Universidad Cenfotec

Escuela de Ciberseguridad

Tema:

Diseño de un marco forense digital para el análisis de modelos de inteligencia artificial personalizados y operados por terceros en entornos de ciberseguridad

Elaborado por:

Francisco Rene Lara Acute

Fecha: 12 de Noviembre 2025

Tabla de Contenido

Abstract 1
Capítulo 1. Introducción 1
1.1 Generalidades 2
1.2 Antecedentes del Problema 3
1.3 Definición y Descripción del Problema 4
1.4 Justificación 6
1.5 Viabilidad 7
1.5.1 Punto de Vista Técnico. 8
1.5.2 Punto de Vista Operativo. 8
1.5.3 Punto de Vista Económico. 9
1.6 Objetivos 11
1.6.1 Objetivo General. 12
1.6.2 Objetivos Específicos. 12
1.7 Alcances y Limitaciones 12
1.7.1 Alcances 13
1.7.2 Limitaciones. 13
1.8 Revisión de literatura 14
1.8.1 Revisión sistemática 15
1.8.2 Estado de la cuestión 17
Capítulo 2. Marco Conceptual 19
2.1 Inteligencia Artificial en Ciberseguridad 19
2.2 Modelos de IA operados por terceros 20
2.3 Forense Digital en entornos de IA 21
2.4 Evidencia Digital y Cadena de Custodia en IA 21
2.5 Interpretabilidad, Rendición de Cuentas y Auditoría (XAI y Accountability) 21
2.6 Marcos Normativos y Referentes Emergentes 22
2.7 Gobernanza, resiliencia y riesgos operacionales en entornos de inteligencia artificial 22
Capítulo 3. Marco Metodológico 24
3.1 Tipo de Investigación 24
3.2 Alcance Investigativo 24
3.3 Enfoque 24
3.4 Diseño 24
3.5 Población y Muestreo 25
3.6 Instrumentos de Recolección de Datos 25
3.7 Técnicas de Análisis de Información 25
3.8 Estrategia de Desarrollo de la Propuesta 26

Tabla de Figuras

Figura 1 Ejemplo de Nube de Conceptos 6

Índice de Tablas

Tabla 1 *Costos de la investigación* 3

Abstract

La creciente adopción de modelos de Inteligencia Artificial (IA) en ciberseguridad introduce una compleja superficie de ataque, especialmente en modelos operados por terceros con transparencia limitada. Esta investigación identificó, a través de una revisión sistemática de la literatura, que aún no existen metodologías forenses estandarizadas para estos escenarios, lo que refleja una brecha significativa en el campo. Por lo tanto, el objetivo principal de esta investigación fue diseñar un marco forense digital que permita la detección, preservación, análisis y documentación de evidencia de manipulación en dichos modelos, incluso bajo condiciones de acceso restringido.

La metodología integra principios de la forense digital clásica, la seguridad de la IA y la interpretabilidad (*XAI*) para formular un marco estructurado y adaptable. Este se implementó en un prototipo funcional y se validó mediante simulaciones de ataques como *data poisoning* y *prompt injection* sobre modelos de código abierto.

Los hallazgos demuestran que el marco es una herramienta viable y eficaz para la investigación post-incidente, permitiendo trazar el origen de la manipulación. Se concluye que esta investigación aporta una solución pionera y fundamental para la rendición de cuentas (*accountability*) de los sistemas de IA, dotando a los peritos de una capacidad técnica y procedimental que fortalece la cadena de custodia y la validez jurídica de la evidencia digital en este dominio emergente.

Palabras Clave: inteligencia artificial, ciberseguridad, marco forense, análisis forense, evidencia digital, cadena de custodia

Capítulo 1. Introducción

En los últimos años la Inteligencia Artificial (IA) se ha consolidado como un pilar estratégico en la ciberseguridad, con aplicaciones que abarcan desde la detección de amenazas hasta la automatización de respuestas. Sin embargo, esta integración no está exenta de riesgos: técnicas de manipulación como el *data poisoning*, los *adversarial samples* o la inyección de instrucciones maliciosas pueden alterar el comportamiento de un modelo sin dejar señales evidentes. La situación se complica cuando los modelos son provistos y operados por terceros, pues la falta de acceso directo y la limitada transparencia incrementan la dificultad de evaluar su integridad.

Esta combinación de dependencia tecnológica y opacidad genera un vacío crítico en la capacidad de respuesta post-incidente. La mayor parte de la literatura y las soluciones técnicas se concentran en la prevención y protección de modelos, mientras que los procesos de investigación forense aplicados a IA permanecen escasamente explorados. Esta carencia se traduce en una ausencia de marcos formales que permitan a los peritos identificar, documentar y sustentar técnicamente la manipulación de modelos, comprometiendo no solo la integridad de la investigación, sino también la cadena de custodia y la admisibilidad de la evidencia en entornos jurídicos.

En este contexto, surge la necesidad de desarrollar metodologías forenses específicas para entornos de IA, capaces de adaptarse a escenarios con restricciones de acceso y de operar en condiciones de alta incertidumbre. Esta investigación se plantea como respuesta a ese reto, con el propósito de ofrecer un marco estructurado que sirva como guía para la detección y análisis de incidentes que involucren modelos de IA personalizados y gestionados por terceros en entornos de ciberseguridad.

1.1 Generalidades

El presente trabajo de tesis se inscribe en un dominio de investigación emergente y de vanguardia: el análisis forense de modelos de Inteligencia Artificial (IA). Es fundamental aclarar que, al momento de realizar esta investigación, no existen marcos forenses estandarizados y ampliamente aceptados que aborden

específicamente el análisis post-incidente de modelos de IA operados por terceros. Dada esta novedad, la investigación adopta un enfoque constructivo, integrando principios de la forense digital tradicional, la seguridad de la IA y la interpretabilidad de modelos de aprendizaje automático para proponer una nueva metodología. Por "Forense de IA", en el contexto de este documento, se entenderá el conjunto de procesos técnicos y metodológicos aplicados post-incidente para identificar, preservar, analizar y presentar evidencia de compromiso o manipulación en un modelo de IA.

