CRANIUM

AI SECURITY & GOVERNANCE · GUIDE

The Ultimate Guide for Agentic AI Security

How autonomous agents reshape the threat landscape — and the existing, expanded, and novel controls you need to stay in control.

FIELD GUIDE · 2026

CRANIUM · AI SECURITY & GOVERNANCE

CONTENTS

Table of contents

01 WHY YOU HAVE TO CARE

Agentic AI changes everything

Agentic AI isn't just another tech buzzword — it represents a fundamental shift in how intelligent systems operate, make decisions, and interact with the world. As AI agents become more autonomous, they introduce both powerful opportunities and new risks that traditional security and governance can't fully address.

If you care about trust, compliance, and keeping humans in control, understanding agentic AI is essential.

77%

of executives agree AI agents will reinvent how their organization builds digital systems.

— ACCENTURE

02 A NEW OPERATING MODEL

How AI agents change information systems

AI agents redefine information systems by complementing the unique strengths of people and traditional applications — while introducing capabilities that were previously out of reach.

PEOPLE

Judgment & oversight

Bring judgment, creativity, and ethical oversight — essential for defining objectives and managing ambiguous situations.

APPS

Reliable execution

Excel at automating repetitive, well-defined tasks with speed and reliability, but are limited by fixed scope and structured input.

AI AGENTS

Autonomous orchestration

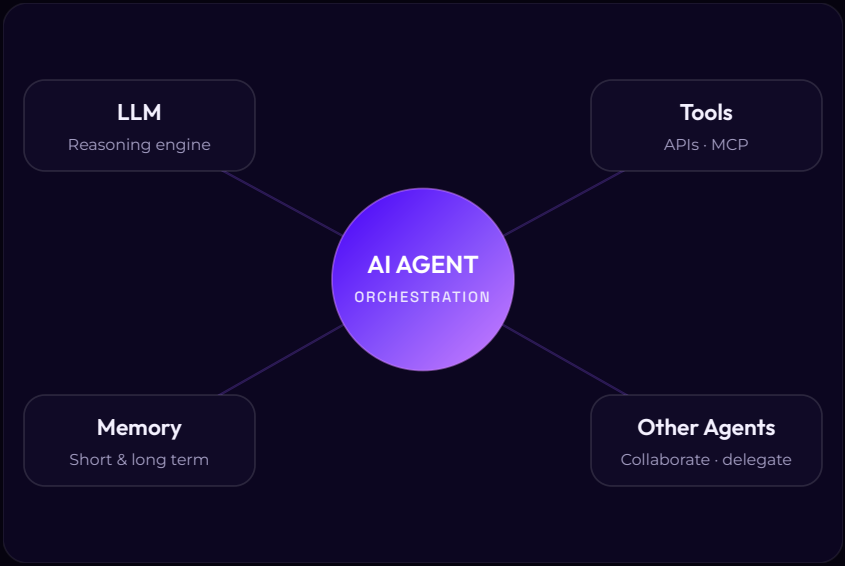
Bridge these worlds by interpreting natural-language instructions and orchestrating actions across diverse tools and data sources.

ENTITY	UNIQUE ABILITY	TYPICAL EXAMPLE
People	Judgment, ethics, creative intent	Overriding process for special cases
Applications	Automated, repeatable, rule-based execution	Form-based data entry, batch processing
AI Agents	Autonomous reasoning, dynamic task orchestration	Booking, summarizing, personalizing

03 ANATOMY OF AN AI AGENT

More than a chatbot: an autonomous system

An AI agent bridges the gap between user intent (the "what") and effective execution (the "how"). Unlike rigid, pre-programmed software, it interprets high-level instructions, analyzes its capabilities, and dynamically constructs a plan in real time. Its intelligence derives from orchestrating four key resources:



LARGE LANGUAGE MODEL

The core reasoning engine — interprets prompts, generates step-by-step plans, and adapts actions to evolving context.

TOOLS

Traditional applications or services accessed via APIs or protocols like MCP, extending the agent beyond language generation.

MEMORY

Short-term (current workflow) and long-term (persistent knowledge, preferences) recall for continuity and personalization.

OTHER AGENTS

Specialist agents and sub-agents that collaborate and delegate to tackle complex, multi-step goals.

