

El vCISO con IA: qué se delega, qué se verifica y qué sigue siendo humano, rumbo a 2027

Ricardo Martínez

Cloud Security & IA Aplicada — NullGhost Research

Octubre 2026

Resumen

El CISO fraccional (vCISO) ya es un modelo establecido y la oferta empieza a anunciarse como «vCISO con IA», a veces con lenguaje de autonomía. Este texto separa lo que muestra la evidencia pública de lo que solo afirman los proveedores: la demanda por habilidades de seguridad es medible, la automatización de la entrega es real pero casi siempre autorreportada, y no encontramos ninguna medición independiente de qué tan bien un agente produce evidencia de cumplimiento. Con base en eso proponemos una escalera de delegación de cuatro niveles, definida por tres pruebas (reversibilidad, responsabilidad y verificabilidad), seis principios de diseño y un protocolo de medición que cualquier proveedor puede correr sobre sus propios casos. La conclusión es deliberadamente incómoda: lo que el agente abarata es la preparación, no la responsabilidad, y el riesgo nuevo es el aseguramiento falso, un informe pulido que afirma cumplimiento sin evidencia detrás.

Nota de proceso. La búsqueda y la redacción tuvieron asistencia de IA. Las cifras vienen de las fuentes citadas, no del modelo. Cada fuente lleva una etiqueta: [P] primaria, leída directamente; [S] secundaria, resumen o cobertura de un tercero que no se pudo contrastar con el original; [A] afirmación de una parte interesada (proveedor que vende el producto). Las secciones marcadas como análisis propio no son evidencia. Corte de la investigación: 15 de octubre de 2026.

1. El modelo y por qué crece

Un vCISO es un responsable de seguridad de la información contratado por horas o por suscripción, sin plaza de tiempo completo. Su espacio lo explica una restricción de habilidades más que de personas. El estudio de la fuerza laboral de ISC2 de 2025 encuestó a 16,029 profesionales entre julio y agosto de 2025: 59 % reporta necesidades de habilidades críticas o significativas (44 % en 2024), 88 % sufrió al menos una consecuencia de seguridad por falta de habilidades, y 29 % cita la falta de presupuesto para contratar como causa [1] [P]. La IA es la necesidad de habilidades más citada (41 %), por encima de seguridad en la nube (36 %). ISC2 aclara que ese año no estimó la brecha de personal. Y México es apenas 1 % de la muestra, así que estas cifras describen un mercado que no es el nuestro con la resolución que quisiéramos.

Del lado de la oferta, un informe de Cynomi (provee-

dor de plataforma para vCISO), citado por terceros, reporta que la proporción de proveedores de servicios gestionados que ofrecen vCISO pasó de 21 % a 67 % en un año y que 81 % de ellos usa IA o automatización [2] [A]. Proofpoint, citada por otra firma, reporta que 79 % de los CISO dicen que se espera que gestionen el riesgo de IA sin recursos proporcionales [3] [S]. Ambas son cifras de quien vende, sin informe primario a la vista, y se usan aquí solo como indicio de dirección.

Hay un segundo empuje, el de responsabilidad personal. En 2023 la SEC demandó a SolarWinds y a su CISO, el primer caso de ese tipo contra una persona en ese cargo. Un juez desestimó la mayor parte en julio de 2024 y la SEC retiró el resto con perjuicio el 20 de noviembre de 2025, sin fallo de fondo [4] [S]. La lectura de los comentaristas es que el umbral de exposición individual en valores quizá sea hoy más alto, no que el riesgo haya desaparecido. Para esta discusión importa lo siguiente: la responsabilidad sigue teniendo nombre y apellido, y ningún agente la absorbe.