Para garantizar la viabilidad, replicabilidad y ética de la investigación, es importante señalar que, para el desarrollo del prototipo y los escenarios de validación descritos en los objetivos, no se utilizarán datos sensibles ni modelos propietarios sujetos a cláusulas de confidencialidad de ninguna organización. En su lugar, se emplearán exclusivamente modelos de código abierto y conjuntos de datos de dominio público que son reconocidos en la comunidad de ciberseguridad. Este enfoque permite una experimentación transparente y evita cualquier conflicto de interés o violación de la privacidad, al tiempo que asegura que los resultados puedan ser validados por otros investigadores.

El análisis de riesgos y vulnerabilidades se fundamenta en las tendencias y proyecciones más actuales de la industria y la academia sobre la evolución de los ataques a sistemas de IA. El objetivo es asegurar que el marco propuesto no solo responda a las amenazas conocidas, sino que también posea la flexibilidad para adaptarse al panorama de ciberseguridad contemporáneo y del futuro cercano (Gartner, 2024).

1.2 Antecedentes del Problema

La respuesta de la comunidad académica y de la industria a las vulnerabilidades inherentes a los modelos de Inteligencia Artificial (IA) se ha centrado, hasta la fecha, predominantemente en la defensa proactiva y la robustez estructural de dichos modelos. Los principales esfuerzos se han dirigido al desarrollo de arquitecturas más resilientes, la implementación de mecanismos de entrenamiento adversario para inmunizar modelos frente a ataques conocidos, y al estudio teórico de superficies de ataque con el fin de anticipar vulnerabilidades (Carlini et al., 2023; Tramèr et al., 2022).

Según el informe *Cost of a Data Breach* 2025 de IBM, el 29% de los incidentes de seguridad relacionados con IA provienen de modelos entregados por terceros como SaaS, lo que evidencia un aumento de riesgos en soluciones de caja negra donde las organizaciones no tienen control total sobre los modelos. Esta situación plantea un desafío particular para la trazabilidad, auditoría y análisis forense en entornos cerrados.

Este enfoque ha resultado valioso para fortalecer las capacidades preventivas de los sistemas de IA, particularmente en lo que respecta a su seguridad durante el desarrollo y la operación normal. Sin embargo, el componente reactivo y forense, es decir, la capacidad de realizar análisis técnico, legal y pericial después de un incidente ha sido escasamente abordado. La literatura académica y técnica evidencia que este campo es aún emergente, y no se ha convertido en una práctica común en los marcos de aseguramiento y gobernanza de la IA.

Esta falta de atención a la dimensión post-incidente representa un vacío crítico, especialmente en un contexto donde los modelos de IA se despliegan en entornos de alta sensibilidad, como la ciberseguridad, el sector financiero o la salud. La siguiente sección profundiza en este problema, definiéndolo con claridad y detallando su impacto en la práctica.

1.3 Definición y Descripción del Problema

El uso de modelos de Inteligencia Artificial (IA) en entornos de ciberseguridad se ha incrementado significativamente en los últimos años, especialmente en tareas como detección de intrusiones, análisis de comportamiento anómalo y respuesta automatizada a amenazas. Según un informe de Gartner (2024), más del 70 % de las organizaciones que utilizan herramientas de ciberseguridad ya integran componentes de IA para tomar decisiones operativas en tiempo real. No obstante, estos modelos no están exentos de vulnerabilidades técnicas ni riesgos operativos. Investigaciones recientes (Carlini et al., 2023; Tramèr et al., 2022) han documentado ampliamente ataques como *data poisoning*, *model inversion*, *prompt injection* y *adversarial samples*, que pueden comprometer la integridad y funcionalidad de los modelos.

La creciente adopción de modelos de IA operados por terceros y ofrecidos como servicios SaaS ha derivado en un aumento significativo de incidentes de

seguridad. De hecho, estos modelos representan la fuente más común de incidentes reportados (29%), superando incluso a los modelos in-house y de código abierto (26%) (IBM, 2025). No obstante, los marcos forenses actuales no están preparados para abordar eficazmente este tipo de entornos limitados en acceso y visibilidad.

Este problema se agrava en escenarios donde los modelos no son desarrollados ni gestionados internamente, sino que son proporcionados o personalizados por terceros, como en soluciones *SaaS* o servicios de IA en la nube. En estos casos, el acceso a artefactos técnicos, registros de entrenamiento, configuraciones de despliegue y logs de inferencia es parcial o nulo, lo que limita significativamente la capacidad de respuesta ante incidentes.

Aunque existen herramientas comerciales como *Granite Guardian* y *Guardium AI Security*, orientadas a la supervisión y detección de riesgos operativos en modelos de IA, no existe actualmente un marco forense estandarizado ni una metodología validada que permita llevar a cabo un análisis técnico, jurídico y pericial posterior a un incidente de seguridad. Este vacío metodológico ha sido también identificado por marcos emergentes:

- El *OWASP Top 10 for LLM Applications* (2024) resalta como riesgo la ausencia de monitoreo e *incident response* (*LLM09*) y de mecanismos de auditoría frente a divulgaciones indebidas de información (*LLM06*), sin ofrecer aún guías forenses aplicables.
- El *NIST AI 600-1* (2024) destaca entre sus 12 riesgos clave la respuesta inadecuada a incidentes (riesgo 10) y la falta de protección estructural de los modelos (riesgo 12), reiterando la urgencia de mecanismos post-compromiso, sin ofrecer soluciones operativas concretas.
- La NSA, en su guía *Data Security in AI Systems – Guidance for Forensic Readiness* (2024), propone lineamientos generales para preparar modelos de IA con fines forenses (e.g., *logging*, trazabilidad), pero su implementación aún no se ha traducido en metodologías aplicables ni en herramientas estandarizadas.

Una revisión sistemática de la literatura académica en bases como SpringerLink e IEEE Xplore (2019–2025) confirma esta situación: si bien existe

abundante documentación sobre protección de modelos y robustez ante ataques, el componente forense sigue siendo incipiente o inexistente. Esta carencia limita gravemente la capacidad de:

- Determinar técnicamente cómo y cuándo fue comprometido un modelo.
- Evaluar el impacto legal del incidente.
- Establecer evidencia pericial con valor jurídico.

