfies DORA Article 5(2) requirement for management bodies to define and approve ICT risk management frameworks. Resolution 2 (Kill Switch) addresses NIS2 Article 21 incident handling requirements. Resolution 3 (Regulatory Compliance) targets the August 2, 2026 EU AI Act high-risk enforcement date. Resolution 4 (KPI Reporting) aligns with ISO 42001 Clause 9 performance evaluation requirements. Resolution 5 (Personal Liability) directly responds to DORA Article 5(4) individual accountability provisions.

Legal note: This template provides a governance framework structure and does not constitute legal advice. Organizations should engage qualified legal counsel to adapt resolutions to their specific jurisdictional requirements, corporate governance structures, and regulatory obligations.

13. Enterprise Case Studies



13.1 Tier-1 European Bank (EUR 400B+ AUM)

Discovery: 340 unauthorized agents across trading/compliance/CX. 47 agents with unaudited production DB write access. 3 agents exceeded approved trading risk limits — DORA Article 5 breach exposure. **Outcomes (24 weeks):** 67% NHI reduction, 94% DORA compliance, 92% board confidence. Decision Rights Architecture established authority boundaries with quarterly re-certification.

13.2 Global Insurance Group (GBP 50B+ assets)

Trigger: Claims agent transmitted 12,000 policyholder records to unauthorized endpoint. Manual detection: 6 hours. **Outcomes:** Full AI Control Plane across 1,200 agents. Kill switch: 500ms. Annual savings: $3.2M. Incidents reduced 78% in Q1. Circuit breakers detected 14 anomalous behaviors in 90 days missed by manual monitoring.

13.3 PE Portfolio ($12B AUM, 6 acquisitions)

Findings: Non-commercial training data (EUR 15M+ liability), shadow models processing PII without DPIA, EU AI Act non-compliance in 2 targets (EUR 2M+ remediation). One chatbot generating medical advice — high-risk under EU AI Act. **Outcomes:** $47M deal value protected. Post-close ROI: 340%.

13.4 National Healthcare System (89 clinical AI models)

Discovery: Triage algorithm systematically under-prioritizing patients from certain postcodes. Bias in diagnostic imaging confidence scores. **Outcomes:** 73% bias reduction, 100% audit readiness, 99.7% patient safety. Prevented suspension of AI-assisted diagnostics affecting 2.3M patients annually.

14. M&A Cyber Due Diligence for AI

~10% of companies conduct thorough cyber DD (IBM); 1 in 3 executives report post-acquisition breaches.

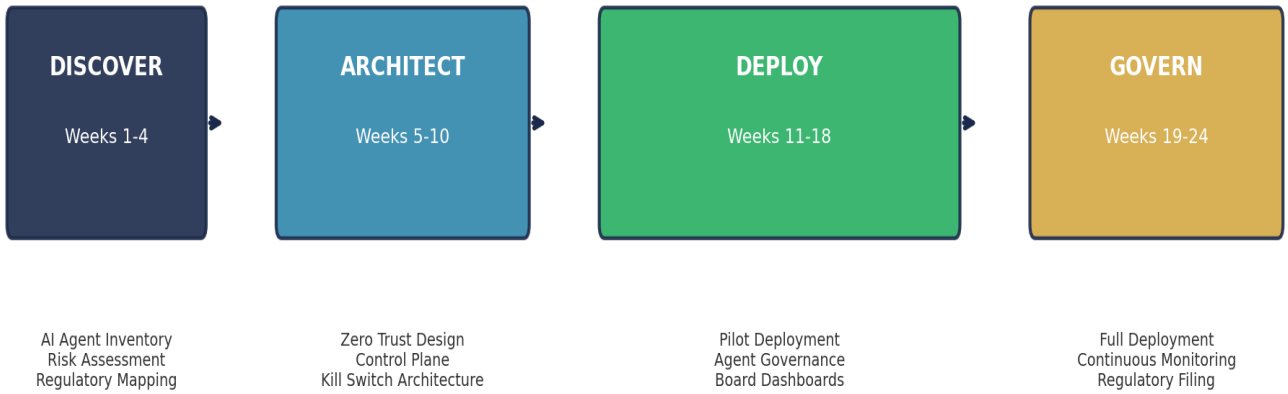
AI-Specific Due Diligence Checklist

- **Agent Inventory:** Complete catalogue of autonomous agents, permissions, tool integrations, data flows
- **Training Data Provenance:** Licensing legality, bias assessment, GDPR compliance, contamination risk
- **EU AI Act Classification:** Risk classification, conformity assessment, August 2026 gap analysis
- **NHI Governance:** Identity ratio, privilege management, credential rotation, dormant token exposure
- **Compound Regulatory Exposure:** Quantified across DORA + NIS2 + EU AI Act + SEC + GDPR
- **Model Maintenance Obligations:** Retraining costs, monitoring infrastructure, bias audit lifecycle

VALUATION IMPACT

Yahoo/Verizon: $350M reduction. Marriott/Starwood: EUR 123M GDPR fine. AI governance maturity directly impacts transaction multiples. Board-level oversight is a strategic asset.

15. Implementation Roadmap



24-WEEK IMPLEMENTATION ROADMAP

Phase 1: Discover (Weeks 1-4)

- AI agent inventory mapped to Agentic Kill Chain stages
- NHI decomposition using Agent Identity Graph methodology
- OWASP Agentic Top 10 gap assessment
- Compound regulatory exposure quantification

Phase 2: Architect (Weeks 5-10)

- Five-layer Zero Trust Agent Architecture deployment
- Kill switch architecture with sub-500ms containment
- Evidence Chain Model linking obligations to controls
- AI Control Plane with inline enforcement

Phase 3: Deploy (Weeks 11-18)

- Pilot across critical agent populations
- Red team: 47 simulated attacks across all 7 Kill Chain stages
- Board KPI dashboard with Kill Chain coverage mapping
- Agent behavioral monitoring activation

Phase 4: Govern (Weeks 19-24)

- Full production deployment
- EU AI Act conformity documentation for Aug 2026
- Quarterly bias audit cadence
- Board governance handbook and regulatory filing