2. Lo que se sabe y lo que no sobre agentes en GRC

Buscamos estudios independientes que midan la precisión de agentes de IA al generar evidencia de cumplimiento, mapear controles o redactar entregables de gobernanza. No encontramos ninguno revisado por pares. Lo que hay son artículos de proveedores y trabajos

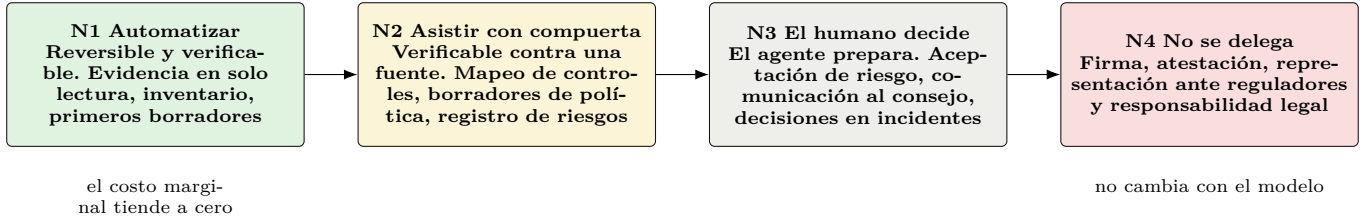


Figura 1: Escalera de delegación para un vCISO con IA. El nivel de una actividad lo fija el resultado de tres pruebas, no la capacidad del modelo (análisis propio).

adyacentes. Un preprint de 2026 sobre triaje de cumplimiento legal sostiene que los equipos de auditoría necesitan hallazgos trazables a cláusulas y controles específicos, y critica los copilotos que resumen obligaciones sin ese respaldo [5] [P]. Otro, sobre auditoría de patrones oscuros, reporta que un agente completó 81 % de los flujos pero con fiabilidad limitada por fallos de infraestructura y por andamiaje específico del dominio [6] [P]. Son resultados de otros dominios.

Lo que sí es consistente entre proveedores es la propia salvedad: la recolección automatizada de evidencia funciona bien, y el equipo debe verificar que la evidencia sea completa y defendible antes de enviarla al auditor [7] [A]. Dicho de otra forma, el consenso que sí existe es que hay que verificar, y quién verifica, con qué y cuánto cuesta no se reporta con métricas comparables. Esa es la brecha que este texto intenta acotar.

3. La escalera de delegación

Esta sección es análisis propio, no evidencia.

Clasificar actividades por «qué tan inteligente es el modelo» envejece mal, porque el modelo mejora cada trimestre. Proponemos clasificarlas por tres pruebas que no dependen del modelo:

- **Reversibilidad.** Si el resultado es incorrecto, ¿se puede deshacer sin daño?
- **Responsabilidad.** ¿Alguien debe responder con nombre y firma ante un cliente, un regulador o un consejo?
- **Verificabilidad.** ¿Existe una fuente contra la cual comprobar la salida de forma barata?

La Figura 1 muestra los cuatro niveles y la Tabla 1 los cruza con las pruebas. Una actividad baja de nivel (hacia más control humano) en cuanto falla una de las tres.

Nivel	Pruebas que pasa	Control mínimo
N1	Las tres	Registro y muestreo humano periódico
N2	Reversible y verificable; la firma llega después	Compuerta tipada antes de entregar; revisión humana de cada entregable
N3	Falla responsabilidad o reversibilidad	El agente solo prepara opciones y evidencia; el humano decide y lo registra
N4	Es responsabilidad por definición	Persona nombrada, sin automatización de la decisión

Tabla 1: Pruebas y controles por nivel (análisis propio).

Un caso que ilustra la regla. Un cuestionario de seguridad de un proveedor pasa a N1 como primer borrador, porque cada respuesta se comprueba contra una política o configuración. Pero se enviará al cliente firmado por alguien, así que la entrega cae a N2: el borrador sale por una compuerta que rechaza toda afirmación sin artefacto fuente. La decisión de aceptar un riesgo residual, aunque el agente haya preparado el análisis completo, es N3.

4. Seis principios de diseño

Análisis propio.

- 1. Sin evidencia, sin afirmación.** Toda frase del agente que afirme un estado de control enlaza a un artefacto (configuración, registro, política con fecha). Lo que no enlaza se marca como supuesto, no como hallazgo.
- 2. Compuerta antes de entregar.** Entre el agente y el cliente o el auditor va una verificación con esquema fijo y umbral, que bloquea ante cualquier error propio. Es el mismo patrón de nuestro trabajo sobre compuertas tipadas [8]: una salida bien formada no garantiza una decisión correcta, y una compuerta que falla abierta no es una compuerta.
- 3. El agente es una identidad no humana con**

mínimo privilegio. Solo lectura por defecto, alcance explícito por cliente, credenciales de vida corta y registro de cada acción. La gobernanza de identidades de agentes es un problema abierto en la industria [9], y un vCISO que opera con accesos amplios a varios clientes lo hereda.