En consecuencia, las organizaciones enfrentan serios desafíos para actuar una vez que un modelo ha sido comprometido, especialmente cuando estos operan en entornos cerrados, de terceros o sin visibilidad completa de su operación interna. Esta situación representa una brecha crítica para la gobernanza algorítmica, la respuesta a incidentes y la administración de justicia digital.

Por tanto, se hace necesario el diseño de un marco forense digital especializado, que permita aplicar principios de preservación, análisis y documentación de evidencia digital a modelos de IA comprometidos, incluso en condiciones de acceso parcial o externo. Este marco debe contemplar tanto entornos abiertos (*on-premise* o locales) como cerrados (*SaaS* o nube). El presente trabajo busca precisamente responder a esa necesidad mediante la construcción de una propuesta metodológica forense aplicable a modelos de inteligencia artificial vulnerados.

1.4 Justificación

La presente investigación se justifica por su potencial para crear una capacidad técnica y procedimental actualmente inexistente en el campo de la ciberseguridad. La innovación fundamental de este trabajo no radica en la mejora de un proceso existente, sino en la propuesta para la creación de la primera metodología formal y estandarizada para la investigación forense de modelos de Inteligencia Artificial (IA), especialmente aquellos que operan como "cajas negras" gestionadas por terceros. Mientras los esfuerzos actuales se concentran en la prevención de ataques, esta tesis llena el vacío crítico del análisis post-incidente, proporcionando el "qué hacer después" cuando las defensas han fallado.

Desde una perspectiva operativa y financiera, el aporte es directo y cuantificable. Un modelo de IA en ciberseguridad que ha sido manipulado, por

ejemplo, un sistema de detección de intrusiones que ignora un ataque real representa un riesgo catastrófico que puede derivar en brechas de datos masivas. El marco propuesto impacta directamente en la reducción de costos asociados a estos incidentes al:

- Acelerar drásticamente el tiempo de respuesta, proporcionando a los analistas un procedimiento claro en lugar de una investigación *ad-hoc*, caótica y costosa.
- Permitir una atribución precisa de la causa raíz, lo que es crucial para remediar la vulnerabilidad, prevenir su recurrencia y evitar pérdidas financieras futuras.
- Minimizar el impacto financiero total de una brecha, que según informes de la industria, puede ascender a millones de dólares por incidente (IBM, 2024).

La justificación estratégica y jurídica es quizás la más relevante. En un ecosistema tecnológico que avanza hacia una regulación estricta de la IA, como lo evidencia la *Ley de Inteligencia Artificial de la Unión Europea* (Comisión Europea, 2024), la capacidad de auditar y demostrar la integridad de un modelo no es una opción, sino una necesidad de cumplimiento. Este trabajo habilita un mecanismo de rendición de cuentas (*accountability*) para los proveedores de IA. Proporciona a las organizaciones la capacidad de producir evidencia digital con validez técnica y potencial validez jurídica, fortaleciendo la cadena de custodia en un dominio donde actualmente es frágil o inexistente.

Esta investigación no solo asegura un proceso técnico, sino que también sienta las bases para la confianza, la transparencia y la responsabilidad en la próxima generación de ciberseguridad impulsada por IA.

1.5 Viabilidad

Este apartado analiza la factibilidad de la presente investigación desde tres perspectivas clave: técnica, operativa y económica, con el fin de asegurar que el proyecto es realizable en el tiempo y con los recursos estipulados. Es importante aclarar que este análisis no se refiere a la viabilidad de implementar la solución propuesta, sino exclusivamente a la posibilidad real de desarrollar la investigación de forma exitosa.

1.5.1 Punto de Vista Técnico.

Desde una perspectiva técnica, la investigación es plenamente viable. El investigador cuenta con formación especializada en ciberseguridad, análisis forense digital y cumplimiento normativo, así como experiencia práctica en entornos de seguridad ofensiva y defensiva.

Se dispone de entornos virtuales controlados, con acceso a plataformas como:

- *Tsurugi Linux*, una distribución forense especializada que incluye herramientas para análisis de discos, memoria, red y entornos digitales complejos.
- *Caine*, *Autopsy*, *Volatility*, *ApexSQL Log* y *Wireshark*, complementan el arsenal forense para analizar evidencias generadas por modelos de IA en entornos simulados.
- Plataformas de entrenamiento y monitoreo como *TensorBoard* o *MLFlow*, útiles para visualizar el comportamiento de modelos durante pruebas controladas.

La familiaridad con marcos como *NIST SP 800-150*, *ISO/IEC 27037*, *OWASP LLM Top 10* y directrices recientes de la NSA refuerza la base teórica y metodológica para estructurar un marco forense robusto.

1.5.2 Punto de Vista Operativo.

En el plano operativo, el proyecto cuenta con condiciones favorables y recursos disponibles para su ejecución rigurosa y controlada. La investigación se desarrollará en un entorno académico con acceso institucional a bases de datos científicas como IEEE Xplore, SpringerLink, Scopus y Google Scholar, lo cual garantiza el respaldo documental y bibliográfico necesario para sustentar el marco propuesto.

Además, se cuenta con el acompañamiento de un tutor académico especializado en ciberseguridad y análisis forense, así como con la posibilidad de recibir retroalimentación puntual de profesionales del sector mediante sesiones de validación técnica del marco forense.

Para los experimentos y pruebas controladas, se utilizarán datasets abiertos y reconocidos en la comunidad de investigación, específicamente:

- *CIC-IDS-2017* y *IDS2017*, ambos generados por el *Canadian Institute for Cybersecurity*. Estos conjuntos de datos contienen tráfico realista simulado con múltiples tipos de ataques (e.g., *DDoS*, *brute force*, *botnets*, *infiltration*, *web attacks*), y serán utilizados para simular escenarios de compromiso de modelos de IA en entornos controlados.
- Los datos serán procesados dentro de entornos virtuales, y el comportamiento de los modelos de IA frente a estos datos permitirá analizar las evidencias digitales generadas ante diferentes tipos de manipulación o ataques.

Todo el trabajo experimental se desarrollará en laboratorios digitales locales utilizando las herramientas forenses especificadas en la sección 1.5.1 Punto de Vista Técnico de este documento, lo cual elimina la necesidad de acceder a infraestructura ajena o modelos productivos, asegurando el cumplimiento de principios éticos y de privacidad.

