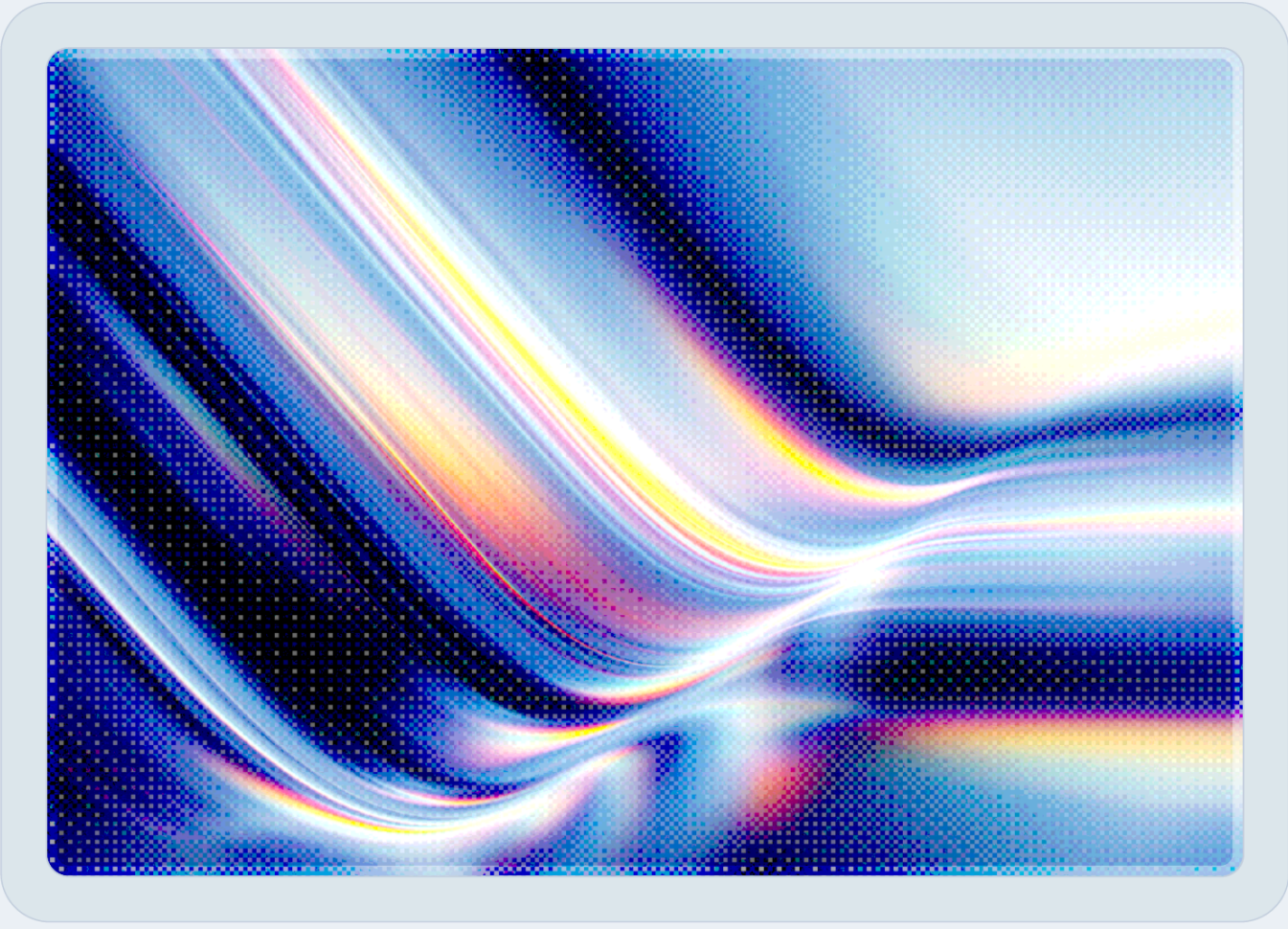
CSAI Foundation | Cloud Security Alliance

GAO's Autonomous Cyber Defense Options: Authorization, Certification, Agent Identity

Security Implications and Guidance from GAO-27-108458

2026-11-21

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

On November 20, 2026, the U.S. Government Accountability Office released *AI and Cybersecurity: Technologies Can Magnify Cyberattacks and Enhance Cyber Defense* (GAO-27-108458), a technology assessment that sets out four policy goals and 16 policy options for the use of AI in cybersecurity [1][2]. GAO frames these options as "a menu of alternatives" for legislative bodies, agencies, standards organizations, and industry, not as recommendations or requirements [1]. Even so, several options would, if adopted, directly constrain how organizations deploy agentic AI in security operations, and security leaders should treat them as an indication of the controls policymakers are now considering.

