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

Contenidos

Prefacio **4**
Resumen Ejecutivo **5**
Introducción **6**

1. Gobernanza Estratégica de la Inteligencia Artificial en las Ciberoperaciones **7**
1.1 Marcos Éticos, Normativos y de Gobernanza para la IA en Contextos Cibernéticos 11
1.2 Modelo de Gobernanza Adaptativa para CSIRT Impulsados por IA 18
1.3 Gobernanza Responsable y Sostenible de Tecnologías Emergentes y Disruptivas en Entornos de la Alta Criticidad 21

2. Desarrollo Tecnológico de la IA en Entornos de Respuesta Cibernética **24**
2.1 Tecnologías emergentes habilitadoras: algoritmos, automatización y capacidades predictivas 29
2.2 Sostenibilidad e Innovación en el ciclo de vida del desarrollo del desarrollo de soluciones de IA 38
2.3 Desarrollo Centrado en la Ciberresiliencia: IA como Herramienta de Anticipación y Mitigación de Amenazas. 39

3. Adopción Responsable y Adaptativa de la IA en los CSIRTs **41**
3.1 Factores Críticos de Éxito para la Adopción de la IA en CSIRT: Cultura, Competencias y Liderazgo 42
3.2 Evaluación de Riesgos y Toma de Decisiones en la Adopción de IA Ciberdefensiva 45
3.3 Modelos Adaptativos y Resilientes de Operaciones Basados en IA para Entornos CSIRTs 46

Conclusiones 48
Recomendaciones 49
Perspectivas de Futuro 52
Llamado de Acción 54
Reflexiones Finales 55
Apéndice: Contexto de Investigación 60
Apéndice: Metodología 61
Colaboradores 63
Agradecimientos 63
Referencias Bibliográficas 64

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.

Prefacio

En la era de la rápida transformación digital, la ciberseguridad se ha convertido en una piedra angular de la resiliencia nacional y organizativa. América Latina y el Caribe (ALC) se enfrenta a retos únicos en este ámbito, caracterizados por la diversidad de capacidades organizativas, las amenazas emergentes y la evolución del panorama normativo. La integración de la inteligencia artificial (IA) en las operaciones de ciberseguridad representa tanto una oportunidad estratégica como una responsabilidad compleja, que exige una gobernanza rigurosa, una visión ética y disciplina operativa.

Esta investigación aborda una laguna de conocimiento crítica al examinar cómo los Equipos de Respuesta a Incidentes de Ciberseguridad (CSIRT, por sus siglas en inglés) pueden aprovechar la IA para fortalecer la ciberdefensa, al tiempo que se enfrentan a limitaciones de recursos, escasez de talento e incertidumbres normativas. Más allá de las consideraciones operativas y técnicas, el estudio hace hincapié en la naturaleza sistémica de la ciberseguridad, donde las interdependencias organizativas y las redes de colaboración determinan la eficacia de las defensas mejoradas por la IA.

La importancia de este trabajo va más allá de la región inmediata. Proporciona un marco para comprender cómo las tecnologías emergentes pueden implementarse de manera responsable en entornos caracterizados por una madurez desigual y contextos sociotécnicos diversos. Al destacar la intersección entre la estrategia, la gobernanza, la capacidad humana y la tecnología, la investigación contribuye a un discurso más amplio sobre la construcción de ecosistemas de ciberseguridad resilientes, adaptables y responsables a nivel mundial.

No solo informa a los profesionales y responsables políticos, sino que también fomenta el diálogo entre el mundo académico, la industria y las instituciones públicas. Esta investigación ejemplifica el tipo de investigación crítica y basada en pruebas que es esencial para traducir el potencial tecnológico en prácticas de ciberseguridad sostenibles, éticas y operativamente eficaces.

Estevenson Solano

Resumen Ejecutivo

Los CSIRT de ALC se encuentran en una etapa crucial de su transformación, impulsada cada vez más por la IA. Este cambio refleja una fuerte ambición estratégica, pero contrasta con la desigual preparación operativa en toda la región. Si bien la mayoría de los equipos reconocen la urgencia de una gobernanza formal y están adoptando marcos técnicos probados, persisten importantes deficiencias en la infraestructura específica de la IA, la gobernanza de los datos, el capital humano cualificado y la gestión de riesgos.

