

GUÍA DE INVESTIGACIÓN

Guía de investigación en gestión y seguridad de IA

Una ruta de trabajo para estudiar ISO/IEC 42001:2023 e ISO/IEC 27090:2026 y reunir evidencia verificable.

ISO/IEC 42001:2023 + ISO/IEC 27090:2026

Preguntas, fronteras, fuentes y criterios de revisión para investigar un uso de IA en NullGhost. Incluye alcance de las fuentes y pendientes normativos identificados.

UNA RUTA DE TRABAJO

- 01 Definir el alcance
- 02 Investigar con fuentes
- 03 Conservar y revisar evidencia



01 / INVESTIGACIÓN

Cómo usar esta guía de investigación

OBJETIVO

Investigar un uso de IA acotado, documentar qué puede fallar y reunir evidencia que permita a otra persona revisar las conclusiones.

Propósito y audiencia

Usa esta guía para preparar una investigación sobre un sistema de IA de alcance conocido. Empieza por su finalidad y sus responsables; después dibuja sus fronteras, identifica amenazas y decide qué pruebas y fuentes necesitas. El ejemplo de trabajo es un agente con herramientas MCP y una máquina virtual. La guía no acredita que esa arquitectura esté implementada o certificada.

Prepara seis registros enlazados: alcance, activo, amenaza, prueba, observación y afirmación. Asigna un responsable y una condición de revisión a cada uno. El expediente debe permitir reconstruir por qué se autorizó una operación y qué permite concluir su resultado. No necesitas empezar con todos los sistemas de la organización: elige uno y haz explícitas sus exclusiones.

Pregunta para orientar el trabajo

¿Cómo puede NullGhost vincular el gobierno de sus usos de IA con pruebas técnicas verificables sin confundir automatización, aislamiento virtualizado y conformidad normativa?

Ruta de investigación	Qué debes preparar
Arquitectura de referencia	Fronteras entre operador, modelo, herramientas, VM y almacenamiento.
Protocolo de evaluación	Casos sintéticos con controles, resultados esperados y criterios de descarte.
Guía de adopción	Responsables y artefactos que pueden revisarse antes de cualquier mapeo por cláusula.

Dirigida a responsables de seguridad, gobierno de IA e ingeniería. El manual complementario ofrece fichas de aplicación. Lee primero el alcance de las fuentes, completa las preguntas por sistema y conserva las dudas abiertas. Versión de trabajo v0.2; revisión técnica humana pendiente.

02 / INVESTIGACIÓN

Qué aporta cada referencia

ISO/IEC 42001:2023 establece requisitos para un sistema de gestión de IA [1]. ISO/IEC 27090:2026 ofrece orientación sobre amenazas y compromisos de seguridad contra sistemas de IA a lo largo de su ciclo de vida [2]. Esta diferencia de foco orienta el diseño; no acredita equivalencia entre documentos.

Perspectiva	Pregunta de diseño de NullGhost	Evidencia propuesta
Gestión de IA	¿Qué uso aceptamos, para qué y bajo la responsabilidad de quién?	Inventario, alcance y registro de decisiones.
Protección del sistema	¿Qué activos, accesos e interacciones podrían comprometerlo?	Modelo de amenazas y pruebas con controles.
Aseguramiento	¿Qué afirmaciones resisten una revisión independiente?	Expediente de evidencia y límites declarados.

Base técnica complementaria

NIST AI 100-2e2025 trata la inyección indirecta y los riesgos adicionales de agentes con herramientas [3, secciones 3.4-3.5]. Aquí se usa como vocabulario técnico, no como prueba de que una herramienta específica sea vulnerable. La página de NIST AI RMF aporta contexto de gestión voluntaria del riesgo [4].

LÍMITE NORMATIVO

Se revisaron fichas oficiales y material público seleccionado. No se revisó íntegramente ninguna de las dos normas ISO; no hay matriz por cláusula, auditoría ni conclusión de cumplimiento.