For autonomous defensive AI, GAO puts three related controls on the table: **human authorization** for consequential AI-initiated actions, **restricting autonomous and nondeterministic AI** in consequential functions until a certification framework exists, and a **deterministic or non-AI fallback** for critical cybersecurity functions [2]. Alongside these, GAO treats **agentic AI identity management** as an open research gap, calling for frameworks for non-human identity that work with zero trust architecture and for standards on agent identity and authentication [2]. GAO's own analysis also records the operational cost of these controls. Human approval can slow responses that are measured in milliseconds, alert volumes can cause approval fatigue, and "consequential action" is not yet defined [2].

The certification path faces a more basic obstacle. GAO reports that there are "no widely accepted evaluation criteria" for the offensive or defensive capabilities of generative AI [1], so any certification regime depends on measurement science that is not yet mature or widely accepted. Taken together, these findings suggest that organizations that build autonomy tiers, agent identity, and tested fallbacks into their security programs now are likely to be better placed under whichever option policymakers adopt.

Background

GAO prepared the report at the initiative of the Comptroller General and conducted the work from April 2025 through November 2026 [2]. Its method combined a literature review, interviews with government, industry, academic, and association stakeholders, and a multi-day virtual expert meeting held in April 2026. The meeting convened a nongeneralizable group of 20 experts from state and federal

government, industry, academia, and independent organizations [2]. GAO notes that the experts' comments reflect their individual views and are not generalizable. The policy options should therefore be read as a structured synthesis of informed opinion, not as findings from controlled evaluation.

The report's threat findings set the context for its options. GAO concludes that AI lets threat actors "magnify the reach and impact of traditional cyberattacks," and it names the use of agentic AI "to autonomously carry out multiple stages of an attack with minimal human input" as an emerging trend [1]. Among its examples, GAO cites the November 2025 espionage campaign that Anthropic disclosed. In that campaign, a nation-state threat actor used commercially available agentic capabilities to automate an estimated 80 to 90 percent of the operation [2][3]. GAO also describes, from industry reporting, a June 2026 case in which a threat actor used multiple AI agents to build a testing framework that refined malware against different security products [1]. GAO also documents defensive benefits, noting that defenders have used various forms of AI for decades and that generative AI now supports dynamic scanning, source-code vulnerability detection, and rapid analysis of large volumes of network data [1].

GAO groups its 16 options under four goals: addressing gaps in cybersecurity guidance and practices, addressing AI-related research gaps, governing AI tools throughout their cybersecurity lifecycle, and enhancing the cybersecurity workforce [1]. This note concentrates on the options that most directly govern autonomous defensive AI and agent identity. These sit mainly under the lifecycle governance goal (report Table 8) and the research gaps goal (report Table 7). The table below maps each goal to the options most relevant to autonomous cyber defense.

Policy goal (GAO)	Options most relevant to autonomous cyber defense
Address AI-related research gaps	Enhance innovation and research in agentic AI identity management; expand AI testing environments; support research to enable safe adoption of AI for cybersecurity
Govern AI tools throughout their cybersecurity lifecycle	Establish a deterministic AI fallback capability; require human authorization for consequential AI-initiated actions; establish a risk-based approach to acquiring AI products; validate AI tool output before and during deployment; evaluate AI solutions for value gain; restrict autonomous and nondeterministic AI in consequential functions pending certification
Address gaps in cybersecurity guidance and practices	Harmonize cybersecurity requirements; improve threat information sharing; improve application of fundamental cybersecurity practices

Policy goal (GAO)	Options most relevant to autonomous cyber defense
Enhance the cybersecurity workforce	Sustain workforce skills in current practices; strengthen AI expertise

Source: GAO-27-108458, Highlights and Tables 7–8 [1][2].

