

The AI-Enabled vCISO: What Gets Delegated, What Gets Verified and What Stays Human, Heading into 2027

Ricardo Martínez

Cloud Security & Applied AI — NullGhost Research

October 2026

Abstract

The fractional CISO (vCISO) is an established model, and the offer is starting to be marketed as an «AI vCISO», sometimes in the language of autonomy. This text separates what public evidence shows from what only vendors claim: demand for security skills is measurable, automation of delivery is real but mostly self-reported, and we found no independent measurement of how well an agent produces compliance evidence. From that we propose a four-level delegation ladder defined by three tests (reversibility, accountability and verifiability), six design principles and a measurement protocol any provider can run on its own cases. The conclusion is deliberately uncomfortable: what the agent makes cheaper is preparation, not accountability, and the new risk is false assurance, a polished report that asserts compliance with no evidence behind it.

Process note. Search and drafting were AI-assisted. Figures come from the cited sources, not from the model. Each source carries a tag: [P] primary, read directly; [S] secondary, a third party's summary or coverage that could not be checked against the original; [A] claim by an interested party (a vendor selling the product). Sections marked as our own analysis are not evidence. Research cutoff: October 15, 2026.

1 The model and why it is growing

A vCISO is an information security lead engaged by the hour or by subscription, without a full-time position. Its space is explained by a skills constraint more than a headcount one. ISC2's 2025 workforce study surveyed 16,029 professionals in July and August 2025: 59% report critical or significant skill needs (44% in 2024), 88% suffered at least one security consequence from a skills shortage, and 29% cite lack of budget to hire as a cause [1] [P]. AI is the most cited skill need (41%), ahead of cloud security (36%). ISC2 notes it did not estimate the workforce gap that year. And Mexico is only 1% of the sample, so these figures describe a market that is not ours at the resolution we would like.

On the supply side, a report from Cynomi (a vCISO platform vendor), cited by third parties, says the share

of managed service providers offering vCISO went from 21% to 67% in a year and that 81% of them use AI or automation [2] [A]. Proofpoint, cited by another firm, reports that 79% of CISOs say they are expected to manage AI risk without proportional resources [3] [S]. Both are vendor figures with no primary report in sight, and they are used here only as an indication of direction.

There is a second push, personal liability. In 2023 the SEC sued SolarWinds and its CISO, the first such case against an individual in that role. A judge dismissed most of it in July 2024 and the SEC withdrew the rest with prejudice on November 20, 2025, with no ruling on the merits [4] [S]. Commentators read it as the threshold for individual securities exposure perhaps being higher today, not as the risk having gone away. What matters here: accountability still has a name and surname, and no agent absorbs it.

2 What is and is not known about agents in GRC

We looked for independent studies measuring the accuracy of AI agents at generating compliance evidence, mapping controls or drafting governance deliverables. We found no peer-reviewed one. What exists is vendor articles and adjacent work. A 2026 preprint on legal compliance triage argues that audit teams need

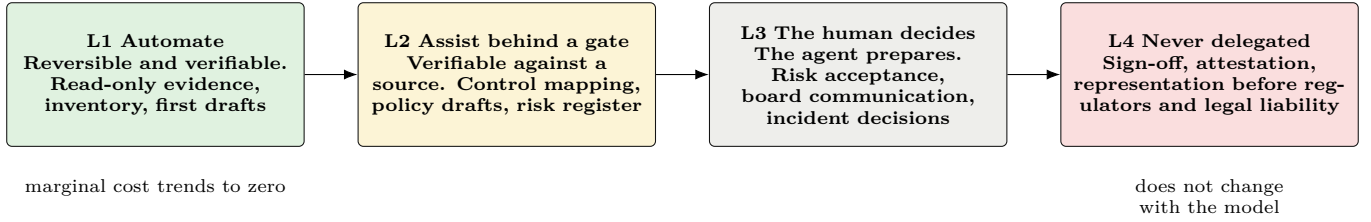


Figure 1: Delegation ladder for an AI-enabled vCISO. An activity’s level is set by the outcome of three tests, not by model capability (own analysis).

findings traceable to specific clauses and controls, and criticizes copilots that summarize obligations without that backing [5] [P]. Another, on dark-pattern audits, reports that an agent completed 81% of workflows but with reliability limited by infrastructure failures and domain-specific scaffolding [6] [P]. These are results from other domains.

