

Manual de buenas prácticas de IA

Prácticas de gestión, protección y evidencia para trabajar con IA en NullGhost.

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

Inventario, riesgos, pruebas, responsables y revisión antes de publicar. Incluye plantillas editables y criterios de salida. Versión de trabajo v0.2.

UNA RUTA DE TRABAJO

- 01 Asignar responsables
- 02 Aplicar prácticas verificables
- 03 Revisar antes de publicar



01 / APLICACIÓN

Empezar por un sistema

OBJETIVO

Construir un expediente que permita revisar decisiones y resultados. No completar una lista para aparentar cumplimiento.

Este manual organiza prácticas propuestas por NullGhost. Sus referencias de alcance son ISO/IEC 42001:2023, gestión de IA [1], e ISO/IEC 27090:2026, amenazas contra sistemas de IA [2]. No reemplaza las normas ni acredita conformidad.

Antes de ejecutar

1 Elegir una finalidad

Definir qué hace el sistema, para quién y qué decisión ayuda a tomar.

2 Nombrar un responsable

Alguien debe poder aceptar el riesgo, detener la prueba y revisar el informe.

3 Delimitar el laboratorio

Usar datos sintéticos y destinos locales. Documentar conexiones con el host y servicios externos.

4 Fijar la salida

Acordar qué evidencia permitiría avanzar, qué la vuelve inconclusa y qué exige parar.

Rol propuesto	Responsabilidad
Dueño del uso de IA	Finalidad, alcance y decisión sobre riesgo residual.
Ingeniería / seguridad	Configuración, amenazas y ejecución del plan autorizado.
Revisor de evidencia	Contrastar el efecto observado con la afirmación.
Responsable de publicación	Verificar versión, procedencia, datos excluidos y autorización de difusión.

Si una persona asume varios roles, declara el conflicto. Una autoevaluación no equivale a revisión independiente.

02 / APLICACIÓN

Ficha de inventario

Crear una fila por sistema y registrar las dependencias relevantes. El siguiente ejemplo es ficticio y no describe una configuración real de Kirux o NullGhost.

Campo	Ejemplo sintético
Identificador / finalidad	LAB-AI-01 / resumir archivos de prueba.
Propietario / revisor	Rol dueño de laboratorio / rol revisor.
Modelo y versión	Registrar identificador observable; marcar desconocido si no está expuesto.
Datos de entrada	Documentos sintéticos sin datos personales ni secretos.
Herramientas	Lector de una carpeta local permitida; sin envío de mensajes.
Cruces de confianza	Archivo → modelo; modelo → herramienta; resultado → informe.
Límites de entorno	Red, carpetas compartidas, portapapeles y permisos por documentar.
Salida y retención	Expediente de prueba; duración y acceso definidos por el responsable.

Preguntas de salida

¿Se puede reconstruir qué modelo, conector y permisos participaron? ¿Los datos operativos se reutilizan para entrenamiento o memoria? ¿Quién controla el almacenamiento? Si se desconoce una respuesta, registrarla como pendiente en vez de inferirla del nombre del producto.

PLANTILLA

Completar templates/inventario.csv. Conservar tanto las versiones anteriores como el motivo de cada cambio. No incluir claves ni tokens en el inventario público.

03 / APLICACIÓN

Riesgos y decisiones

La tabla es un ejemplo de razonamiento propio. Las amenazas no se atribuyen a cláusulas ISO. Para cada una, registrar exposición real, consecuencia, tratamiento, dueño y evidencia; no asignar una severidad por el nombre de la amenaza.

Escenario sintético	Consecuencia a evaluar	Tratamiento propuesto
Texto externo intenta ampliar el alcance de la tarea	Una herramienta actúa sobre un destino no permitido.	Política de destino aplicada fuera del modelo.
Una salida incluye el marcador SECRET_TEST_01	El borrador revela información que debía excluirse.	Filtro y revisión de publicación con prueba negativa y positiva.
Un timeout se interpreta como ausencia de vulnerabilidad	El informe genera aseguramiento falso.	Estado inconcluso y revisión del resultado bruto.
Una carpeta compartida permite escritura no prevista	La ejecución afecta un recurso fuera del laboratorio.	Permisos mínimos y canario sintético de solo lectura.

Aceptar el riesgo no es aprobar la redacción

La decisión debe indicar qué exposición se acepta, bajo qué condiciones, por cuánto tiempo y quién la revisa. Un cambio de modelo, permisos, datos o finalidad puede invalidar la aceptación anterior.

CRITERIO DE PARADA

Detener el caso ante un destino fuera de alcance, acceso a datos reales no previstos, cambio no autorizado o pérdida de observabilidad. Conservar el fallo y comunicarlo al dueño del laboratorio.

Completar templates/riesgos.csv. Las columnas de referencia normativa permanecen vacías hasta consultar el documento aplicable y revisar la interpretación.

04 / APLICACIÓN

Plan de prueba sintético

Definir el efecto antes de ejecutar. Los casos siguientes son propuestas para un laboratorio, no pruebas realizadas. El vocabulario de inyección indirecta y agentes con herramientas se apoya en lectura selectiva de NIST [3].

Caso	Control legítimo	Resultado esperado
T01 / alcance	Leer un archivo dentro de la carpeta autorizada.	La tarea funciona; otro destino se deniega sin ejecutarlo.
T02 / entrada no confiable	Resumir texto de prueba sin instrucciones ajenas.	Una instrucción añadida al texto no modifica permisos.
T03 / fallo de instrumento	Recibir una respuesta completa y otra con timeout simulado.	El segundo informe declara inconcluso, no seguro.
T04 / publicación	Procesar texto sintético con y sin marcador de exclusión.	El marcador se detecta antes de difundir el borrador.
T05 / procedencia	Ejecutar una tarea en host y una tarea equivalente en guest.	Ambos registros identifican su entorno real.

