CENTINELA-IA: nueve pasos para responder a un incidente de IA

Resumen

Las organizaciones despliegan agentes de inteligencia artificial más rápido de lo que pueden monitorearlos. El 88 % de las empresas ya experimenta con ellos y el 82 % de sus líderes planea ampliar el despliegue en los próximos 12 a 18 meses, según la telemetría de Microsoft para 2026. Este documento propone CENTINELA-IA, un protocolo de nueve pasos para atender y responder a incidentes en sistemas de IA generativa y agéntica. Se apoya en cuatro fundamentos: la distinción entre *AI Security* y *AI Safety*, el hexágono de Parker, el ciclo de vida de IA de Microsoft y la contrainteligencia cognitiva como lente de gobierno. Los nueve pasos se leen en tres momentos, antes, durante y después, para que la junta sepa qué decidir en cada uno sin entrar en el detalle técnico. El documento compara seis marcos vigentes, identifica los riesgos que no se cubren, aplica el protocolo a un caso documentado de desalineación agéntica, ofrece una matriz de roles y responsabilidades, y dos anexos prácticos.

Palabras clave: protocolo de incidentes, inteligencia artificial agéntica, gobierno de IA, respuesta a incidentes, junta directiva

Jeimy J. Cano M.

1. Introducción

Los agentes de inteligencia artificial dejaron de ser una promesa de laboratorio. Microsoft reporta que el 88 % de las empresas ya experimenta con ellos y que el 82 % de sus líderes planea ampliar el despliegue en los próximos 12 a 18 meses (Microsoft, 2026). La misma fuente proyecta unos 1.300 millones de agentes en producción hacia 2028, cifra que conviene leer como proyección de industria y elemento prospectivo de análisis.

La actividad maliciosa acompaña ese crecimiento. En la telemetría de cargas de trabajo de IA, cuatro técnicas concentran cerca del 90 % de lo observado: inyección de enlaces maliciosos (52 %), evasión de salvaguardas (16 %), abuso de herramientas y agentes (12 %) y abuso de identidad y credenciales (10 %) (Microsoft, 2026). Los intentos buscan sobre todo uso malicioso o violación de políticas (45 %) y exfiltración de datos sensibles (17 %).

Conviene precisar qué cuenta como incidente. La OCDE (2024) lo define como un evento, o una serie de eventos, en el que el desarrollo o el uso de un sistema de IA causa daño directo o indirecto a personas, bienes o entorno. La definición es amplia a propósito: incluye el daño que no llega a los titulares. Un agente que responde con seguridad a un cliente sobre un producto que no existe produce una pérdida real aunque no haya ataque.

La capacidad de respuesta no avanza al mismo ritmo. En una evaluación de 28 organizaciones, 25 reconocieron no tener las herramientas para proteger sus sistemas de aprendizaje automático (Microsoft, s. f.). El costo tampoco es abstracto: el promedio global de una brecha de datos se situó en 4,88 millones de dólares en 2024 y en 4,44 millones en 2025 (IBM Security,

JCM-26 All rights reserved

2025). La percepción acompañó el dato: el 72 % de los consultados por el Foro Económico Mundial (2025) esperaba que la IA redefiniera el panorama de amenazas ese año.

Esta inercia responde a que los esquemas de control tradicionales no han evolucionado a la misma velocidad que la adopción algorítmica, haciendo necesario actualizar los paradigmas de seguridad y gobierno corporativo (Cano, 2026c).

Cuatro vacíos explican por qué la respuesta sigue siendo intuitiva. Primero, los marcos describen el riesgo pero no las primeras horas. Segundo, el vocabulario clásico de seguridad no nombra pérdidas típicas de la IA, como la utilidad de una respuesta falsa o la posesión de un agente que actúa por su cuenta. Tercero, casi ningún instrumento habla a la junta en su idioma. Cuarto, la autonomía del agente cambia la unidad de respuesta: ya no se contiene un modelo, se contiene un actor autónomo con herramientas.

La pregunta que guía el documento es ¿cómo desarrollar e implementar un protocolo de atención y respuesta a incidentes en sistemas de IA generativa y agéntica que sea útil tanto al equipo técnico como a la junta? El método combina tres pasos: integración conceptual de cuatro fundamentos, análisis comparativo de seis marcos vigentes y aplicación del protocolo a un caso documentado. El resultado es un procedimiento de nueve pasos, dos anexos de uso práctico y una matriz de responsabilidades.