What is consistent across vendors is their own caveat: automated evidence collection works well, and the team must check that evidence is complete and defensible before sending it to the auditor [7] [A]. Put differently, the consensus that does exist is that verification is needed, while who verifies, with what, and at what cost is not reported with comparable metrics. That is the gap this text tries to bound.

3 The delegation ladder

This section is our own analysis, not evidence.

Classifying activities by «how smart the model is» ages badly, because the model improves every quarter. We propose classifying them by three tests that do not depend on the model:

- **Reversibility.** If the result is wrong, can it be undone without harm?
- **Accountability.** Must someone answer by name and signature to a client, a regulator or a board?
- **Verifiability.** Is there a source against which the output can be checked cheaply?

Figure 1 shows the four levels and Table 1 crosses them with the tests. An activity moves down a level (toward more human control) as soon as one of the three fails.

Level	Tests passed	Minimum control
L1	All three	Logging and periodic human sampling
L2	Reversible and verifiable; sign-off comes later	Typed gate before delivery; human review of every deliverable
L3	Fails accountability or reversibility	The agent only prepares options and evidence; the human decides and records it
L4	Accountability by definition	Named person, no automation of the decision

Table 1: Tests and controls by level (own analysis).

A case that illustrates the rule. A vendor security questionnaire is L1 as a first draft, because each answer is checked against a policy or configuration. But it will be sent to the client signed by someone, so delivery falls to L2: the draft goes through a gate that rejects any claim without a source artifact. The decision to accept a residual risk, even if the agent prepared the full analysis, is L3.

4 Six design principles

Own analysis.

1. No evidence, no claim. Every agent sentence that asserts a control state links to an artifact (configuration, log, dated policy). What does not link is marked as an assumption, not a finding.

2. A gate before delivery. Between the agent and the client or auditor sits a verification with a fixed schema and a threshold, which blocks on any error of its own. It is the same pattern as our work on typed gates [8]: a well-formed output does not guarantee a correct decision, and a gate that fails open is not a gate.

3. The agent is a non-human identity with least privilege. Read-only by default, explicit scope per client, short-lived credentials and a log of every action. Governance of agent identities is an open problem in the industry [9], and a vCISO operating with broad

access across several clients inherits it.

4. Client documents are untrusted input. An agent that reads policies, emails and tickets can receive instructions hidden in them. Evidence is treated as data, never as orders, and the agent carries no write tools while processing third-party material.

5. Isolation and confidentiality per client. One client’s information does not feed another’s work, neither as an example nor as context. In Mexico, the personal data law in force since March 21, 2025 widened the definition of controller to include anyone who processes data even without deciding on the processing, and processors [10] [S]. We found no article specific to AI, and the updated regulation was not confirmed at the cutoff date, so any processing of personal data with a model should be reviewed by a lawyer. Operationally: minimize and redact before sending to a model, and prefer local or contracted models when sensitive data is involved.

6. A named human signs. Every deliverable carries the name of the person who reviewed it and what they reviewed. It is what separates an AI-enabled vCISO from a subscription to a report generator.

5 Likely failures

Own analysis.

False assurance. The new risk is not that the agent errs, but that the report looks more secure than it is. A fluent document in an auditor’s tone conveys a confidence the evidence does not back. It is the governance version of the «silent failure»: what hurts most is what does not warn.

Reviewer automation bias. When 95% of drafts are right, review becomes a formality and the remaining 5% slips through. The mitigation is process: sampling, drafts with seeded errors, and measuring review time per deliverable.

Price compression without risk compression. If L1 and L2 get cheaper, pressure lands on the hourly price. The risk is that the provider cuts exactly the human review that carries accountability. This is a hypothesis, not a measurement.

Cross-client drift. An agent with shared memory or mixed context can cite one client while working for another. It is a confidentiality failure and a reputational one.

6 Measurement protocol

A proposal, with no figures of our own.

Whoever offers an AI-enabled vCISO should be able to answer with numbers from their own cases, not from the model vendor:

- **Control-mapping precision** against a blind human review of the same material.
- **Rate of claims without evidence** that the gate catches before delivery.
- **Escalation rate** to L3 and what share turned out to be necessary.
- **Evidence freshness:** average age of the cited artifacts.
- **Human review time** per deliverable and its trend.
- **False-assurance incidents:** cases where a client or auditor found an unsupported claim.

