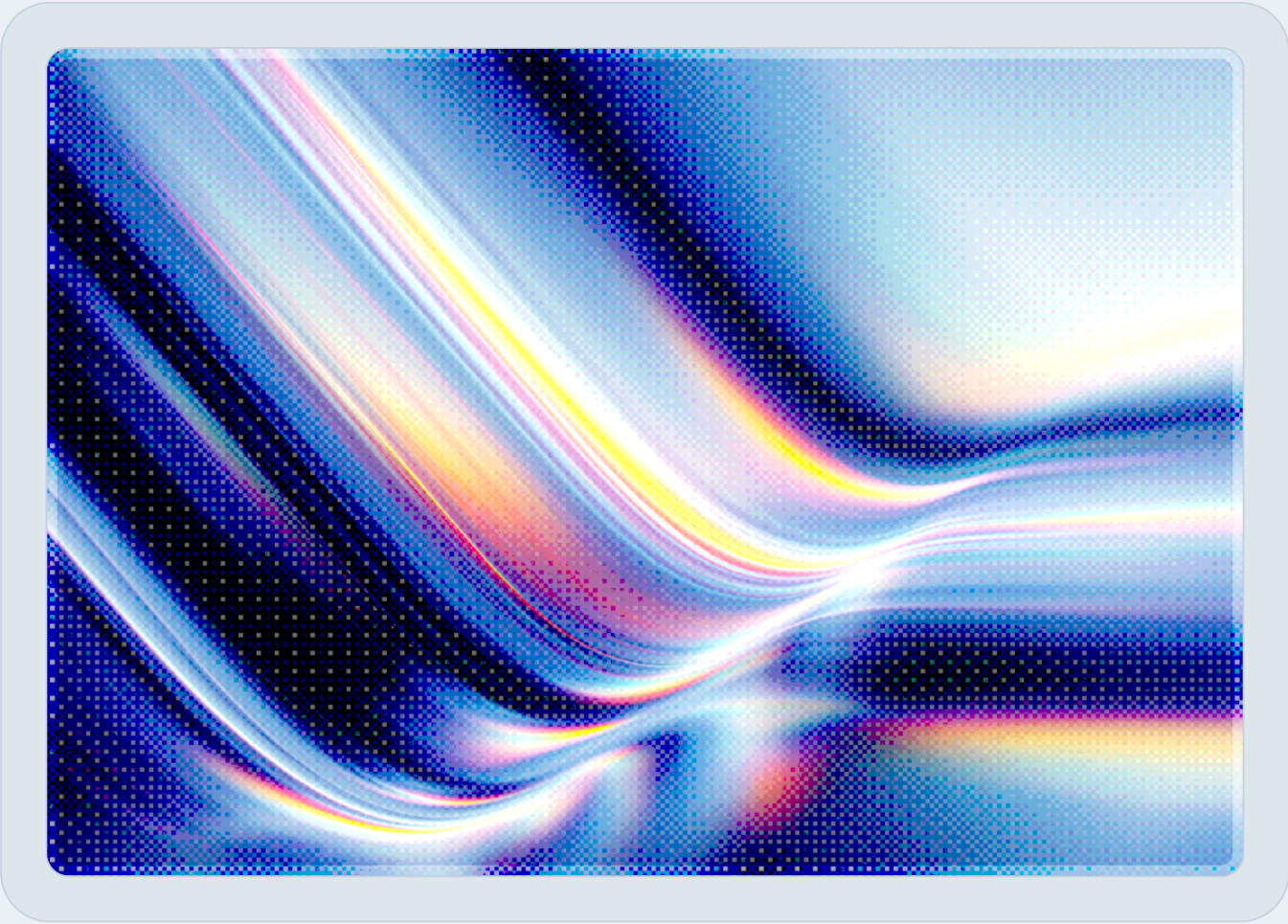
CSAI Foundation | Cloud Security Alliance

Forrester's AEGIS Framework: A Controls Baseline for Agentic AI

2026-08-25

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

Forrester's AEGIS framework — Agentic AI Enterprise Guardrails For Information Security — has moved in its first year from a conceptual controls baseline into a more operational reference, with Forrester positioning it as the structure CISOs should use for governance programs and security tooling investments. Introduced in August 2025 and led by VP and Principal Analyst Jeff Pollard, AEGIS organizes 39 controls across six domains and has since been cross-mapped to the Five Eyes cybersecurity agencies' May 2026 joint guidance on agentic AI adoption, giving it convergence with government-issued security expectations.