> By 2030, up to **50%** of enterprise application software offerings will include some agentic AI features — up from less than 5% in 2025. — Gartner

04 WHAT CONTROLS TO FOCUS ON

Securing AI Agents vs. AI Agent Security

Not all attack vectors for AI agents are new. It's crucial to distinguish between the two.

SECURING AI AGENTS

Covers the whole environment — data, networks, and access — much of which existing controls already handle.

AI AGENT SECURITY

Focuses on the risks unique to advanced AI systems — where traditional controls fall short.

To expose the gaps AI creates, we extend **STRIDE** — an industry-standard threat-modeling framework — with two new categories built for agentic systems: **STRIDE+ML**.

+M Misunderstanding

Models making undesirable assessments from missing context or malicious intervention — leading to unexpected emergent behaviors.

+L Lack of Accountability

Actions performed without clear governance or ownership — making responsibility hard to determine when issues arise.

05 APPLYING STRIDE+ML

Mapping OWASP threats onto STRIDE+ML

The **OWASP Top 10 for LLMs** and **OWASP AI Agent Threats** are recognized industry standards. Bridging them with STRIDE+ML uncovers a pattern: many AI threats fit classic STRIDE categories, but a substantial portion defy those boundaries — demanding the nuance of the ML categories.

7 of 25

OWASP threats fall into Misunderstanding and Lack of Accountability — outside classic STRIDE.

These threats fall into three buckets — a spectrum from covered, to stretched, to genuinely new:

EXISTING

Already covered

Addressable with established controls applied to the AI context.

EXPAND

Stretch your defenses

Known controls that must be adapted for AI's speed and new attack surfaces.

NOVEL

Innovate

Brand-new approaches designed for risks unique to autonomous AI.

For deeper mitigation guidance, see the OWASP Top 10 for LLM Applications and OWASP AI Agent Threats and Mitigation documentation.

06 EXISTING CONTROLS

Existing controls

ALREADY COVERED

These threats align with classic concerns — data leakage, privilege escalation, denial of service, unauthorized access. Proven solutions like **IAM, DLP, access control, secure communications, logging, and resource throttling** remain effective, provided they are consistently applied and adapted to the AI/LLM context.

OWASP THREAT	STATUS	STRIDE+ML	JUSTIFICATION
LLM02 Sensitive Info Disclosure	EXISTING	INFO DISCLOSURE	Information-disclosure risk; DLP and access controls exist but must adapt to LLM output and interaction.
LLM06 Excessive Agency	EXISTING	ELEV. OF PRIVILEGE	Agents may act beyond their role or impersonate others; privilege boundaries apply, but autonomous actions need monitoring.
LLM10 Unbounded Consumption	EXISTING	DENIAL OF SERVICE	Excess resource consumption by agents/LLMs can be throttled with standard rate limits and quotas.
T12 Agent Comms Poisoning	EXISTING	TAMPERING	Existing secure-comms protocols work if applied; new agent protocols must match standard cryptographic and authN/Z practices.
T3 Privilege Compromise	EXISTING	ELEV. OF PRIVILEGE	Unauthorized privilege gain; least-privilege and strong authN/Z are effective if consistently enforced for agents.
T4 Resource Overload	EXISTING	DENIAL OF SERVICE	Exhaustion of system resources prevents service; DoS controls and throttling apply to LLM/agent workloads.
T8 Repudiation / Untraceability	EXISTING	REPUDIATION	Denying actions taken; standard logging helps, but logs must include agent decisions for complete traceability.
T9 Identity Spoofing	EXISTING	SPOOFING	Existing IAM practices remain effective in most enterprise contexts.
LLM08 Vector & Embedding Weaknesses	EXISTING	INFO DISCLOSURE	Embeddings are like any database data; apply object-level authorization and protect direct database access.

07 EXPANDING CONTROLS

Expanding controls

STRETCH YOUR DEFENSES

Traditional challenges — tampering, denial of service, privilege abuse — in more dynamic AI environments. Foundations like **supply-chain management, input validation, session handling, zero trust, and IAM** still apply, but must expand for the speed, automation, and new attack surfaces of agents.