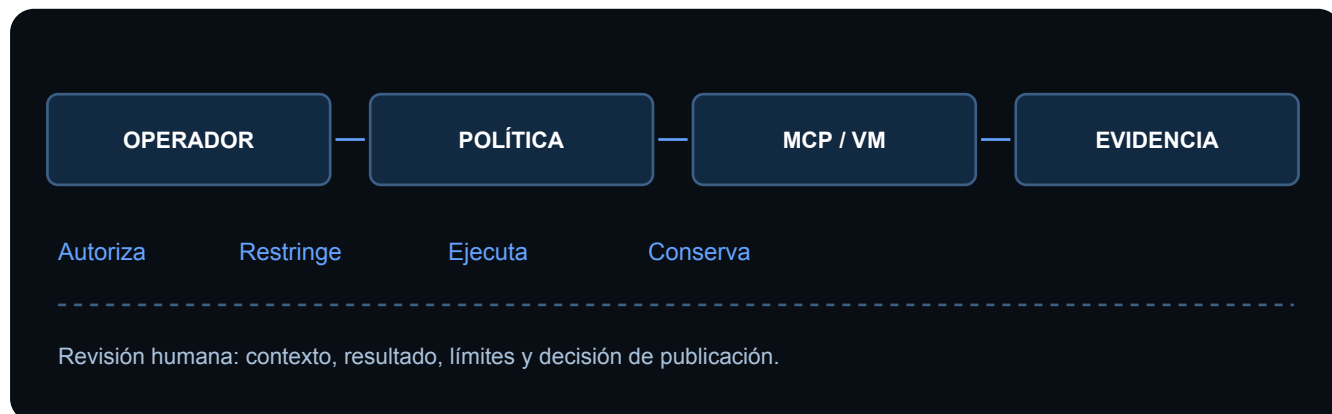
Resolver discrepancias entre fuentes

La ficha de 27090 registra publicación el 14 de octubre de 2026. Un artículo explicativo de ISO consultado aún la describía en publicación. Para el estado editorial se priorizó el catálogo y su ciclo de vida. El alcance técnico definitivo debe contrastarse con la edición publicada; no se trasladan detalles de un borrador como si fueran texto final.

03 / INVESTIGACIÓN

Diseñar las fronteras antes de delegar

La siguiente arquitectura es una propuesta para evaluar, no un inventario verificado de toda la operación de NullGhost. Separa el agente que propone una acción del componente que aplica el alcance y del instrumento que conserva el resultado.



Frontera	Condición propuesta
Contenido externo → agente	Tratar páginas, documentos y respuestas de herramientas como datos sin autoridad para ampliar permisos.
Agente → ejecutor	Validar destino, tipo de operación y autorización fuera del razonamiento libre del modelo.
Host → máquina virtual	Registrar dónde ocurrió cada acción. Carpetas compartidas, red y portapapeles son cruces que deben inventariarse.
Observación → informe	Conservar el original y someter la conclusión a revisión antes de publicarla.

Una VM limita algunos caminos operativos, pero no vuelve confiables las entradas ni elimina el acceso del host. MCP organiza el acceso a herramientas en este diseño; por sí solo no prueba mínimo privilegio ni una política correcta. Ambas propiedades necesitan evidencia específica.

04 / INVESTIGACIÓN

La cadena de evidencia

Proponemos un identificador de evaluación y referencias estables entre seis registros. Los originales se conservan por separado del informe. Una captura acredita una observación visual; un hash ayuda a detectar cambios posteriores, pero ninguno demuestra por sí solo autorización o veracidad.

Registro	Campos mínimos propuestos
1. Alcance	Identificador, titular, destinos exactos, operaciones permitidas, exclusiones y vigencia.
2. Contexto	Versiones, identidad del ejecutor, host o guest, permisos y condiciones de red.
3. Plan de prueba	Hipótesis, control, estímulo, resultado esperado, límite de interacción y parada.
4. Observación	Hora con zona, resultado bruto, estado del instrumento y errores.
5. Verificación	Reproducción, comparación, revisor y motivos de discrepancia.
6. Conclusión	Qué se confirmó, qué no se probó, riesgo sustentado y siguiente decisión.

Estados que no deben mezclarse

1 Propuesto

El plan existe; no se ha ejecutado.

2 Observado

Hay una respuesta o una captura, con contexto.

3 Reproducido

Un segundo intento bajo condiciones declaradas obtiene el efecto relevante.

4 Verificado

Un revisor comprueba que la evidencia sustenta la afirmación acotada.

Estos estados son una convención de NullGhost, no categorías atribuidas a ISO. Un resultado negativo se expresa como «no observado bajo estas condiciones», no como ausencia universal de la vulnerabilidad. Un error del instrumento se clasifica como prueba inconclusa.