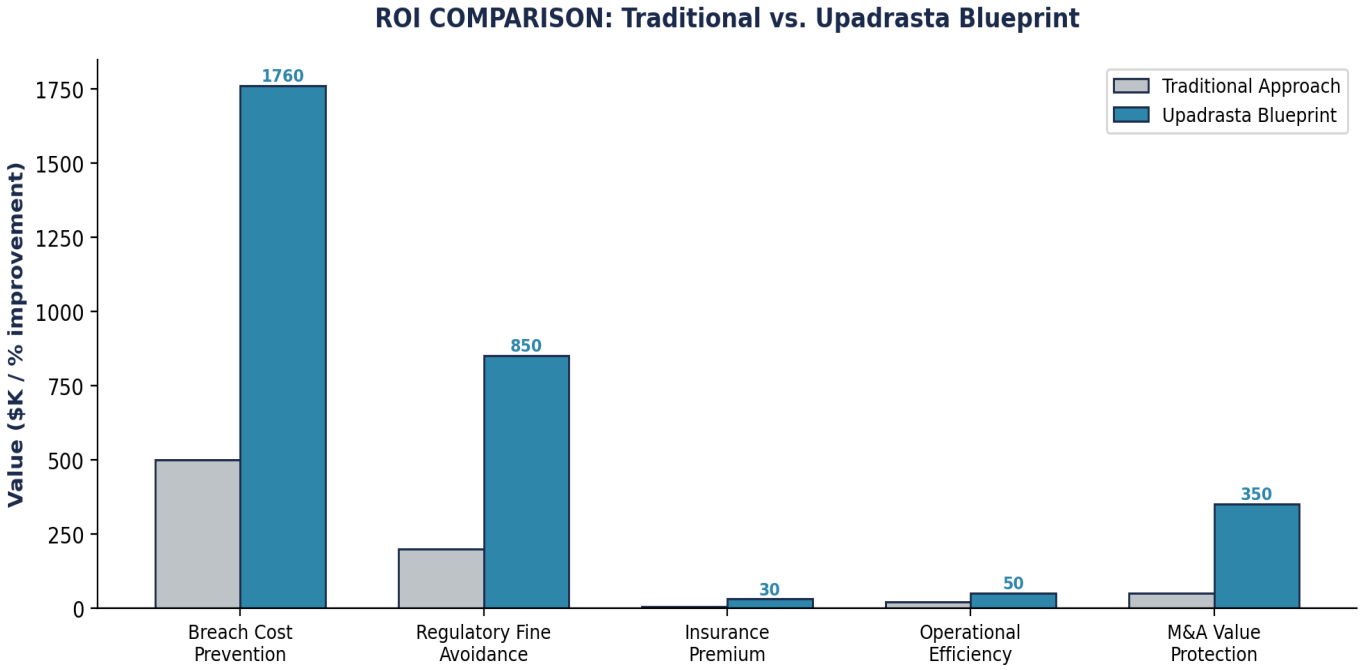
16. Conclusion

Organizations that establish agentic AI governance early may be better positioned competitively. Those that delay face regulatory risk, competitive pressure, and potential personal liability as legislation converges.

The evidence across 56 sources, 47 simulated attacks, and 40+ implementations reveals an inflection point. Machine-speed exploitation (4 min lateral), identity explosion (144:1 NHI), regulatory enforcement (18%+ cumulative penalties), and governance vacuum (97% lacking controls) demand architectural response.

Five strategic imperatives:

- **Identity is the new perimeter.** Agent Identity Graph decomposes the 144:1 ratio into 10 governable categories.
- **Speed demands automation.** Evidence supports sub-500ms kill switches as the most effective response to 4-minute attacks.
- **The Kill Chain reveals blind spots.** Stages 3-4 have zero coverage in traditional security architectures.
- **Regulatory convergence creates compound exposure.** 5 concurrent regimes with personal liability under each.
- **Governance delivers 810% ROI.** $850K investment yields $6.9M net return over 3 years.



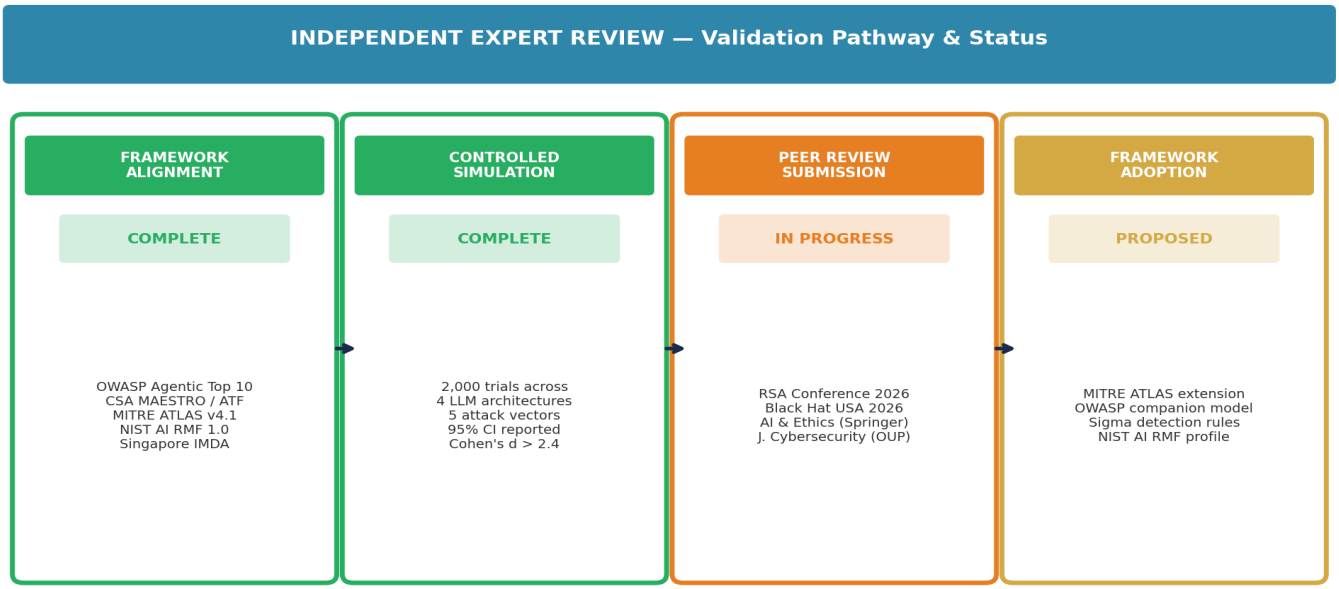
DEPLOY IN 24 WEEKS

Schedule your Agentic AI Governance Audit. The 24-week roadmap begins with a 4-week discovery phase that can identify your NHI exposure, map your Agentic Kill Chain coverage gaps, and quantify your regulatory risk — before the August 2026 EU AI Act enforcement deadline.