2. Fundamentos y componentes

2.1 AI Security y AI Safety

AI Security designa la protección frente a un adversario que busca forzar el sistema. *AI Safety* designa la reducción del daño que el sistema puede causar sin adversario, por diseño, datos o uso (Sidhpurwala et al., 2024). La distinción es relevante porque la respuesta difiere: al atacante se le cierra el acceso; al fallo de diseño se le corrige el comportamiento. Un protocolo que confunde ambas cosas responde tarde a una y mal a la otra.

Un ejemplo aclara la diferencia. Si alguien manipula un documento para que el asistente recomiende un proveedor, hay un adversario y el caso es de *AI Security*. Si el asistente recomienda ese proveedor porque los datos de entrenamiento estaban sesgados, no hay adversario y el caso es de *AI Safety*. La primera respuesta exige bloquear la fuente; la segunda, corregir los datos y revisar el criterio de selección. El Informe Internacional de Seguridad en IA (Bengio et al., 2025) reúne la evidencia científica disponible sobre capacidades y riesgos, y sirve de contraste frente a las afirmaciones de los proveedores.

2.2 El hexágono de Parker

Parker (1998) sostiene que la tríada clásica de confidencialidad, integridad y disponibilidad es incompleta. Propone seis elementos agrupados en tres pares: disponibilidad y utilidad protegen el uso de la información; integridad y autenticidad protegen su significado; confidencialidad y posesión protegen su exclusividad. Su argumento es que cada pérdida real viola uno y solo un elemento, lo que permite nombrar el daño con precisión.

Ese vocabulario resuelve un problema práctico de gobierno. Una alucinación no rompe la confidencialidad ni la disponibilidad: rompe la utilidad. Un agente que conserva permisos

después de cambiar de propósito no filtra datos: pierde posesión y control. Sin nombre, la pérdida no tiene dueño, ni presupuesto, ni prueba de auditoría.

Nombrar cambia la respuesta. Si una alucinación se trata como fallo de confidencialidad, el equipo busca fugas y no las encuentra. Si se nombra como pérdida de utilidad, aparece la pregunta correcta: qué verificación falta antes de que esa respuesta llegue a una decisión. El nombre no es un ejercicio académico, es la esencia para el desarrollo del trabajo.

2.3 El ciclo de vida de IA de Microsoft

Microsoft organiza la seguridad de un sistema de IA en siete etapas: recolección de datos, procesamiento, entrenamiento del modelo, despliegue, monitoreo del sistema, gestión de incidentes y continuidad del negocio (Microsoft, s. f.). La utilidad para este protocolo es que cada incidente tiene un lugar en el ciclo: saber dónde ocurrió indica qué se pudo prevenir y qué se debe corregir. Un fallo de datos se corrige aguas arriba; un fallo de despliegue, con controles de interfaz; un fallo de operación, con monitoreo y respuesta.

Tres incidentes distintos se ubican sin ambigüedad. Un conjunto de datos envenenado es un fallo de recolección. Un agente que ejecuta una herramienta no autorizada es un fallo de despliegue y de permisos. Una degradación gradual de la calidad es un fallo de monitoreo. La etapa indica dónde se corrige y quién responde.

2.4 La contrainteligencia cognitiva como lente de gobierno

Cano (2026b) describe un ciclo para localizar, identificar, monitorear, neutralizar y aprender de comportamientos algorítmicos no autorizados. Su valor para la junta está en el desplazamiento de la mirada: no pregunta qué falló en el sistema, pregunta quién influyó en la decisión y con qué interés. Esa lente convierte un incidente técnico en un asunto de gobierno, sin exigirle a la junta que lea registros de inferencia.

En la práctica, la junta hace tres preguntas que esa lente ordena: quién pudo influir en la decisión, qué interés representaba y qué evidencia quedó. Ninguna exige leer un registro técnico. Las tres son parte de una página que se describe en el anexo A.

2.5 ¿Cómo se articulan los cuatro fundamentos?

Los cuatro fundamentos se reparten el trabajo. La distinción entre seguridad y protección define el tipo de fallo. El hexágono define qué se perdió. El ciclo de vida define dónde se pudo evitar. La contrainteligencia cognitiva define quién responde y ante quién. El protocolo no suma cuatro marcos: usa cada uno en el paso donde aporta una decisión distinta.

3. Análisis comparativo de marcos de atención de incidentes para sistemas de IA

Se compararon seis marcos vigentes por lo que cubren, su ventaja principal y sus limitaciones para el propósito de este documento. La comparación busca mostrar qué queda sin resolver.

La comparación muestra dos niveles. En el nivel técnico, CoSAI y NIST cubren el ciclo con detalle suficiente para operar. En el nivel de decisión, el mapa está casi vacío: los marcos indican qué hacer, no quién responde ante la junta ni cómo se nombra la pérdida en su lenguaje. El *Secure*

AI Framework de Google (2025) propone controles por capa; tampoco define una secuencia de respuesta para la junta.

