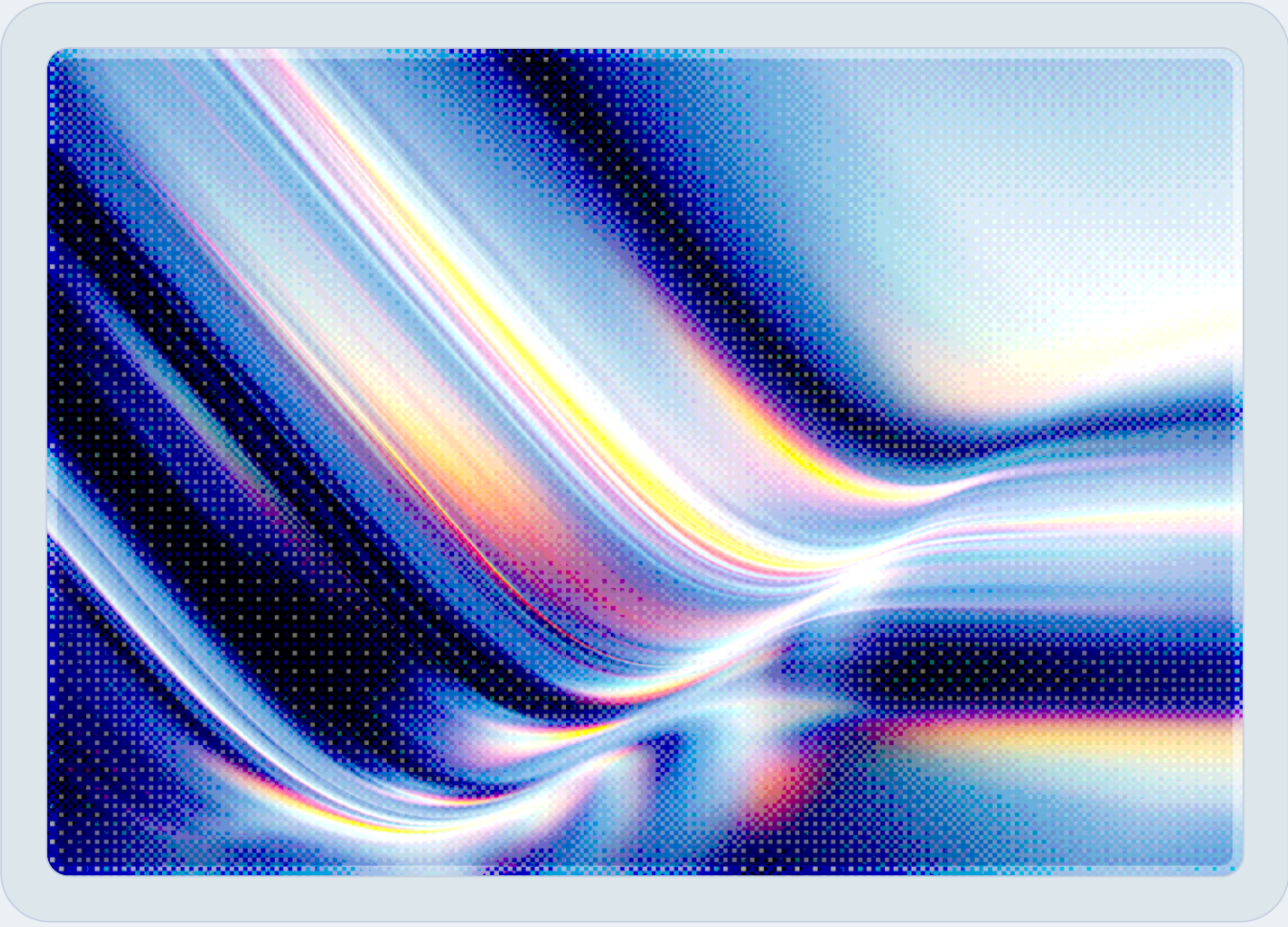
CSAI Foundation | Cloud Security Alliance

