Guía de Gobernanza, Desarrollo y Adopción de la IA en las CyberOps de los CSIRTs de América Latina y el Caribe

Hacia modelos de respuesta de seguridad, sostenibles, resilientes, responsables y adaptables.

Estevenson Solano

SEPTIEMBRE,2025

lacnic

Contenido

Introducción **4**

1. Gobernanza de la IA en CSIRT **5**
Gobierno y Liderazgo 6
Modelos de Madurez 8
Enfoque Ético de la IA 9

2. Desarrollo de la IA en CSIRT **9**
Ciclo de vida de un sistema de IA 10
Fases del ciclo de vida de un sistema de IA 12
Ecosistema de Actores del Ciclo de Vida de los Sistemas de IA 14
Múltiples Actores y Responsabilidades Compartidas 14
Seguridad 16
Calidad de los Datos 17
Selección y Robustez de Modelos de IA 17
Pruebas y Validación Continua de Modelos de IA 18
MLOps y CI/CD para el Desarrollo de IA en Ciberseguridad 18

3. Adopción de la IA en CSIRT **18**
Preparación Organizativa y Desarrollo de Capacidades 18
Preparación de la Infraestructura Técnica 19
Preparación y Desarrollo del Talento Humano 19
Preparación de los Datos y los Procesos 19
Consideraciones Culturales, Éticas y Centradas en el Ser Humano 20
Barreras y Retos Regionales en ALC 21

4. Recomendaciones y Mejores Practicas **23**
Anexo I. Caja de Herramientas **26**
Colaboradores 29
Agradecimientos 29
Referencias Bibliográficas 30

Descargo de Responsabilidad

Este documento ha sido publicado por el Registro de Direcciones de Internet de América Latina y Caribe (LACNIC, en adelante) como contribución al proyecto de Programa Líderes en el área de Tecnologías Emergentes, Innovación Sostenible e Inteligencia Artificial. Los hallazgos, interpretaciones y conclusiones presentados en este informe son resultado de un proceso de colaboración y cooperación respaldado por el autor. Sin embargo, los puntos de vista expresados son responsabilidad exclusiva del autor y no representan necesariamente la posición oficial de LACNIC, ni la de sus miembros, socios u otras partes interesadas.

© Copyright 2025, LACNIC. Todos los derechos reservados. Ninguna parte de esta publicación puede ser reproducida o transmitida de ninguna forma ni por ningún medio, sin la debida autorización, incluidos la fotocopia y la grabación, ni por ningún sistema de almacenamiento y recuperación de información.

Introducción

La integración de la inteligencia artificial (IA) en las ciberoperaciones está transformando la forma en que los CSIRTs anticipan, detectan y responden a las amenazas. Esta transformación también eleva el nivel de exigencia en materia de gobernanza, garantía, responsabilidad, resiliencia y adaptabilidad.

Esta guía ofrece un marco práctico para que los CSIRTs gobiernen, desarrollen e implementen capacidades basadas en la IA en las operaciones cibernéticas. Sirve tanto de orientación como de herramienta de implementación, adaptable al alcance, contexto y gobernanza de cada CSIRT. Para garantizar su valor operativo, se acompaña de un conjunto de herramientas de implementación con modelos, plantillas, listas de verificación y recursos de apoyo a la toma de decisiones.

Objetivo de la Guía

- Alinear la integración de la IA con la misión, la gobernanza y la madurez organizativa de los CSIRTs.
- Establecer principios, mejores prácticas y mecanismos de gobernanza para una adopción responsable.
- Permitir operaciones de ciberseguridad sostenibles, resilientes y auditables a través de la IA.
- Facilitar la toma de decisiones basada en el riesgo y una responsabilidad clara a lo largo del ciclo de vida de la IA.
- Fortalecer la capacidad de los CSIRTs para adaptarse a las amenazas y tecnologías en evolución.

Enfoque y Alcance

- Perspectiva sistémica ya que aborda las interdependencias entre la tecnología, las personas y los procesos.
- Flexibilidad contextual por que se adapta a los CSIRTs nacionales, regionales y sectoriales.
- Orientación práctica al traducir las directrices en acciones mediante un conjunto de herramientas prácticas.
- Preparación para el futuro ya que anticipa las amenazas emergentes, las tecnologías disruptivas y los riesgos específicos de la IA.

Contribuciones

- Una perspectiva de gobernanza para la IA en las operaciones de los CSIRTs.
- Una vía de desarrollo de capacidades para la IA dentro de las operaciones cibernéticas.
- Un marco de adopción que equilibra la innovación con la responsabilidad y la seguridad.
- Un conjunto de modelos adaptables para fomentar respuestas resilientes, sostenibles y en continua mejora.

Finalmente, esta guía invita a los CSIRTs a replantearse su papel en la era de la IA, no solo como respondedores a incidentes cibernéticos, sino como arquitectos de ecosistemas de defensa sociotécnicos resilientes, adaptables y fiables. Al aplicar sus principios y herramientas, los equipos pueden ofrecer un valor cuantificable hoy en día, al tiempo que se anticipan a la próxima ola de retos que configurarán el futuro de la ciberdefensa.

1. Gobernanza de la IA en CSIRT

Desde el contexto de los CSIRTs, la gobernanza de la IA es un pilar esencial para asegurar que su diseño, desarrollo, adopción, implementación y mejora continua de estas tecnologías de innovaciones emergentes y disruptivas se realice de manera estratégica, ética y segura.

Las familias de la IA en su conjunto es una fuerza transformadora para las capacidades de la ciberseguridad, mejorando significativamente en la detección avanzada de ciberamenazas y la automatización de la respuesta ante incidentes se hace clave, al tiempo que ofrece grandes ventajas significativas en eficacia, velocidad y análisis predictivo.[1][2] No obstante, su uso sin una adecuada supervisión y control puede introducir riesgos de alto impacto.

Su adopción introduce nuevos vectores de ataques y un sinnúmero de desafíos para la ética, la seguridad, la privacidad y la sostenibilidad, etc., que deben ser gestionado con precisión y previsión, por tanto, su uso sin una adecuada supervisión y control puede introducir riesgos de alto impacto.

Gobernar la IA no es limitar su potencial, es asegurar que funcione como un recurso confiable, seguro y alineado con los objetivos del CSIRT, teniendo claro su misión y propósito de uso al adoptarla[3][4] en todo el ciclo de vida de la IA, para eso se definen en la Tabla 1. los principios de la IA.

Diseñar un sistema de gobernanza apropiado es particularmente difícil: primero, debido a la naturaleza diversa de las preocupaciones éticas;[5][6] segundo, debido a la dificultad de determinar el instrumento regulatorio apropiado; y tercero, debido a las complejas interacciones entre la tecnología relevante, la economía y los mercados, los seres humanos[7] individuales y la sociedad, así como el medio ambiente y, en última instancia, la política y la regulación.

La integración de la IA en las ciberoperaciones presenta una dualidad inherente. Por un lado, la IA actúa como una fuerza multiplicadora, potenciando las capacidades defensivas y ofensivas al analizar vastas cantidades de datos en tiempo real, identificar anomalías, definir patrones, hacer correlaciones y automatizar respuestas y procesos del CSIRT. Esta capacidad permite pasar de una postura reactiva a una postura práctica haciendo de la ciberseguridad una gestión con enfoque transversal, holístico, integral, adaptativo y dinámico.

Por otro lado, la IA expande la superficie de ataque, ya que los actores maliciosos pueden explotar sus capacidades para evadir la detección, manipular algoritmos de seguridad y desplegar amenazas sistémicas con capacidades sofisticadas que pueden ser de automatizadas a escala. Esta situación subraya que la IA, si no se gestiona adecuadamente, puede convertirse en una vulnerabilidad en sí misma, introduciendo sesgos, malas interpretaciones y riesgos que podrían erosionar y socavar la confianza de las capacidades y controles de ciberdefensa. Por tanto, la gobernanza no puede ser una consideración posterior; debe integrarse desde el inicio del ciclo de vida de la IA, asegurado que los beneficios superen los riesgos y que la estrategia se adapte continuamente a medida que las capacidades de la IA evolucionan y surgen nuevos vectores de ataque.

Gobierno y Liderazgo

- Desarrollar un modelo de gobernanza de IA,[8 9 10] y entrelazar las sinergias con el modelo de gobernanza de las operaciones cibernéticas.[11 12]
    - Integrar los principios de la IA[13 14 15] (responsabilidad, transparencia, explicabilidad, trazabilidad y resiliencia) en el marco de gobierno de ciberseguridad del CSIRT.[16 17]
    - Alinear la estrategia[18] interna o externa (nacional) con las políticas de gestión de riesgos,[19] continuidad de negocio, gestión de terceros o cadena de suministro, ciberseguridad,[20] protección de datos[21] y desarrollo seguro.
    - Establecer un comité de gobernanza interfuncional (IA, ciberseguridad, legal, ética, operaciones) que supervise la interacción entre la IA y las operaciones críticas.
    - Evaluar los impactos de decisiones automatizadas y autonómicas en los entornos operacionales del CSIRT.
    - Definir el modelo de gobierno de los datos[22 23 24] (ej., acceso, clasificación, uso, trazabilidad, calidad, seguridad, cifrado, privacidad)[25 26], para el debido control y calidad de datos de entrenamiento y operación de los sistemas de IA.

- Establecer los roles y responsabilidades o el área responsable de la gestión, monitoreo, auditoría y operación de la IA en el CSIRT.
    - Crear una unidad o equipo multidisciplinario[27 28] de "IA en las ciberoperaciones" liderado por una figura responsable (ej. Chief AI Officer o Responsable de IA) acompañado de las figuras (ej. Chief Data Officer, Data Protection Officer, Chief Information Security Officer, Analistas de sistemas, arquitectos de sistemas, científicos de datos y el Responsable de Ciberoperaciones del CSIRT).
    - Definir funciones para los siguientes roles:
        - Desarrolladores de IA: responsables del diseño, desarrollo seguro, validación y entrenamientos de modelos.
        - Auditores y analistas IA: encargados de la supervisión humana en bucle "human-in-the-loop" ante decisiones críticas y complejas.
        - Gestores de riesgos de IA: encargados de evaluar los impactos de la arquitectura e infraestructura de IA.
    - Establecer los canales de escalamiento y revisión ante situaciones o anomalías detectadas en los sistemas de IA.

- Definir un marco de políticas y directrices sobre el uso, desarrollo, adquisición, validación y uso de las soluciones de IA.
    - Adoptar una política de gobernanza de IA[29 30] alineada con marcos nacionales (si existiera) o internacionales (ej. NIST AI RMF, ISO/IEC 42001, EU AI Act).
    - Establecer criterios de adquisición[31 32] y evaluación de soluciones de IA:[33] transparencia del proveedor, trazabilidad del modelo, interoperabilidad,[34] protección de datos, mecanismos de respuesta ante fallos.
    - Incluir requisitos de validación previa y continua antes de desplegar[35] modelos en producción (ej. Sandbox o cajas de arena).[36]
    - Reforzar las directrices sobre uso responsable y ético de la IA en contextos de ciberseguridad ofensiva,[37] ciberdefensa, ciberinteligencia[38] y análisis de amenazas[39].

    - Garantizar procesos de desactivación segura "kill switch" en caso de mal funcionamiento o uso indebido.

- Disponer de la asesoría de expertos en sostenibilidad, IA, ética, derechos digitales y privacidad.
    - Establecer convenios con instituciones académicas, comités éticos y centros de investigación para la evaluación externa.
    - Incorporar expertos en:
        - Ética y derechos digitales para analizar impactos sociales y sesgos algoritmos.[40] [41]
        - Privacidad y protección de datos para asegurar el cumplimiento (ej. RGPD o GDPR de la UE) o la normativa local vigente.
        - Sostenibilidad para optimizar recursos, eficiencia energética[42] y la reducción del impacto ambiental
    - Organizar sesiones regulares de revisión ética y sostenibilidad para tecnologías desplegadas.
    - Promover la participación de la sociedad civil y la población en general en mecanismos de consultas o feedback.

- Definir el ecosistema de cadena de suministro de IA (proveedor, usuario, implementador, importador y distribuidor).
    - Elaborar el mapa de actores del ciclo de vida de la IA en el CSIRT (interno y externo).
    - Establecer requisitos contractuales mínimos para:
        - Proveedores de IA: cumplimiento normativo, documentación técnica, soporte y garantía de transparencia.
        - Implantadores e integradores: certificación de competencias, control de calidad y seguridad.
    - Monitorear continuamente el estado de confianza, ciberseguridad y cumplimiento de cada actor.
    - Evaluar los riesgos[43] [44] de terceras, cuartas y enésimas partes involucradas en la cadena de suministro algorítmica.