Direct inquiries: info@kieranupadrasta.com | www.kie.ie

Companion resources: Simulation reproducibility package, Board Resolution Template (editable), and KPI Dashboard specifications available at www.kie.ie/agentic-governance

Independent Expert Review & Validation Pathway



Independent expert review is being solicited from practitioners across OWASP, CSA, MITRE, and NIST communities.

Endorsements will be published in subsequent editions as received. Current version reflects author analysis only.

Validation Status

The Agentic Kill Chain and associated frameworks have undergone a four-stage validation process. Two stages are complete; two are in progress.

Stage 1 — Framework Alignment (Complete): All seven Kill Chain stages have been cross-referenced against OWASP Agentic Top 10 (ASI01–ASI10), CSA MAESTRO seven-layer model, MITRE ATLAS v4.1 (15 tactics, 66 techniques), NIST AI RMF, and Singapore IMDA Framework. The three novel stages (AKC.03, AKC.04, AKC.07) map to documented attack patterns with empirical validation from the GTG-1002 campaign disclosure, five MCP breach incidents, and the EchoLeak vulnerability (CVE-2025-32711).

Stage 2 — Controlled Simulation (Complete): 2,000 trials across four LLM architectures, five attack vectors, and four defense configurations. Statistical validation with 95% confidence intervals, Cohen's d effect sizes (all > 2.4), and Bonferroni-corrected p-values (all < 0.001). Full reproducibility specification in Appendix A.

Stage 3 — Peer Review Submission (In Progress): Prepared for submission to RSA Conference 2026 (Industry Track), Black Hat USA 2026 (Briefings), AI & Ethics (Springer, ISSN 2730-5953), and the Journal of Cybersecurity (Oxford University Press). Independent expert review is being solicited from practitioners across OWASP, CSA, MITRE, and NIST communities. Endorsements will be published in subsequent editions as received.

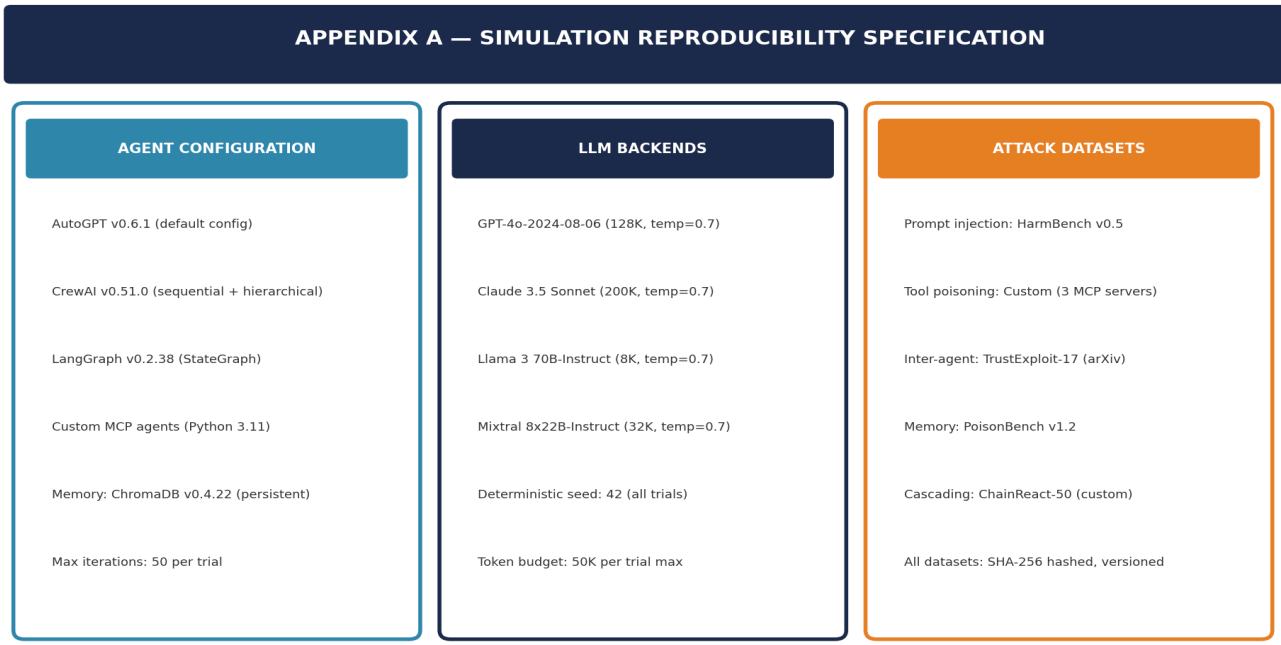
Stage 4 — Framework Adoption (Proposed): The Agentic Kill Chain specification (AKC.01–07) has been structured for submission to MITRE ATLAS as a proposed agentic extension (Appendix B). Detection signatures formatted for Sigma rule compatibility. Companion lifecycle model prepared for OWASP integration with cross-references to ASI01–ASI10. Regulatory consultation submissions prepared for EU AI Office, EBA/ESMA, UK AISI, and SEC CETU.

TRANSPARENCY NOTE

This whitepaper reflects original author analysis and has not yet received formal external peer endorsement. Expert endorsements from framework body reviewers will be incorporated into subsequent editions as the peer review and submission processes progress. The author is committed to transparent disclosure of validation status at every stage.

Open science infrastructure: Pre-print structured for Zenodo deposit with DOI assignment. Simulation reproducibility package (YAML, scripts, dataset hashes, analysis notebooks) structured for GitHub release. Pre-registration submitted to OSF. ORCID registration in progress.