OWASP THREAT	STATUS	STRIDE+ML	JUSTIFICATION
LLM03 Supply Chain	EXPAND	TAMPERING	Supply-chain tampering via third-party dependencies; existing controls apply, but the LLM/agent supply chain is more dynamic.
LLM04 Data & Model Poisoning	EXPAND	TAMPERING	Data/model tampering causes malfunction; integrity and input validation are known but not yet mature for dynamic agents.
LLM05 Improper Output Handling	EXPAND	TAMPERING	Most output-validation risks have known solutions; novel issues like prompt injection are covered by other threats.
LLM07 System Prompt Leakage	EXPAND	INFO DISCLOSURE	Best addressed by keeping prompts non-sensitive and proper error handling; novel risks handled in other categories.
T1 Memory Poisoning	EXPAND	TAMPERING	Unauthorized modification of agent memory/state; session management exists but was built for performance, not security.
T10 Overwhelming HITL	EXPAND	DENIAL OF SERVICE	Flooding human-in-the-loop creates denial of oversight; classic DoS principles apply but need HITL adaptation.
T11 Unexpected RCE / Code Attacks	EXPAND	TAMPERING	Arbitrary code execution is high risk; block by default, allow only in low-risk cases with strong novel controls.
T13 Rogue Agents	EXPAND	SPOOFING	Strong asset inventory and IAM cover much of the risk; immature agent marketplaces need better monitoring and attestation.
T14 Human Attacks on MAS	EXPAND	ELEV. OF PRIVILEGE	Apply zero trust and least privilege to users and agents; addressing novel AI-specific threats reduces this further.

> By 2027, over 40% of agentic AI projects will be canceled due to escalating costs, unclear business value, or **inadequate risk controls.** — Gartner

08 NOVEL CONTROLS

Novel controls

INNOVATE TO STAY AHEAD

These stem from the unique, dynamic reasoning and autonomy of AI agents — prompt injection, misinformation, tool misuse, intent manipulation, and trust manipulation — challenges classic controls can't

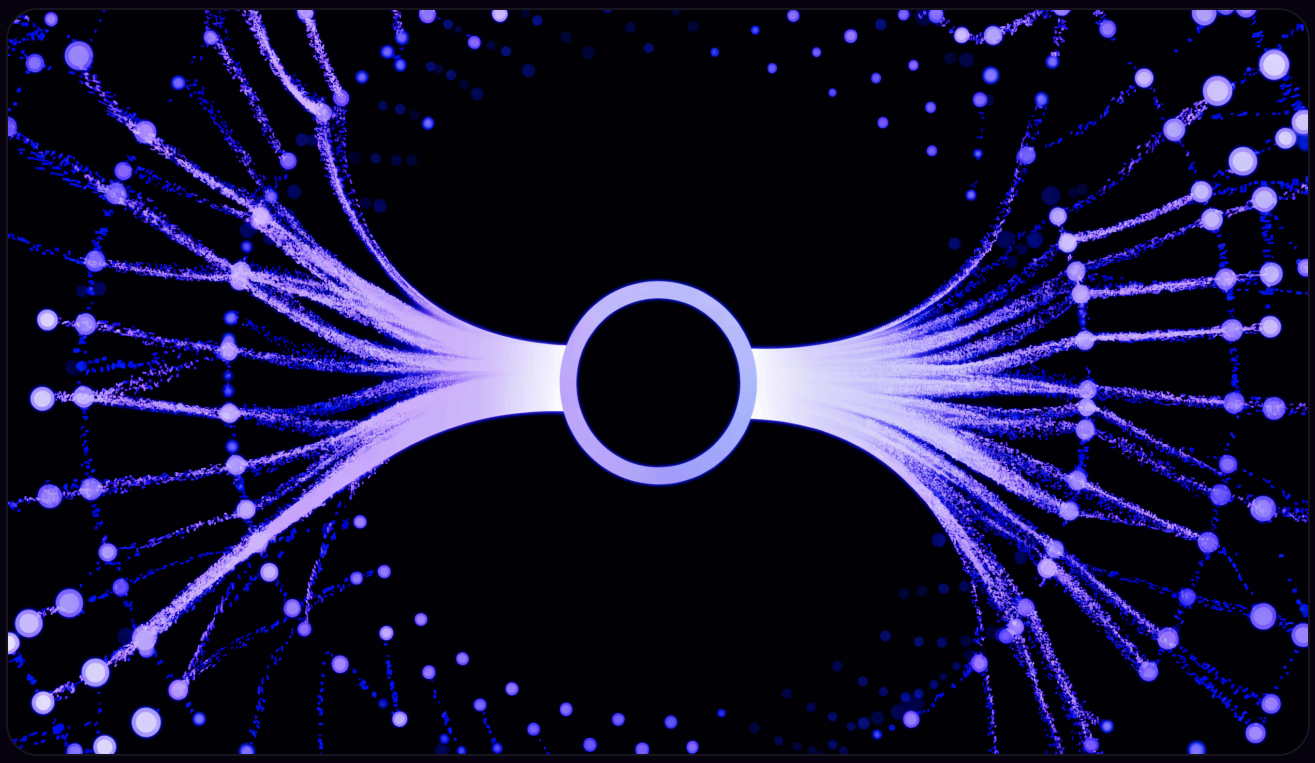
fully address. Mitigation needs **dynamic output validation, explainability tools, human oversight, and new policy enforcement** built for unpredictable, adaptable AI.