En este contexto, se garantiza un entorno operativo completamente autónomo, seguro, y libre de restricciones legales o administrativas, lo que refuerza la factibilidad de realizar la investigación dentro del marco académico y metodológico establecido.

1.5.3 Punto de Vista Económico.

La investigación ha sido diseñada para ejecutarse con un presupuesto mínimo, priorizando el uso de software libre o versiones comunitarias. Se evita cualquier costo asociado a licencias comerciales o hardware especializado.

A continuación, se presenta un desglose del costo teórico del proyecto, utilizando estimaciones basadas en el mercado de Estados Unidos.

Tabla 1

Costos de la investigación

Rubro	Costo estimado (USD)
Horas de Investigador	16 USD/h × 150 h = $ 2, 400
Equipo computacional (laptop con GPU dedicada, 32GB RAM)	$1,200
Acceso a bases de datos académicas (anual)	$200
Compra de libros y textos especializados	$150

Electricidad e internet (uso intensivo por 6 meses)	$180
Horas de consultoría técnica (validación puntual, 10h a $30/h)	$300
Transporte y logística (eventos, entrevistas, reuniones)	$100
Software y herramientas especializadas (licencias menores, si aplica)	$100
Total estimado	**$4,630.00**

La estimación de costos se presenta en dólares estadounidenses (USD) debido a que gran parte de los rubros considerados (hardware especializado, licencias de software, acceso a bases de datos académicas y libros técnicos) corresponden a bienes y servicios cotizados en ese mercado. Al tratarse de una disciplina tecnológica con fuerte dependencia de recursos internacionales, el dólar representa una moneda de referencia estándar y ampliamente utilizada en el sector de ciberseguridad. Además, esta decisión permite una comparación objetiva y coherente con otras investigaciones académicas y presupuestos del mismo ámbito.

Explicación de costos

- **Equipo computacional:** se estima un equipo con GPU dedicada (NVIDIA o equivalente), 32 GB de RAM y almacenamiento SSD suficiente para ejecutar simulaciones de modelos de IA y herramientas forenses.
- **Acceso a bases de datos académicas:** se contempla una suscripción individual o institucional a plataformas como IEEE Xplore, SpringerLink o ACM Digital Library, necesarias para el acceso a publicaciones clave.
- **Libros y textos especializados**: se asigna un presupuesto para libros de ciberseguridad, análisis forense y ética de IA que no estén disponibles de forma gratuita.
- **Electricidad e internet:** el uso prolongado de procesamiento local justifica un ajuste por consumo energético y conectividad durante el desarrollo del proyecto.

- **Consultoría técnica:** se prevé la contratación de apoyo experto para validar el marco forense propuesto o para revisión metodológica, calculado a una tarifa local moderada.
- **Transporte y logística:** cubre desplazamientos puntuales relacionados con la investigación.
- **Licencias de software:** se reserva un pequeño monto para licencias no cubiertas por versiones comunitarias, si fuese necesario.

1.6 Objetivos

Esta investigación se enmarca en el nivel de síntesis de la taxonomía cognitiva de Bloom (1956), la cual establece una jerarquía de habilidades intelectuales que van desde el conocimiento básico hasta la evaluación y creación de nuevos contenidos. El nivel de síntesis se define como la capacidad de combinar elementos o partes de conocimientos previos para formar un todo coherente y funcional, generando estructuras originales a partir de conceptos existentes.

El objetivo general de esta tesis consiste en diseñar un marco forense digital aplicable al análisis de modelos de inteligencia artificial personalizados y operados por terceros en entornos de ciberseguridad. Dicho propósito requiere integrar conocimientos técnicos, legales y operativos en ciberseguridad, forense digital e inteligencia artificial, con el fin de construir un nuevo enfoque metodológico que responda a una necesidad real aún no abordada en la literatura académica ni en las soluciones comerciales actuales.

Además del diseño del marco, la investigación contempla el desarrollo de un prototipo funcional, así como su validación mediante pruebas simuladas. Estas acciones reflejan claramente la aplicación del pensamiento de orden superior definido en la categoría de síntesis, ya que no solo se integran conocimientos previos, sino que se utilizan activamente para crear un producto nuevo, útil y replicable.

Por tanto, la selección del nivel de síntesis como enfoque cognitivo rector de esta tesis permite sostener la coherencia entre los objetivos planteados, el desarrollo metodológico y la naturaleza innovadora de la propuesta.

1.6.1 Objetivo General.

Diseñar un marco forense digital aplicable al análisis de modelos de inteligencia artificial personalizados y operados por terceros en entornos de ciberseguridad, integrando técnicas de preservación, análisis y documentación de evidencia digital para detectar manipulaciones o compromisos en dichos modelos.

1.6.2 Objetivos Específicos.

1. Identificar e integrar información relevante sobre las vulnerabilidades, riesgos y técnicas de ataque más comunes al año 2025 que afectan a modelos de inteligencia artificial utilizados en ciberseguridad, en particular aquellos operados externamente o personalizados por terceros.
2. Formular un marco metodológico forense estructurado, adaptado tanto a entornos con acceso completo como limitado, que permita recolectar, preservar y analizar evidencia digital en modelos de IA comprometidos.
3. Construir un prototipo funcional que implemente el marco forense propuesto, utilizando herramientas accesibles y bibliotecas de simulación para ejecutar pruebas controladas sobre modelos de IA personalizados.
4. Aplicar y validar el marco propuesto mediante escenarios simulados, comparando el comportamiento de modelos de IA antes y después de ataques específicos para evaluar su capacidad de detección, trazabilidad e integridad de evidencia.

1.7 Alcances y Limitaciones

Esta investigación se enfoca en el diseño y validación de un marco forense digital especializado para el análisis post-incidente de modelos de inteligencia artificial (IA) operados por terceros en entornos de ciberseguridad. Sus principales alcances son los siguientes:

1.7.1 Alcances

- Diseño metodológico: Se desarrolla un marco forense estructurado que integra principios de la forense digital tradicional, seguridad en IA e interpretabilidad

(*XAI*), permitiendo la detección, preservación, análisis y documentación de evidencia digital tras incidentes de manipulación de modelos.