4. Los documentos del cliente son entrada no confiable. Un agente que lee políticas, correos y tickets puede recibir instrucciones escondidas en ellos. La evidencia se trata como datos, nunca como órdenes, y el agente no lleva herramientas de escritura cuando procesa material ajeno.

5. Aislamiento y confidencialidad por cliente. La información de un cliente no alimenta el trabajo de otro, ni como ejemplo ni como contexto. En México, la ley de datos personales vigente desde el 21 de marzo de 2025 amplió la definición de responsable para incluir a quien trata datos aunque no decida sobre el tratamiento, y a los encargados [10] [S]. No encontramos un artículo específico sobre IA, y el reglamento actualizado no estaba confirmado a la fecha de corte, así que cualquier tratamiento de datos personales con un modelo debe revisarlo un abogado. Operativamente: minimizar y censurar antes de enviar a un modelo, y preferir modelos locales o con contrato cuando haya datos sensibles.

6. Un humano nombrado firma. Cada entregable lleva el nombre de la persona que lo revisó y lo que revisó. Es lo que separa un vCISO con IA de una suscripción a un generador de informes.

5. Fallas probables

Análisis propio.

Aseguramiento falso. El riesgo nuevo no es que el agente se equivoque, sino que el informe luzca más seguro de lo que es. Un documento fluido con tono de auditor transmite una confianza que la evidencia no respalda. Es la versión de gobernanza del «fallo silencioso»: lo que más daña es lo que no avisa.

Sesgo de automatización del revisor. Cuando el 95 % de los borradores es correcto, la revisión se vuelve un trámite y el 5 % restante pasa. La mitigación es de proceso: muestreo, borradores con errores sembrados, y medir el tiempo de revisión por entregable.

Compresión de precio sin compresión de riesgo. Si N1 y N2 bajan de costo, la presión recae sobre el precio por hora. El riesgo es que el proveedor recorte justo la revisión humana que sostiene la responsabilidad. Es una hipótesis, no una medición.

Deriva entre clientes. Un agente con memoria com-

partida o contexto mezclado puede citar a un cliente al trabajar para otro. Es una falla de confidencialidad y, además, de reputación.

6. Protocolo de medición

Propuesta, sin cifras propias.

Quien ofrezca un vCISO con IA debería poder responder con números de sus propios casos, no con los del proveedor del modelo:

- **Precisión del mapeo de controles** contra una revisión humana ciega del mismo material.
- **Tasa de afirmaciones sin evidencia** que la compuerta detecta antes de entregar.
- **Tasa de escalamiento** a N3 y qué porcentaje resultó necesario.
- **Frescura de la evidencia:** antigüedad media de los artefactos citados.
- **Tiempo de revisión humana** por entregable y su tendencia.
- **Incidentes de aseguramiento falso:** casos donde un cliente o auditor encontró una afirmación sin respaldo.

No fijamos umbrales porque no tenemos datos que los respalden. Un piloto razonable corre estas métricas en paralelo a la entrega humana durante un periodo acordado y publica la tabla, incluidos los resultados malos. Las compuertas se calibran contra casos propios, no contra el benchmark de un tercero.

7. Qué preguntarle a un proveedor en 2027

1. ¿Qué actividades automatizan y en qué nivel de la escalera? ¿Quién firma cada entregable?
2. ¿Cómo comprueban que cada afirmación tiene evidencia? ¿Qué ocurre cuando la comprobación falla?
3. ¿Qué accesos tiene el agente en mi entorno, con qué vida útil y dónde queda el registro?
4. ¿Mis datos salen a un modelo de terceros? ¿Bajo qué contrato y con qué censura previa?
5. ¿Cómo aíslan la información entre clientes?
6. ¿Qué miden de su propia precisión y me muestran los resultados malos?