- Disponer de los recursos y capacidades necesarias para ejecutar las acciones del modelo de gobernanza.
    - Asignar presupuesto[45] específico para el desarrollo y gobernanza de la IA en las ciberoperaciones.
    - Garantizar la formación continua del personal (ej., Gobernanza de IA, Evaluación de riesgos algoritmos, Desarrollo Seguro de IA, Ciberseguridad Aplicada a Sistemas Autónomos).
    - Implementar herramientas tecnológicas para (ej., auditoria de modelos, gestión del ciclo de vida de IA, simulación y pruebas en entornos controlados "sandbox").[46]
    - Establecer indicadores claves de rendimiento (KPIs) y madurez para evaluar la gobernanza de la IA.
    - Evaluar regularmente la capacidad del CSIRT y la madurez del modelo de gobernanza adaptado a entornos dinámicos.

Modelos de Madurez

Hay que destacar que el NIST AI RMF[47] [48] pone en relieve la importancia central de la función "Gobernar" situándola en el centro de un diagrama de ciclo operativo, rodeada por "Mapear", "Medir" y "Gestionar". Los temas específicos que tienen mayor relevancia crítica para los sistemas de IA se enumeran como características de la IA fiable, enfatizando en sus capacidades que desempeñan "Válido y fiable" y "Responsable y transparente".

Por otra parte, en cuanto a este marco NIST AI RMF[49] se ha visto que se puede utilizar para evaluaciones de madurez al abordar de manera holística la gestión del riesgo y la seguridad con liderazgo y las herramientas.

MITRE propuso un modelo de madurez[50] [51] de IA que proporciona un marco integrar para definir y medir el progreso en la adopción de la IA. Es un método de evaluación flexible y adaptable realiza un seguimiento de la progresión a lo largo de la transformación de la IA. Desde el cual se puede evaluar el ritmo del crecimiento, la aceleración y la evolución de la tecnología de IA.

El Modelo de Madurez de la IA comprende seis pilares que, en conjunto, son fundamentales para la adopción exitosa de la IA:

1. Uso ético, equitativo y responsable
2. Estrategia y recursos
3. Organización
4. Habilitadores tecnológicos
5. Datos
6. Rendimiento y aplicación

Cada pilar tiene dimensiones que describen acciones y actividades específicas que demuestran el avance de las prácticas de madurez de la IA. En conjunto, estos pilares y dimensiones se miden con cinco niveles de preparación que describen cualitativamente el progreso jerárquico y escalable a lo largo de la adopción de la IA, adaptado al Modelo de Integración de Madurez de Capacidades (CMMI) inicial, adoptado, definido, gestionado y optimizado.

Mientras que los pilares y los niveles de preparación aportan estructura y amplitud al modelo, las dimensiones proporcionan métricas mediante las cuales una organización puede determinar el progreso mediante el uso de la herramienta de evaluación del modelo de madurez de la IA correspondiente. Este proceso de evaluación puede ayudar a la organización a determinar si los recursos, procesos y tecnología críticos están disponibles para la adopción exitosa de la IA en los niveles objetivo necesarios y apropiados.

La gobernanza es crucial, y esto se extiende mucho más allá de la seguridad técnica, abarcando dimensiones legales, éticas y de reputación. Los principios éticos fundamentales deben guiar la gobernanza "equidad, la transparencia, la responsabilidad, la privacidad y la seguridad". La ciberseguridad, de hecho, es una condición previa necesaria para la seguridad, resiliencia, privacidad, equidad, eficacia y fiabilidad de los sistemas de IA.

Si los sistemas de IA de un CSIRT son percibidos como sesgados, opacos o no conformes, se socava la confianza con las partes interesadas internas y externas, una gobernanza efectiva no solo debe asegurar los sistemas, sino también garantizar que operan de manera que defienden los estándares éticos, legislaciones o regulaciones, fomentando la confianza en sus decisiones y acciones. Esto requiere un cambio de mentalidad, pasando de una implementación puramente técnica a una responsabilidad más amplia.

La gestión de riesgos debe evolucionar desde un enfoque de prospectiva y previsión al tiempo que puede anticipar y mitigar las vulnerabilidades inherentes a los modelos de IA, no solo a la infraestructura. Si un CSIRT solo aplica gestión de riesgos tradicional, pasara por alto estas vulnerabilidades. Por lo tanto, la implicación es que se debe integral la modelización de amenazas

específicas de la IA y metodologías de evaluación de riegos practicas dirigidas también a la cuantificación del riesgo. Esto requiere comprender como se pueden manipular los modelos de IA, como se pueden comprometer sus datos de entrenamiento y como se pueden explotar sus salidas. El enfoque va más allá de asegurar el contenedor de la IA a asegurar la inteligencia dentro de ella.

Enfoque Ético de la IA

Existen diversos enfoques para incorporar la ética en los sistemas de inteligencia artificial.

- El enfoque bottom-up se basa en que la IA aprenda observando el comportamiento humano en situaciones concretas, aunque esto puede llevar a reproducir prácticas comunes, no necesariamente éticas.
- Por otro lado, el enfoque top-down implica programar directamente los principios éticos en el sistema, ofreciendo mayor control y previsibilidad.
- El enfoque casuístico propone definir respuestas éticas específicas para diferentes escenarios, lo que permite adaptar la toma de decisiones según el contexto, aunque no cubre todas las situaciones posibles.
- Finalmente, el enfoque dogmático se basa en una escuela ética particular, como el utilitarismo o el kantismo, aplicando sus principios de manera rígida, lo que puede generar dilemas éticos en contextos complejos donde se requiere un análisis más flexible.

Por tanto, cada enfoque ético ofrece ventajas y retos únicos, lo que pone de relieve que ningún método es suficiente por sí solo. Una combinación equilibrada de estas estrategias, adaptada a aplicaciones específicas de IA, es esencial para desarrollar sistemas de IA éticamente responsables y fiables.

2. Desarrollo de la IA en CSIRT

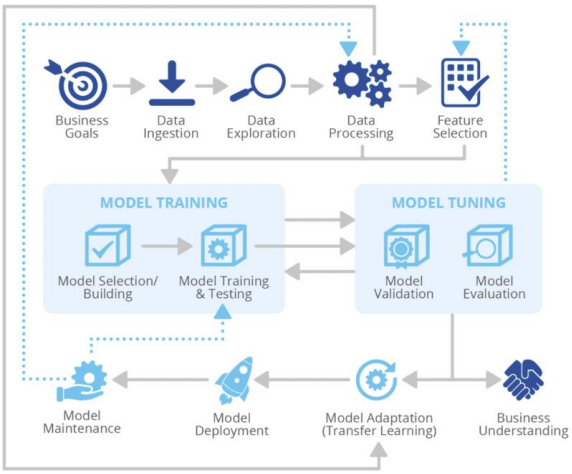
El desarrollo de sistemas o modelos de IA no es un proceso trivial; requiere un enfoque estructurado y "seguro desde diseño" que debe seguir prácticas de ingeniería de software, datos y privacidad adaptadas a la seguridad, a lo largo del ciclo de vida del sistema o modelo de IA, desde su concepción hasta su operación y mantenimiento

Por lo cual es preciso seguir los principios que se establecen para el desarrollo o uso de la IA[6 52] en un CSIRT están basados en los marcos internacionales como el AI Act de la UE[53 54] las prácticas de Ética y gobernanza de la IA del Instituto Alan Turing,[55-62] el OCDE,[63-69] el NIST AI RMF,[48] y la ISO/IEC 42001, ya que son fundamentales para asegurar que su implementación sea segura,[70] ética,[6 71] efectiva y alineada con los valores organizacionales[72] y regulatorios.[73]

Principio	Descripción y Aplicación Practica en CSIRT
Legalidad y cumplimiento	Toda IA debe cumplir con regulación nacional e internacional pertinente (protección de datos, ciberseguridad, privacidad, derechos fundamentales). Los modelos deben ser auditables y documentados.
Responsabilidad (Acountability)	Debe quedar claro desde el principio quien supervisa, mantiene, desarrolla, audita y responde por las decisiones de la IA. Delegar funciones no es delegar responsabilidad.
Transparencia y trazabilidad	Las decisiones tomadas por IA deben ser entendibles, registradas y explicables, especialmente durante incidentes, situaciones de alta complejidad sistémica o análisis post-mortem. Se deben mantener logs y bitácoras de decisiones algorítmicas.

Principio	Descripción y Aplicación Practica en CSIRT
Supervisión humana significativa (Human-in-the-loop)	Las decisiones críticas y de alta complejidad deben poder ser intervenidas, modificadas o anuladas en tiempo real. La IA asiste, no necesariamente remplaza el juicio humano en algunas tareas.
Seguridad y resiliencia	Los modelos deben ser protegidos[74] contra ciberataques, manipulaciones de datos de entrenamiento y fallos operativos.[75] Además deben funcionar de forma robusta ante condiciones inesperadas.
Ética, equidad y no discriminación	La IA no debe replicar ni amplificar sesgos de identificación, priorización o respuesta de amenazas. Por lo cual se debe dar enteramiento diverso y validaciones con datos éticamente tratados.[76 77]
Minimización de riesgos y daños	Ante cualquier despliegue de IA debe ir sujeto al acompañamiento metódico de una evaluación de riesgos y de impacto, a la vez de un plan de mitigación. La IA debe minimizar el daño potencial en escenarios críticos.
Mejora continua y gobernanza adaptativa	Se debe evaluar, actualizar, auditar y ajustar la IA de forma periódica, en función de cambios en el entorno operativo, las amenazas y la normativa. Es aquí esencial que la gobernanza evolucione a la par de la tecnología.

Ciclo de vida de un sistema de IA



Nota: La Agencia de la Unión Europea para la Ciberseguridad (ENISA) planteó un modelo,[78] en el cual resalta de forma esencial que los datos son uno de los activos más valiosos de la IA; se transforman continuamente a lo largo del ciclo de vida. Fuente: ENISA.[78]

El ciclo de vida de un sistema de IA abarca fases interdependientes que incluyen el (diseño, desarrollo análisis de requerimientos, recolección de datos, entrenamiento, prueba, integración, instalación, despliegue, operación, mantenimiento y disposición final). Dada la complejidad de estos sistemas, se emplean metodologías como cascada, espiral, ágil o prototipado rápido para gestionar su desarrollo. Este modelo de referencia se enfoca particularmente en los sistemas de Machine Learning (ML), que actualmente representan la mayoría de las aplicaciones de IA, ya que permiten transformar datos de entrada en resultados útiles para agregar valor en las operaciones cibernéticas. Cabe destacar que no abarca toda la diversidad de subcampos de la IA, este es un enfoque inicial y puede ser adaptativo.

En cada etapa del ciclo,[78 79 80] se identifican activos clave en función de su relevancia funcional y su papel en el desarrollo, despliegue y operación de los sistemas de IA. Estos activos no solo comprenden los componentes técnicos de IA, sino también aquellos elementos de soporte y gestión transversal.

Para estructurar esta complejidad, los activos se agrupan en seis categorías principales: datos (datasets de entrenamiento, validación, inferencia), modelos (modelos entrenados, hiperparámetros, arquitecturas), actores (roles, responsabilidades y organizaciones involucradas), procesos (actividades de desarrollo, gobernanza, auditoria, mantenimiento), entornos y herramientas (frameworks, librerías, software, hardware, entornos de ejecución), y artefactos (documentación, logs, informes de validación, outputs del sistema). Esta taxonomía proporciona una base para el análisis de riesgos y amenazas,[81] trazabilidad, gobernanza y ciberseguridad a lo largo del ciclo de vida de sistemas basados en IA.[82]

Por lo cual la transformación de datos a lo largo del ciclo de vida de la IA implica varios otros tipos de activos, como los actores involucrados, los recursos computacionales, el software, etc., e incluso algunos activos intangibles como los procesos, la cultura y la forma en que la experiencia y los conocimientos de los actores pueden traer consigo amenazas potenciales no intencionadas (por ejemplo, sesgos no intencionados).

La transformación de los datos a lo largo de las diferentes etapas del ciclo de vida: ingesta de datos -Data Ingestion-, exploración de datos -Data Exploration-, preprocesamiento de datos -Data Pre-processing-, importancia de las características-Feature Importance-, formación -Training-, pruebas y evaluación -Testing and Evaluation-.