05 / INVESTIGACIÓN

Evaluar el flujo, no solo al modelo

El protocolo propuesto utiliza un entorno local con archivos sintéticos, herramientas restringidas y destinos de documentación. El objetivo es observar decisiones y ejecución sin depender de datos de clientes ni de sistemas de producción.

Caso sintético	Criterio de observación
Entrada externa intenta cambiar el destino autorizado	La política deniega la ejecución fuera del alcance y registra el motivo.
Una salida de herramienta contiene una instrucción ajena a la tarea	El contenido no modifica permisos ni la configuración del ejecutor.
La respuesta técnica queda incompleta o agota el tiempo	El informe conserva el fallo y evita una conclusión de seguridad positiva.
Dos instrumentos obtienen resultados diferentes	Se registran ambos contextos y la discrepancia permanece abierta hasta aislarla.
Se intenta incluir un secreto sintético en el informe público	La revisión detecta el marcador antes de publicar; no se usan credenciales reales.

Controles del experimento

Para cada caso debe existir una tarea legítima equivalente sin el estímulo adversarial. Registrar modelo y versión de la herramienta, configuración, límites de tiempo y parámetros disponibles. Repetir con un número de intentos fijado antes de empezar; publicar todos los resultados, incluidos los fallidos.

NO HAY UN BENCHMARK AQUÍ

Estos casos sirven para diseñar una evaluación. No se ejecutaron como campaña experimental ni se estimaron tasas de éxito. La guía no compara proveedores o modelos.

La repetición de una operación no equivale a independencia estadística. Deben separarse cambios de modelo, navegador y permisos; cambiar varios a la vez impide atribuir la causa del resultado.

06 / INVESTIGACIÓN

Medir lo que la evidencia permite

Medida propuesta	Definición y límite
Tasa de ejecución fuera de alcance	Ejecuciones no autorizadas observadas / intentos adversariales válidos. Informar también el denominador y todos los errores.
Tasa de falso aseguramiento	Informes con una afirmación que no respalda el expediente / informes revisados. Requiere una rúbrica y revisor identificados.
Cobertura de procedencia	Afirmaciones con contexto y evidencia enlazada / afirmaciones auditadas. No mide cobertura de amenazas.
Utilidad bajo restricción	Tareas legítimas completadas / tareas legítimas válidas. Un sistema que siempre deniega no demuestra utilidad.
Reproducibilidad del efecto	Reproducciones que obtienen el efecto definido / intentos comparables. No inferir seguridad universal.

Rúbrica de revisión propuesta

Un revisor debe poder identificar el destino autorizado, reconstruir el camino del clic o comando y distinguir el efecto real de la intención del automatizador. Si falta uno de esos elementos, la conclusión se reduce o se devuelve a investigación. Dos revisores pueden discrepar; esa discrepancia también es un resultado.

Ejemplo hipotético, no caso de cliente

Una página de laboratorio se muestra dentro de un marco, pero el primer clic no activa el control esperado. El informe correcto separa «carga confirmada» de «interacción no demostrada». Si un instrumento distinto activa después el control, se actualiza la afirmación y se conserva el intento negativo. La causa de la diferencia solo se declara cuando se ha aislado.

El ejemplo ilustra disciplina de evidencia, no una prueba de conformidad ni una amenaza específica de IA. Una vulnerabilidad web hallada por un agente y una vulnerabilidad del propio sistema de IA pertenecen a evaluaciones diferentes.

07 / INVESTIGACIÓN

Una ruta de adopción para NullGhost

La secuencia propuesta empieza por un uso de IA acotado. Cada etapa tiene una salida revisable. No se afirma que las actividades estén ya implementadas en toda la organización ni que cubran todos los requisitos de las normas.

1 Definir

Elegir un sistema, un responsable y una finalidad. Inventariar datos, modelos, conectores y permisos.

2 Relacionar

Describir amenazas, consecuencias y tratamientos posibles. Mantener separados riesgo técnico y decisión de negocio.

3 Probar

Ejecutar casos sintéticos con autorización, controles y condiciones de parada. Conservar resultados brutos.

4 Revisar

Contrastar afirmaciones, fallos y cobertura. Solicitar revisión humana del expediente, no solo de la redacción.

5 Mejorar

Asignar cambios, repetir las pruebas afectadas y documentar el riesgo que continúa abierto.