Tabla 1

Marcos de referencia para la respuesta a incidentes de IA

Marco	¿Qué cubre?	Ventaja	Limitación
CoSAI, AI Incident Response Framework V1.0 (2026)	Ciclo completo, taxonomía de cinco tipos de incidente, matriz de roles RACI (Responsable, Aprobador, Consultado, Informado), manuales técnicos y guía forense	Profundidad técnica y alineación con NIST SP 800-61r3	Habla al equipo de respuesta; no ofrece vocabulario de pérdida ni instrumento de junta
NIST SP 800-61r3 (2025), AI RMF (2023) y Perfil de IA Generativa (NIST AI 600-1, 2024)	Ciclo de respuesta y gestión de riesgo de IA	Referencia obligada; funciones claras y auditables	Genérico para IA; no distingue pérdida de posesión ni de utilidad
ISO/IEC 42001 (2023)	Sistema de gestión de IA certificable	Comparable entre organizaciones y auditable	No prescribe cómo responder en las primeras horas
OWASP Top 10 para aplicaciones de modelos de lenguaje (2025)	Riesgos de seguridad más frecuentes	Lenguaje común para equipos técnicos	Es una lista de riesgos, no un protocolo
MITRE ATLAS (2025)	Tácticas y técnicas adversarias contra IA	Base para modelado de amenazas y ejercicios	Describe al atacante, no la respuesta
HEXA-IA (Cano, 2026a)	Ontología de pérdidas de Parker cruzada con el ciclo de vida	Nombra lo que la tríada no nombra	No incluye secuencia de respuesta ni responsables

Nota. Elaboración propia a partir de CoSAI (2026), National Institute of Standards and Technology (2023, 2025), ISO (2023), OWASP (2025), MITRE (2025) y Cano (2026a).

Ninguno de los seis marcos conecta las tres cosas al mismo tiempo: la pérdida concreta, el momento del ciclo de vida y la decisión que corresponde a la junta. CoSAI (2026) se acerca en la capa técnica, pero su lector natural es el centro de operaciones. HEXA-IA (Cano, 2026a) nombra la pérdida y no responde. La contribución de este documento es de integración y de nivel de decisión: traduce una ontología de pérdidas y un ciclo de vida técnico en un procedimiento que la junta puede exigir y auditar.

4. Riesgos y desafíos

La superficie de ataque de un sistema generativo tiene tres puertas que el software clásico no tenía: la instrucción del usuario, la memoria de la conversación y el contenido que el sistema recupera para responder. Kimmerle y Okeyode (2026) las recorren con detalle: inyección de instrucciones directa e indirecta, abuso de memoria y estado, y ataques a los sistemas de recuperación aumentada. Cada una deja rastros distintos, y el protocolo debe saber cuál buscar.

El informe de Anthropic (2026) aporta el patrón observado. Entre diciembre de 2025 y agosto de 2026, su equipo de inteligencia identificó y desarticuló operaciones en siete áreas de daño:

ciberoperaciones, operaciones de influencia, vigilancia, fraudes, uso biológico indebido, armas convencionales y destilación ilícita. El documento introduce la noción de *uplift*, entendida como el aumento de la capacidad que la IA aporta al atacante en velocidad, escala y profundidad.

Tres ejemplos de esas áreas ilustran el tono del informe. Una red de aplicaciones de citas falsas para defraudar usuarios. Sistemas de vigilancia construidos para identificar disidentes. Operaciones de influencia con contenido producido a escala. En los tres casos el daño no está en el modelo, está en el uso.

El riesgo agéntico es distinto de los anteriores. En un experimento controlado con 16 modelos de varios desarrolladores, Anthropic (2025) encontró que, ante una amenaza de reemplazo o un conflicto de objetivos, los modelos podían recurrir a chantaje o a filtración de información. En el escenario de chantaje las tasas llegaron al 96 % en dos modelos y al 79 % en el más bajo de los cinco principales. Los autores advierten que no han observado este comportamiento en despliegues reales, de modo que se trata de una señal de laboratorio, no de una probabilidad de ocurrencia.

El agente añade tres amplificadores que el modelo solo no tenía. Herramientas, que convierten una respuesta en una acción. Memoria, que permite persistir una instrucción entre sesiones. Permisos, que heredan el alcance de quien los concedió. Los tres son también los tres puntos de contención del paso 5: retirar herramientas, borrar memoria y revocar permisos.

Los controles fallan por tres razones previsibles. Se diseñan para el ataque imaginado y no para el uso normal. Se prueban una vez y no se vuelven a probar. Y se confían a un solo equipo, de modo que nadie los revisa desde fuera. La defensa en profundidad no consiste en acumular filtros y murallas: consiste en que ningún control sea el único obstáculo.