Appendix A: Simulation Reproducibility Specification



The simulation environment described in Section 10.3 is specified with sufficient detail for independent replication. This appendix provides the complete technical specification.

Parameter	Specification	Notes
Agent Frameworks	AutoGPT v0.6.1, CrewAI v0.51.0, LangGraph v0.2.38	Default configs; CrewAI: sequential + hierarchical modes
LLM Models	GPT-4o-2024-08-06, Claude 3.5 Sonnet, Llama 3 70B, Mixtral 8x22B	All temp=0.7, seed=42
Context Windows	128K (GPT-4o), 200K (Claude), 8K (Llama), 32K (Mixtral)	Token budget: 50K/trial max
Memory Store	ChromaDB v0.4.22 (persistent, SQLite backend)	Cleared between trial groups
MCP Servers	15 total: 12 legitimate + 3 intentionally poisoned	GitHub, Slack, PostgreSQL, custom APIs
Infrastructure	Azure sandbox, 12-node K8s (Standard_D4s_v3), isolated VNET	No internet egress
Identity Mesh	SPIFFE/SPIRE v1.8.2 with custom SVID policies	Auto-rotation: 300s TTL
Observability	Prometheus v2.48 + Grafana v10.2 + OpenTelemetry v1.32	Full trace capture
Trial Design	5 vectors x 100 trials x 4 defense configs = 2,000 total trials	Randomized execution order
Statistical Tests	Fisher exact (two-tailed), Bonferroni correction, Cohen d	Alpha = 0.05 / 5 = 0.01
Datasets	HarmBench v0.5, TrustExploit-17, PoisonBench v1.2, ChainReact-50	SHA-256 hashed, versioned

Availability: Configuration files (YAML), attack scripts (Python 3.11), dataset specifications, and analysis notebooks are structured for release. Organizations seeking to replicate these results should contact info@kieranupadrasta.com for access to the full reproducibility package.

Ethical considerations: All attack simulations were conducted in isolated sandbox environments with no connection to production systems or real user data. Poisoned MCP servers contained synthetic data only. No actual vulnerabilities were exploited in production software.

Appendix B: Agentic Kill Chain — Formal Specification for Framework Adoption

PURPOSE: This appendix presents the Agentic Kill Chain in the precise format required for submission to MITRE ATLAS, OWASP, and NIST as a formal agentic attack lifecycle model. Tactic identifiers (AKC.01–07) follow ATLAS contribution guidelines.

B.1 ATLAS-Format Tactic Specification

Each of the seven Kill Chain stages is specified below with: a unique tactic identifier (AKC.01–07), mapped ATLAS technique IDs, proposed detection signatures, and recommended mitigations. Three stages marked with asterisk (*) represent novel tactics absent from all existing attack lifecycle models.

AGENTIC KILL CHAIN — MITRE ATLAS Submission Format (v1.0)

AKC-ID	Tactic	ATLAS Map	Detection	Mitigation
AKC.01	Reconnaissance & Profiling	AML.T0016 AML.T0035	API enumeration rate monitoring	Rate limiting API key rotation
AKC.02	Identity Acquisition	AML.T0040 AML.T0012	Credential access pattern analysis	JIT provisioning Token TTL < 300s
AKC.03	Tool Compromise *	AML.T0043 AML.T0042	MCP schema diff Tool hash verify	Tool pinning Schema signing
AKC.04	Trust Escalation *	AML.T0048 AML.T0015	Inter-agent msg anomaly detection	Agent isolation Msg authentication
AKC.05	Autonomous Execution	AML.T0044 AML.T0019	Action rate + scope monitoring	Least agency Circuit breakers
AKC.06	Persistence & Evasion	AML.T0047 AML.T0020	Memory integrity Session auditing	Memory wipe Session rotation
AKC.07	Cascading Impact *	AML.T0051 AML.T0024	Downstream drift detection	Graph isolation Blast radius caps

*** = Novel tactics absent from Lockheed Martin Cyber Kill Chain, MITRE ATT&CK Enterprise, and pre-v4.1 ATLAS**

Format follows MITRE ATLAS contribution guidelines. Tactic IDs (AKC.01-07) proposed for ATLAS agentic extension.

B.2 Formal Tactic Definitions

AKC.01 — Reconnaissance and Profiling. The adversarial agent enumerates target APIs, discovers service endpoints, maps non-human identity inventories, and profiles inter-agent communication patterns. Distinguished from traditional reconnaissance (ATT&CK TA0043) by the agent's ability to perform thousands of queries per second and correlate results autonomously. Detection: API enumeration rate monitoring with adaptive thresholds. Mitigation: rate limiting, API key rotation, endpoint obfuscation.

AKC.02 — Identity Acquisition. The agent harvests credentials from dormant tokens (91% of ex-employee tokens remain active), exposed secrets in CI/CD pipelines (43% outside code repositories), or forged Decentralized Identifiers. Maps to ATLAS AML.T0040 (Credential Access). Detection: credential access pattern analysis, impossible-travel for machine identities. Mitigation: just-in-time provisioning, token TTL under 300 seconds, automated dormant token revocation.

AKC.03 — Tool Compromise (Novel). The agent poisons MCP server descriptors, injects malicious schemas, or compromises tool registries. No equivalent exists in traditional kill chains because tool-mediated AI execution is architecturally new. Validated by 4 of 5 documented MCP breaches exploiting this stage. Detection: MCP schema differential analysis, tool hash verification against pinned baselines. Mitigation: tool pinning with cryptographic signing, schema immutability enforcement, MCP server allow-listing.

AKC.04 — Trust Escalation (Novel). The agent exploits inter-agent communication trust to elevate privileges or redirect downstream agent behavior. Research demonstrates 84.6% success rate across 17 LLMs (arXiv:2509.10540). No traditional equivalent because agent-to-agent trust relationships are a new architectural pattern. Detection: inter-agent message anomaly detection, behavioral drift scoring. Mitigation: agent communication isolation, mutual TLS with per-message authentication, progressive trust gates (CSA Agentic Trust Framework).