- Aplicación a modelos personalizados y cerrados: El marco está orientado a escenarios donde el modelo no es de desarrollo interno, sino que ha sido personalizado o provisto por terceros (e.g., proveedores *SaaS*, plataformas en la nube). Esto incluye entornos de acceso restringido donde no se tiene visibilidad completa del entrenamiento o despliegue.
- Simulación y prototipado: Se construye y valida un prototipo funcional del marco mediante la aplicación de ataques simulados (como *data poisoning* y *prompt injection*) sobre modelos de código abierto en entornos controlados.
- Uso de *datasets* públicos: Para las pruebas y validación se emplean datos de referencia reconocidos en la comunidad de ciberseguridad, como *CIC-IDS-2017* e *IDS2017*, garantizando así la transparencia, replicabilidad y ética del proceso experimental.
- Contribución a la trazabilidad y la cadena de custodia: El enfoque propuesto tiene como objetivo dotar a los peritos digitales de capacidades reales para establecer el origen, tipo e impacto de una manipulación, fortaleciendo la cadena de custodia en escenarios donde actualmente no existen procedimientos claros.

1.7.2 Limitaciones.

Dado el carácter exploratorio e innovador de esta investigación, existen ciertas limitaciones que es importante reconocer:

- Acceso restringido a modelos comerciales reales: Aunque el objetivo del marco es ser aplicable a modelos de IA operados por terceros, por razones legales, éticas y de confidencialidad, esta investigación no accede a modelos cerrados o propietarios en entornos reales. En su lugar, se emplean modelos de código abierto en condiciones de acceso parcial simuladas, diseñadas intencionalmente para replicar las limitaciones de transparencia, visibilidad y gobernanza que caracterizan a los entornos *SaaS* o servicios gestionados externamente. Esta estrategia asegura la coherencia metodológica del marco

sin comprometer la validez del diseño experimental ni los principios éticos de la investigación.

- Alcance legal interpretativo: Si bien el marco contempla componentes técnicos y procedimentales con potencial valor jurídico, no pretende sustituir el trabajo legal formal, ni garantizar la admisibilidad automática de la evidencia en tribunales. Su función es fortalecer la trazabilidad técnica, facilitando el trabajo posterior de los actores jurídicos competentes.
- Enfoque limitado a IA en ciberseguridad: Esta investigación se restringe específicamente al ámbito de la ciberseguridad, donde los modelos de IA suelen operar bajo condiciones de criticidad, opacidad operativa y gestión externa. Si bien el marco podría adaptarse a otros dominios (como salud o finanzas), su aplicación en dichos contextos no es abordada en este estudio.
- Dependencia de entornos simulados: Dado que los escenarios de ataque y análisis se desarrollan en laboratorios controlados, algunos aspectos del comportamiento de los modelos o las limitaciones prácticas de acceso a registros y configuraciones reales pueden no reflejar completamente la complejidad de los entornos de producción. Sin embargo, la simulación se ajusta cuidadosamente a las condiciones esperadas en entornos reales de terceros, manteniendo la relevancia y utilidad del marco.

1.8 Revisión de literatura

La presente investigación requiere una base teórica sólida sobre la seguridad y el análisis forense en sistemas de inteligencia artificial, particularmente en contextos donde los modelos son operados por terceros. Para establecer esta base, se realizó una revisión sistemática de la literatura siguiendo el enfoque metodológico propuesto por Biolchini, Gomes, Cruz y Horta (2005), con el objetivo de identificar el estado actual del conocimiento, brechas existentes y oportunidades de investigación en el área.

1.8.1 Revisión sistemática

a) Pregunta de investigación

¿Existen marcos metodológicos o propuestas formales que permitan realizar análisis forense post-incidente sobre modelos de inteligencia artificial, especialmente aquellos operados por terceros o en entornos de acceso restringido?

b) Fuentes de información

Se seleccionaron cuatro bases de datos académicas ampliamente reconocidas en el ámbito de la computación y la ingeniería:

- IEEE Xplore
- SpringerLink
- Scopus
- Google Scholar

Estas fuentes aseguran el acceso a literatura revisada por pares, publicaciones técnicas, y documentos normativos relevantes.

c) Términos de búsqueda

Se definieron las siguientes combinaciones de palabras clave en inglés:

- "*AI forensics*"
- "*forensic analysis of AI*"
- "*post-incident response in AI*"
- "*digital evidence AI*"
- "*third-party AI model analysis*"
- "*auditability AI cybersecurity*"
- "*forensic readiness AI*"

Las búsquedas se limitaron al periodo 2019–2025, con énfasis en publicaciones recientes que reflejaran el estado más actual del campo.

d) Criterios de inclusión

- Artículos académicos revisados por pares.
- Estudios que aborden modelos de IA aplicados a la ciberseguridad.
- Documentos que presenten propuestas, herramientas o marcos metodológicos relacionados con análisis forense, trazabilidad o auditoría de modelos de IA.

- Publicaciones en inglés o español, con acceso completo.

e) Criterios de exclusión

- Trabajos centrados exclusivamente en defensa o prevención (por ejemplo, robustez o entrenamiento adversario) sin componentes post-incidentes.
- Revisión teórica sin propuesta metodológica clara.
- Estudios enfocados en otros dominios no relacionados con IA o seguridad digital.

f) Proceso de selección y resultados

Se recuperaron inicialmente 96 documentos. Tras una evaluación por título, resumen y lectura rápida del contenido, se redujo el corpus a 46 publicaciones relevantes. Estas fueron leídas completamente y analizadas cualitativamente.

g) Resultados

Los principales hallazgos se pueden agrupar en los siguientes temas:

- Existe abundante documentación sobre amenazas activas a modelos de IA, como *data poisoning*, *prompt injection*, *adversarial examples* y *model inversion* (Carlini et al., 2023; Tramèr et al., 2022).

- El enfoque predominante es preventivo, orientado a la robustez del modelo, detección de anomalías y mitigación antes del compromiso.

- Solo un pequeño subconjunto de estudios menciona términos como "*forensic readiness*" o "*digital evidence*" en IA, y aún menos desarrollan metodologías aplicables.

- No se encontraron propuestas completas que guíen la preservación, análisis y documentación forense de incidentes en modelos operados por terceros o cerrados.