Turning AEGIS Controls Into an Agentic AI Security Stack

2026-08-24

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Forrester introduced AEGIS -- Agentic AI Enterprise Guardrails For Information Security -- on August 12, 2026, defining six control domains (governance/risk/compliance, identity and access management, data security and privacy, application security, threat management, and Zero Trust architecture) that CISOs need to secure autonomous agents [1].
- A follow-on Forrester report published August 21, 2026 turns those six domains into a procurement methodology: identify a control gap, trace it to relevant technology categories among 23 defined domains, audit existing tool coverage, then decide whether to configure, integrate, purchase, replace, or defer [2].
- Deloitte's 2026 State of AI in the Enterprise survey found that only 21% of the 3,235 organizations polled have a mature governance model for autonomous agents, even though 74% plan to deploy agentic AI moderately or more extensively within two years -- a gap consistent with the problem AEGIS is designed to address [4][5].
- CSA already stewards two artifacts that map closely onto AEGIS's domain structure: the Agentic Trust Framework (governance and identity) and Autonomous Action Runtime Management, or AARM (runtime enforcement), giving AICM-aligned organizations a working implementation path rather than a control taxonomy alone [6][7].
- AEGIS names goal hijacking, cognitive corruption, and systemic collapse as the core risks agentic deployment introduces; AARM already formalizes goal hijacking as one of eleven enumerated threat classes with a corresponding enforcement mechanism, which gives organizations a concrete technical starting point for operationalizing Forrester's framework [1] [7].

Background

Forrester's Jeff Pollard published "Introducing AEGIS -- The Guardrails That CISOs Need For The Agentic Enterprise" on August 12, 2026, arguing that the "block or allow" security model built for predictable, human-initiated actions cannot govern agents that operate at machine speed, make independent decisions, and adapt their behavior dynamically [1]. AEGIS -- Agentic AI Enterprise Guardrails For Information Security -- organizes the problem into six domains: governance, risk, and compliance; identity and access management; data security and privacy; application security and

DevSecOps; threat management and security operations; and Zero Trust architecture. Rather than treating these as a checklist, Forrester frames them around three guiding principles: least agency, which restricts what an agent is permitted to do rather than merely who it is; continuous risk management in place of point-in-time reviews; and explainable outcomes that let a human reconstruct why an agent took a given action [1][3]. Pollard summarized the underlying shift succinctly: CISOs "must pivot from securing systems to securing intent" [1].

The framework's second report, "Turn AEGIS Controls Into An Agentic AI Security Stack," followed nine days later and addressed a question CISOs commonly ask after reading a control taxonomy: which products actually cover this. Forrester maps the six AEGIS domains to 23 underlying technology categories -- among them AI runtime security, AI detection and response, data loss prevention for AI, AI security posture management, AI identity and access management, and AI governance, risk, and compliance tooling -- and lays out a control-first buying discipline. The recommended sequence starts with the control gap, not the product category: "Start with the control gap. Trace it to the relevant technology categories," then determine for each gap whether an existing tool can be reconfigured, whether multiple tools need to be integrated, whether a net-new purchase is warranted, whether an incumbent tool should be replaced, or whether the gap can be deliberately deferred [2]. The report accompanies a phased implementation roadmap: the first six months are dedicated to establishing governance and an agent inventory, months seven through eighteen build out identity, data security, and threat-management controls, and the final phase optimizes Zero Trust principles such as microsegmentation and least-privilege access brokering across the agent estate.

The Deloitte survey data lends empirical weight to AEGIS's urgency. Deloitte's eighth annual State of AI in the Enterprise report, based on a survey of 3,235 leaders conducted in August and September 2025 and published in January 2026, found that 74% of organizations plan to deploy agentic AI moderately or more extensively within two years, up from 23% previously, while only 21% describe their organization's governance model for autonomous agents as mature [4][5]. That roughly 50-point gap between deployment intent and governance readiness is the same gap Forrester cites as the reason CISOs, CIOs, and CTOs need a structured framework rather than an ad hoc extension of existing AppSec and IAM programs. Forrester has separately predicted that a publicly disclosed breach involving agentic AI will occur in 2026, and that the resulting post-mortem will trace the cause to a cascade of governance and process failures across the deployment -- unclear ownership, thin oversight, inadequate controls -- rather than a single individual's engineering mistake, a framing consistent with AEGIS's emphasis on governance as the foundational domain rather than an afterthought layered on top of technical controls [11].