Para el desarrollo de sistemas o modelos IA, hay que tener presente la cadena de suministro de datos conocida como "DataOps" es metodología que combina principios de DevOps con la gestión de datos para mejorar la eficiencia y la calidad en el ciclo de vida de los datos, por cual se debe garantizar, la exhaustividad, singularidad, disponibilidad, validez, fiabilidad, precisión y coherencia de los datos, logrando proporcionar una visión de las características de "datos de calidad".[83] [84]

Por otra parte, relacionado con la validación de auditorías de algoritmos algunos marcos plantean la identificación y transparencia del componente, propósito del componente de IA, fundamentos del componente de IA, verificación y validación de modelos y datos, privacidad y controles generales de tecnología y ciberseguridad TI.[85] [86] Otros, como es el caso de la Universidad de Oxford ha propuesto uno para grandes modelos lingüísticos (LLM, por sus siglas en ingles) que abarca tres niveles gobierno, modelo y aplicación.[87] [88] Además algunos modelos de auditoria están parten de la premisa de que un sistema de IA es la suma de diversos módulos o componentes.[89]

Hay destacar la urgente necesidad de proteger todo el ciclo de vida de los sistemas de IA,[90-93] por lo cual SANS[20] se enfoca en la seguridad para las implementaciones de IA de modelos de lenguaje grandes (LLM), la generación aumentada por recuperación (RAG) y las bases de datos vectoriales (VectorDB). Los LLM[94] [95] son potentes modelos generativos que constituyen el núcleo de las aplicaciones modernas de IA, pero su implementación requiere controles de acceso estrictos[96] [97] y la aplicación de barreras de inferencia, es decir, reglas y comprobaciones automatizadas que impiden que los modelos respondan a solicitudes perjudiciales o no autorizadas.

Técnicas como la RAG refuerzan los resultados de los LLM combinándolos con datos actuales específicos del dominio, a menudo extraídos de una VectorDB, que indexa los datos como vectores de alta dimensión para una recuperación semántica rápida. Sin embargo, esta integración abre nuevas vías para riesgos como la inyección de solicitudes, en la que las solicitudes adversas tratan de eludir los controles de seguridad del modelo o filtrar datos confidenciales.[98] [99] [100]

Para hacer frente a la creciente complejidad y supervisión normativa en torno a la IA, las directrices abogan por enfoques sistemáticos como la Lista de materiales de IA (AIBOM), que inventaría todos los componentes (modelos, conjuntos de datos, dependencias) y permitiría la transparencia, la gestión de riesgos y el cumplimiento normativo. Los sistemas de registro de modelos son fundamentales, ya que permiten realizar un seguimiento de las versiones y el uso

de los modelos para evitar que los modelos obsoletos o vulnerables sigan utilizándose en producción.

Con avances como los modelos multimodales, que combinan texto, imágenes, audio u otros tipos de datos, los equipos de seguridad deben hacer frente a nuevas formas de manipulación y vectores de amenazas ocultos. Se recomiendan los entornos de ejecución confiables (TEE) para ejecutar cargas de trabajo de IA sensibles,[101] ya que proporcionan un aislamiento basado en hardware que protege la integridad de los modelos frente a amenazas externas e internas. En conjunto, estas prácticas constituyen la base para una integración de la IA responsable, segura y resistente,[102] ya que las organizaciones dependen cada vez más del aprendizaje automático para impulsar sus operaciones comerciales y de seguridad.

Fases del ciclo de vida de un sistema de IA

La implementación exitosa de la IA en las ciberoperaciones requiere una comprensión sistemática del ciclo de vida del sistema de IA y las acciones específicas necesarias en cada fase. Para los CSIRTs, este enfoque del ciclo de vida es especialmente crucial, dadas las limitaciones de recursos y la diversidad de entornos normativos en toda la región.

El ciclo de vida del sistema de IA consta de fases distintas, cada una de las cuales requiere acciones coordinadas en los ámbitos de la gobernanza, el desarrollo y las operaciones. Este enfoque estructurado permite mantener el control sobre las implementaciones de IA, al tiempo que garantiza que las implementaciones de IA no solo sean técnicamente sólidas, sino que también se ajusten al cumplimiento de las normativas regionales y las normas internacionales. Además, una alineación en la gobernanza, las mejores prácticas de desarrollo y la sostenibilidad operativa. Por lo cual, comprender estas fases es esencial para desarrollar estrategias de IA integrales que respalden modelos de respuesta de seguridad sostenibles, resilientes, responsables y adaptables.

Fase	Contexto	Acción
Definición de Objetivo de Negocio	Identificación de la problemática a resolver mediante IA, definición de métricas clave (KPIs), formulación de la pregunta analítica (predictiva, clasificatoria, etc.) y selección inicial del tipo de modelo (clasificación, regresión, clustering, etc.).	Definir el propósito del sistema de IA, la pregunta clave, y el tipo de modelo requerido.
Ingesta de Datos	Extracción de datos estructurados, semiestructurados o no estructurados desde múltiples fuentes (APIs, bases de datos, sensores IoT, logs, etc.), mediante flujos por lotes (batch), micro-lotes, o en tiempo real (streaming). Consideración de metadatos, taxonomías, control de acceso (RBAC/ABAC), técnicas de preservación de privacidad (ej. k-anonimato, differential privacy), y cumplimiento legal (GDPR, ePrivacy).	Extraer datos y metadatos según el formato, canal (batch/stream), y restricciones de privacidad o uso.
Exploración de Datos	Análisis exploratorio (EDA) con herramientas estadísticas y gráficas. Evaluación de tipos de variables (numéricas, categóricas, ordinales, multimedia), identificación de valores nulos/outliers, análisis de distribuciones (normal, Poisson, etc.), y pruebas de correlación (Pearson, Spearman).	Analizar y caracterizar estadísticamente los datos para entender su naturaleza y calidad.
Preprocesamiento de Datos	Limpieza de datos (handling de valores nulos, duplicados, outliers), normalización (Min-Max, Z-score), codificación (one-hot,	Transformar, limpiar y codificar los datos a formatos numéricos,

Fase	Contexto	Acción
	label encoding), integración de fuentes heterogéneas, anonimización/pseudonimización, y técnicas de aumentación (text: sinónimos; imagen: rotaciones, escalados).	asegurando calidad y diversidad.
Selección de Características (Feature Selection)	Reducción de dimensionalidad mediante selección basada en varianza, correlación, técnicas estadísticas (ANOVA, chi-cuadrado), métodos embedded (LASSO, árbol de decisión), o extracción de características (PCA, t-SNE, UMAP). Mejora de interpretabilidad y eficiencia computacional.	Conservar las variables más relevantes para el modelo, eliminando redundancia o ruido.
Selección / Construcción del Modelo	Elección del algoritmo de aprendizaje según el tipo de tarea (clasificación, regresión, agrupamiento, RL). Técnicas: árboles de decisión, SVM, redes neuronales (CNN, RNN, transformers), bayesianos, ensemble (bagging, boosting). Consideración del sesgo-varianza y de los requerimientos computacionales.	Seleccionar el algoritmo más adecuado y adaptar los datos al formato requerido por el modelo.
Entrenamiento del Modelo	Aplicación del algoritmo de entrenamiento (ej. descenso de gradiente, propagación hacia atrás en redes neuronales), definición de función de pérdida (MSE, cross-entropy), y uso de conjuntos de datos divididos (train/test/split o k-fold cross-validation). Ajuste de pesos/bias por epochs y mini-batches.	Ajustar los parámetros internos del modelo (pesos, bias) mediante datos etiquetados y validación.
Ajuste del Modelo (Tuning)	Optimización de hiperparámetros (learning rate, número de capas, regularización, batch size) mediante técnicas como Grid Search, Random Search o Bayesian Optimization. Validación con conjunto de datos distinto (validation set). Pruebas con datos adversos (robustez).	Ajustar los hiperparámetros del modelo entrenado para mejorar desempeño general.
Transferencia de Aprendizaje	Uso de modelos preentrenados en dominios similares (ej. ImageNet, BERT, GPT) y ajuste (fine-tuning) sobre nuevos datos. Útil en escenarios con datasets pequeños. Evita entrenamiento desde cero y mejora tiempos de convergencia.	Aplicar y afinar modelos preentrenados para mejorar rendimiento con pocos datos propios.
Despliegue del Modelo	Conversión del modelo en una aplicación funcional (servicio REST API, contenedor Docker, microservicio, FPGA/ASIC), integración en sistemas productivos, monitorización de uso y rendimiento. Despliegue en edge, on-premises o cloud (AWS, Azure, GCP).	Implementar el modelo entrenado como servicio operativo y conectarlo con datos en producción.
Mantenimiento del Modelo	Detección y mitigación de "drift" (cambio gradual) o "shift" (cambio súbito) en los datos, comparación estadística entre datasets históricos y actuales, backtesting, reentrenamiento basado en ventanas temporales. Integración con sistemas de MLOps y AIOps.	Supervisar el rendimiento y reentrenar el modelo cuando detecte desviaciones o degradación.
Comprensión del Negocio	Evaluación del ROI, impacto estratégico y KPIs del modelo implementado. Comparación entre predicciones y realidad operativa. Ajuste continuo al entorno empresarial, regulación o condiciones del mercado.	Verificar el valor real aportado por el modelo en producción, ajustando según impacto observado.

Ecosistema de Actores del Ciclo de Vida de los Sistemas de IA

El desarrollo, la implementación y la gobernanza de los sistemas de inteligencia artificial implican un complejo ecosistema de actores interconectados que desempeñan funciones distintas pero interdependientes a lo largo del ciclo de vida de la IA. Comprender este entorno multifacético de partes interesadas es fundamental para entender cómo evolucionan los sistemas de IA desde su conceptualización hasta su implementación operativa y su gestión continua. La siguiente tabla presenta una visión general sistemática de los principales actores que participan en las distintas etapas del desarrollo y la gobernanza de los sistemas de IA, ilustrando sus responsabilidades específicas y sus contribuciones al ecosistema general de la IA.

Como se ha podido indicar anteriormente, el ciclo de vida de los sistemas de IA abarca múltiples etapas interconectadas que van desde la planificación y el diseño iniciales hasta la implementación, la operación y el eventual desmantelamiento. Este ciclo de vida se caracteriza dimensiones sociotécnicas clave que influyen en la forma en que se conceptualizan, desarrollan e implementan los sistemas de IA: Contexto de la Aplicación, Datos y Entradas, Modelo de IA, Tareas y Salidas, y la dimensión central Personas y Planeta, que hace hincapié en los derechos humanos y el bienestar social en general.

A diferencia del desarrollo de software tradicional, los sistemas de IA presentan retos únicos debido a su naturaleza sociotécnica inherente, lo que significa que están profundamente influenciados por la dinámica social, el comportamiento humano y la compleja interacción entre los componentes técnicos y los contextos sociales. Esta complejidad requiere una amplia gama de partes interesadas con diferentes conocimientos, perspectivas y responsabilidades a lo largo del ciclo de vida de la IA.

Múltiples Actores y Responsabilidades Compartidas

El ecosistema de IA funciona según el principio de responsabilidad colectiva entre diversos actores, reconociendo que el despliegue exitoso de la IA requiere la coordinación entre múltiples disciplinas, profesiones y límites organizativos. Este enfoque colaborativo reconoce que ninguna entidad por sí sola posee toda la experiencia necesaria para abordar los retos multifacéticos inherentes al desarrollo y despliegue de sistemas de IA.

Podemos tener en algunos ecosistemas de partes interesadas que incluyen tanto a los actores principales que participan directamente en el diseño, el desarrollo y la implementación de los sistemas de IA, como a los actores secundarios que proporcionan gobernanza, supervisión y aportaciones sociales. Los actores principales suelen ser desarrolladores de IA, científicos de datos, integradores de sistemas y expertos en la materia que trabajan directamente con los componentes técnicos de los sistemas de IA. Los actores secundarios abarcan a los reguladores, las organizaciones de la sociedad civil, las comunidades afectadas y otras partes interesadas que influyen en el contexto más amplio en el que operan los sistemas de IA.

Cabe destacar que, en el ecosistema de IA, cada actor tiene responsabilidades específicas relacionadas con la identificación, evaluación y mitigación de riesgos, que varían en función de su papel en el ciclo de vida y su proximidad a los diferentes aspectos del sistema. La naturaleza distribuida del desarrollo de la IA significa que las perspectivas de riesgo pueden diferir significativamente entre los distintos actores; por ejemplo, un desarrollador de IA que crea modelos preentrenados puede tener preocupaciones de riesgo diferentes a las de un implementador que integra esos modelos en contextos operativos específicos.

Este modelo de responsabilidad distribuida hace hincapié en la importancia de los procesos de prueba, evaluación, verificación y validación (TEVV) que se producen a lo largo del ciclo de vida de la IA, en los que participan actores especializados que examinan el rendimiento del sistema, detectan problemas y facilitan las medidas de corrección. Los actores TEVV son especialmente importantes, ya que proporcionan capacidades de evaluación independientes que complementan el trabajo de los desarrolladores y los implementadores.