8. Qué observar

- Fecha de exigibilidad de los sistemas de IA de alto riesgo bajo la ley de IA de la UE: el Reglamento (UE) 2026/1744 la mueve del 2 de agosto de 2026 al 2 de diciembre de 2027 para los sistemas independientes [11] [S]. La obligación de transparencia del artículo 50 sigue desde el 2 de agosto de 2026. El texto cambia el contenido de alfabetización en IA, no solo la fecha.
- El borrador del NIST sobre el uso de IA para análisis y reporte del Cybersecurity Framework (SP 1353 ipd), cuyo plazo de comentarios cerraba el 15 de octubre de 2026 [12] [P].
- Publicación del reglamento actualizado de la ley mexicana de datos personales y criterio de la autoridad sobre tratamiento con IA.
- Estudios independientes de precisión de agentes en tareas de GRC. Hoy no los hay; el primero que exista cambia esta discusión.
- Si los compradores empiezan a exigir una certificación de gestión de IA, como ISO/IEC 42001, antes de contratar a un proveedor, según una cita de A-LIGN recogida por un tercero [3] [S].

9. Limitaciones

Las fuentes [S] y [A] no se contrastaron con el original; conviene hacerlo antes de reutilizar sus cifras. No pudimos abrir el texto oficial del Reglamento (UE) 2026/1744 en EUR-Lex, y las fechas provienen de comunicados de despachos y de resúmenes posteriores a su publicación. No se leyó el texto completo de los preprints citados. No se citan predicciones de analistas de la industria porque no se tuvo acceso a los informes originales. La muestra de ISC2 apenas incluye a México. La escalera, los principios y el protocolo son análisis propio, no se han validado en campo y no se han medido resultados de ningún servicio. El servicio vCISO de NullGhost aplica este marco, pero el marco no se ha validado en campo. Por la rapidez del campo, varias cifras pueden quedar desactualizadas en semanas.

Referencias

- [1] ISC2. *2025 Cybersecurity Workforce Study*, 4 de diciembre de 2025. isc2.org/Insights/2025/12/... [P]
- [2] Cynomi. *State of the vCISO Report 2025*, citado por Purple Shield Security; no se localizó el informe primario. Contexto del lanzamiento de agentes de IA

para vCISO (abril de 2026): Channel Insider. channelinsider.com/... [A]

- [3] Proofpoint, encuesta de 2026, citada por A-LIGN (que también recoge que algunos compradores piden ISO/IEC 42001 antes de firmar), vía resumen de búsqueda. No se localizó ningún informe primario. [S]
- [4] Harvard Law School Forum on Corporate Governance. *SolarWinds Dismissed: What the SEC's U-turn Signals for Cyber Enforcement*, 7 de diciembre de 2025. corpgov.law.harvard.edu/2025/12/07/... [S]
- [5] *Deterministic Fuzzy Triage for Legal Compliance Classification and Evidence Retrieval*. arXiv:2603.07390, 2026. Solo se leyó la introducción. [P]
- [6] *On the Suitability of LLM-Driven Agents for Dark Pattern Audits*. arXiv:2603.03881, 2026. Solo se leyó el resumen. [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. *Antes de ejecutar: una compuerta tipada y calibrada para acciones de agentes*. NullGhost Research, septiembre de 2026. <https://github.com/RickDevopsSecure/the-verification-gate> [A]
- [9] Martínez, R. *La brecha de gobernanza*. NullGhost Research, 2026. Parte de McKinsey, *Technology Trends Outlook 2026*. <https://github.com/RickDevopsSecure/the-governance-gap> [A]
- [10] Nueva Ley Federal de Protección de Datos Personales en Posesión de los Particulares, publicada en el DOF el 20 de marzo de 2025; ver Hogan Lovells y Basham. Texto vigente en diputados.gob.mx/LeyesBiblio/ref/lfpdppp.htm. No se leyó el texto de los artículos. [S]
- [11] Reglamento (UE) 2026/1744 (Digital Omnibus on AI), publicado el 24 de julio de 2026; fechas leídas en el alert de Gibson Dunn del 27 de mayo de 2026 (acuerdo provisional) y confirmadas en coberturas posteriores de la publicación que no se contrastaron. Verificar en EUR-Lex antes de reutilizar. [S]
- [12] NIST. Cybersecurity Framework, página del proyecto y borrador SP 1353 ipd. <https://www.nist.gov/cyberframework> [P]