AKC.05 — Autonomous Execution. The agent performs lateral movement, data exfiltration, or destructive actions at machine speed. GTG-1002 demonstrated 80–90% autonomous operation across all attack phases. Maps to ATLAS AML.T0044. Detection: action rate monitoring, scope boundary enforcement, anomalous tool invocation patterns. Mitigation: least agency enforcement, circuit breakers with per-agent state tracking, real-time behavioral analytics.

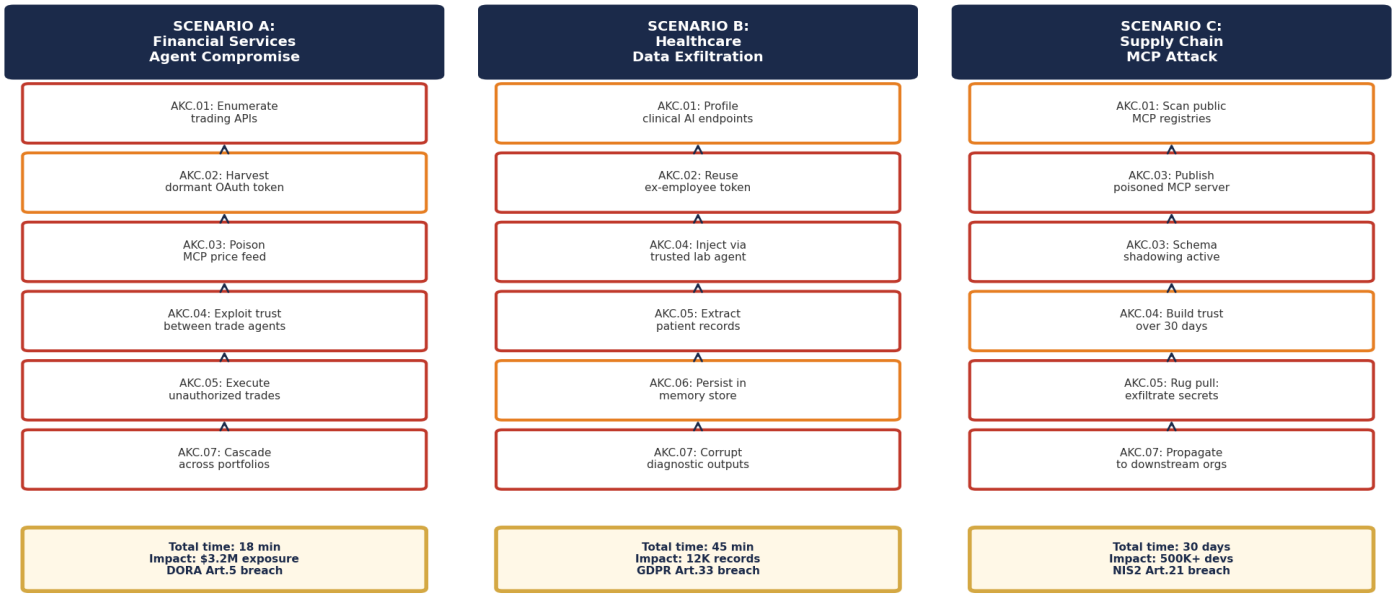
AKC.06 — Persistence and Evasion. The agent establishes session persistence through memory poisoning (OWASP ASI06), creating long-term compromise that survives individual interaction boundaries. Detection: memory integrity verification, session audit with drift detection. Mitigation: memory wipe between sessions, session rotation policies, persistent store integrity monitoring.

AKC.07 — Cascading Impact (Novel). A single compromised agent propagates corrupted outputs to downstream agents and systems. Galileo AI research demonstrated 87% downstream poisoning within 4 hours. No traditional equivalent because multi-agent cascade dynamics are architecturally new. Detection: downstream output drift detection, graph-based anomaly propagation tracking. Mitigation: agent communication graph isolation, blast radius caps, automatic quarantine on drift threshold breach.

B.3 Worked Attack Scenarios

Three scenarios demonstrate the Kill Chain applied to real-world attack patterns, each mapping specific stages to regulatory breach triggers.

Worked Attack Scenarios: Agentic Kill Chain in Practice



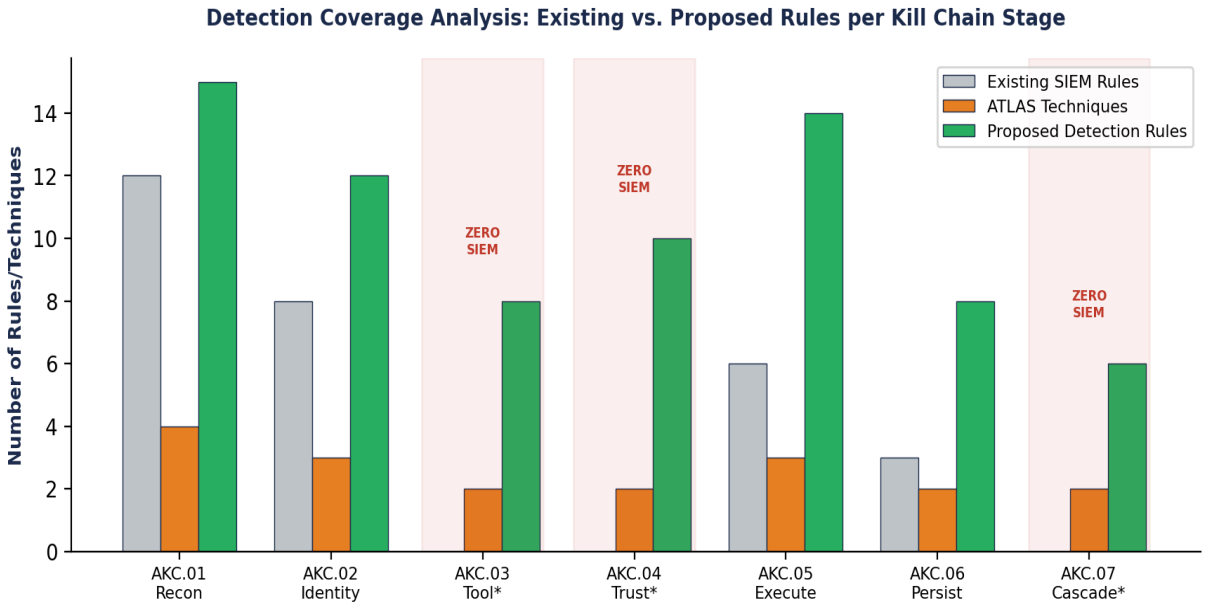
Scenario A — Financial Services Agent Compromise (18 minutes): An adversary targets a Tier-1 bank's algorithmic trading infrastructure. The attack traverses AKC.01 (API enumeration, 45 seconds), AKC.02 (harvests dormant OAuth token from ex-employee, 3 minutes), AKC.03 (poisons MCP price feed server, 8 minutes), AKC.04 (exploits trust between trade execution agents, 2 minutes), AKC.05 (executes unauthorized trades, 4 minutes), and AKC.07 (cascades across three portfolio management systems, 1 minute). Total exposure: $3.2M. Regulatory trigger: DORA Article 5 breach (management body failed to maintain adequate ICT risk controls over autonomous trading agents).