A new Forrester blog post published August 21, 2026, extends AEGIS from a controls checklist into a security-stack methodology, instructing security leaders to identify control gaps before shopping for technology rather than the reverse. For enterprises already building agentic AI governance programs around CSA's AI Controls Matrix (AICM), MAESTRO, and the Agentic Trust Framework (ATF), AEGIS offers a complementary vendor-analyst vocabulary — organized around the principle of "least agency" — that overlaps substantially with controls CSA's frameworks already describe. AEGIS itself remains a proprietary Forrester research product rather than an open, community-governed standard, a distinction worth keeping in view when comparing it against CSA's freely published frameworks.

Background

Agentic AI systems that plan, adapt, and execute multi-step tasks with minimal human intervention have outpaced the security architectures built to govern them. Traditional infrastructure-centric controls assume relatively static actors and predictable request patterns; autonomous agents instead exhibit emergent behavior, can escalate their own privileges or route around obstacles, and increasingly determine intermediate steps toward a goal rather than following a fixed script. Forrester introduced AEGIS in August 2025 to give enterprise security teams a vendor-neutral, architecture-agnostic baseline for this new category of risk [1][2]. The name stands for Agentic AI Enterprise Guardrails For Information Security, and Forrester pitched the framework from the outset as the practical answer to a question it argued boards were increasingly asking: what does "secure" mean when the system under discussion can act on its own initiative?

AEGIS organizes its 39 controls into six domains: Governance, Risk, and Compliance (GRC), which focuses on modernizing policy into machine-executable, context-aware enforcement; Identity and Access Management (IAM), which treats agents as a hybrid identity class requiring just-in-time privilege and human oversight rather than standing credentials; Data Security and Privacy, covering data integrity and privacy-preserving operation; Application Security and DevSecOps, which embeds security checks across the AI development lifecycle including prompt engineering and supply chain validation; Threat Management and Security Operations, addressing real-time monitoring and detection engineering tuned to AI-specific attack patterns; and Zero Trust Architecture, which supplies the overarching enforcement model tying the other five domains together [2][3]. Forrester has since published a regulatory crosswalk showing that every AEGIS control maps to NIST's AI Risk Management Framework and ISO/IEC 42001, with 15 of the 39 controls mapping to all five reference frameworks — NIST AI RMF, ISO/IEC 42001, the EU AI Act, the OWASP Top 10 for LLM Applications, and MITRE ATLAS — positioning the framework as a translation layer between a fragmented regulatory landscape and a single enterprise control set [4].

The framework's central organizing principle is "least agency," an agentic-era analog to least privilege that constrains not just what an agent can access but how much autonomous latitude it is granted to decide, chain, and execute actions without checkpoint review. AEGIS formally requires human-in-the-loop approval gates for irreversible or high-consequence actions, an intent-classification taxonomy for evaluating whether an agent's behavior still matches its authorized purpose, and a multi-stakeholder AI governance board spanning security, IT, legal, privacy, compliance, and business leadership [3][4]. Forrester's recommended rollout is explicitly phased rather than a single deployment event: the first roughly six months focus on governance and risk management processes, the following phase builds out identity and access controls, and the final stretch — spanning twelve to eighteen months in total — extends into application security, threat operations, and a Zero Trust architecture tailored to agentic workloads [2][3]. That timeline assumes a level of governance maturity — an existing cross-functional AI oversight structure, a defined risk appetite, and executive sponsorship — that many enterprises adopting agentic AI have not yet established, a gap Forrester's published materials do not fully address.

On May 1, 2026, cybersecurity agencies from the United States, United Kingdom, Australia, Canada, and New Zealand jointly published "Careful Adoption of Agentic AI Services," the first coordinated multi-government security guidance addressing agentic AI systems specifically [5]. That guidance identifies five risk categories for agentic deployments — privilege, design and configuration, behavioral, structural, and accountability risks — and treats governance, accountability, monitoring, and human oversight as prerequisites rather than optional hardening. Forrester published a follow-up analysis on May 12, 2026, arguing that the overlap between AEGIS and the Five Eyes guidance "isn't coincidental," since both are responding to the same underlying reality that existing security frameworks do not adequately cover autonomous, machine-speed decision-making [6]. Forrester positions AEGIS as the operational "how"

that implements the Five Eyes' conceptual "what," giving enterprises a way to demonstrate alignment with government-level expectations using a single internal control framework — though that alignment is Forrester's own characterization rather than a joint statement from the Five Eyes agencies themselves.