We set no thresholds because we have no data to support them. A reasonable pilot runs these metrics in parallel with human delivery for an agreed period and publishes the table, bad results included. Gates are calibrated against your own cases, not a third party’s benchmark.

7 What to ask a provider in 2027

1. Which activities do you automate and at what level of the ladder? Who signs each deliverable?
2. How do you check that every claim has evidence? What happens when the check fails?
3. What access does the agent have in my environment, with what lifetime, and where is the log?
4. Does my data go to a third-party model? Under what contract and with what prior redaction?
5. How do you isolate information between clients?
6. What do you measure about your own accuracy, and will you show me the bad results?

8 What to watch

- Enforcement date for high-risk AI systems under the EU AI Act: Regulation (EU) 2026/1744

moves it from August 2, 2026 to December 2, 2027 for standalone systems [11] [S]. The Article 50 transparency obligation applies from August 2, 2026. The text changes the AI literacy duty, not just the date.

- The NIST draft on using AI for Cybersecurity Framework analysis and reporting (SP 1353 ipd), whose comment period closed on October 15, 2026 [12] [P].
- Publication of the updated regulation under the Mexican personal data law and the authority’s position on AI processing.
- Independent studies of agent accuracy on GRC tasks. There are none today; the first one changes this discussion.
- Whether buyers start requiring an AI management certification, such as ISO/IEC 42001, before hiring a provider, as reported through a quote from A-LIGN [3] [S].

9 Limitations

[S] and [A] sources were not checked against the original; do so before reusing their figures. We could not open the official text of Regulation (EU) 2026/1744 on EUR-Lex, and the dates come from law-firm alerts and summaries published afterwards. The full text of the cited preprints was not read. Industry analyst predictions are not cited because the original reports were not accessible. The ISC2 sample barely includes Mexico. The ladder, the principles and the protocol are our own analysis, have not been validated in the field, and no service results have been measured. NullGhost’s vCISO service applies this framework, but the framework has not been validated in the field. Given how fast the field moves, several figures may be out of date within weeks.

References

- [1] ISC2. *2025 Cybersecurity Workforce Study*, December 4, 2025. isc2.org/Insights/2025/12/... [P]
- [2] Cynomi. *State of the vCISO Report 2025*, cited by Purple Shield Security; the primary report was not located. Context on the launch of AI agents for vCISO (April 2026): Channel Insider. channelinsider.com/... [A]
- [3] Proofpoint, 2026 survey, cited by A-LIGN (which also records that some buyers ask for ISO/IEC 42001 before signing), via a search summary. No primary report was located. [S]
- [4] Harvard Law School Forum on Corporate Governance. *SolarWinds Dismissed: What the SEC’s U-turn Signals for Cyber Enforcement*, December 7, 2025. corpgov.law.harvard.edu/2025/12/07/... [S]
- [5] *Deterministic Fuzzy Triage for Legal Compliance Classification and Evidence Retrieval*. arXiv:2603.07390, 2026. Only the introduction was read. [P]
- [6] *On the Suitability of LLM-Driven Agents for Dark Pattern Audits*. arXiv:2603.03881, 2026. Only the abstract was read. [P]
- [7] SureCloud. *5 Agentic AI Compliance Tasks That Work in 2026*. <https://www.surecloud.com/blog-hub/agentic-ai-compliance-tasks> [A]
- [8] Martínez, R. *Before Execution: A Typed, Calibrated Gate for Agent Actions*. NullGhost Research, September 2026. <https://github.com/RickDevopsSecure/the-verification-gate> [A]
- [9] Martínez, R. *The Governance Gap*. NullGhost Research, 2026. Builds on McKinsey, *Technology Trends Outlook 2026*. <https://github.com/RickDevopsSecure/the-governance-gap> [A]
- [10] New Federal Law on the Protection of Personal Data Held by Private Parties (Mexico), published in the DOF on March 20, 2025; see Hogan Lovells and Basham. Current text at diputados.gob.mx/LeyesBiblio/ref/lfpdppp.htm. The text of the articles was not read. [S]
- [11] Regulation (EU) 2026/1744 (Digital Omnibus on AI), published July 24, 2026; dates read in Gibson Dunn’s alert of May 27, 2026 (provisional agreement) and confirmed in later coverage of the publication that was not cross-checked. Verify on EUR-Lex before reuse. [S]
- [12] NIST. Cybersecurity Framework, project page and draft SP 1353 ipd. <https://www.nist.gov/cyberframework> [P]