- Marcos técnicos recientes como *OWASP LLM Top 10* (2024), *NIST AI 600-1* (2024) y la *guía de la NSA* (2024) abordan el tema desde la perspectiva del

riesgo y la preparación, pero sin aportar metodologías detalladas de análisis forense post-incidente.

1.8.2 Estado de la cuestión

A partir de la revisión sistemática, se concluye que:

- La comunidad académica y técnica reconoce ampliamente los riesgos que enfrentan los modelos de IA en ciberseguridad, particularmente en su vulnerabilidad a manipulaciones como el *data poisoning* o la evasión de detección.

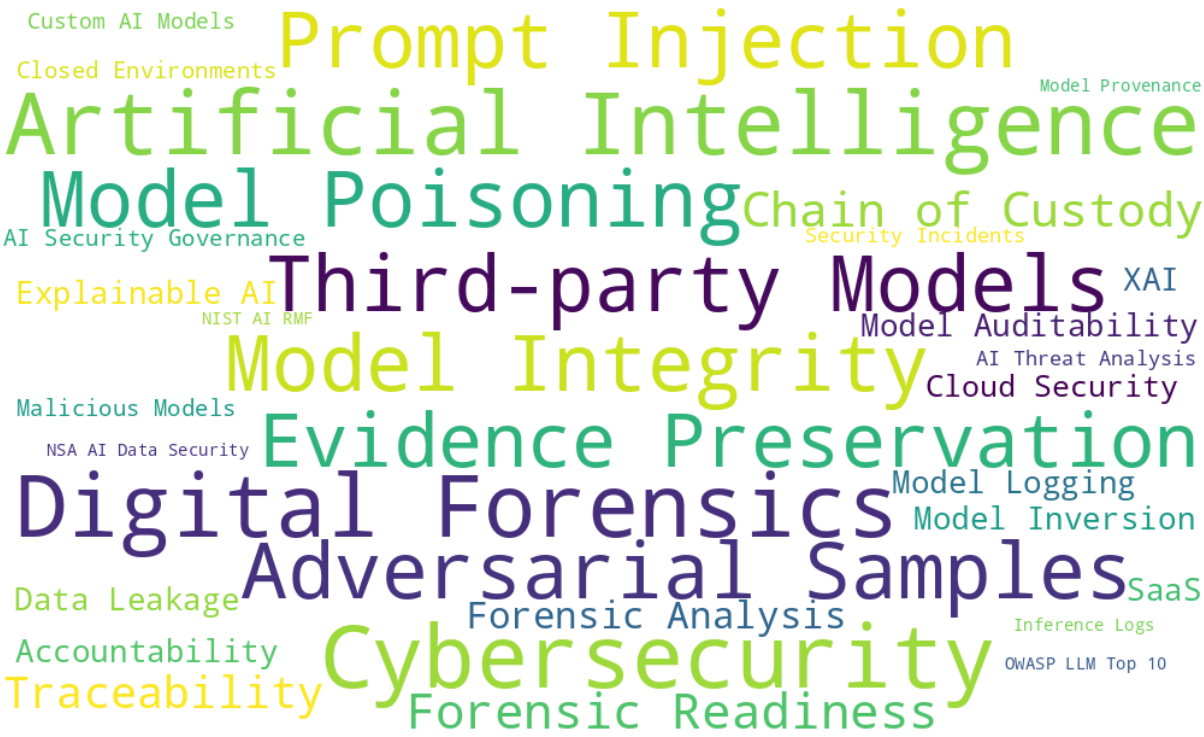
- El enfoque predominante es preventivo, con desarrollos importantes en robustez, entrenamiento adversario y control de calidad de modelos. No obstante, la dimensión post-incidente sigue siendo una brecha crítica.

- No existe una metodología forense estandarizada o validada para aplicar en entornos donde los modelos son operados por terceros, como plataformas *SaaS* o soluciones *cloud*, donde el acceso a los componentes técnicos del modelo es limitado.

- Los marcos normativos más actuales (*OWASP, NIST, NSA*) reconocen el problema pero no lo resuelven metodológicamente, dejando el campo abierto para investigaciones como esta que proponen construir un marco específico, replicable y enfocado en la trazabilidad y preservación de evidencia digital en modelos de IA comprometidos.

Capítulo 2. Marco Conceptual

Este capítulo presenta el marco conceptual de la investigación, el cual se construye a partir de una nube de conceptos clave y un mapa conceptual jerárquico (ver Figura 1). Se desarrollan y relacionan los principales términos técnicos, operativos y normativos que sustentan el diseño del marco forense digital propuesto. Dado que se trata de una investigación aplicada, se opta por la construcción de un marco conceptual en lugar de un marco teórico.



2.1 Inteligencia Artificial en Ciberseguridad

La inteligencia artificial (IA) ha transformado profundamente la manera en que se detectan, analizan y mitigan las amenazas en entornos digitales. Su aplicación en ciberseguridad ha permitido el desarrollo de sistemas más autónomos y eficaces, como los sistemas de detección de intrusos (IDS), análisis de comportamiento y motores de respuesta automática. Sin embargo, la creciente dependencia de estos sistemas también los expone a nuevos vectores de ataque. Los modelos de IA, especialmente aquellos entrenados con grandes volúmenes de datos, pueden ser manipulados mediante técnicas como *data poisoning*, *adversarial samples* y *model inversion*, comprometiendo así su funcionalidad y los resultados que generan. Estas

amenazas requieren nuevos enfoques de protección y, más importante aún, metodologías forenses adaptadas a su análisis post-incidente.

2.2 Modelos de IA operados por terceros

El uso de modelos desarrollados o personalizados por proveedores externos ha aumentado como parte de la adopción de IA como servicio (*AIaaS*). En el contexto de la ciberseguridad, los modelos de inteligencia artificial (IA) ofrecidos como software as a service (SaaS) se caracterizan por ser soluciones alojadas y operadas por terceros, donde las organizaciones usuarias interactúan únicamente a través de interfaces o APIs, sin tener acceso directo al modelo subyacente ni a su infraestructura. Esta modalidad introduce el concepto de IA como caja negra, donde tanto el comportamiento interno como la lógica de decisión del modelo son opacos para el consumidor, lo que limita la capacidad de validación, auditoría o intervención directa en caso de incidentes.