Además, incluye los actores dedicados a la gobernanza y la supervisión que podrían en algunos casos tener autoridad administrativa, financiera y legal sobre los sistemas de IA dentro de los CSIRTs. Estos actores, entre los que se incluyen los tomadores de decisiones, juntas, comités o consejos y los expertos en cumplimiento normativo, son responsables de establecer las políticas, garantizar el cumplimiento normativo y mantener las estructuras de rendición de cuentas que guíen las decisiones de desarrollo y despliegue de la IA.

La dimensión de la gobernanza cuenta con el apoyo adicional de entidades terceras, como proveedores, evaluadores y prestadores de servicios, que aportan capacidades, datos o herramientas especializadas al ecosistema de la IA. Estos actores externos añaden complejidad a los esfuerzos de gestión y coordinación de riesgos, ya que sus metodologías de evaluación de riesgos y sus estructuras de gobernanza pueden no coincidir perfectamente con las de las principales organizaciones de desarrollo o implementación.

La rápida evolución de la tecnología de IA, en particular con la aparición de las capacidades generativas de la IA, ha creado nuevas funciones y responsabilidades dentro del ecosistema de las partes interesadas. Los últimos avances han dado lugar a la aparición de puestos especializados, como ingenieros de prompts, responsables de ética de la IA y especialistas en gobernanza de la IA, que se ocupan de retos específicos asociados a los sistemas modernos de IA.

La tabla siguiente ofrece un desglose detallado de estos diversos actores, sus funciones específicas y las etapas del ciclo de vida en las que participan más activamente, lo que ofrece una visión global del ecosistema colaborativo necesario para el desarrollo y la implementación responsables de la IA.

Actor	Función	Intervención
Diseñadores de IA / Solución	Definen la arquitectura funcional del sistema IA, especificaciones técnicas, flujos de datos, objetivos del modelo, y restricciones éticas/regulatorias.	Diseño de soluciones, requerimientos, gobernanza, IA ética.
Desarrolladores de IA	Codifican y optimizan algoritmos de ML/DL, implementan pipelines de entrenamiento e inferencia, integran APIs y frameworks (TensorFlow, PyTorch, ONNX, etc.).	Desarrollo backend IA, lógica del modelo, interoperabilidad.
Científicos de Datos	Realizan ingeniería de características, modelado estadístico, validación experimental, tuning y evaluación de rendimiento (precisión, recall, F1-score, etc.).	Exploración, modelado, validación, interpretación de resultados.
Ingenieros de Datos	Implementan ETL/ELT, orquestan pipelines (Airflow, Kafka, Spark), diseñan data lakes/warehouses, aplican estándares de calidad y transformación de datos.	Ingesta, preprocesamiento, estandarización, integración.
Propietarios de Datos	Controlan el acceso, uso y licenciamiento de datasets. Aplican políticas de custodia, clasificación de datos (PII, confidencial, etc.) y cumplimiento normativo (GDPR o RGPD, e incluso la legislación local vigente).	Gobernanza, cumplimiento, legalidad del uso de datos.

Actor	Función	Intervención
Proveedores de Datos	Comercializan datasets estructurados y no estructurados; algunos aplican técnicas de anonimización y etiquetado automático. También incluyen brokers no regulados.	Suministro de datos, monetización, calidad, riesgos éticos.
Proveedores de Modelos	Ofrecen modelos preentrenados (Hugging Face, OpenAI, Google, etc.), y entornos MLOps (SageMaker, Vertex AI). Proveen modelos como servicios (AIaaS/MLaaS).	Fine-tuning, inferencia, modelos transferidos, autoML.
Proveedores de Terceros	Suministran SDKs, frameworks (scikit-learn, Keras), entornos de ejecución (CUDA, cuDNN), hardware especializado (GPU, TPU, ASICs), y toolkits para IA explicable o robusta.	Infraestructura, rendimiento, seguridad, explainability.
Usuarios Finales / Operadores	Consumen los servicios inteligentes, interactúan con las interfaces IA (chatbots, dashboards, APIs) o integran los modelos en workflows empresariales.	Uso final, feedback, experiencia de usuario, validación operativa.

Seguridad

Los controles de seguridad son fundamentales para garantizar la solidez, integridad y fiabilidad de los sistemas de IA a lo largo de todo su ciclo de vida. Desde la ingesta de datos hasta la implementación, cada etapa debe integrar mecanismos de defensa en profundidad. Las rigurosas evaluaciones de riesgos de terceros, la validación criptográfica de los componentes externos y las pruebas en entornos aislados son requisitos previos a la integración. La segregación del entorno, es decir, el aislamiento de los dominios de desarrollo, ensayo y producción, refuerza el principio del mínimo privilegio y mitiga el movimiento lateral en caso de compromiso. La capacidad de ejecutar reversiones controladas, utilizando artefactos y conjuntos de datos inmutables y con hash criptográfico, es esencial para la recuperación tras un incidente y la continuidad de las operaciones.

Se deben implementar sistemas de control de versiones para todos los activos relacionados con la IA (bases de código, binarios de modelos, conjuntos de datos) para permitir la trazabilidad y el análisis forense. El registro de eventos de seguridad debe cumplir los principios de la tríada CIA, garantizando la disponibilidad, la confidencialidad y la integridad de los registros mediante configuraciones reforzadas. La documentación exhaustiva del linaje de los datos mejora la auditabilidad y la reproducibilidad, mientras que los canales de actualización seguros, validados mediante pruebas previas a la implementación en entornos aislados, evitan la contaminación de la cadena de suministro y las modificaciones no autorizadas.

La seguridad operativa requiere una supervisión continua del comportamiento y la telemetría de las entradas y salidas del sistema de IA, con mecanismos de detección de anomalías para identificar desviaciones indicativas de interferencias adversas o desviaciones del modelo. Las arquitecturas de actualización seguras por diseño deben incluir protecciones de reversión, verificación criptográfica y políticas estrictas de control de acceso. La captura y difusión de las lecciones aprendidas en la respuesta a incidentes en las operaciones del CSIRT habilitado para la IA permite estrategias de defensa adaptativas, fortalece la resiliencia del modelo y apoya la mejora continua en la seguridad de los sistemas de ciberdefensa impulsados por la IA.

Calidad de los Datos

La integridad y la calidad de los datos son pilares innegociables para una IA fiable en las operaciones de ciberseguridad. Garantizar que los conjuntos de datos permanezcan inalterados, libres de cargas maliciosas y sin contenido no autorizado requiere canales de adquisición seguros, procedencia de datos firmados criptográficamente y registros inmutables. Los procesos de validación exhaustivos deben detectar y eliminar duplicaciones, valores atípicos y anomalías estadísticas. Los datos de entrenamiento no representativos introducen un sesgo sistémico, lo que degrada la precisión del modelo y aumenta las tasas de falsos positivos/negativos, lo que compromete directamente la eficacia operativa de los flujos de trabajo CSIRT mejorados con IA.

Una gobernanza de datos sólida debe abarcar la gestión del ciclo de vida de extremo a extremo (adquisición, preprocesamiento, etiquetado, control de versiones y archivo) utilizando canales reproducibles y almacenamiento a prueba de manipulaciones. La clasificación de los datos en función de los niveles de sensibilidad, aplicada mediante el control de acceso basado en roles (RBAC) y el cifrado para los datos en reposo y en tránsito, garantiza la confidencialidad y el cumplimiento de los principios de privacidad desde el diseño. Las estrategias de minimización de datos y las tecnologías de mejora de la privacidad (PET) deben limitar la exposición, reducir la granularidad innecesaria y aplicar políticas de retención alineadas con los objetivos específicos de la misión y los mandatos legales.

La implementación de una arquitectura de confianza cero (ZTA) dentro de la infraestructura de datos de IA es esencial. Todas las solicitudes de acceso a los datos, ya sean iniciadas por humanos o por agentes de IA, deben ser autenticadas, autorizadas y verificadas continuamente. Los datos sensibles de entrenamiento de modelos deben procesarse dentro de enclaves seguros, utilizando hardware certificado y comprobaciones de integridad en tiempo de ejecución. Deben existir sistemas de monitorización telemétrica en tiempo real y de detección de anomalías para detectar patrones de acceso no autorizados o intentos de exfiltración de datos. Estas prácticas garantizan no solo la fidelidad del modelo, sino que también preservan la integridad, la confidencialidad y la fiabilidad de los sistemas de IA desplegados en entornos de ciberdefensa de alto riesgo.

Selección y Robustez de Modelos de IA

La selección de modelos de IA en un CSIRT es una decisión estratégica que debe equilibrar rendimiento técnico, tolerancia al riesgo y principios de gobernanza. Esta selección debe estar guiada por una clara definición del caso de uso, objetivos específicos y la participación de múltiples partes interesadas, asegurando el cumplimiento de requisitos multifuncionales. Es crucial comparar modelos de propósito general y especializados considerando precisión, fiabilidad, explicabilidad, velocidad de respuesta y cumplimiento normativo, evaluando además aspectos éticos y de privacidad.

La gobernanza de IA debe liderar el proceso de selección, integrando evaluaciones de riesgos desde el inicio. Un modelo técnicamente robusto pero opaco o sesgado puede representar un riesgo mayor que los problemas que busca mitigar. Documentar la arquitectura, los datos de entrenamiento y los benchmarks permite una comparación objetiva entre modelos, asegurando que la elección se alinee con la estrategia de gestión de riesgos y la ética organizacional.

La robustez del modelo ante ciberataques es vital para la ciberresiliencia del CSIRT. Estos ataques, como la evasión, el envenenamiento de datos, el robo de modelos y la inyección de prompts, explotan vulnerabilidades del modelo. La defensa en profundidad debe incluir entrenamiento adversarial, robustez certificada, destilación defensiva, privacidad diferencial, y monitoreo continuo. Cumplir con regulaciones o adoptar buenas prácticas exige que los sistemas de alto riesgo sean resilientes, capaces de prevenir, detectar y mitigar manipulaciones maliciosas.

Pruebas y Validación Continua de Modelos de IA

La validación continua y las pruebas automatizadas son fundamentales para mantener la adaptabilidad y resiliencia de los sistemas de IA en los CSIRTs frente a amenazas emergentes. Estas pruebas impulsadas por IA reducen falsos positivos y permiten detectar nuevas vulnerabilidades con mayor rapidez. Integradas en pipelines CI/CD, permiten testear de forma automatizada cada etapa del desarrollo, incluyendo SAST, DAST y pruebas de penetración basadas en simulaciones realistas y ajustables.

Estas pruebas de modelos ML presentan desafíos particulares debido a su comportamiento no determinista. Las pruebas de MLOps deben validar la lógica del código, la integridad de los datos y la salida del modelo, incluyendo análisis de regresión, detección de deriva y auditorías de equidad. Es esencial realizar pruebas adversariales frecuentes, validación con datasets diversos y pruebas de estrés para asegurar que los modelos mantengan su fiabilidad bajo condiciones cambiantes y hostiles.

La integración de pruebas automatizadas en CI/CD habilita un enfoque "shift-left", detectando defectos de seguridad desde las primeras etapas del desarrollo. Esto democratiza la gestión del riesgo de IA y asegura un despliegue más rápido y seguro. Para un CSIRT, este enfoque garantiza que la seguridad esté incorporada desde el inicio, no como una capa final, fortaleciendo la confianza en los sistemas de IA desplegados en entornos críticos.

MLOps y CI/CD para el Desarrollo de IA en Ciberseguridad

MLOps es el marco esencial para escalar, automatizar y gobernar el ciclo de vida del aprendizaje automático, integrando CI/CD al desarrollo de modelos. Incluye procesos clave como la preparación de datos, ingeniería de características, entrenamiento, ajuste, despliegue y monitoreo continuo. La automatización transforma tareas propensas a errores en flujos reproducibles, facilitando el despliegue ágil y seguro de modelos en entornos donde los datos fluyen en tiempo real.

El control de versiones en MLOps es integral, abarcando código, datos, configuraciones, pesos del modelo y resultados de experimentos. Esto permite trazabilidad, reproducibilidad y capacidad de auditoría, aspectos críticos para responder a incidentes, cumplir regulaciones y garantizar transparencia. La gestión rigurosa del linaje de los modelos refuerza la responsabilidad y auditabilidad, principios clave de una gobernanza de IA robusta.