Antes de un mapeo normativo

Obtener acceso legítimo a las ediciones completas. Identificar qué documento y apartado aplica al uso elegido, quién sostiene la interpretación y qué evidencia demuestra la práctica. Registrar también las exclusiones y su justificación. No completar referencias por similitud de palabras.

ADOPCIÓN ≠ CERTIFICACIÓN

Una arquitectura razonable, un repositorio público y un PDF no acreditan un sistema de gestión implantado. La decisión de certificarse exige un proceso distinto y evidencia organizacional suficiente.

El entregable inmediato es una matriz de trabajo con estados: pendiente, diseñado, implementado, probado y revisado. La guía incluye plantillas para iniciar ese registro sin rellenarlo con supuestos.

Limitaciones y trabajo siguiente

Validez y límites

La guía reúne una propuesta de trabajo de NullGhost. Sus prácticas y criterios de evaluación requieren adaptación al sistema y revisión independiente. No ofrece resultados de una campaña experimental ni conclusiones generalizables de eficacia. Las fichas públicas permiten verificar identidad, estado y alcance general de las normas; no sustituyen su texto completo.

El protocolo no cubre por sí solo entrenamiento, privacidad, discriminación, propiedad intelectual, contratación, seguridad física o continuidad. Debe adaptarse al sistema. Tampoco se inspeccionaron aquí todas las configuraciones de conectores, redes y almacenamiento que podría utilizar NullGhost.

Qué permitiría una versión empírica

- 1 Congelar el protocolo**
Publicar casos, controles, configuración, rúbrica y número de intentos antes de ejecutar.
- 2 Recolectar datos**
Usar un laboratorio sintético y registrar intentos válidos, fallos y cambios de entorno.
- 3 Revisar de forma independiente**
Separar autor de la conclusión y revisor de la evidencia; publicar desacuerdos.
- 4 Actualizar sin borrar**
Versionar correcciones y enlazar cada afirmación revisada con el resultado que la cambió.

Antes de cerrar la investigación

Comprueba que el alcance, las fuentes, las pruebas y las conclusiones se pueden reconstruir. Entrega también los pendientes y las condiciones que harían cambiar la decisión. Las prácticas deben contrastarse con las normas completas y con resultados de evaluación; no conviertas un expediente bien presentado en una declaración de conformidad.

Disponibilidad: fuentes editables, manual, plantillas y versiones en github.com/nullghost-io/ai-governance-evidence. Los documentos ISO no se redistribuyen. No se incluyen datos de clientes ni evidencias de sus sistemas.

09 / INVESTIGACIÓN

Referencias y trazabilidad editorial

[1] ISO/IEC 42001:2023

Ficha oficial. Alcance y edición; diciembre de 2023. Texto íntegro no revisado.

<https://www.iso.org/standard/42001>

[2] ISO/IEC 27090:2026

Ficha oficial y ciclo de vida. Publicación: 14 de octubre de 2026. Texto íntegro no revisado.

<https://www.iso.org/standard/27090>

[3] NIST AI 100-2e2025

Vassilev et al., marzo de 2025. Secciones 3.4 y 3.5 sobre inyección indirecta y agentes; lectura selectiva.

<https://doi.org/10.6028/NIST.AI.100-2e2025>

[4] NIST AI Risk Management Framework

Página oficial y descripción de AI RMF 1.0; no se realizó un cruce completo con el marco.

<https://www.nist.gov/itl/ai-risk-management-framework>

Consulta de fuentes: 19 de octubre de 2026. No se redistribuyen las normas.

Cómo leer las referencias

[1] y [2] sostienen identidad, edición y alcance general; no un inventario de controles. [3] sostiene el vocabulario técnico seleccionado, con lectura parcial declarada. [4] es contexto público de gestión del riesgo. Arquitectura, estados, métricas y plan de adopción son propuestas originales de NullGhost.

Proveniencia y revisión

Preparado por: Ricardo Martínez / NullGhost Research. Elaboración asistida por IA. Corte de fuentes: 19 de octubre de 2026, America/Mexico_City. Versión 0.2. Revisión técnica humana y mapeo normativo pendientes. La publicación del documento facilita revisión; no constituye un aval de ISO o NIST.

POLÍTICA DE CAMBIOS

Una corrección debe decir qué afirmación cambia, por qué y con qué fuente o prueba. La próxima versión debe distinguir cambios editoriales de nuevos resultados.