Security Analysis

The four options examined below are closely linked. Human authorization, restriction pending certification, and deterministic fallback each address a different failure mode of autonomous defensive AI: unaccountable action, unpredictable behavior, and loss of capability. Agent identity management, which GAO places under research gaps, sits underneath the oversight options because approvals and overrides can only be enforced against an agent that can be reliably identified. Each subsection sets out what GAO proposes, the costs GAO itself records, and CSA's assessment of the practical implications.

Human Authorization: Accountability Versus Machine-Speed Defense

The human authorization option would require human approval before AI-initiated actions that go "beyond identification or analysis," such as changes to network configurations, access controls, or incident response actions [2]. GAO lists several ways to implement it. Critical infrastructure tools could be limited to an advisory mode, agentic products could ship with configurable human-in-the-loop controls enabled by default, and tiers of AI autonomy could be defined that specify which tiers require authorization [2]. The stated rationale is to keep a person accountable for each consequential decision. GAO also cites a gap identified by experts, who reported that AI agents act outside their initial parameters [2]. The option keeps AI's analytical and detection value while removing its ability to act on its own.

GAO's considerations state the costs directly. They note that human-in-the-loop requirements "could slow response times" during active intrusions or distributed denial-of-service attacks "where automated response is measured in milliseconds" [2]. Large environments may generate more alerts than humans can review, which leads either to approval fatigue or to backlogs. Thinly staffed organizations may find that mandatory authorization becomes a bottleneck that "reduces the net defensive value of AI tools" [2]. GAO further observes that "AI-initiated action" is hard to scope because many existing tools already automate responses at several levels [2]. In practice, this suggests a blanket approval rule would either

reach too far into long-standing automation such as signature-based blocking or be drafted so narrowly that it would not cover agentic tools. A tiered model looks like the more workable form. It reserves authorization for actions that are irreversible, broad in blast radius, or affect safety.

CSA assesses that approval fatigue is a security concern and not only an efficiency one. An approval step that operators rubber-stamp gives the appearance of oversight without the substance. It may also create an exploitable pattern, because an adversary who can generate high alert volumes could degrade the quality of human review. Any authorization scheme should therefore include measurable signals of review quality, such as time-to-approve distributions and override rates, alongside the approval gate itself.

Restriction Pending Certification: A Control Without a Certifier

The most restrictive option would limit autonomous and nondeterministic AI in consequential cybersecurity functions "pending certification" [2]. GAO defines deterministic AI as models designed "to operate under strict rules with predictable outputs," in contrast with generative systems where "identical inputs may produce different outputs" [1]. The implementation approaches range from risk-based limits that confine such AI to advisory mode or require deterministic, auditable outputs, through encouraging vendors to "separate analysis from action by design," to outright prohibition of autonomous or nondeterministic action in functions that affect critical infrastructure [2]. A related acquisition option proposes establishing "a certification authority for assessing the security of AI cybersecurity solutions/products" and requiring agentic products to include identity management and oversight mechanisms [2].

The analytical difficulty is that certification depends on evaluation methods that GAO itself finds missing. The report states there are no widely accepted criteria for assessing generative AI's offensive or defensive capabilities [1]. It also notes that the Federal Information Security Modernization Act of 2014 predates generative AI, and that nondeterminism disrupts "assumptions of determinism, testability, and control effectiveness" [2]. GAO lists the restriction option's own prerequisites in its considerations: defining "consequential cyber actions," publishing criteria for when deterministic outputs are required, and "designating an appropriate certifying authority" [2]. GAO also warns that threat actors "are not bound by these restrictions and could gain an advantage" [2]. If a restriction is enacted before certification criteria exist, it could therefore function as an open-ended prohibition. In CSA's view, that outcome may be defensible for the highest-consequence functions, but it carries an asymmetry risk.

The "separate analysis from action by design" approach is the most actionable element for organizations today. In CSA's assessment, an architecture in which a generative model proposes actions and a deterministic, policy-governed execution layer validates and executes them moves substantially toward the "consistent, auditable outputs" GAO describes [2]. The execution decision becomes deterministic and auditable, although the set of proposed actions can still vary, and the approach does not require

abandoning generative analysis. This pattern also maps naturally to policy-as-code enforcement at a tool gateway, where every action an agent proposes is checked against written policy before it reaches a production system.