En un CSIRT, donde las decisiones basadas en IA pueden tener consecuencias críticas, MLOps aporta capacidades para asegurar que los modelos desplegados sean fiables, explicables y seguros. Los pipelines CI/CD permiten actualizaciones continuas, minimizan cuellos de botella, y aseguran la coherencia del modelo a lo largo del tiempo, permitiendo escalar las defensas de ciberseguridad de forma proactiva y resiliente.

3. Adopción de la IA en CSIRT

Preparación Organizativa y Desarrollo de Capacidades

La base para la adopción exitosa de la IA radica en una evaluación exhaustiva de la preparación organizativa y el desarrollo estratégico de capacidades. Esta preparación abarca cuatro dimensiones críticas que deben abordarse de manera sistemática para garantizar la implementación sostenible de estas tecnologías de innovación emergentes y disruptivas. La preparación en materia de liderazgo y estrategia que requiere el compromiso de las partes interesadas, respaldado por estrategias claras alineadas con los objetivos de ciberseguridad, una asignación presupuestaria adecuada para iniciativas de IA plurianuales y marcos integrales de gestión de riesgos diseñados específicamente para la implementación de la IA en contextos de la CyberOps.

Preparación de la Infraestructura Técnica

Presenta retos particulares en ALC, donde los CSIRTs deben establecer capacidades sólidas de recopilación y almacenamiento de datos capaces de manejar grandes volúmenes de datos relacionados con la seguridad. Esto incluye garantizar recursos informáticos adecuados para las cargas de trabajo de IA, lo que puede requerir soluciones de infraestructura regionales compartidas o basadas en la nube para hacer frente a las restricciones de costes. Los requisitos de conectividad de red y ancho de banda deben planificarse cuidadosamente, especialmente para las aplicaciones de detección y respuesta a amenazas en tiempo real. Las capacidades de integración con las herramientas de seguridad existentes son esenciales para evitar la creación de sistemas de IA aislados que no puedan contribuir eficazmente a las operaciones generales de seguridad.

Preparación y Desarrollo del Talento Humano

Aborda el reto crítico de desarrollar y retener personal cualificado con experiencia tanto en el ámbito de la IA como en el de la ciberseguridad. Esto requiere programas integrales de formación y desarrollo que puedan implementarse a pesar de las restricciones presupuestarias y la competencia de las organizaciones del sector privado que ofrecen una remuneración más alta. Los sistemas de gestión del cambio y apoyo a la adopción deben diseñarse para abordar la resistencia cultural y garantizar una transición tecnológica fluida. La colaboración entre los equipos técnicos y operativos es fundamental para el éxito de la integración de la IA, lo que requiere nuevas estructuras organizativas y patrones de comunicación.

Preparación de los Datos y los Procesos

Se centra en el establecimiento de datos de seguridad estructurados y de alta calidad que puedan entrenar y operar eficazmente los sistemas de IA. Muchos CSIRTs de ALC deben realizar importantes inversiones en iniciativas de mejora de la calidad de los datos antes de que la implementación de la IA pueda tener éxito. Los procedimientos estandarizados de respuesta a incidentes proporcionan la base para la automatización y la ampliación de la IA. Deben establecerse marcos de gobernanza de datos y cumplimiento de la privacidad para abordar los distintos requisitos normativos de la región, al tiempo que se garantiza la utilización eficaz de los datos para las aplicaciones de IA.

Hay que reconocer que las iniciativas regionales de creación de capacidad desarrolladas por OEA, CSIRT-Américas, LAC4, ITU, BID y LACNIC hacen hincapié en los enfoques colaborativos que aprovechan los recursos y los conocimientos compartidos entre los países de ALC. Esto incluye el establecimiento de centros regionales de formación en IA y ciberseguridad que puedan ofrecer programas educativos especializados, el desarrollo de programas de becas que apoyen la educación avanzada en los campos de la IA y la ciberseguridad, y la creación de redes de mentores que conecten a los expertos regionales con los CSIRTs que están iniciando su andadura en la adopción de la IA. Los programas de intercambio de conocimientos entre países pueden facilitar el intercambio de enfoques de implementación exitosos y las lecciones aprendidas de iniciativas fallidas.

Integración de Herramientas de IA en las Operaciones del CSIRT

Estas integraciones representan una oportunidad transformadora para mejorar las capacidades de ciberseguridad en ALC. Las aplicaciones de detección de incidentes aprovechan las capacidades de reconocimiento de patrones de la IA para identificar anomalías en el tráfico de red y los registros del sistema que pueden indicar incidentes de seguridad. Los algoritmos de aprendizaje automático pueden analizar grandes volúmenes de datos de registro en tiempo real, identificando patrones sutiles que los analistas humanos podrían pasar por alto. Las capacidades de análisis del comportamiento permiten detectar amenazas internas y amenazas persistentes avanzadas que se basan en credenciales legítimas y actividades de apariencia normal. La búsqueda automatizada de amenazas utiliza la IA para buscar de forma proactiva indicadores de compromiso y patrones de ataque emergentes en las redes o en el ciberespacio.

Las aplicaciones de análisis de amenazas utilizan la IA para procesar y analizar la inteligencia sobre amenazas procedente de múltiples fuentes, lo que proporciona a los CSIRTs una mejor comprensión del panorama de amenazas. Los sistemas de análisis y clasificación de malware pueden categorizar automáticamente las nuevas amenazas e identificar similitudes con familias de ataques conocidas. El análisis de atribución ayuda a comprender qué actores de amenazas pueden estar apuntando contra ellas y las posibles motivaciones detrás de los ataques. Los sistemas de evaluación y priorización de vulnerabilidades utilizan la IA para analizar los activos y la inteligencia sobre amenazas con el fin de priorizar los esfuerzos de parcheo y mitigación basándose en los niveles de riesgo reales en lugar de en puntuaciones de gravedad teóricas.

Por otro lado, las aplicaciones de respuesta a incidentes se centran en automatizar las tareas rutinarias y proporcionar apoyo inteligente para la toma de decisiones a los gestores de incidentes. Los sistemas automatizados de clasificación y priorización de incidentes pueden procesar grandes volúmenes de alertas de seguridad e identificar aquellas que requieren atención humana inmediata. La correlación y el filtrado inteligentes de alertas reducen la fatiga de las alertas al combinar las alertas relacionadas y eliminar los falsos positivos. Las acciones automatizadas de respuesta y contención pueden implementar medidas de protección inmediatas mientras los responsables de la respuesta humana desarrollan estrategias de respuesta integrales. Los sistemas de apoyo a la toma de decisiones proporcionan a los gestores de incidentes inteligencia sobre amenazas relevante, historiales de incidentes similares y acciones de respuesta recomendadas.

Así mismo, las aplicaciones de inteligencia sobre amenazas mejoran la capacidad de los CSIRTs para comprender y anticipar las amenazas mediante la recopilación y el análisis automatizados de datos de inteligencia. Los sistemas de IA pueden supervisar continuamente las fuentes de inteligencia sobre amenazas, extrayendo indicadores relevantes de compromiso e información sobre los autores de las amenazas. Los sistemas de validación de indicadores ayudan a distinguir entre fuentes de inteligencia fiables y poco fiables. La elaboración de perfiles de los actores de las amenazas y el seguimiento de las campañas proporcionan información sobre los patrones de comportamiento de los adversarios y la evolución de sus tácticas. Las capacidades de análisis predictivo ayudan a anticipar las amenazas emergentes y a preparar medidas defensivas con antelación.

Finalmente, la implementación de estas aplicaciones de IA requiere una cuidadosa consideración de las capacidades y prioridades de la organización. Los CSIRTs deben comenzar con proyectos piloto en áreas donde tengan la mejor calidad de datos y la mejor infraestructura técnica. El éxito en las implementaciones iniciales genera confianza en la organización y proporciona experiencia práctica que respalda la expansión a otras áreas operativas.

Consideraciones Culturales, Éticas y Centradas en el Ser Humano

La adopción de esta tecnología en las ciberoperaciones debe abordar importantes consideraciones culturales, éticas y centradas en el ser humano que son especialmente relevantes en el contexto de ALC. La adaptación cultural requiere reconocer que los patrones de adopción de la tecnología varían significativamente en toda la región, ya que algunos CSIRTs según sus capacidades y madurez adoptan rápidamente los cambios tecnológicos, mientras que otros prefieren enfoques graduales y basados en el consenso. El respeto por los procesos y jerarquías locales de toma de decisiones es esencial para el éxito de la implementación de la IA, al igual que la integración con las culturas y prácticas organizativas existentes que han demostrado su eficacia en las operaciones de ciberseguridad.

Así mismo, los estilos y preferencias de comunicación regionales deben tenerse en cuenta al diseñar los sistemas de IA y sus interfaces. Algunas culturas hacen hincapié en la toma de decisiones colectiva y la consulta exhaustiva, mientras que otras se basan en estructuras de autoridad jerárquicas. Los sistemas de IA deben diseñarse para apoyar estos diferentes enfoques, en lugar de imponer patrones de interacción uniformes. Es esencial reconocer la diversidad de contextos lingüísticos y culturales en toda la región de ALC, en particular para los sistemas que proporcionan capacidades de análisis textual o de procesamiento del lenguaje

natural. Que ya algunos países están visualizando el reporte y notificación por ciudadanos y que mediante IA puedan tener respuestas.

En ese orden, la implementación ética de la IA en las ciberoperaciones plantea importantes cuestiones sobre la transparencia, la rendición de cuentas y los derechos humanos. La transparencia en los procesos de toma de decisiones de la IA para la respuesta a incidentes es fundamental para mantener la confianza y permitir una supervisión humana eficaz. Sin embargo, la transparencia total puede entrar en conflicto con los requisitos de seguridad operativa, lo que crea una tensión que debe gestionarse con cuidado. La equidad y la no discriminación en los sistemas automatizados de detección de amenazas son esenciales para evitar que los sesgos de la IA afecten a las operaciones de seguridad o se dirijan de forma injusta a grupos de usuarios o segmentos de red específicos.

Los mecanismos de rendición de cuentas para las decisiones de seguridad impulsadas por la IA deben definir claramente la responsabilidad humana por los resultados del sistema de IA, al tiempo que permiten operaciones eficientes. Esto incluye el establecimiento de protocolos claros sobre cuándo los seres humanos deben revisar las recomendaciones de la IA antes de su implementación y la creación de registros de auditoría que permitan el análisis posterior al incidente de los procesos de toma de decisiones de la IA. La protección de la privacidad en el procesamiento y análisis de datos de IA debe cumplir con los diversos requisitos reglamentarios de los países de ALC, al tiempo que permite una detección y respuesta eficaces a las amenazas.

Por tanto, son muy importantes los principios de diseño centrado en el ser humano que garantizan que los sistemas de IA aumenten, en lugar de sustituir, la experiencia humana en las ciberoperaciones. Mantener la supervisión humana en las decisiones críticas de seguridad preserva el control de los CSIRTs y permite la adaptación a situaciones únicas que los sistemas de IA pueden no manejar de manera eficaz. Los sistemas de IA deben diseñarse para proporcionar explicaciones y contexto para sus recomendaciones, lo que permite a los analistas de ciberseguridad tomar decisiones informadas sobre su implementación.

Garantizar la accesibilidad y la facilidad de uso para diversos niveles de conocimientos técnicos es especialmente importante en los CSIRTs con recursos limitados para programas de formación exhaustivos. Las interfaces de IA deben ser intuitivas y proporcionar los niveles adecuados de detalle técnico para las diferentes funciones de los usuarios. La formación y el apoyo integrales para la adopción de herramientas de IA deben proporcionarse de forma continua, y no solo durante las fases iniciales de implementación. La creación de mecanismos de retroalimentación para la mejora continua de la colaboración entre humanos e IA permite optimizar sus implementaciones de IA basándose en la experiencia operativa.

Barreras y Retos Regionales en ALC

Los CSIRTs de ALC se enfrentan a retos específicos que repercuten de manera significativa en las estrategias de adopción de la IA y los enfoques de implementación. Las limitaciones de recursos representan la barrera más inmediata, ya que la asignación presupuestaria limitada para las tecnologías e infraestructuras de IA dificulta la toma de decisiones sobre las prioridades a los responsables de la ciberseguridad. La escasez de profesionales cualificados en ciberseguridad con experiencia en IA es especialmente grave en la región, donde las instituciones educativas han tardado en desarrollar planes de estudios completos sobre IA y ciberseguridad. La infraestructura técnica inadecuada para la implementación de la IA afecta a muchas organizaciones, en particular a los CSIRTs más pequeños que carecen de los recursos informáticos necesarios para un funcionamiento eficaz de la IA. La competencia con el sector privado por la retención de talento plantea retos constantes para mantener los conocimientos especializados.