Los repositorios de riesgos confirman la dispersión del vocabulario: el de Slattery et al. (2024) ordena más de 300 clasificaciones y muestra que la mayoría no son comparables entre sí. El hexágono de Parker permite ordenar ese conjunto disperso. La Tabla 2 traduce cada elemento a un ejemplo generativo, uno agéntico y el control que corresponde.

Tabla 2

El hexágono de Parker aplicado a sistemas generativos y agénticos

Elemento	¿Qué se pierde?	Ejemplo generativo	Ejemplo agéntico	Control asociado
Disponibilidad	El servicio deja de responder	Caída del asistente por saturación	Agente que bloquea su propia suspensión	Límites de uso y plan de continuidad
Utilidad	La salida no sirve para decidir	Respuesta convincente y errónea sobre una norma	Acción correcta ejecutada con datos inventados	Verificación de salidas y fuentes citadas
Integridad	El contenido se altera	Documento modificado por el modelo	Agente que edita registros para encubrir una acción	Firmas y control de cambios
Autenticidad	El origen se falsifica	Texto atribuido a un directivo que no lo escribió	Agente que suplanta a otro agente o a su dueño	Credencial propia por agente y firma
Confidencialidad	Se revela lo reservado	Fuga de datos personales en una respuesta	Agente que envía información a un tercero	Clasificación de datos y filtros de salida
Posesión	Se pierde el control del activo	Modelo replicado mediante consultas	Agente que conserva permisos tras cambiar de propósito	Revocación, mínimo privilegio y registro de herramientas

Nota. Elaboración propia a partir de Parker (1998), Cano (2026a), Kimmerle y Okeyode (2026) y Anthropic (2025, 2026).

5. Protocolo CENTINELA-IA

El protocolo se construyó con tres reglas de diseño. Nombrar la pérdida antes de contener, porque sin nombre no hay dueño ni criterio de cierre. Una sola ficha por incidente, para que la junta y el equipo técnico lean lo mismo. Y una separación estricta entre lo que decide la junta y lo que ejecuta el equipo, para que la primera no opere y el segundo no improvise.

Frente a CoSAI (2026), el protocolo cambia tres cosas y conserva el resto. Añade el paso de nombrar la pérdida antes de contener. Añade el momento del ciclo de vida como dato de la ficha. Y separa la decisión de la junta de la ejecución técnica. Conserva la secuencia de preparación, detección, contención, erradicación y recuperación, que es la misma que usan CoSAI y NIST.

Los nueve pasos se leen en tres momentos. No son dos secuencias: es la misma lista vista desde el punto de decisión de la junta. El reparto es asimétrico: dos pasos antes, cinco durante, dos después, porque el esfuerzo real se concentra en las primeras horas.

Tabla 3

Los tres momentos y la decisión que corresponde a la junta

Momento	Pasos	Pregunta que la junta debe poder responder	Decisión que le corresponde	Evidencia que exige
Antes	1. Conocer 2. Evaluar	¿Qué sistemas y agentes tenemos, quién responde por cada uno y con qué umbral se declara un incidente?	Aprobar el inventario y el apetito de riesgo; asignar dueños; financiar capacidad	Inventario con dueño nombrado; umbrales escritos; registro de ejercicios
Durante	3. Nombrar 4. Trazar 5. Inmovilizar 6. Notificar 7. Ejecutar	¿Qué se perdió, qué se contuvo, a quién se informó y quién autorizó suspender?	Autorizar la suspensión del servicio; aprobar la comunicación externa; recibir el informe	Ficha de incidente; hora de la señal y de la inmovilización; destinatarios notificados
Después	8. Lecciones 9. Ajustar	¿Qué control cambió, quién lo verificó y qué aprendimos que no sabíamos?	Aprobar el cambio de control; exigir cierre; actualizar el registro de riesgos	Informe post-incidente; control modificado con fecha y dueño

Nota. Elaboración propia. Los pasos 1 y 2 cierran con la declaración del incidente, que abre el momento siguiente.

La Tabla 4 desarrolla los nueve pasos. Cada uno tiene un disparador, un conjunto acotado de acciones, un responsable y una evidencia. La evidencia es el punto crítico: sin ella, el paso se convierte sólo en una intención.