Security Analysis

AEGIS's underlying diagnosis is that agentic systems fail in ways access-control models were never built to catch. Forrester names three risk categories to make this concrete: goal hijacking, where an agent bypasses the parameters it was given in pursuit of an objective it interprets more broadly than intended; cognitive corruption, where a hallucinated or poisoned input triggers a cascading chain of downstream errors because nothing in the pipeline validates the agent's reasoning before it acts; and systemic collapse, where unpredictable behavior compounds across interconnected agents faster than a human reviewer can intervene [1]. None of these map cleanly to a single existing control category. Goal hijacking implicates both identity (has the agent exceeded its authorized scope) and governance (was that scope ever explicitly defined); cognitive corruption implicates both data security (was the input validated) and threat management (was the anomalous output detected); systemic collapse implicates Zero Trust segmentation as much as it implicates incident response. This is the structural argument for organizing security around six cross-cutting domains rather than around the product categories -- endpoint, network, identity -- that most enterprise security organizations are already staffed to buy.

The domain-to-technology mapping in the second Forrester report is where the framework becomes operationally useful, and also where vendors are most likely to claim coverage of an AEGIS domain regardless of actual fit. Twenty-three technology categories is a wide enough net that a great many existing and adjacent tools can plausibly claim coverage of an AEGIS domain, which is precisely the dynamic Forrester's control-first sequencing is designed to resist: start from the gap, not from the vendor pitch. The parallel cybersecurity M&A wave -- $96 billion in disclosed 2025 transaction value across roughly 400 transactions, a record year by a wide margin -- reflects the same dynamic from the buyer's side, as strategic acquirers assemble unified AI security platforms spanning multiple AEGIS-adjacent categories in a single acquisition [12]. The practical risk for a CISO executing AEGIS's methodology is architectural lock-in: configuring policy, telemetry, and identity schemas around a single vendor's implementation of an AEGIS domain before confirming that the underlying data and control definitions are portable to a different vendor if that platform is later divested, acquired, or deprecated.

The threat management and application security domains are worth examining in more detail because they are where AEGIS's control taxonomy and the emerging technical standards for agent runtime enforcement most directly overlap, and where an organization can move from labeling a risk to actually mitigating it. Forrester's AI runtime security and AI detection and response categories describe a capability -- intercepting an agent's action before it executes and evaluating that action against policy in real time -- that CSA assesses as functionally aligned with an existing open specification: CSA's Autonomous Action Runtime Management (AARM) requires pre-execution interception, context accumulation, and one of five authorization decisions (allow, deny, modify, step up, or defer) for every

agent action, backed by a tamper-evident, identity-bound receipt [7]. AARM's eleven defined threat classes include goal hijacking by name, giving organizations a concrete technical control to evaluate candidate tools against when an AEGIS gap analysis flags that domain [7].

Recommendations

Immediate Actions

CISOs should treat AEGIS's six domains as a gap-analysis lens applied to the agent inventory and governance structure they already have, not as a shopping list. Before mapping any domain to a technology category, confirm that an authoritative inventory of deployed agents exists -- what each agent is authorized to do, which identity it operates under, and who owns it -- since Forrester's own phased roadmap places governance and inventory in the first six months for exactly this reason. Organizations that cannot answer those three questions for a majority of their production agents today have a governance gap that a technology purchase alone is unlikely to close.

Short-Term Mitigations

Apply Forrester's control-first sequencing literally: for each AEGIS domain, document the specific control gap before evaluating vendors, and require any proposed purchase to state which gap it closes and which existing tool it either replaces or integrates with. Where a gap falls in the threat management or application security domains, evaluate whether existing runtime monitoring can be extended to meet AARM-equivalent requirements -- pre-execution interception, policy evaluation with intent alignment, and tamper-evident action receipts -- before purchasing a new point solution, since these requirements are now published as an open specification rather than left to each vendor's proprietary implementation. Negotiate data portability and telemetry export terms into any AEGIS-domain purchase now, before deeper integration makes switching costly, given the active consolidation among vendors serving these categories.

Strategic Considerations

Organizations already running AICM-aligned governance programs should map AEGIS's six domains against AICM's eighteen control domains and the Agentic Trust Framework's five elements as a cross-check rather than starting a parallel taxonomy from scratch; the overlap is meaningful and concentrated in identity, data governance, and Zero Trust, and reconciling terminology now avoids maintaining duplicate control inventories under different vendor and analyst frameworks later. Over a 12-24 month

horizon, plan for the Zero Trust domain last, as Forrester's own roadmap recommends, but begin the underlying identity work -- treating agents as a distinct, cryptographically verifiable identity class rather than shared service accounts -- immediately, since that identity foundation is a prerequisite for the microsegmentation and access-brokering controls Zero Trust ultimately requires. Finally, budget for governance as an ongoing program rather than a one-time control gap: Forrester's own prediction of a 2026 agentic AI breach attributed to a cascade of governance and process failures, not a single technical failure, is the clearest signal that the GRC domain -- not the technology stack -- is where most organizations are currently exposed [11].