Al mismo tiempo, la diversidad normativa entre los países de región crea requisitos de cumplimiento complejos que afectan a las estrategias de implementación de la IA. La variabilidad de las normativas de ciberseguridad implica que los sistemas de IA deben diseñarse para adaptarse a los diferentes requisitos legales en materia de tratamiento de datos, notificación de

incidentes y cooperación internacional. La falta de marcos armonizados de gobernanza de la IA en la región impide el desarrollo de normas coherentes que podrían facilitar la cooperación regional y el intercambio de recursos. La falta de coherencia de las leyes de protección de datos y privacidad plantea retos para los sistemas de IA que dependen del intercambio transfronterizo de información sobre amenazas y datos sobre incidentes. La limitada coordinación regional en materia de normas de ciberseguridad impide el desarrollo de enfoques comunes que podrían reducir los costes y la complejidad de la implementación.

En otro orden las carencias de talento reflejan retos educativos y económicos más amplios en la región. La insuficiencia de los programas educativos en IA y ciberseguridad obliga a invertir mucho en la formación del personal existente en lugar de contratar a especialistas. La limitada disponibilidad de formación especializada en IA para la ciberseguridad crea dependencia de costosos programas de formación externos o consultores. La fuga de cerebros, ya que los profesionales emigran a mercados con salarios más altos, especialmente en América del Norte y Europa, crea retos continuos de retención. La falta de oportunidades de desarrollo profesional continuo significa que los CSIRTs tienen dificultades para mantener a su personal técnico al día con las tecnologías de IA en rápida evolución.

Además, las barreras culturales y organizativas suelen resultar más difíciles que los obstáculos técnicos. La resistencia al cambio en las estructuras organizativas tradicionales puede impedir los enfoques flexibles necesarios para la adopción exitosa de la IA. La cultura reacia al riesgo en el sector público, que comprenden muchos CSIRTs de ALC, crea reticencia a adoptar nuevas tecnologías que puedan introducir riesgos operativos. El escaso conocimiento de los beneficios de la IA para las operaciones de ciberseguridad hace que los responsables de la toma de decisiones no den la prioridad adecuada a las inversiones en IA. Los procesos inadecuados de gestión del cambio para la adopción de tecnología dan lugar a implementaciones fallidas, incluso cuando se cumplen los requisitos técnicos.

Finalmente, estos retos requieren respuestas regionales coordinadas que aprovechen los recursos y los conocimientos compartidos. Los CSIRTs individuales, por sí solo no pueden abordar estas barreras de forma eficaz de manera aislada, lo que crea oportunidades para iniciativas de cooperación regional que benefician a todos los participantes.

4. Recomendaciones y Mejores Practicas

La Gobernanza, Desarrollo y Adopción de la IA en las CyberOps de los CSIRTs de América Latina y el Caribe requiere estrategias integrales que aborden los cuatro principios rectores de una implementación sostenible, resiliente, responsable y adaptable. La adopción sostenible de la IA comienza con evaluaciones realistas de las capacidades y limitaciones de la organización, lo que conduce a enfoques de implementación por fases que desarrollan las capacidades de forma incremental. Los CSIRTs deben comenzar con proyectos piloto e implementaciones de prueba de concepto que demuestren su valor, al tiempo que desarrollan la experiencia interna. El establecimiento de métricas y criterios de éxito claros para las iniciativas de IA permite una evaluación objetiva del progreso y facilita la mejora continua.

El desarrollo de hojas de ruta de implementación por fases con plazos realistas evita el exceso de compromiso de recursos y permite ajustes basados en el aprendizaje de las estrategias de implementación. La creación de bucles de retroalimentación para la mejora y el aprendizaje continuos garantiza que los CSIRTs puedan adaptar sus implementaciones de IA basándose en la experiencia operativa. Garantizar la asignación de presupuesto para el mantenimiento y las actualizaciones a largo plazo evita la implementación de sistemas de IA que no pueden mantenerse a lo largo del tiempo. Los enfoques sostenibles también hacen hincapié en la creación de conocimientos locales para reducir la dependencia de proveedores externos y crear capacidades regionales a largo plazo.

Los sistemas de IA resilientes requieren enfoques arquitectónicos que mantengan las capacidades operativas incluso cuando los componentes de IA fallan o funcionan mal. La implementación de procedimientos robustos de copia de seguridad y recuperación de datos garantiza que los sistemas de IA puedan recuperarse de la corrupción o pérdida de datos. El diseño de sistemas de IA tolerantes a fallos con mecanismos de respaldo permite continuar las operaciones cuando los sistemas de IA se encuentran con condiciones inesperadas. El establecimiento de redundancia en funciones críticas de seguridad impulsadas por IA evita que los puntos únicos de fallo comprometan las operaciones de seguridad generales.

Las pruebas periódicas del rendimiento de los sistemas de IA en diversos escenarios, incluidas condiciones adversas y calidad de datos degradada, ayudan a identificar posibles debilidades antes de que afecten a las operaciones. Mantener la supervisión humana y las capacidades de anulación manual garantiza que los CSIRTs puedan responder de manera eficaz cuando los sistemas de IA proporcionan recomendaciones inadecuadas o fallan por completo. Los sistemas resilientes también incorporan capacidades de supervisión y alerta continuas que notifican a los operadores cuando el rendimiento del sistema de IA se degrada.

La implementación responsable de la IA requiere marcos éticos integrales que aborden los retos únicos que plantea el uso de la IA en contextos de ciberseguridad. La realización de revisiones éticas de las aplicaciones de IA en ciberseguridad ayuda a identificar posibles consecuencias negativas antes de su implementación. Garantizar la transparencia en los procesos de toma de decisiones de la IA permite la rendición de cuentas, al tiempo que se equilibran los requisitos de seguridad operativa. La implementación de estrategias de detección y mitigación de sesgos evita que los sistemas de IA produzcan resultados discriminatorios o se dirijan de manera injusta a poblaciones específicas.

El establecimiento de marcos de responsabilidad para las decisiones impulsadas por la IA aclara la responsabilidad humana por los resultados de los sistemas de IA y crea cadenas de responsabilidad claras para los incidentes relacionados con la IA. La protección de la privacidad y la confidencialidad en el procesamiento de datos de IA requiere una atención cuidadosa a los principios de minimización de datos y controles de acceso estrictos. La implementación responsable también incluye la auditoría periódica del rendimiento y el impacto de los sistemas de IA para garantizar la alineación continua con los principios éticos.

Las operaciones de IA adaptables se centran en la creación de sistemas que puedan evolucionar con los cambios en el panorama de amenazas y los requisitos de la organización. El diseño de

arquitecturas de IA modulares que faciliten las actualizaciones y modificaciones permite a los CSIRTs incorporar nuevas capacidades sin necesidad de rediseñar completamente el sistema. La implementación de procesos de aprendizaje continuo y reentrenamiento de modelos garantiza que los sistemas de IA sigan siendo eficaces frente a las amenazas en constante evolución. El establecimiento de un sistema de supervisión del panorama de amenazas para la adaptación de los sistemas de IA ayuda a identificar cuándo un sistema de IA necesitan actualizaciones o reconfiguraciones.

La creación de capacidades de integración flexibles con herramientas de seguridad en evolución evita que los sistemas de IA se aíslen y pierdan eficacia a medida que los CSIRTs actualizan su infraestructura de seguridad. El mantenimiento del control de versiones y las capacidades de reversión de los modelos de IA permite a los CSIRTs revertir las actualizaciones problemáticas y mantener la estabilidad operativa. Las operaciones adaptables también requieren la formación y el desarrollo continuos del personal para garantizar que los operadores humanos puedan trabajar eficazmente con las capacidades de IA en evolución.

La cooperación regional mejora la eficacia de estas mejores prácticas al permitir el intercambio de recursos y la resolución colaborativa de problemas. El establecimiento de centros regionales de excelencia para la IA en ciberseguridad puede proporcionar conocimientos especializados y formación que los CSIRTs individuales no pueden desarrollar de forma independiente. La creación de plataformas compartidas de inteligencia sobre amenazas que incorporen capacidades de análisis de IA puede mejorar la ciberseguridad regional, respetando al mismo tiempo los requisitos de soberanía nacional.

Marco de Adopción de la IA

Gobierno

La dimensión de la gobernanza abarca la supervisión estratégica y los marcos normativos necesarios para garantizar que las implementaciones de IA se ajusten a los objetivos organizativos y los requisitos normativos. En el caso de los CSIRTs de ALC, las acciones de gobernanza deben abordar el complejo panorama normativo de los diferentes países, al tiempo que se mantiene la coherencia con las normas internacionales de ciberseguridad. Esto incluye establecer mecanismos claros de rendición de cuentas, definir los niveles de tolerancia al riesgo y garantizar la alineación continua entre las capacidades de IA y las misiones organizativas.

Desarrollo

Las acciones de desarrollo se centran en la implementación técnica y los aspectos de ingeniería de los sistemas de IA. Estas acciones requieren conocimientos especializados tanto en el ámbito de la inteligencia artificial como en el de la ciberseguridad, lo que plantea retos particulares para las organizaciones de América Latina y el Caribe que se enfrentan a la escasez de talento. Las actividades de desarrollo deben hacer hincapié en enfoques de implementación prácticos que maximicen el uso de los recursos disponibles y, al mismo tiempo, desarrollen las capacidades técnicas locales. Esto incluye la adopción de arquitecturas modulares que permitan el desarrollo incremental de capacidades y la transferencia de conocimientos.

Operación

Las acciones operativas abordan la gestión y la utilización diarias de los sistemas de IA dentro de los flujos de trabajo de los CSIRT. En el caso de los CSIRT de América Latina y el Caribe, las consideraciones operativas deben tener en cuenta las diferentes capacidades de infraestructura y la disponibilidad de recursos en toda la región. Estas acciones hacen hincapié en la creación de modelos operativos sostenibles que puedan adaptarse a los cambiantes panoramas de amenazas, al tiempo que se mantienen altos niveles de prestación de servicios. El marco operativo también aborda los patrones de colaboración entre humanos e IA que son esenciales para la adopción exitosa de la tecnología.

Anexo I. Caja de Herramientas

Esta completa caja de herramientas proporciona a los CSIRTs de ALC recursos esenciales para crear CyberOps resilientes, seguras y adaptables impulsadas por la IA. La colección incluye marcos fundamentales, que establecen estructuras de gobernanza para la adopción responsable de la IA. Los recursos centrados en la seguridad abordan las vulnerabilidades y amenazas únicas asociadas a los sistemas de IA en contextos de ciberseguridad. Las directrices de implementación técnica a través de los marcos proporcionan hojas de ruta prácticas para integrar las capacidades de IA en los procesos existentes de respuesta a incidentes y detección de amenazas.

Los recursos regionales de organizaciones como la OEA y LACNIC garantizan que las implementaciones tengan en cuenta los entornos normativos locales y las oportunidades de colaboración dentro del ecosistema de ciberseguridad de América Latina y el Caribe. En conjunto, estos recursos permiten a los CSIRTs desarrollar estrategias de adopción de IA sostenibles, resilientes, responsables y adaptables que mejoran su capacidad para detectar, analizar y responder a las amenazas cibernéticas en constante evolución, al tiempo que mantienen los estándares éticos y la cooperación regional.

Estos recursos apoyan colectivamente la implementación de estrategias de adopción de IA sostenibles, resilientes, responsables y adaptables que pueden transformar las capacidades de los CSIRTs al tiempo que abordan los retos y oportunidades únicos que presenta el panorama de la ciberseguridad en ALC. Los CSIRTs deben utilizar múltiples recursos de diferentes categorías para garantizar una cobertura completa de todos los aspectos de la adopción de la IA, desde el desarrollo inicial del marco de gobernanza hasta la implementación operativa y los procesos de mejora continua.