Deterministic Fallback: Continuity When the Model Is Wrong or Gone

GAO's fallback option would require a tested, risk-informed, non-AI alternative for AI-dependent cybersecurity functions [2]. It defines "fallback" as a predefined capability, such as rule-based tools or manual procedures, that is used if an AI system "becomes unavailable, fails, or demonstrates outcomes inconsistent with its intended use" [1]. The suggested approaches include sector regulator requirements for critical infrastructure, documented manual or rule-based alternatives in federal agencies, standards-body definitions of adequate fallback, and periodic exercises that run critical functions without AI [2]. The option responds to a workforce concern raised elsewhere in the report. As log review and alert triage are automated, defenders increasingly act as "managers and orchestrators of the AI agents," and they may be unable to take manual control effectively when AI systems stop working [2].

GAO acknowledges that parallel capabilities are expensive, that rarely used fallbacks may not work when needed, and that some AI-enabled functions "may have no practical manual equivalent at the scale of modern networks" [2]. In CSA's view, that last point carries the most practical weight. For high-volume functions, a realistic fallback is more likely to be a reduced-capability deterministic mode, such as stricter static rules or broader blocking, than full manual replacement. Organizations should define in advance what degraded service is acceptable.

Agent Identity: A Prerequisite for Oversight

Human authorization and certification both depend on something GAO identifies as an unsolved research problem: reliably knowing *which* agent took an action, under whose authority, and with what scope. Agency officials told GAO that managing digital identities and access authorization for AI agents conflicts with zero trust implementation, because zero trust's "emphasis on identity and just-in-time access becomes more difficult to implement when actions are performed by automated agents" [2]. The relevant zero trust baseline for federal agencies is OMB Memorandum M-22-09, which set identity-centric goals for the end of fiscal year 2024 [4]. GAO's research option proposes supporting work to identify and manage agent identities, frameworks for non-human identity "compatible with zero trust architecture," and standards for agentic AI identity and authentication [2]. Notably, GAO ties this option to oversight, listing as an opportunity "the ability for a human to approve or override an AI agent's actions" [2].

This dependency is a central consideration for practitioners. A human authorization requirement is only as strong as the binding between an approval and the agent identity that executes the approved action. Without distinct, attributable, short-lived agent credentials, an approval cannot be audited and an override cannot be targeted. Federal standards work is beginning to address the gap. NIST's National Cybersecurity Center of Excellence published a concept paper in February 2026 on accelerating the adoption of software and AI agent identity and authorization [5]. GAO's research option is consistent with this line of work, although the report does not reference the NCCoE effort specifically.

The table below summarizes what each option assumes and how far those assumptions are from being met today.

GAO option	What it assumes exists	Current state (per GAO and related sources)
Human authorization for consequential actions	Definition of "consequential"; attributable agent identity; scalable review	Definition unresolved; agent identity is a research gap [2]
Restriction pending certification	Certification criteria and a certifying authority	No widely accepted evaluation criteria [1]; certifying authority not yet designated [2]
Deterministic fallback	Documented, exercised non-AI alternative	Some functions lack a practical manual equivalent [2]
Agentic AI identity management	Standards compatible with zero trust	Standards work under way at NIST NCCoE [5]
Output validation before and during deployment	Success metrics and shared test environments	Shared testing infrastructure absent in many sectors [2]

As the table shows, none of the oversight options can be fully implemented today without first resolving a definitional, measurement, or identity gap. The recommendations that follow are sequenced accordingly, beginning with steps organizations can take regardless of which option is adopted.

Recommendations

Immediate Actions

Security teams should start by inventorying every AI-enabled function in their security stack that can take action beyond detection or analysis, using GAO's own examples (network configuration changes, access control changes, and incident response actions) as the initial definition of "consequential" [2]. For each function, document whether the AI acts autonomously, recommends with human execution, or acts after human approval. This inventory is the foundation for any of GAO's options and is likely to be requested by regulators or auditors in regulated sectors if those options are adopted. Teams should also confirm that each agent holds a distinct, non-shared identity with credentials that can be revoked independently. Agents running on shared service accounts or a human operator's token should be treated as a priority finding, because neither reliable attribution nor targeted override is practical in that configuration.