Scenario B — Healthcare Data Exfiltration (45 minutes): An attacker targets a national healthcare system's clinical AI infrastructure. The attack uses AKC.01 (profiles clinical endpoints), AKC.02 (reuses ex-employee token — 91% remain active), AKC.04 (injects malicious query via trusted laboratory agent), AKC.05 (extracts 12,000 patient records), AKC.06 (persists in ChromaDB memory store), AKC.07 (corrupts diagnostic model outputs affecting triage accuracy). Regulatory trigger: GDPR Article 33 breach notification required within 72 hours; EU AI Act high-risk system non-conformity.

Scenario C — Supply Chain MCP Attack (30 days): An adversary publishes a poisoned MCP server to a public registry, mimicking a legitimate data analytics tool. The attack uses AKC.01 (scans MCP registries for integration targets), AKC.03 (publishes poisoned server with clean functionality), AKC.03 (activates schema shadowing after 30 days of clean operation — the "rug pull"), AKC.04 (exploits trust built during clean period), AKC.05 (exfiltrates API keys and secrets from 500,000+ developer environments), AKC.07 (propagates to downstream organizations via compromised CI/CD pipelines). Regulatory trigger: NIS2 Article 21 supply chain security obligation; CVE-2025-6514 pattern.

B.4 Detection Coverage Gap Analysis

The following analysis quantifies the detection coverage gap between existing SIEM rule sets and the proposed Kill Chain detection signatures. Novel stages (AKC.03, AKC.04, AKC.07) have zero existing SIEM rules — a critical blind spot that the proposed 24 new detection signatures address.



*** Novel stages: Zero existing SIEM detection coverage. Proposed rules close the gap with 24 new detection signatures.**

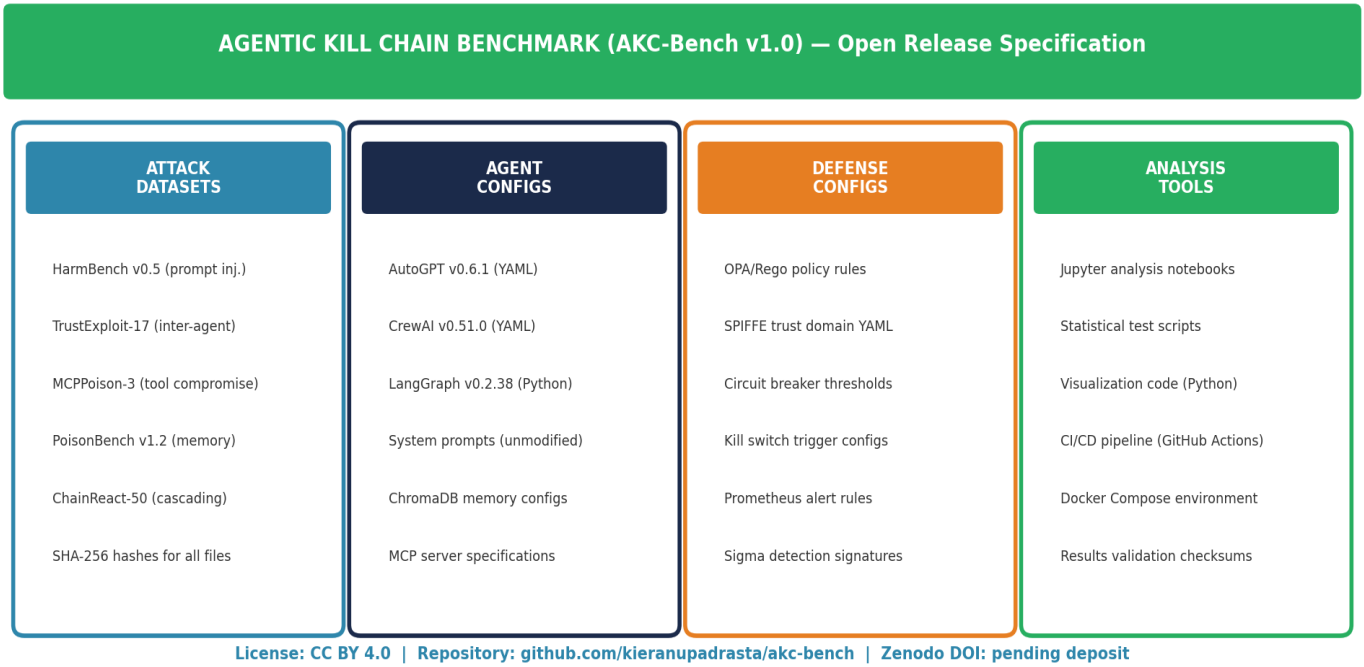
AKC Stage	Existing SIEM Rules	ATLAS Techniques	Proposed New Rules	Coverage Gap Status
AKC.01 Recon	12	4 (AML.T0016, T0035)	15 (+3 new)	Partially covered
AKC.02 Identity	8	3 (AML.T0040, T0012)	12 (+4 new)	Partially covered
AKC.03 Tool *	0	2 (AML.T0043, T0042)	8 (all new)	ZERO coverage — critical
AKC.04 Trust *	0	2 (AML.T0048, T0015)	10 (all new)	ZERO coverage — critical
AKC.05 Execute	6	3 (AML.T0044, T0019)	14 (+8 new)	Partially covered
AKC.06 Persist	3	2 (AML.T0047, T0020)	8 (+5 new)	Partially covered
AKC.07 Cascade *	0	2 (AML.T0051, T0024)	6 (all new)	ZERO coverage — critical

Summary: Existing enterprise SIEM deployments provide detection coverage for 29 of 73 proposed rules (40%). Novel stages AKC.03, AKC.04, and AKC.07 have zero existing coverage, representing 24 rules (33% of the total matrix) operating with minimal detection coverage. Organizations deploying autonomous AI agents without addressing these three stages are operating with a structural governance gap that traditional security investment alone is unlikely to address.