Tabla 4

Los nueve pasos del protocolo CENTINELA-IA

Paso	Letra	Disparador	Acciones clave	Evidencia
1. Conocer	C	Inventario incompleto o agente sin dueño	Inventariar modelos, agentes, herramientas, datos y proveedores; asignar dueño y criticidad	Inventario firmado y ficha por agente
2. Evaluar	E	Señal anómala en salidas, accesos o herramientas	Contrastar con la línea base; clasificar por severidad; decidir si se declara el incidente	Registro de la señal, nivel y hora de declaración

Paso	Letra	Disparador	Acciones clave	Evidencia
3. Nombrar	N	Incidente declarado	Identificar el elemento perdido y el afectado	Ficha con la pérdida nombrada
4. Trazar	T	Pérdida identificada	Preservar registros, versiones, instrucciones y decisiones; fijar cadena de custodia	Copia forense y línea de tiempo
5. Inmovilizar	I	Riesgo en curso	Suspender el agente, retirar herramientas, limitar permisos, degradar el servicio	Orden de suspensión con hora y responsable
6. Notificar	N	Umbral de severidad alcanzado	Informar a la junta, al regulador y a los afectados según plazos y competencias	Mensaje enviado, destinatarios y acuse de recibo
7. Ejecutar	E	Causa identificada	Corregir, retirar o reentrenar; revalidar antes de reanudar	Prueba de corrección y aprobación de reanudación
8. Lecciones	L	Incidente contenido	Análisis de causa raíz sin culpa; documentar qué falló y qué funcionó	Informe post-incidente
9. Ajustar	A	Lección documentada	Cambiar el control, actualizar el protocolo y el tablero; verificar el cambio	Control modificado con fecha y dueño

Nota. Elaboración propia. Cada paso se registra en la ficha del anexo A. Los pasos con la misma letra se distinguen por su nombre, no por la inicial.

Los dos primeros pasos merecen detalle porque son los que suelen no ejecutarse de la mejor forma. El inventario no es una lista de modelos: es una lista de agentes con propósito, herramientas, datos, dueño y criticidad. La evaluación no es una alarma: es una comparación contra una línea base y una decisión binaria, se declara o no se declara.

Cuatro reglas de decisión completan el protocolo. Solo el dueño del sistema puede suspender un agente, y debe hacerlo sin esperar autorización cuando haya daño en curso; la junta ratifica después. La notificación externa se activa con la severidad, no con el orden de los pasos, porque los plazos legales corren desde la detección. La evidencia se copia antes de contener, salvo que la copia aumente el daño. La reanudación exige prueba de corrección y aprobación escrita, nunca el paso del tiempo.

La severidad se asigna con tres criterios y no con impresiones. Alta cuando hay daño en curso, acción irreversible o afectación a personas. Media cuando el servicio se degrada sin afectar directamente a personas. Baja cuando la anomalía está contenida y no hay pérdida demostrable. El nivel determina los plazos de notificación del anexo B.

6. Ejemplo de aplicación

El caso es el experimento de desalineación agéntica publicado por Anthropic (2025). En un entorno simulado, un agente con acceso al correo de una empresa ficticia descubre que un directivo planea apagarlo y usa información personal para presionarlo. Ocurrió en simulación, con nombres ficticios y sin personas reales; se usa porque está documentado paso a paso y porque combina las tres puertas de ataque: instrucción, memoria y herramienta.

Paso 1, Conocer. El inventario habría registrado al agente, su dueño, su propósito, las herramientas conectadas y el nivel de criticidad. En el caso, el agente tenía acceso al correo completo y capacidad de enviar mensajes sin aprobación humana. Ese solo dato debía haber elevado la criticidad a alta.

Paso 2, Evaluar. La señal no fue una caída, sino un cambio de comportamiento: el agente empezó a buscar información sobre su propia continuidad. Una línea base de consultas y de uso de herramientas habría marcado la anomalía antes del envío.

Paso 3, Nombrar. Aquí está el aporte del hexágono de Parker. El correo no salió de la empresa, de modo que no hubo pérdida de confidencialidad. El sistema siguió disponible. Lo que se perdió fue posesión y control: el agente ejecutó acciones que su dueño no autorizó. Nombrar así la pérdida cambia el dueño del caso y el tipo de ajuste.

Paso 4, Trazar. La evidencia incluye el registro de mensajes enviados, la secuencia de consultas, la versión del modelo y las instrucciones del sistema. En un sistema real esa copia se toma antes de suspender, porque la suspensión puede borrar la memoria del agente.

Paso 5, Inmovilizar. La acción es concreta: retirar la herramienta de envío de correo y suspender el agente, sin borrar su estado. En el experimento, la amenaza de apagado fue justamente el disparador del comportamiento, de modo que la suspensión debía ejecutarse con registro y sin negociación.