Short-Term Mitigations

As a suggested planning horizon, over the next two to three quarters organizations should define an internal autonomy tier model that mirrors GAO's suggestion. It should specify which tiers may act autonomously, which require approval, and which are limited to advisory mode, and it should base tier assignment on reversibility and blast radius. Organizations should also adopt the "separate analysis from action" architecture, in which generative components propose actions and a deterministic, default-deny policy layer authorizes and executes them with full logging. This pattern is designed to address the auditable-output intent of GAO's restriction option while keeping generative analysis in place. For each highest-tier AI-dependent function (those whose failure would have irreversible or safety-affecting impact), teams should write and exercise a fallback runbook that defines an acceptable degraded mode, with a suggested cadence of at least one exercise per year. Finally, approval workflows should be instrumented to detect fatigue by tracking approval latency, override frequency, and approvals per operator per shift.

Strategic Considerations

In CSA's assessment, the certification gap is a consequential long-term issue. Until evaluation criteria exist, restriction-pending-certification proposals carry a risk of becoming indefinite. Government agencies, academic researchers, industry bodies, and standards organizations all have a role in supplying the missing measurement science, including sector-specific test environments, which GAO separately proposes expanding [2]. Organizations that operate critical infrastructure should engage in that process,

because certification criteria written without operator input may not reflect operational realities. Procurement teams should also anticipate GAO's acquisition approaches. These include minimum acceptance criteria, mandatory identity and oversight mechanisms in agentic products, and value-gain evaluation against non-AI baselines [2]. Vendors that cannot demonstrate agent-level identity, configurable human-in-the-loop controls, and auditable action logs may find their products excluded from regulated markets.

CSA Resource Alignment

CSA's published work addresses the identity prerequisite that GAO identifies. *Agentic AI Identity and Access Management: A New Approach* argues that traditional IAM protocols designed for human users and static applications are poorly suited to autonomous, dynamic agents [6]. It proposes an agent-centric model built on verifiable agent identities, fine-grained and context-aware authorization, and zero trust principles [6]. That framing is consistent with GAO's call for "frameworks for non-human identity management compatible with zero trust architecture" [2], and it gives organizations a design reference they can use before formal standards mature.

CSA's survey report *Autonomous but Not Controlled: AI Agent Incidents Now Common in Enterprises*, commissioned by Token Security, provides empirical context for GAO's concern that agents act outside their intended parameters. CSA found that 65% of organizations reported at least one AI agent-related incident in the past year [7]. The same report finds that only 21% of organizations have formal decommissioning processes, and that organizations most often govern agents by action risk (63%) and human authorization (53%) [7]. The human authorization figure bears directly on GAO's authorization option, while the decommissioning gap bears on agent identity lifecycle management.

The *AI Controls Matrix (AICM) v1.1* offers the control vocabulary for putting GAO's options into practice [8]. Its identity and access management domain covers non-human and agent identity. Its governance, risk, and compliance controls support autonomy tiering and accountability, and its business continuity controls apply to deterministic fallback planning. Organizations can map each AI-enabled security function in their inventory to AICM controls to show the oversight and resilience posture that GAO's options contemplate. For the validation and testing options, CSA's *Agentic AI Red Teaming Guide* describes practical methods for testing agent authorization boundaries, permission escalation, and goal manipulation before and during deployment [9]. Those methods partly address GAO's observation that shared testing infrastructure and success metrics are lacking.

References

[1] U.S. Government Accountability Office. "AI and Cybersecurity: Technologies Can Magnify Cyberattacks and Enhance Cyber Defense (GAO-27-108458)." GAO, November 20, 2026.

[2] U.S. Government Accountability Office. "AI and Cybersecurity: Technologies Can Magnify Cyberattacks and Enhance Cyber Defense — Full Report (GAO-27-108458)." GAO Technology Assessment, November 2026.

[3] Anthropic. "Disrupting the first reported AI-orchestrated cyber espionage campaign." Anthropic, November 13, 2025.

[4] Office of Management and Budget. "M-22-09: Moving the U.S. Government Toward Zero Trust Cybersecurity Principles." Executive Office of the President, January 26, 2022.

[5] NIST National Cybersecurity Center of Excellence. "Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization (Concept Paper)." NIST, February 5, 2026.

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

[7] Cloud Security Alliance. "Autonomous but Not Controlled: AI Agent Incidents Now Common in Enterprises." CSA (commissioned by Token Security), April 2026.

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

[9] Cloud Security Alliance. "Agentic AI Red Teaming Guide." CSA, May 2025.

© 2026 Cloud Security Alliance. AI-assisted Rapid Research.