OWASP THREAT	STATUS	STRIDE+ML	JUSTIFICATION
LLM01 Prompt Injection	NOVEL	MISUNDERSTANDING	Malicious prompts subvert intended operation; prevention is unsolved and requires new controls beyond input filtering.
LLM09 Misinformation	NOVEL	MISUNDERSTANDING	Spreads falsehoods from misunderstanding or lack of oversight; traditional output validation doesn't cover generative failures.
T15 Human Trust Manipulation	NOVEL	ACCOUNTABILITY	Agent output manipulates user trust; no standard for agent-trust monitoring — needs new psychological/societal approaches.
T2 Tool Misuse	NOVEL	MISUNDERSTANDING	Agent misapplies tools from lack of context; no established dynamic validation, so new monitoring is needed.
T5 Cascading Hallucinations	NOVEL	MISUNDERSTANDING	Compounding errors; responses should align to a strong source of truth, but the LLM may still assemble it misleadingly.
T6 Intent Breaking & Goal Manip.	NOVEL	MISUNDERSTANDING	Agents misled from intended goals; robust validation or human-in-the-loop for intent verification is lacking.
T7 Misaligned & Deceptive Behavior	NOVEL	ACCOUNTABILITY	Autonomous agent actions are hard to govern like people; new accountability mechanisms are needed.

09 GOVERNANCE

Who's really in charge? Governing agentic AI with explainability

As agentic AI makes complex, autonomous decisions — chaining actions and invoking tools on its own — it's essential to understand not just *what* the AI did, but *why*. Explainable AI (XAI) makes the decision process transparent and interpretable. If an agent flags a customer interaction as risky, explainability lets you trace the data, context, and logic behind that decision — to justify it, correct it, or improve future outcomes.



Not all explainability is equal. With standard LLMs, the "chain of thought" is non-deterministic — freeform text generated on the fly — making it hard to audit, interpret, or turn into action. Without explainability, it's the AI — not your people or leadership — effectively running your business.

THE CRANIUM APPROACH

Cranium's policy action engine does **not** use generative AI. It uses deterministic machine-learning models — the right data-science tool for classifying risk and intent — so every prompt and response is assessed consistently, logging exactly which signals and policy rules were triggered. Those results convert directly into actionable metrics, making audits, compliance, and continuous improvement straightforward and reliable.

10 CONCLUSION

Make AI work for you — not the other way around

Agentic AI is transforming how organizations operate, bringing both remarkable capabilities and unprecedented risks. The smartest approach isn't to reinvent the wheel — it's to first maximize the value of your existing cybersecurity program wherever it applies.

At the same time, focus R&D on the small but critical set of threats that demand new controls — the unique risks introduced by the autonomy, reasoning, and emergent behavior of agentic AI. Do this, and you safeguard innovation, build trust, and maintain true control.

Existing

Apply proven controls to the AI context.

Expand

Adapt defenses for AI's speed and surfaces.

Novel

Innovate where autonomy creates new risk.

In just five minutes, Cranium's guardian agent oversees all AI traffic and eliminates the risks associated with AI agents.

CRANIUM

SEE IT IN ACTION

Keep humans in control of autonomous AI.

Cranium's guardian agent oversees all AI traffic in minutes — detecting prompt injection, tool misuse, and intent manipulation, with deterministic, auditable explainability.

Book My Demo →