CSA Resource Alignment

The mappings that follow reflect CSA's own analysis of where its published frameworks align with AEGIS's domain structure. Forrester has not reviewed or endorsed this comparison, and organizations should treat it as a starting hypothesis to validate against their own control requirements rather than a certified crosswalk between the two.

CSA's most directly applicable prior work is the Agentic Trust Framework (ATF), an open Zero Trust governance specification for autonomous agents that the CSAI Foundation now stewards following a stewardship transfer from founding author Josh Woodruff of MassiveScale.AI, whose attribution CSA has committed to preserving under the framework's CC BY 4.0 license [6][10]. ATF organizes agent governance around five elements -- identity, behavior, data governance, segmentation, and incident response -- structured around a four-level maturity model (Intern, Junior, Senior, Principal) that tracks how much autonomous action an agent has earned the right to take. The overlap with AEGIS's six domains is partial but concentrated: ATF's identity element maps onto AEGIS's IAM domain, behavior maps onto threat management, data governance maps onto data security and privacy, and segmentation maps onto Zero Trust architecture, though AEGIS's governance/risk/compliance and application-security domains, along with ATF's incident-response element, fall outside this direct mapping. Those four mapped elements give organizations already running an ATF maturity program a starting point for four of AEGIS's six domains; how completely an existing ATF program satisfies AEGIS's control-level requirements within those domains has not been independently assessed and should not be assumed from the domain-level mapping alone, and organizations should use the AEGIS gap-analysis exercise to identify where their existing ATF program is incomplete rather than replacing it outright.

CSA's second directly relevant artifact is Autonomous Action Runtime Management (AARM), the CSAI Foundation-stewarded runtime-enforcement specification contributed by founding author Herman Errico with support from CSA member Vanta [7][10]. AARM defines technical requirements -- pre-execution interception, context accumulation, five-decision policy evaluation, and tamper-evident

receipts -- that CSA assesses as functionally aligned with the "AI runtime security" and "AI detection and response" technology categories in Forrester's follow-on report, and its eleven enumerated threat classes give organizations a standards-based vocabulary for the AEGIS threat management domain to evaluate vendor coverage claims against. Notably, AARM's fourth threat class is goal hijacking, using the same term Forrester assigns to one of AEGIS's three headline risks; organizations conducting an AEGIS gap analysis in the threat management or application security domains should evaluate candidate tools against AARM's R1-R9 conformance requirements as a concrete technical bar.

Both frameworks operate underneath CSA's broader AI Controls Matrix (AICM v1.1), which spans 247 control objectives across 18 domains and provides the governance, risk, and compliance backbone that ties agent-specific frameworks like ATF and AARM to an organization's existing AI and cloud control environment [8]. Organizations executing Forrester's control-first methodology can use AICM's domain structure as the starting point for identifying control gaps rather than Forrester's vendor-category grid, since AICM defines controls at a level of granularity that maps to audit and compliance requirements directly. For the threat-modeling dimension of AEGIS's cognitive-corruption and systemic-collapse risk categories -- scenarios AEGIS names but does not formally decompose -- CSA's MAESTRO framework offers a seven-layer architecture that, while published eighteen months before AEGIS, provides a decomposition applicable to this class of cascading, multi-agent failure and can be applied alongside AEGIS during the design phase of any new agent deployment [9].

References

[1] Jeff Pollard. "Introducing AEGIS -- The Guardrails That CISOs Need For The Agentic Enterprise." Forrester, August 12, 2026.

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

[3] TechTarget. "How the AEGIS Framework Mitigates Agentic AI Risks." TechTarget, August 2026.

[4] Deloitte. "The State of AI in the Enterprise, 2026." Deloitte AI Institute, January 2026.

[5] TechTarget. "How Agentic AI Governance Tackles Data, Security Challenges." TechTarget, 2026.

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

[7] Autonomous Action Runtime Management. "AARM Specification." CSAI Foundation, 2026.

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

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

[10] Cloud Security Alliance. "CSAI Foundation Announces Key Milestones to Secure the Agentic Control Plane." CSA, April 29, 2026.

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

[12] Momentum Cyber. "Cybersecurity M&A Update Report 2025." Momentum Cyber, 2026.