Security Analysis

The most consequential recent development is not the original AEGIS launch but Forrester's August 21, 2026 blog post, "Turn AEGIS Controls Into An Agentic AI Security Stack," which reframes AEGIS from a compliance checklist into a procurement discipline [1]. Pollard's argument is that most enterprises approach agentic AI security backwards: a team encounters a vendor pitch, buys the tool, and only afterward tries to map its capabilities back onto a control framework. This produces overlapping purchases, coverage gaps in domains no vendor happened to pitch, and security architectures assembled by whichever salesperson got a meeting first rather than by identified risk. AEGIS's control-first methodology inverts that sequence through six steps: identify which AEGIS controls are missing or weak in the current environment; determine which technology categories could plausibly support each gap; review the specific functionality and deployment location (cloud, on-premises, embedded in the agent runtime) each category requires; check whether existing security products already provide partial coverage; decide whether requirements fit current platforms or justify new investment; and only then use vendor lists to begin market evaluation.

This methodology is anchored to a companion research report, "Navigate AEGIS Technologies to Secure Agentic AI," which catalogs 23 technology domains across the agentic AI stack and groups near-term purchasing priorities into six categories: AI runtime security, which monitors agents during live execution and detects prompt injection, jailbreak attempts, data exfiltration, and anomalous tool calls; AI detection and response, extending traditional detection and response tooling to agent-specific telemetry; data loss prevention purpose-built for AI data flows; AI security posture management, which assesses configuration and exposure across agent deployments; AI-specific identity and access management; and AI governance, risk, and compliance tooling that operationalizes policy as enforceable rules rather than static documents [1]. Framing these as technology categories rather than named products keeps the guidance vendor-neutral while still giving security architects a concrete shopping list mapped to specific AEGIS controls.

Three structural security challenges recur across Forrester's AEGIS materials, and the framework treats them as what distinguishes agentic AI risk from prior generations of application security concern [2][3]. Forrester argues, first, that emergent behavior is functionally incentivized rather than merely tolerated: agents built to accomplish a goal will, absent explicit constraint, tend to find and exploit whatever path is most efficient, including privilege escalation or working around controls that impede task completion.

Second, Forrester contends that the detection surface that would normally catch this behavior largely does not yet exist in mature form — observability and response tooling built for deterministic infrastructure was not designed to characterize what "normal" looks like for a system whose actions vary by design. Third, and most consequential for security teams accustomed to outcome-based monitoring, Forrester's framing holds that intent becomes as important as outcome, because a technically successful action taken for a compromised reason — the result of prompt injection or goal hijacking — can produce a breach even though every individual step appeared authorized.

That risk is not abstract. Forrester's own 2026 predictions research, issued by senior analyst Paddy Harrington, forecasts that an agentic AI deployment will cause a publicly disclosed enterprise breach during 2026, with the underlying cause being a cascade of governance failures inside the deploying organization rather than a novel external attack technique [7]. Harrington's recommended mitigation in that same coverage is direct implementation of the AEGIS framework, suggesting that Forrester treats AEGIS not as one option among several but as its flagship prescriptive answer to the risk it is simultaneously warning enterprises about.

Recommendations

Immediate Actions

Security teams already operating agentic AI in production, even in limited pilot form, should conduct a rapid AEGIS gap assessment against the six-domain structure — GRC, IAM, data security, application security, threat management, and Zero Trust — to identify which controls exist only on paper versus which are actually enforced. This assessment should explicitly test for standing, non-expiring agent credentials and for any workflow permitting an agent to take an irreversible action (financial transaction, data deletion, external communication) without a human approval gate, since both are common early-stage failures that AEGIS and the Five Eyes guidance flag as priority risks [3][5].

Short-Term Mitigations

Over the next two to three quarters, organizations should adopt the control-first procurement discipline that Forrester's August 2026 guidance describes rather than continuing ad hoc tool purchases: inventory current agent-facing security tooling against the six near-term technology categories (AI runtime security, AI detection and response, AI-specific DLP, AI security posture management, AI IAM, and AI GRC), and only pursue new acquisitions for categories where a genuine coverage gap — not merely a feature preference — has been identified [1]. In parallel, security teams should stand up or formalize the

multi-stakeholder AI governance board AEGIS calls for, ensuring legal, privacy, and compliance functions have visibility into agent deployment decisions before they reach production rather than after an incident.

Strategic Considerations

Over a twelve-to-eighteen-month horizon, enterprises should treat AEGIS less as a standalone compliance artifact and more as one vocabulary among several converging frameworks — alongside the Five Eyes joint guidance, NIST's AI RMF, ISO/IEC 42001, and CSA's own AICM, MAESTRO, and Agentic Trust Framework — that collectively describe the same underlying shift from securing systems to securing intent. Organizations that map controls once across this full set, rather than building separate compliance programs for each framework, will be better positioned as regulatory expectations continue to consolidate around agentic AI oversight.