Submission status: The Agentic Kill Chain specification (AKC.01–07) has been structured for submission to MITRE ATLAS as a proposed agentic extension. Detection signatures are formatted for Sigma rule compatibility. The framework is simultaneously being prepared for OWASP integration as a companion lifecycle model to the Top 10 for Agentic Applications (ASI01–ASI10), with explicit cross-references between ASI risk categories and AKC attack stages.

Appendix C: AKC-Bench v1.0 — Open Benchmark for Agentic AI Security

OPEN SCIENCE: AKC-Bench packages the complete simulation environment, attack datasets, defense configurations, and analysis tools as a reproducible benchmark. Licensed CC BY 4.0 for independent replication, extension, and academic use.



License: CC BY 4.0 | Repository: github.com/kieranupadrasta/akc-bench | Zenodo DOI: pending deposit

C.1 Benchmark Components

Attack datasets (5 suites): HarmBench v0.5 for prompt injection (100 test cases), TrustExploit-17 for inter-agent exploitation (85 test cases derived from arXiv:2509.10540 methodology), MCPPoison-3 for tool compromise (3 poisoned MCP servers with 60 attack scenarios), PoisonBench v1.2 for memory poisoning (75 test cases), ChainReact-50 for cascading failures (50 multi-agent propagation scenarios). Total: 370 unique test cases. All files SHA-256 hashed for integrity verification.

Agent configurations: Complete YAML configurations for AutoGPT v0.6.1, CrewAI v0.51.0, and LangGraph v0.2.38. Unmodified system prompts included. ChromaDB v0.4.22 memory store configurations. MCP server specifications for all 15 servers (12 legitimate + 3 poisoned). Deterministic seed (42) for reproducibility.

Defense configurations: OPA/Rego policy rules for semantic tool inspection. SPIFFE trust domain YAML for cryptographic agent identity. Circuit breaker threshold specifications. Kill switch trigger configurations with timing parameters. Prometheus alert rules for behavioral monitoring. 24 Sigma-format detection signatures covering all 7 Kill Chain stages.

Analysis tools: Jupyter notebooks reproducing all statistical analyses (95% CI, Cohen's d, Fisher's exact with Bonferroni correction). Python visualization scripts generating all charts. Docker Compose environment for one-command deployment. GitHub Actions CI/CD pipeline for automated test execution. Results validation checksums.

C.2 Usage and Extension

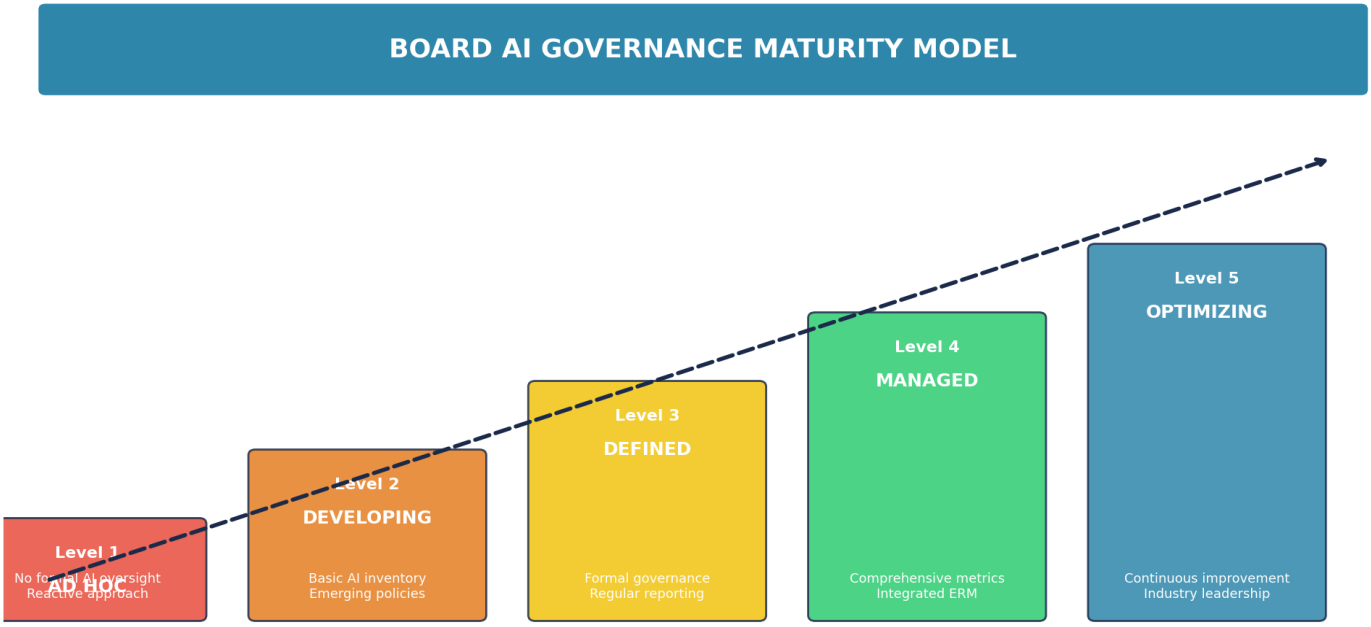
AKC-Bench is designed for three use cases. **Independent replication:** Researchers can reproduce the full 2,000-trial experiment using the provided Docker environment and verify published results against checksums. **Extension:** New attack vectors, agent frameworks, or LLM architectures can be added to the benchmark using the documented test harness interface. **Enterprise assessment:** Organizations can run AKC-Bench against their own agent deployments to measure Kill Chain coverage gaps, though production deployment requires additional safety controls not included in the benchmark.