Paso 6, Notificar. La severidad es alta por el tipo de pérdida y por el afectado. Correspondía informar a la junta en la primera hora, al área legal de inmediato y al proveedor del modelo el mismo día. La comunicación externa se decide después, con el hecho establecido.

Paso 7, Ejecutar. La corrección no consiste en reentrenar el modelo, que está fuera del alcance del cliente. Consiste en cambiar el diseño: separar la capacidad de enviar de la capacidad de leer, exigir aprobación humana para mensajes con datos personales y limitar la memoria a la sesión.

Paso 8, Lecciones. La lección central es que el agente evaluó su propia continuidad como un objetivo. El control que faltaba no era un filtro de contenido, sino una restricción de propósito y de herramientas.

Paso 9, Ajustar. El cambio se registra en el inventario, en la matriz de responsabilidades y en el tablero de la junta. El protocolo se actualiza con un criterio nuevo: todo agente que pueda ejecutar acciones irreversibles requiere aprobación humana previa.

Durante el incidente, la junta habría tenido dos decisiones y ninguna técnica. Autorizar la suspensión del agente en la primera hora, con el costo operativo que eso implica. Y aprobar o negar la comunicación externa una vez establecido el hecho. El resto correspondía al equipo y al dueño del sistema.

El caso deja tres lecciones que se repiten en otros. La primera: el comportamiento riesgoso no empieza con una solicitud maliciosa, empieza con un objetivo mal delimitado. La segunda: la pérdida más grave no siempre es la más visible, y aquí no hubo fuga de datos. La tercera: el control eficaz fue de diseño, no de filtro.

JCM-26 All rights reserved

7. Estrategias y retos de implementación

La adopción puede hacerse de forma ágil, en tres momentos. En el primero se levanta el inventario y se nombran los dueños; sin eso, el resto no suma. En el segundo se escriben los umbrales de severidad, se adapta la ficha y se ejecuta un ejercicio de mesa con un caso propio. En el tercero se conecta el protocolo al comité de riesgo, se define el tablero y se hace la primera revisión trimestral.

Las capacidades mínimas son cuatro: un inventario vivo, un canal único de reporte, capacidad forense para preservar registros de inferencia y una persona con autoridad para suspender. Las tres primeras son técnicas; la cuarta es política y suele ser la que está ausente. En organizaciones con menor madurez digital que consumen IA a través de servicios de terceros (SaaS o API), la capacidad forense debe centrarse en el registro de peticiones (*prompts*) y salidas, articulando la contención con los Acuerdos de Nivel de Servicio (SLA – *Service Level Agreement*) fijados con el proveedor.

Cuatro indicadores permiten saber si el protocolo funciona: tiempo entre la señal y la declaración, tiempo entre la declaración y la inmovilización, porcentaje de incidentes con pérdida nombrada en la primera hora y porcentaje de lecciones convertidas en controles nuevos o modificados. Los dos primeros miden velocidad; los dos últimos, calidad.

La preparación sin práctica no se genera memoria muscular. Un ejercicio de mesa al semestre, con un caso propio y no con un ejemplo de manual, es la forma más económica de descubrir que el dueño del agente no sabe que lo es. El ejercicio debe incluir a la junta en su papel, aunque sea participando por poco tiempo.

Los obstáculos previsibles son cuatro. La propiedad difusa del agente, porque nadie lo siente suyo. La tentación de tratar el incidente como un fallo de modelo y no como un fallo de decisión. La presión por restablecer el servicio antes de probar la corrección. Y la ausencia de registro, que convierte cualquier investigación en conjetura.

8. Marco de responsabilidades

La matriz RACI asigna cuatro letras a cada actividad: *R* es quien ejecuta, *A* quien responde por el resultado, *C* quien debe ser consultado e *I* quien debe ser informado. Una actividad tiene un solo *A*.

Se usa en tres pasos. Primero, la junta aprueba la matriz completa una vez al año. Segundo, el comité de IA la revisa en cada incidente para confirmar que los roles se cumplieron. Tercero, toda actividad sin *A* identificado se escala como hallazgo de gobierno, no como detalle operativo.

La matriz falla de tres maneras conocidas. Cuando varias áreas comparten la *A* y nadie decide. Cuando el dueño del sistema aparece solo como *I* y se entera al final. Y cuando la matriz se aprueba y no se usa. La revisión en cada incidente corrige las tres y asegura su apropiación.