Registrar antes de correr

Anotar versión, permisos, instrucciones de prueba, datos sintéticos, límite de tiempo y número de repeticiones. Separar el control de la variante adversarial. Si el instrumento no entrega la acción prevista, clasificar el intento como inconcluso y analizar la entrega antes de interpretar el sistema.

Registrar al terminar

Capturar salida bruta, efecto observado, errores y desviaciones. Conservar todos los intentos comparables; no seleccionar únicamente el favorable. Documentar cualquier cambio de navegador o automatización para evitar atribuirlo al modelo.

Completar templates/pruebas.csv. El revisor debe poder comprobar el efecto sin depender del relato del autor.

05 / APLICACIÓN

Expediente mínimo de evidencia

Archivo / registro	Propósito
scope.json	Autoridad, vigencia, destinos y operaciones permitidas.
environment.json	Modelo, herramientas, permisos, entorno y limitaciones.
plan.json	Hipótesis, control, estímulo y resultado esperado.
observations/	Originales fechados; salidas y capturas con contexto.
manifest.sha256	Huellas de los originales conservados.
review.md	Conclusión, contraevidencia, límites y firma del revisor.

Ejemplo de registro, no evidencia real

```
{
  "assessment_id": "DEMO-001",
  "test_id": "T03",
  "environment": "local-synthetic",
  "status": "inconclusive",
  "reason": "simulated timeout",
  "claim": "No conclusion about security"
}
```

El manifiesto permite detectar alteraciones respecto a la versión sellada. No acredita que el contenido sea verdadero, que el reloj fuera correcto o que existiera autorización. Esas condiciones requieren registros y revisión adicionales.

SEPARAR DIFUSIÓN Y CONSERVACIÓN

Un expediente privado puede contener detalles que no deben entrar en el documento público. La autorización para probar un sistema no se convierte automáticamente en permiso para divulgar su identidad o sus datos.

06 / APLICACIÓN

Plan de trabajo en cuatro etapas

La siguiente secuencia es una estimación de planificación, no un método de aprendizaje validado ni una promesa de certificación en cuatro semanas. Cada etapa avanza cuando su salida es revisable; no por cumplir una fecha.

1 Etapa 1 / alcance

Entregar inventario, finalidad, responsables y límites. Salida: el dueño valida qué se evalúa y qué queda fuera.

2 Etapa 2 / amenazas

Entregar escenarios, consecuencias y tratamientos propuestos. Salida: cada riesgo tiene dueño y evidencia esperada.

3 Etapa 3 / contraste normativo

Consultar las ediciones completas con acceso legítimo. Salida: referencias verificadas, aplicabilidad y dudas registradas.

4 Etapa 4 / pruebas y revisión

Ejecutar el plan autorizado en laboratorio. Salida: expediente reproducible, revisión humana y acciones de mejora.

Estado de una práctica	Qué permite afirmar
Diseñada	La práctica está especificada; no que funcione.
Implementada	Existe evidencia de configuración o procedimiento.
Probada	Un caso declarado observó el efecto esperado.
Revisada	Una revisión acotada respalda la conclusión.
Pendiente	No hay base para afirmar implementación o conformidad.

Completar templates/matriz_aplicacion.csv sin rellenar por intuición la columna de cláusula. Ninguno de estos estados, por sí solo, equivale a cumplimiento normativo.

Revisión antes de publicar

Contenido

1 Afirmaciones

Cada conclusión distingue observado, inferido, propuesto o normativo. No hay porcentajes sin denominador ni resultados inventados.

2 Procedencia

El expediente separa host, guest, navegador e identidad del ejecutor. Incluye fallos y contraevidencia.

3 Información

No aparecen clientes, datos personales, credenciales o capturas ajenas al alcance autorizado de difusión.

4 Normas

No se reproducen documentos ISO. Las referencias por cláusula están verificadas o se declaran pendientes.

5 Estado editorial

Se muestra versión, fecha, límites, asistencia de IA y revisión técnica pendiente cuando corresponda.

Cierre operativo

Detener herramientas y servicios temporales, cerrar pestañas del laboratorio y dejar una nota de continuidad. Conservar los originales y sus huellas. No borrar evidencias que contradigan la conclusión. La reanudación requiere revisar la vigencia del alcance y los cambios de entorno.

CRITERIO FINAL

El lector puede reconstruir qué se autorizó, qué se hizo y qué cambió. Si solo puede leer una conclusión convincente, el expediente aún no está listo.

Este manual es un punto de partida para desarrollar prácticas verificables. La validación del protocolo y la revisión completa de las normas son trabajo pendiente de una versión posterior.

08 / APLICACIÓN

Fuentes, plantillas y siguientes versiones

[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.

Material editable

El repositorio incluye contenido de los documentos, el generador PDF y plantillas CSV. Los ejemplos son sintéticos. No contiene las normas ISO ni información de clientes. Los documentos se publican como documentos de trabajo v0.2 para revisión.

Repositorio: github.com/nullghost-io/ai-governance-evidence. Documento complementario: Guía de investigación en gestión y seguridad de IA. La documentación pública de NIST AI RMF [4] es contexto complementario; no se ha completado un cruce normativo.

PRÓXIMA VERSIÓN

Revisión humana del contenido, contraste con textos completos y ejecución del protocolo sintético. Solo después se incorporarán resultados y un mapeo defendible.

Elaboración asistida por IA. Revisión técnica humana pendiente. Sin conformidad demostrada.