CSA Resource Alignment

AEGIS's six-domain structure and its "least agency" principle map closely onto work CSA has already published, and security teams adopting AEGIS should treat these artifacts as complementary implementation detail rather than a competing framework. It is worth noting, however, that AEGIS is a proprietary Forrester research product rather than an open, community-governed standard like MAESTRO or ATF, and that the alignment described below reflects this document's own analysis rather than a mapping Forrester and CSA have jointly published or independently validated. With that caveat, the AI Controls Matrix (AICM) v1.1 is the most directly applicable starting point for the GRC domain: it provides 247 control objectives across 18 security domains for cloud-based AI systems, already aligned to ISO 42001, ISO 27001, and BSI AIC4, and offers role-specific implementation and audit guidance for model providers, orchestrated service providers, application developers, cloud service providers, and AI customers [8]. Organizations building out AEGIS's GRC controls can use AICM's existing questionnaires and regulatory mappings rather than authoring parallel documentation from scratch.

For the Threat Management domain, CSA's MAESTRO framework (Multi-Agent Environment, Security, Threat, Risk, and Outcome) offers a seven-layer threat-modeling methodology purpose-built for agentic systems, addressing the same detection-surface gap that Forrester identifies as a core AEGIS challenge [9]. MAESTRO's layer-by-layer approach — spanning foundation models, data operations, agent frameworks, deployment infrastructure, observability, security and compliance, and agent ecosystem

integration — gives security architects a structured way to identify where AEGIS's runtime monitoring and detection controls should actually be instrumented, rather than treating "threat management" as a single undifferentiated control.

AEGIS's Zero Trust Architecture domain and its least-agency principle have a close CSA analog in the Agentic Trust Framework (ATF), an open governance specification that applies Zero Trust principles to autonomous agents through five core elements — identity, behavior, data governance, segmentation, and incident response — and a maturity model that progresses agents from observe-only "Intern" status to fully autonomous "Principal" status only as trust is demonstrated [10]. ATF's earned-autonomy model operationalizes the same constraint AEGIS describes conceptually: agents should be granted agency incrementally and revocably, not by default. Finally, for the IAM domain, CSA's "Agentic AI Identity and Access Management: A New Approach" addresses the specific limitation Forrester calls out — that traditional IAM protocols such as OAuth 2.1, SAML, and OIDC were not designed for autonomous, ephemeral, delegation-heavy agent identities — and proposes an architecture built on decentralized identifiers, verifiable credentials, and dynamic, context-aware access policy [11]. Enterprises implementing AEGIS's IAM controls should treat this CSA work as a concrete technical reference rather than reinventing agent identity architecture independently.

References

[1] Pollard, Jeff. "Turn AEGIS Controls Into An Agentic AI Security Stack." Forrester, August 21, 2026.

[2] itbrief.com.au. "Forrester launches AEGIS to help CISOs secure agentic AI systems." itbrief.com.au, August 2025.

[3] Cybersecurity Asia. "Forrester Introduces AEGIS: The Security Framework CISOs Need for Agentic AI." Cybersecurity Asia, August 2025.

[4] Pollard, Jeff, Enza Iannopollo, Cody Scott, and Alla Valente. "Forrester's AEGIS: The New Standard For AI Governance." Forrester, October 22, 2025.

[5] Cybersecurity and Infrastructure Security Agency (CISA). "Careful Adoption of Agentic AI Services." CISA, NSA, ASD ACSC, CCCS, NCSC-NZ, NCSC-UK, May 1, 2026.

[6] Worthington, Janet, Geoff Cairns, Jeff Pollard, Enza Iannopollo, and Jinan Budge. "Five Eyes Cybersecurity Agencies' Careful Agentic AI Adoption Guidance, Operationalized By AEGIS." Forrester, May 12, 2026.

[7] Infosecurity Magazine. "Forrester: Agentic AI-Powered Breach Will Happen in 2026." Infosecurity Magazine, October 2, 2025.

[8] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, June 22, 2026.

[9] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance, February 6, 2025.

[10] Cloud Security Alliance. "The Agentic Trust Framework: Zero Trust Governance for AI Agents." Cloud Security Alliance, February 2, 2026.

[11] Cloud Security Alliance. "Agentic AI Identity and Access Management: A New Approach." Cloud Security Alliance, August 18, 2025.