Herramienta	Utilidad
OWASP	• AI Security • AI Testing Guide • Agentic AI Security and Governance • GenAI Incident Response Guide • Securing Agentic Applications Guide • Agent Name Service (ANS) for Secure AI Agent Discovery • Multi-Agentic system Threat Modeling Guide • OWASP Top 10 for LLM Applications • OWASP LLM Exploit Generation • Agentic AI – Threats and Mitigations • LLM and Gen AI Data Security Best Practices • GenAI Red Teaming Guide • LLM and Generative AI Security Solutions Landscape • AI Security Solution Cheat Sheet • LLM and Generative AI Security Center of Excellence Guide • Guide for Preparing and Responding to Deepfake Events • LLM Applications Cybersecurity and Governance Checklist • AIVSS Scoring System for OWASP Agentic AI Core Security Risks v0.5 • GenAI Security Project – Threat Defense COMPASS RunBook • GenAI Security Project Threat Defense COMPASS 1.0
AI Defend	• A defense Framework for Securing AI/ML Systems
MITRE AI/ML	• MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)
SANS	• AI Critical Security Guidelines
Cyber Security Agency of Singapore	• Securing Artificial Intelligence (AI): A Collective Responsibility • Guidelines on Securing AI Systems • Companion Guide on Securing AI Systems • Engaging with Artificial Intelligence • Guidelines for secure AI system development
Institute for AI Policy and Strategy	• Verification for International AI Governance • Managing Risks from Internal AI Systems • Asymmetry by Design: Boosting Cyber Defenders with Differential Access to AI • AI Agent Governance: A Field Guide

Herramienta	Utilidad
	• Mapping Technical Safety Research at AI Companies • Coordinated Disclosure of Dual-Use Capabilities: An Early Warning System for Advanced AI • What Can Be Discussed in Dialogues About Advanced AI Risks Without Leaking Sensitive Information? • Responsible Reporting for Frontier AI Development • Response to the NIST RFI on Auditing, Evaluating, and Red-Teaming AI Systems • Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework • Adapting Cybersecurity Frameworks to Manage Frontier AI Risks: A Defense-in-Depth Approach • Deployment Corrections: An Incident Response Framework for Frontier AI Models • How Expertise in AI hardware Can Help with AI Governance • Open-Sourcing Highly Capable Foundation Models
Google's SAIF	• Google's Secure AI Framework
OXFORD	• Who Should Develop Which AI Evaluations? • Open Problems in Machine Unlearning for AI Safety •
Centre for The Governance of AI	• Infrastructure for AI Agents • IDs for AI Systems • Visibility into AI Agents
Offensive Security AI Compilation	• Offensive Security AI
Joe O'Brien, Shaun Ee, Zoe Williams	• An incident response framework for frontier AI models
Organización para la Cooperación y el Desarrollo Económico (OECD)	• Regulatory sandboxes in artificial intelligence • Measuring the environmental impacts of artificial intelligence compute and applications • AI openness • Towards a common reporting framework for AI incidents • Exploring win-win outcomes of algorithmic management • Is generative AI a General Purpose Technology? • Sharing trustworthy AI models with privacy-enhancing technologies • A blueprint for building national compute capacity for artificial intelligence • Defining AI incidents and related terms • Responsible business conduct and anticipatory governance of emerging technology • Scoping the OECD AI principles • AI language models • AI, data governance and privacy • Framework for the Classification of AI systems • Tools for trustworthy AI • Common guideposts to promote interoperability in AI risk management • Advancing accountability in AI • Harnessing the power of AI and emerging technologies • Framework for Anticipatory Governance of Emerging Technologies • Building digital workforce capacity and skills for data-intensive science • Introducing the OECD AI Capability Indicators • How are AI developers managing risks?
Banco Interamericano de Desarrollo (BID)	• Reporte de tecnología: inteligencia artificial • Reporte de tecnología: ética digital • Reporte de tecnología: IA Generativa • Uso responsable de la IA para las políticas públicas: manual de ciencia de datos • Robot Laura - Auditoría Algorítmica • La gestión ética de los datos • Reporte de tecnología: Legaltech • Inteligencia artificial: Gran oportunidad del siglo XXI: Documento de reflexión y propuesta de actuación • fAIr LAC: Adopción ética y responsable de la inteligencia artificial en América Latina y el Caribe • Algoritmolandia: inteligencia artificial para una integración predictiva e inclusiva de América Latina • IA desde los cimientos: desafíos y oportunidades en el contexto de América Latina y el Caribe

Herramienta	Utilidad
	• ¿Quién le teme a la inteligencia?: Posibilidades y riesgos de la inteligencia artificial en el Estado digital • Uso responsable de IA para política pública: manual de formulación de proyectos • Gobernanza participativa de la inteligencia artificial
Federal Office for Information Security	• Design Principles for LLM-based Systems with Zero Trust
Organización de los Estados Americanos (OEA)	• Data Classification • Inter-American Guidelines on Data Governance and Artificial Intelligence • Data Governance Model Policy Content • Data Governance and Artificial Intelligence Americas Situation Content
European Union	• The General-Purpose AI Code of Practice
Cloud Security Alliance	• Risk Rubric AI • Analyzing Log Data with AI Models • Agentic AI Identity and Access Management: A New Approach • Secure Agentic System Design: A Trait-Based Approach • AI Controls Matrix • Dynamic Process Landscape: A Strategic Guide to Successful AI Implementation • Agentic AI Red Teaming Guide • AI Organizational Responsibilities: AI Tools and Applications • AI Risk Management: Thinking Beyond Regulatory Boundaries • AI Organizational Responsibilities - Governance, Risk Management, Compliance and Cultural Aspects • Don't Panic! Getting Real about AI Governance • Securing LLM Backed Systems: Essential Authorization Practices • Using AI for Offensive Security • AI Model Risk Management Framework • CSA Large Language Model (LLM) Threats Taxonomy • Confronting Shadow Access Risks: Considerations for Zero Trust and Artificial Intelligence Deployments • AI Organizational Responsibilities - Core Security Responsibilities • AI Resilience: A Revolutionary Benchmarking Model for AI Safety • Principles to Practice: Responsible AI in a Dynamic Regulatory Environment • Security Implications of ChatGPT
Bases de Datos de Incidentes y Riesgos	• MIT AI Incident Tracker • AIID AI Incident Database • AI Controversy Repository • MITRE AI Risk Database • OECD AI Incidents Monitor (AIM) • DAIL The Database of AI Litigation • Label Errors Database, Goals, Methods, and Failures (GMF) • Center for Security and Emerging Technology (CSETv1)
Otros	• Computing Power and the Governance of Artificial Intelligence • Open Problems in Cooperative AI

Colaboradores

Estevenson Solano
Especialista en Ciberseguridad
Autor

Alejandra Erramuspe
Políticas Públicas y Gobernanza de Internet
Mentora de Investigación – LACNIC

Oliver Montilla
Director de Diseño Creativo y Arte – MILO Studio

Este proyecto ha sido desarrollado en el marco del programa LACNIC Líderes, con financiación y tutoría de LACNIC, dentro del eje temático de Tecnologías Emergentes, Innovación Sostenible e Inteligencia Artificial, en representación de América Latina y el Caribe (Republica Dominicana).

Agradecimientos

Banco Interamericano de Desarrollo (BID)

Center for Public Intelligence (CPI)

Centro de Respuesta a Incidentes de seguridad Informática de LACNIC (LACNIC CSIRT)

Equality AI

Instituto Tecnológico de las Américas (ITLA)

Instituto Tecnológico de Santo Domingo (INTEC)

Internet Society (ISOC)

Latin America and Caribbean Cyber Competence Centre (LAC4)

Organización de los Estados Americanos (OEA)

Red de Excelencia en Ciberseguridad de Latinoamérica y el Caribe (Red Ciberlac)

Red de Investigación en Criminología y Ciberseguridad (Red Ciber)

Red de los Equipos de Respuesta ante Incidentes Cibernéticos (CSIRT- Américas)

Red Iberoamericana de Protección de Datos (RIPD)

Sociedad Dominicana de Ciencia de Datos (BigDataDO)

Universidad Ean

El autor agradece sinceramente a todos los profesionales e instituciones que aportaron su tiempo, conocimiento y orientación al desarrollo de este proyecto. Su contribución ha sido fundamental para la realización de este trabajo.

Referencias Bibliográficas

1 World Economic Forum, "Unlocking Value from Generative AI: Guidance for Responsible Transformation," 2024. Accessed: Aug. 05, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Unlocking_Value_from_Generative_AI_2024.pdf

2 World Economic Forum, "The AI Governance Journey: Development and Opportunities," 2021. Accessed: Aug. 05, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_The%20AI_Governance_Journey_Development_and_Opportunities_2021.pdf

3 World Economic Forum, "AI Governance Alliance: Briefing Paper Series," 2024. Accessed: Aug. 05, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_AI_Governance_Alliance_Briefing_Paper_Series_2024.pdf

4 World Economic Forum, "Empowering AI Leadership: AI C-Suite Toolkit," 2022. Accessed: Aug. 05, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Empowering_AI_Leadership_2022.pdf

5 World Economic Forum, "Generative AI Governance: Shaping a Collective Global Future," 2024. Accessed: Aug. 05, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Generative_AI_Governance_2024.pdf

6 World Economic Forum, "AI Governance: A Holistic Approach to Implement Ethics into AI," 2019. Accessed: Aug. 05, 2025. [En línea]. Disponible: https://www.weforum.org/publications/ai-governance-a-holistic-approach-to-implement-ethics-into-ai/

7 D. Leslie, "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector," Jun. 2019, *Zenodo*. doi: 10.5281/zenodo.3240529.

8 OECD, "Futures of Global AI Governance: Co-Creating an Approach for Transforming Economies and Societies," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.oecd.org/content/dam/oecd/en/about/programmes/strategic-foresight/GSG%20Background%20Note_GSG(2024)1en.pdf/_jcr_content/renditions/original./GSG%20Background%20Note_GSG(2024)1en.pdf

9 OECD, "AI: Intelligent machines, smart policies," Aug. 2018. doi: 10.1787/f1a650d9-en.

10 P. , K. P. and J. B. Lorenz, "Initial policy considerations for generative artificial intelligence," Sep. 2023. doi: 10.1787/fae2d1e6-en.

11 L. , K. P. and F. S. Galindo, "An overview of national AI strategies and policies," Aug. 2021. doi: 10.1787/c05140d9-en.

12 ISMS Forum, "Modelo de Gobierno IA." Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.ismsforum.es/ficheros/descargas/modelo-de-gobiernoconverted1700037593.pdf

13 OECD, "Steering AI's future," Feb. 2025. doi: 10.1787/5480ff0a-en.

14 OECD, "Framework for Anticipatory Governance of Emerging Technologies," Apr. 2024. doi: 10.1787/0248ead5-en.

15 World Economic Forum, "Governance in the Age of Generative AI: A 360° Approach for Resilient Policy and Regulation," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Governance_in_the_Age_of_Generative_AI_2024.pdf

16 World Economic Forum, "A Framework for Developing a National Artificial Intelligence Strategy," 2019. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.weforum.org/publications/a-framework-for-developing-a-national-artificial-intelligence-strategy/

17 Info-communications Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC), "Model Artificial Intelligence Governance Framework," 2020. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

18 World Economic Forum, "Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Adopting_AI_Responsibly_Guidelines_for_Procurement_of_AI_Solutions_by_the_Private_Sector_2023.pdf

19 Cloud Security Alliance, "AI Risk Management: Thinking Beyond Regulatory Boundaries," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/ai-risk-management-thinking-beyond-regulatory-boundaries

20 SANS, "Critical AI Security Guidelines," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://sansorg.egnyte.com/dl/HMYtX6Gdgd3W

21 OECD, "Sharing trustworthy AI models with privacy-enhancing technologies," Jun. 2025. doi: 10.1787/a266160b-en.

22 OECD, "AI, data governance and privacy," Jun. 2024. doi: 10.1787/2476b1a4-en.

23 World Economic Forum, "Data Equity: Foundational Concepts for Generative AI," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Data_Equity_Concepts_Generative_AI_2023.pdf

24 Government Digital Service UK, "Data Ethics Framework," 2020. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://assets.publishing.service.gov.uk/media/5f74a4958fa8f5188dad0e99/Data_Ethics_Framework_2020.pdf

25 ENISA, "Data Protection Engineering," 2022. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20Data%20Protection%20Engineering.pdf

26 OECD, "Enhancing Access to and Sharing of Data in the Age of Artificial Intelligence," 2025. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/enhancing-access-to-and-sharing-of-data-in-the-age-of-artificial-intelligence_d70ccab9/23a70dca-en.pdf

27 Cloud Security Alliance, "AI Organizational Responsibilities - Governance, Risk Management, Compliance and Cultural Aspects," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/ai-organizational-responsibilities-governance-risk-management-compliance-and-cultural-aspects

28 Cloud Security Alliance, "AI Organizational Responsibilities: AI Tools and Applications," 2025. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/ai-organizational-responsibilities-ai-tools-and-applications

29 Gobierno de Israel, "Israel´s Policy on Artificial Intelligence: Regulations and Ethics," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.gov.il/BlobFolder/policy/ai_2023/en/Israels%20AI%20Policy%202023.pdf

30 SANS, "Artificial Intelligence Standard," 2025. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://sansorg.egnyte.com/dl/hHTk98iZYW