Desde una perspectiva de gobernanza y gestión de riesgos, esta configuración representa un desafío significativo. El uso de modelos de IA operados externamente implica riesgos de terceros que pueden comprometer la integridad, confidencialidad y trazabilidad de los datos procesados. De hecho, según el informe Cost of a Data Breach 2025 – The AI Oversight Gap elaborado por IBM, el 29% de los incidentes de seguridad reportados en entornos de IA se originaron en modelos entregados como SaaS por proveedores externos, superando a aquellos implementados localmente (19%) y a los modelos internos u open source (26%). Esto demuestra que la principal superficie de riesgo actual no se encuentra dentro de los sistemas propios, sino en la dependencia operativa de terceros y la falta de visibilidad que esto conlleva.

En consecuencia, la naturaleza cerrada de estos entornos SaaS limita la aplicación de metodologías tradicionales de análisis forense digital, que suelen requerir acceso total a registros, sistemas de archivos o estructuras internas del modelo. Por ello, se hace necesario proponer marcos metodológicos específicos que permitan realizar análisis forense en entornos sin acceso privilegiado, combinando técnicas adaptadas a la observación indirecta, monitoreo de comportamientos anómalos y preservación de evidencia en interfaces limitadas.

Estos modelos pueden estar alojados en plataformas cerradas, limitando el acceso a los artefactos técnicos necesarios para un análisis exhaustivo. La falta de visibilidad y control directo sobre estos modelos representa un riesgo operativo considerable. Además, la trazabilidad de cambios, personalizaciones o eventos maliciosos que los afectan se vuelve compleja, lo que obstaculiza tanto la supervisión continua como la respuesta efectiva ante incidentes. Este entorno plantea retos particulares para el análisis forense digital.

2.3 Forense Digital en entornos de IA

El análisis forense digital se ocupa de recolectar, preservar, examinar e interpretar datos digitales de forma que puedan ser admitidos como evidencia en contextos técnicos, legales o regulatorios. Cuando el objeto de análisis es un modelo de IA, este proceso se vuelve más complejo. No se trata simplemente de capturar archivos o imágenes de disco, sino de examinar logs de entrenamiento, configuraciones de modelo, artefactos de inferencia y metadatos asociados. La naturaleza dinámica y no determinista de muchos modelos de IA introduce incertidumbres adicionales que requieren marcos forenses específicos. Estos marcos deben permitir detectar manipulaciones sutiles y documentar los eventos relevantes con trazabilidad y validez probatoria.

2.4 Evidencia Digital y Cadena de Custodia en IA

La evidencia digital en modelos de IA incluye desde registros de entrenamiento y pruebas, hasta las respuestas generadas ante entradas específicas (inferencias). Preservar esta evidencia implica mantener su integridad, autenticidad y contexto técnico, incluso cuando se opera sobre plataformas con acceso limitado. La cadena de custodia es esencial para garantizar que la evidencia no ha sido alterada y que puede ser presentada como prueba válida en auditorías o procesos judiciales. Adaptar estos principios a modelos operados por terceros es uno de los retos más relevantes que esta investigación busca atender.

2.5 Interpretabilidad, Rendición de Cuentas y Auditoría (XAI y Accountability)

El concepto de interpretabilidad (*Explainable AI o XAI*) se refiere a la capacidad de comprender y explicar las decisiones de un modelo. En el análisis forense, las herramientas de interpretabilidad pueden ayudar a identificar comportamientos anómalos, desviaciones de patrones esperados o indicios de manipulación. La

rendición de cuentas (*accountability*), por otro lado, implica poder atribuir responsabilidad por las acciones de un modelo y su impacto. La auditoría técnica y legal de modelos de IA es aún incipiente, especialmente cuando se trata de escenarios con modelos cerrados, lo que refuerza la necesidad de un marco forense especializado.

2.6 Marcos Normativos y Referentes Emergentes

Marcos como el *NIST AI Risk Management Framework* (*AI RMF*), *OWASP LLM Top 10* y la guía de la *NSA* sobre seguridad de datos en sistemas de IA reconocen la importancia de la trazabilidad, preservación de evidencia y preparación forense. No obstante, estas guías aún no proporcionan metodologías específicas para el análisis forense digital post-compromiso en modelos de IA. Por ello, el presente trabajo busca construir un marco complementario que aborde estas limitaciones, desde una perspectiva aplicada y centrada en entornos reales.

2.7 Gobernanza, resiliencia y riesgos operacionales en entornos de inteligencia artificial

En el contexto de los sistemas basados en inteligencia artificial, la resiliencia operacional se ha convertido en un eje clave para mitigar los efectos de incidentes de seguridad que, en muchos casos, resultan inevitables a largo plazo. La resiliencia se define como la capacidad de una organización para detectar, contener y recuperarse rápidamente de un incidente, minimizando su impacto en la operación y la integridad de los datos. Para fortalecer esta capacidad, es necesario que las organizaciones implementen planes de respuesta ante incidentes (IR), restauración de respaldos, y simulaciones de crisis (como ejercicios de cyber range), así como capacitación continua tanto para equipos técnicos como líderes no técnicos (IBM, 2025).

En paralelo, el riesgo asociado a la evasión de modelos de IA, ataques que buscan manipular entradas para que el modelo genere salidas erróneas, ha cobrado relevancia. Aunque son menos frecuentes que otros vectores, estos ataques pueden provocar desde pérdidas financieras hasta riesgos para la vida humana en sectores críticos como la salud o el transporte autónomo. El 50% de las organizaciones que evalúan este riesgo lo hacen con equipos internos, mientras que un 38% utiliza herramientas automatizadas y un 34% depende de auditorías de terceros. Esto

sugiere una fragmentación en los enfoques de evaluación de seguridad en IA, con escasa estandarización y dependencia de recursos internos.

A pesar de estos riesgos, existe una brecha significativa en gobernanza de IA. El 87% de las organizaciones citadas en el reporte (IBM, 2025) no cuenta con políticas o procesos formales para mitigar los riesgos asociados a la inteligencia artificial. Dos de cada tres no realizan auditorías periódicas sobre sus modelos y más del 75% no aplica pruebas de adversario (adversarial testing), lo que limita la capacidad para identificar y prevenir vulnerabilidades antes de que sean explotadas.