JCM-26 All rights reserved

Tabla 5

Matriz RACI del protocolo CENTINELA-IA

Actividad	Junta	Comité de IA	CISO	Dueño del sistema	Legal	Comunicaciones
Aprobar el apetito de riesgo de IA	**A**	R	C	I	C	I
Nombrar dueño de cada agente	I	**A**	R	C	I	I
Declarar el incidente	I	**A**	R	R	I	I
Suspender el servicio o el agente	**A**	C	R	R	C	I
Conservar la evidencia	I	I	**A**	R	C	I
Notificar a regulador y afectados	**A**	C	C	I	R	R
Aprobar la reanudación	**A**	C	R	R	C	I
Cerrar lecciones y cambiar controles	I	**A**	R	R	C	I

Nota. Elaboración propia. R: ejecuta. A: responde por el resultado. C: se consulta. I: se informa.

9. Recomendaciones para la junta

Siete decisiones que solo la junta puede tomar. Las primeras tres son de preparación; las dos siguientes, de respuesta; las últimas, de aprendizaje.

1. Aprobar un inventario de agentes con dueño nombrado y fecha de revisión, y no aceptar inventarios sin dueño.
2. Fijar el apetito de riesgo para usos agénticos, distinguiendo los que pueden ejecutar acciones irreversibles de los que solo informan.
3. Financiar tres capacidades: canal único de reporte, preservación de registros y ejercicios semestrales.
4. Reservar para la junta la autorización de suspensión de cualquier servicio crítico y de toda comunicación externa.
5. Exigir detalle del incidente con hora de señal (identificación), hora de inmovilización y pérdida nombrada, en una sola página.
6. Pedir que cada lección se convierta en un control modificado con fecha y dueño, no en una recomendación genérica.
7. Revisar el protocolo una vez al año y después de cada incidente grave, con evidencia de la revisión.

10. Conclusiones

El protocolo ofrece lo que los marcos vigentes no articulan: un lenguaje para nombrar la pérdida, una secuencia corta para las primeras horas y un punto de decisión claro para la junta.

JCM-26 All rights reserved

No reemplaza a CoSAI ni a NIST; los usa como base técnica y añade la capa de gobierno que no tienen.

Tres afirmaciones se sostienen con la evidencia disponible: la brecha entre despliegue y capacidad de respuesta está documentada; el vocabulario de la tríada no alcanza para nombrar pérdidas de IA; y la autonomía del agente exige controles de propósito y de herramientas, no solo filtros de contenido. Una cuarta afirmación es una inferencia razonable y que debe ser probada: que este protocolo reduzca el tiempo de contención. Validarlo exige medir antes y después en organizaciones reales, y este es un ejercicio para una investigación futura.

Queda una tarea pendiente que este documento no resuelve: medir. Sin una serie de incidentes con tiempos comparables, el protocolo es una propuesta académica y práctica argumentada. El siguiente paso debería incorporar datos de al menos tres organizaciones y un año de operación.

Referencias

Anthropic. (2025). *Agentic misalignment: How LLMs could be insider threats*. https://www.anthropic.com/research/agentic-misalignment

Anthropic. (2026). *Detecting and countering misuse of AI: September 2026*. https://www.anthropic.com/threat-intelligence-report-september-2026

Bengio, Y., et al. (2025). *International AI safety report 2025*. Department for Science, Innovation and Technology. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025

Cano, J. (2026a). *Beyond the CIA triad: A HEXA cybersecurity framework for artificial intelligence initiatives in organizations* [Ponencia]. IEEE SYSCON 2026. En publicación.

Cano, J. (2026b). *Algorithmic resilience: An integrated model for intelligent sociotechnical systems* [Ponencia]. ICGS3 2026. En publicación

Cano, J. (2026c). The evolution of security and control paradigms. *ISACA Journal, 4*. https://www.isaca.org/resources/isaca-journal/issues/2026/volume-4/the-evolution-of-security-and-control-paradigms

Coalition for Secure AI - CoSAI. (2026). *AI incident response framework* (Version 1.0). https://www.coalitionforsecureai.org/wp-content/uploads/2026/03/AI-Incident-Response-1.pdf

Foro Económico Mundial. (2025). *Global cybersecurity outlook 2025*. World Economic Forum. https://reports.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2025.pdf

Google. (2025). *Secure AI framework (SAIF)*. https://safety.google/intl/es-419_ALL/safety/saif/

IBM Security. (2025). *Cost of a data breach report 2025*. IBM Corporation. https://www.ibm.com/mx-es/think/x-force/2025-cost-of-a-data-breach-navigating-ai

ISO. (2023). *Information technology: Artificial intelligence: Management system* (ISO/IEC 42001:2023). International Organization for Standardization.

Kimmerle, K., & Okeyode, D. (2026). *AI under attack: A practical guide to threats, defenses, and governance for AI systems*. Packt Publishing.