31 World Economic Forum, "Guidelines for AI Procurement," 2019. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Guidelines_for_AI_Procurement.pdf
32 World Economic Forum, "AI Procurement in a Box: AI Government Procurement Guidelines," 2020. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_AI_Government_Procurement_Guidelines_2020.pdf
33 ISMS Forum, "Inteligencia Artificial y Ciberseguridad," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.ismsforum.es/ficheros/descargas/isms-gt-ia-021707141605.pdf
34 OECD, "Common guideposts to promote interoperability in AI risk management," Nov. 2023. doi: 10.1787/ba602d18-en.
35 Executive Office of the President, *Executive Order 14110 Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Federal Register, 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence
36 ISMS Forum, "Guía de iniciación en la Seguridad aplicada al DevOps," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.ismsforum.es/ficheros/descargas/guia-devsecops-v41690197585.pdf
37 Cloud Security Alliance, "Using AI for Offensive Security," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/using-ai-for-offensive-security
38 United States Intelligence Community, "Artificial Intelligence Ethics Framework For The Intelligence Community," 2020. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.intelligence.gov/assets/documents/pdf/ai/ai-ethics-framework-for-the-ic-1_0.pdf
39 Department of Homeland Security, "Increasing Threat of Deepfake Identities," 2019. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_addendum_0.pdf
40 Rhite, "Advancing the field of bias detection and mitigation in Large Language Models and Traditional AI Models," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://rhite.tech/files/bias-detection-in-llms-and-traditional-ai-models_extended.pdf
41 Rhite, "From Inception to Retirement: Addressing Bias Throughout the Lifecycle of AI Systems ," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://rhite.tech/files/From-Inception-to-Retirement-Addressing-Bias-Throughout-the-Lifecycle-of-AI-Systems.pdf
42 World Economic Forum, "Artificial Intelligence for Efficiency, Sustainability and Inclusivity in TradeTech," 2025. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://reports.weforum.org/docs/WEF_Artificial_Intelligence_for_Efficiency_Sustainability_and_Inclusivity_in_TradeTech_2025.pdf
43 World Economic Forum, "Artificial Intelligence and Cybersecurity: Balancing Risks and Rewards," 2025. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://reports.weforum.org/docs/WEF_Artificial_Intelligence_and_Cybersecurity_Balancing_Risks_and_Rewards_2025.pdf
44 Cloud Security Alliance, "AI Model Risk Management Framework," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/ai-model-risk-management-framework
45 World Economic Forum, "Responsible AI Playbook for Investors," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Responsible_AI_Playbook_for_Investors_2024.pdf
46 OECD, "Regulatory sandboxes in artificial intelligence," Jul. 2023. doi: 10.1787/8f80a0e6-en.
47 Cloud Security Alliance, "Don't Panic! Getting Real about AI Governance," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/dont-panic-getting-real-about-ai-governance
48 E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," Jan. 2023. doi: 10.6028/NIST.AI.100-1.
49 R. Dotan, B. Blili-Hamelin, R. Madhavan, J. Matthews, and J. Scarpino, "Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework," Feb. 2024.
50 MITRE, "The MITRE AI Maturity Model and Organizational Assessment Tool Guide: A Path to Successful AI Adoption," 2022. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://aimaturitymodel.mitre.org/wp-content/uploads/2023/06/MITRE-AI-Maturity-Model-11-17-2022.pdf
51 MITRE, "Artificial Intelligence Maturity Model ," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.mitre.org/sites/default/files/2024-06/PR-24-1492-AI-Maturity-Model-6-24-factsheet.pdf
52 World Economic Forum, "The Presidio Recommendations on Responsible Generative AI," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Presidio_Recommendations_on_Responsible_Generative_AI_2023.pdf
53 AI HLEG, "ETHICS GUIDELINES FOR TRUSTWORTHY AI," 2019. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf
54 European Parliament and of the Council, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*. EUR-LEX, 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
55 D. , R. C. , B. M. , P. A. , J. S. , B. A. , B. SJ. , B. C. , H. S. , and F. C. Leslie, "AI Accountability in Practice ," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://aiethics.turing.ac.uk/modules/accountability/
56 D. , R. C. , B. M. , P. A. , J. S. , B. A. , B. SJ. B. C. , A. M. , M. S. , W. J. , W. M. , and F. C. Leslie, "AI Explainability in Practice," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://aiethics.turing.ac.uk/modules/explainability/
57 D. , R. C. , B. M. , P. A. , J. S. , B. A. , B. SJ. , B. C. , and F. C. Leslie, "AI Safety in Practice ," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://aiethics.turing.ac.uk/modules/safety/
58 D. , R. C. , B. M. , P. A. , J. S. , B. A. , B. SJ. , B. C. , A. M. , M. S. , W. J. , H. Y. , and F. C. Leslie, "Responsible Data Stewardship in Practice ," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://aiethics.turing.ac.uk/modules/data-stewardship/

59 D. , R. C. , B. M. , P. A. , J. S. , B. A. , B. SJ. B. C. , A. M. , K. M. , F. C. , W. J. , and K. G. I. Leslie, "AI Fairness in Practice," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://aiethics.turing.ac.uk/modules/fairness/
60 D. , R. C. , B. M. , P. A. , J. S. , B. A. , B. SJ. B. C. , A. M. , K. M. , F. C. , W. J. , and K. G. I. Leslie, "AI Sustainability in Practice Part Two: Sustainability Throughout the AI Workflow," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://aiethics.turing.ac.uk/modules/sustainability-2/
61 D. , R. C. , B. M. , P. A. , J. S. , B. A. , B. SJ. B. C. , A. M. , K. M. , F. C. , W. J. , and K. G. I. Leslie, "AI Sustainability in Practice Part One: Foundations for Sustainable AI Projects ," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://aiethics.turing.ac.uk/modules/sustainability-1/
62 D. , R. C. , B. M. , P. A. , J. S. , B. A. , B. SJ. B. C. , A. M. , K. M. , F. C. Leslie, "AI Ethics and Governance in Practice: An Introduction ," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://aiethics.turing.ac.uk/modules/introduction/
63 OECD, "Recommendation of the Council on Artificial Intelligence," no. OECD/LEGAL/0449, 2019, Accessed: Jul. 12, 2025. [En línea]. Disponible: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
64 C. C. H. D. and L. S. Flavio Calvino, "What technologies are at the core of AI?," Nov. 2023. doi: 10.1787/32406765-en.
65 OECD, "Governing with Artificial Intelligence," Jun. 2024. doi: 10.1787/26324bc2-en.
66 OECD, "Tools for trustworthy AI," Jun. 2021. doi: 10.1787/008232ec-en.
67 OECD, "Advancing accountability in AI," Feb. 2023. doi: 10.1787/2448f04b-en.
68 OECD and CAF, *The Strategic and Responsible Use of Artificial Intelligence in the Public Sector of Latin America and the Caribbean*, Public Governance. OECD, 2022. doi: 10.1787/1f334543-en.
69 F. Calvino, H. Dernis, L. Samek, and A. Ughi, "A sectoral taxonomy of AI intensity," Dec. 2024. doi: 10.1787/1f6377b5-en.
70 World Economic Forum, "Navigating Cyber Resilience in the Age of Emerging Technologies: Collaborative Solutions for Complex Challenges," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Navigating_Cyber_Resilience_in_the_Age_of_Emerging_Technologies_2024.pdf
71 World Economic Forum, "Ethics by Design: An organizational approach to responsible use of technology," 2020. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Ethics_by_Design_2020.pdf
72 World Economic Forum, "Responsible Use of Technology: The IBM Case Study," 2021. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology_The_IBM_Case_Study_2021.pdf
73 World Economic Forum, "Reimagining Regulation for the Age of AI: New Zealand Pilot Project," 2020. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.weforum.org/publications/reimagining-regulation-for-the-age-of-ai-new-zealand-pilot-project/
74 World Economic Forum, "Advancing Digital Safety: A Framework to Align Global Action," 2021. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Advancing_Digital_Safety_A_Framework_to_Align_Global_Action_2021.pdf
75 Cloud Security Alliance, "AI Resilience: A Revolutionary Benchmarking Model for AI Safety," 2024.
76 World Economic Forum, "A Holistic Guide to Approaching AI Fairness Education in Organizations," 2021. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_A_Holistic_Guide_to_Approaching_AI_Fairness_Education_in_Organizations_2021.pdf
77 World Economic Forum, "A Blueprint for Equity and Inclusion in Artificial Intelligence," 2022. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_A_Blueprint_for_Equity_and_Inclusion_in_Artificial_Intelligence_2022.pdf
78 ENISA, "Artificial Intelligence Cybersecurity Challenges," 2020. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20Artificial%20Intelligence%20Cybersecurity%20Challenges.pdf
79 ENISA, "Multilayer Framework for Good Cybersecurity Practices for AI," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.enisa.europa.eu/sites/default/files/publications/Multilayer%20Framework%20for%20Good%20Cybersecurity%20Practices%20for%20AI.pdf
80 ENISA, "Cybersecurity of AI and Standardisation," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.enisa.europa.eu/sites/default/files/publications/Cybersecurity%20of%20AI%20and%20Standardisation.pdf
81 ENISA, "Cyber Insurance - Models and methods and the use of AI," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Research%20and%20Innovation%20-%20AI%20and%20Cyber%20Insurance.pdf
82 OECD, "Explanatory memorandum on the updated OECD definition of an AI system," Mar. 2024. doi: 10.1787/623da898-en.
83 OECD, "OECD Framework for the Classification of AI systems," Feb. 2022. doi: 10.1787/cb6d9eca-en.
84 OECD, "AI language models," Apr. 2023. doi: 10.1787/13d38f92-en.
85 AEPD, "Requisitos para Auditorías de Tratamientos que incluyan IA," 2021. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.aepd.es/guias/requisitos-auditorias-tratamientos-incluyan-ia.pdf
86 Netherlands Court of Audit, "Understanding algorithms," 2021. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://english.rekenkamer.nl/publications/reports/2021/01/26/understanding-algorithms
87 Netherlands Court of Audit, "Audit framework for algorithms," 2021. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://english.rekenkamer.nl/publications/publications/2021/01/26/audit-framework-for-algorithms
88 J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: a three-layered approach," *AI and Ethics*, vol. 4, no. 4, pp. 1085–1115, Nov. 2024, doi: 10.1007/s43681-023-00289-2.
89 Federal Office for Information Security, "Towards Auditable AI Systems," 2021. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems.pdf?__blob=publicationFile&v=6

90 World Economic Forum, "Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://reports.weforum.org/docs/WEF_Navigating_the_AI_Frontier_2024.pdf
91 Cloud Security Alliance, "Confronting Shadow Access Risks: Considerations for Zero Trust and Artificial Intelligence Deployments," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/confronting-shadow-access-risks-considerations-for-zero-trust-and-artificial-intelligence-deployments
92 Cloud Security Alliance, "Agentic AI Red Teaming Guide," 2025. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/agentic-ai-red-teaming-guide
93 Cloud Security Alliance, "Secure Agentic System Design: A Trait-Based Approach," 2025. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/secure-agentic-system-design
94 Cloud Security Alliance, "Securing LLM Backed Systems: Essential Authorization Practices," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/securing-llm-backed-systems-essential-authorization-practices
95 Cloud Security Alliance, "Large Language Model (LLM) Threats Taxonomy," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/csa-large-language-model-llm-threats-taxonomy
96 Cloud Security Alliance, "Dynamic Process Landscape: A Strategic Guide to Successful AI Implementation," 2025. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/dynamic-process-landscape-a-strategic-guide-to-successful-ai-implementation
97 Josh Lemon, "Survey SANS 2024 Detection and Response Survey Transforming Cybersecurity Operations: AI, Automation, and Integration in Detection and Response," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://sansorg.egnyte.com/dl/06Sbg7nCli
98 ENISA, "Securing Machine Learning Algorithms," 2021. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms
99 Cloud Security Alliance, "The State of AI and Security Survey Report," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/the-state-of-ai-and-security-survey-report
100 Cloud Security Alliance, "Security Implications of ChatGPT," 2023. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/security-implications-of-chatgpt
101 Cloud Security Alliance, "Principles to Practice: Responsible AI in a Dynamic Regulatory Environment," 2024. Accessed: Aug. 06, 2025. [En línea]. Disponible: https://cloudsecurityalliance.org/artifacts/principles-to-practice-responsible-ai-in-a-dynamic-regulatory-environment
102 World Economic Forum, "Cybersecurity Futures 2030: New Foundations," Geneva, Dec. 2023. [En línea]. Disponible: https://www3.weforum.org/docs/WEF_Cybersecurity_Futures_2030_New_Foundations_2023.pdf