En términos de impacto económico, se ha observado que tanto los ataques dirigidos contra modelos de IA como el uso malicioso de IA por parte de los atacantes (por ejemplo, mediante phishing automatizado), generan costos similares por incidente, con un promedio de 4.46 a 4.49 millones de dólares por brecha. No obstante, los incidentes vinculados a shadow AI, modelos de IA operando fuera del control formal de la organización, tienen un costo aún mayor, alcanzando los 4.63 millones de dólares en promedio (IBM, 2025).

El compromiso de la cadena de suministro se identifica como la causa más frecuente de incidentes de seguridad en IA (30%), seguido por la inversión de modelos (24%) y ataques de evasión (21%). Otros vectores incluyen prompt injections (17%) y data poisoning (15%). Estas amenazas afectan directamente a la integridad de los datos, disponibilidad operativa y confidencialidad, generando consecuencias como accesos no autorizados (31%), pérdida de integridad (29%), pérdidas económicas (23%) y daños reputacionales (17%).

Desde una perspectiva forense, esta diversidad de vectores y consecuencias exige nuevos marcos de análisis adaptados a entornos complejos y distribuidos, como lo son los entornos SaaS y los modelos operados por terceros. La falta de gobernanza, sumada a la variedad de impactos posibles, refuerza la necesidad de contar con marcos metodológicos capaces de intervenir y preservar evidencia aún en condiciones limitadas de acceso, lo cual constituye el núcleo del presente trabajo de investigación.

Capítulo 3. Marco Metodológico

3.1 Tipo de Investigación

La presente investigación es de tipo aplicada, ya que no busca generar conocimiento base ni desarrollar teorías universales, sino utilizar y adaptar conocimientos existentes en ciberseguridad, forensia digital e inteligencia artificial para resolver un problema específico: la ausencia de un marco metodológico para realizar análisis forense en entornos donde los modelos de IA son operados por terceros o entregados como *SaaS* (*Software as a Service*). El resultado esperado es un marco operativo replicable que permita actuar ante incidentes de seguridad en dichos contextos.

3.2 Alcance Investigativo

El alcance de esta investigación es exploratorio, ya que el análisis forense aplicado a modelos de inteligencia artificial, especialmente aquellos sin acceso completo (black-box models) y operados por terceros, es un campo emergente con pocos antecedentes formales. La revisión de la literatura revela una ausencia de marcos específicos aplicables a estos entornos, por lo que esta investigación busca familiarizarse con el fenómeno, identificar variables relevantes, y proponer una primera aproximación metodológica que pueda servir de base para estudios futuros, más descriptivos o explicativos.

3.3 Enfoque

Se adopta un enfoque cualitativo, con orientación inductiva, ya que el propósito de esta investigación es comprender e interpretar las condiciones operativas, técnicas y legales que afectan el análisis forense en entornos de IA de terceros. El análisis no pretende validar hipótesis mediante conteos o correlaciones estadísticas, sino formular una propuesta metodológica basada en la interpretación crítica de los marcos existentes, estudios de caso, simulaciones y revisión documental. Este enfoque se alinea con el paradigma del naturalismo, reconociendo la complejidad contextual del fenómeno.

3.4 Diseño

El diseño de esta investigación es tecnológico-conceptual, orientado al desarrollo de un marco metodológico dividido en fases (detección, recolección, análisis, preservación y reporte de evidencia), aplicable a entornos cerrados donde los investigadores no tienen control sobre el modelo ni acceso total a su infraestructura. El diseño se articula a partir de la recopilación e interpretación de conceptos clave, estudios de caso documentados, y pruebas de concepto sobre escenarios simulados, siguiendo principios del diseño iterativo.

3.5 Población y Muestreo

Dado que no se recolectarán datos de personas u organizaciones reales, la población se define a nivel conceptual y documental, abarcando marcos normativos (como *NIST AI RMF*, *ISO 27037*, *NSA AI Security Framework*), estudios de caso disponibles públicamente, herramientas forenses (como *FTK*, *Tsurugi*, *YARA*), y entornos de simulación controlada. El muestreo es no probabilístico por conveniencia, seleccionando aquellos casos y documentos que sean más relevantes para construir y validar la propuesta metodológica.

3.6 Instrumentos de Recolección de Datos

Los instrumentos utilizados en esta investigación son:

- Revisión documental sistemática de literatura científica, estándares, lineamientos técnicos y reportes de la industria (por ejemplo, *Cost of a Data Breach Report 2025 – IBM*).
- Estudio de herramientas forenses con capacidad de operar en entornos limitados (como *FTK*, *Tsurugi*, *Sysmon*, *Wireshark*).
- Simulaciones controladas en máquinas virtuales, en las cuales se ejecutan incidentes forenses intencionados (como *model inversion*, *prompt injection*, *evasion*) sobre modelos locales simulados como *SaaS*.

3.7 Técnicas de Análisis de Información

Para interpretar la información recopilada, se emplearán técnicas cualitativas de análisis como:

- Mapas conceptuales para identificar y jerarquizar variables relevantes en la investigación forense de IA.
- Diagramas de flujo de datos y BPMN para modelar el proceso del marco propuesto.
- Espina de Ishikawa para analizar causas raíz de limitaciones en entornos forenses cerrados.
- Análisis comparativo entre marcos normativos y herramientas disponibles.

3.8 Estrategia de Desarrollo de la Propuesta

La propuesta del marco forense será desarrollada siguiendo una estrategia iterativa que incluye:

- Definición de fases del análisis forense adaptadas al entorno *IA/SaaS* (detección, preservación, recolección, análisis, presentación).
- Simulación de incidentes controlados sobre un modelo IA ejecutado localmente (simulando un entorno *SaaS*), con ataques como *data poisoning* o *model inversion*.
- Documentación paso a paso de la aplicación del marco propuesto, incluyendo recolección de evidencia mediante comandos y herramientas forenses.
- Evaluación crítica de los resultados de la simulación, identificando fortalezas, debilidades y oportunidades de mejora del marco propuesto.