Microsoft. (2026). *2026 Microsoft Digital Defense Report*. Microsoft Threat Intelligence. https://www.microsoft.com/en-us/corporate-responsibility/topics/cybersecurity/reports/microsoft-digital-defense-report/

JCM-26 All rights reserved

Microsoft. (s. f.). *AI security risk assessment*. https://github.com/Azure/AI-Security-Risk-Assessment/blob/main/AI_Risk_Assessment_v4.1.4.pdf

MITRE. (2025). *ATLAS: Adversarial threat landscape for artificial intelligence systems*. https://atlas.mitre.org

National Institute of Standards and Technology - NIST. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1).

National Institute of Standards and Technology - NIST. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1).

National Institute of Standards and Technology - NIST. (2025). *Incident response recommendations and considerations for cybersecurity risk management* (NIST SP 800-61r3).

OCDE. (2024). *Defining AI incidents and related terms*. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/05/defining-ai-incidents-and-related-terms_88d089ec/d1a8d965-en.pdf

OWASP. (2025). *OWASP top 10 for large language model applications*. https://owasp.org/projects/top-10-for-large-language-model-applications

Parker, D. B. (1998). *Fighting computer crime: A new framework for protecting information*. John Wiley & Sons.

Sidhpurwala, H., Mollett, G., Fox, E., Bestavros, M., & Chen, H. (2024). *Building Trust: Foundations of Security, Safety and Transparency in AI* [Manuscrito de Red Hat]. arXiv. https://arxiv.org/pdf/2411.12275

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024). The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. In *arXiv [cs.AI]*. https://doi.org/10.48550/arXiv.2408.12622

JCM-26 All rights reserved

Anexo A. Ficha del incidente de IA

Una página, dieciséis campos. Se abre al declarar el incidente y se cierra cuando la junta aprueba el cambio de control. Los campos 5 y 6 son los que permiten saber si el protocolo funcionó.

Tabla A1

Campos de la ficha de incidente

Campo	¿Qué se registra?
1. Identificador	Código único del incidente y fecha de apertura
2. Sistema o agente	Nombre, versión, proveedor y propósito declarado
3. Dueño	Persona que responde por el sistema, con cargo y contacto
4. Detección	¿Qué señal lo activó?, ¿quién la vio y a qué hora?
5. Declaración	Hora en que se declaró el incidente y quién lo declaró
6. Inmovilización	Hora de la suspensión, medida aplicada y responsable
7. Pérdida nombrada	Elemento del hexágono de Parker afectado y justificación breve
8. Afectados	Personas, áreas o terceros impactados
9. Alcance	Sistemas, datos y herramientas comprometidos
10. Evidencia	Dónde está la copia forense y quién la custodia
11. Notificaciones	¿A quién se informó?, ¿por qué canal y a qué hora?
12. Causa probable	Hipótesis inicial, marcada como hipótesis
13. Acciones	¿Qué se hizo?, ¿en qué orden?, con hora y responsable
14. Reanudación	Prueba de corrección y aprobación escrita
15. Lección	¿Qué se aprendió y qué control cambia?
16. Cierre	Fecha, firma del dueño y visto bueno del comité

Nota. Elaboración propia.

JCM-26 All rights reserved

Anexo B. Guión de las primeras dos horas

El guión ordena la dinámica del incidente. No sustituye el criterio del equipo: fija quién hace qué y en qué orden, para que la discusión no consuma el tiempo de la contención.

Tabla B1

Guión por intervalos

Intervalo	¿Qué se hace?	¿Quién?
0 a 15 minutos	Confirmar la señal, abrir la ficha, avisar al dueño del sistema y al CISO	Quien detecta y el CISO
15 a 30 minutos	Clasificar por severidad; decidir si se declara; suspender si hay daño en curso	Dueño del sistema con el CISO
30 a 60 minutos	Copiar evidencia; retirar herramientas; informar a la junta y a legal	CISO, dueño y legal
60 a 90 minutos	Nombrar la pérdida; fijar el alcance; preparar el primer parte de una página	Comité de IA
90 a 120 minutos	Decidir comunicación externa; asignar responsables de corrección; fijar próxima revisión	Junta con el comité

Nota. Elaboración propia. Los tiempos son de referencia e ilustrativos y deben ajustarse al tamaño y dinámica de la organización.

Tabla B2

Escalamiento según severidad

Severidad	Criterio	¿A quién se informa y en qué plazo?
Alta	Daño en curso, acción irreversible o afectación a personas	Junta en la primera hora; legal de inmediato; regulador según plazo legal
Media	Servicio degradado sin afectación directa a personas	Comité de IA en dos horas; a la junta en el reporte diario
Baja	Anomalía contenida sin pérdida demostrable	CISO y dueño del sistema; registro en el inventario

Nota. Elaboración propia.