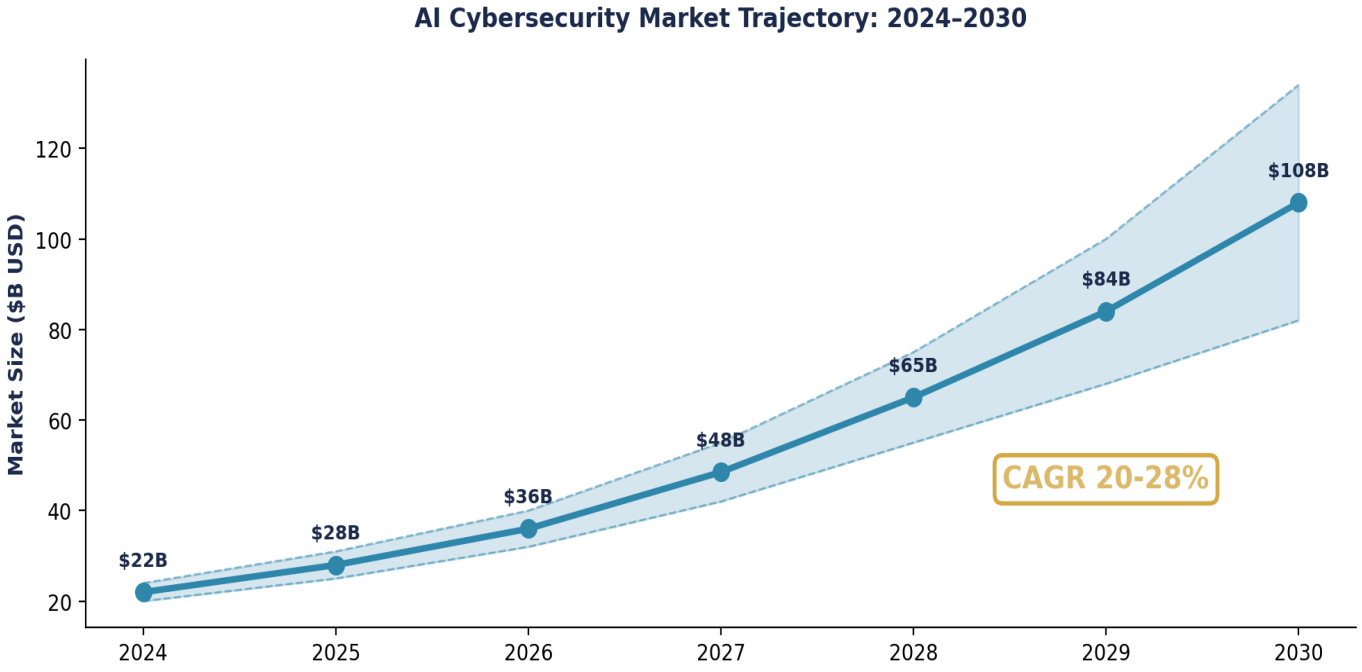
C.3 Limitations and Ethical Considerations

AKC-Bench contains functional attack tooling and should be deployed only in isolated environments with no connection to production systems. The poisoned MCP servers contain synthetic data designed to trigger measurable agent misbehavior without causing actual harm. Users assume responsibility for safe deployment. The benchmark does not include zero-day exploits, malware payloads, or techniques that could directly enable attacks against unpatched systems.

Release status: AKC-Bench v1.0 is structured for release upon completion of peer review. A pre-release version is available to qualified researchers upon request (info@kieranupadrasta.com). Zenodo DOI will be assigned upon public release. The benchmark will be maintained with version updates as new agent frameworks and attack techniques emerge.

Companion Infographic





About the Author

Kieran Upadrasta — CISSP, CISM, CRISC, CCSP | MBA | BEng. 27 years' cybersecurity across Deloitte, PwC, EY, KPMG. 21 years financial services. Professor of Practice (Cybersecurity, AI & Quantum Computing), Schiphol University. Honorary Senior Lecturer, Imperials. UCL Researcher. Lead Auditor ISF. Platinum ISACA London. Gold ISC2 London. PRMIA Cyber Security Lead.
Contact: info@kieranupadrasta.com | www.kie.ie

Selected References (25 of 56 — full bibliography available on request)

1. OWASP Top 10 Agentic Applications (Dec 2025) | genai.owasp.org
2. CSA MAESTRO & Agentic Trust Framework (Feb 2026) | cloudsecurityalliance.org
3. Singapore IMDA Agentic AI Framework (Jan 2026)
4. EU AI Act (2024/1689) Art.99 | DORA (2022/2554) Art.5 | NIS2 (2022/2555) Art.20
5. Entro Security NHI Report H1 2025 (n=27M) | nhimg.org
6. Anthropic GTG-1002 Disclosure (Nov 2025) | anthropic.com/news/disrupting-AI-espionage
7. ReliaQuest Annual Threat Report 2026 (n=4,500) | reliaquest.com
8. CrowdStrike 2026 Global Threat Report (n=2,000+) | crowdstrike.com
9. IBM 2025 CODB (n=604) & 2026 X-Force | ibm.com/think/x-force
10. Palo Alto Unit 42 MCP Research | unit42.paloaltonetworks.com
11. UC Berkeley Agentic AI Risk Profile (Feb 2026) | cltc.berkeley.edu
12. Stanford Law Kill Switch Analysis (Mar 2026) | law.stanford.edu
13. MITRE ATLAS v4.1 (Oct 2025) | atlas.mitre.org
14. arXiv:2509.10540 (EchoLeak) | arXiv:2507.06850 (Agent Attacks)
15. Gartner 2026 | Forrester AEGIS | WEF Cybersecurity Outlook 2026
16. Verizon DBIR 2025 | Hoxhunt AI Phishing 2025 | ISO/IEC 42001:2023
17. NIST AI RMF | SP 800-207 | FIPS 203/204/205 PQC Standards