El impulso normativo está creciendo en torno a la IA, la ciberdelincuencia y la protección de datos. Sin embargo, la legislación emergente a menudo sigue anclada en estatutos obsoletos, con dificultades para abordar las amenazas cibernéticas en rápida evolución y las complejidades de la responsabilidad algorítmica. Esta desalineación crea tanto oportunidades para la innovación como retos para una implementación eficaz.

En el ámbito operativo, la IA ya ha demostrado su valor al mejorar la detección de amenazas y acelerar los tiempos de respuesta. Muchos CSIRTs están pasando de las pruebas de concepto a la dependencia diaria de playbooks basados en la IA y la supervisión continua de modelos. Sin embargo, los retos de integración, la financiación inconsistente vinculada a los presupuestos públicos y la dependencia de la experiencia externa contribuyen a niveles desiguales de madurez. Las aplicaciones avanzadas, como la búsqueda predictiva de amenazas y la respuesta autónoma, aún están en fase incipiente, obstaculizadas por problemas de calidad de los datos y la escasez de talento especializado, en particular ingenieros de seguridad de IA, expertos en operaciones de aprendizaje automático (MLOps, por sus siglas en inglés) y responsables de riesgos de modelos.

La ambición estratégica suele superar la capacidad de crear infraestructuras resilientes y equipos especializados. Para abordar estas deficiencias, son esenciales los modelos de gobernanza por niveles, los mecanismos de servicios compartidos y una mayor armonización entre la legislación emergente y las normas técnicas. Estas medidas fomentarían una ciberdefensa sostenible, resiliente y responsable, mejorada por la IA, en diversos contextos organizativos.

Una dimensión adicional y a menudo pasada por alto es la sostenibilidad medioambiental. La formación y el despliegue de modelos de IA conllevan una huella de carbono significativa, especialmente cuando la tecnología a gran escala depende de fuentes de energía no renovables. La incorporación de algoritmos eficientes desde el punto de vista energético, la optimización de los ciclos de vida de los modelos y la adopción de recursos informáticos ecológicos compartidos serán fundamentales para equilibrar los resultados en materia de seguridad con los compromisos climáticos.

De cara al futuro, el camino a seguir requiere una colaboración regional decisiva. Las iniciativas lideradas por la OEA a través de CSIRT-Américas, LAC4, BID, ITU, LACNIC y la Unión Europea ilustran el potencial de la acción colectiva. Al poner en común sus conocimientos especializados, desarrollar conjuntos de herramientas compartidas basadas en la IA, crear infraestructuras ecológicas y armonizar los marcos jurídicos y técnicos, los CSIRTs de ALC pueden superar la fragmentación. Este enfoque reducirá los costos unitarios, reforzará la rendición de cuentas y cerrará las brechas de madurez, al tiempo que se salvaguardarán los derechos fundamentales.

Si se mantienen, estos esfuerzos permitirán a la región aprovechar la IA no solo como un conjunto de herramientas operativas, sino como la piedra angular de un ecosistema de ciberresiliencia sostenible, adaptable y responsable.

Introducción

Ante la convergencia de la IA con las operaciones cibernéticas representa uno de los paradigmas más transformadores y complejos del panorama tecnológico contemporáneo. Los CSIRTs se enfrentan a una demanda creciente de capacidades adaptativas, automatizadas y predictivas para confrontar las ciberamenazas cada vez más sofisticadas.[1] [2] [3] Esta convergencia no solo amplifica las capacidades operacionales, sino que también introduce nuevas dimensiones de riesgos que requieren marcos de gobernanza estratégicas comprensivos.

En un mundo cada vez más interconectado e impulsado por el crecimiento y desarrollo exponencial de las tecnologías de innovación emergentes y disruptivas, donde las sociedades pretenden aprovechar el papel de las tecnologías para impulsar el crecimiento económico, mejorar los servicios públicos y aumentar el nivel de calidad de vida, se enfrentan a mayores riesgos asociados a las amenazas cibernéticas. En ese contexto, esta obra demuestra que la ciberseguridad es esencial para el progreso socioeconómico de las naciones.[4] Por tanto, papel de la ciberseguridad en el panorama actual no debe subestimarse, debido su complejidad, la constante evolución, la sofisticación de las crecientes ciberamenazas y capacidades de los actores maliciosos, lo que exige estrategias innovadoras para combatir ataques cada vez más sofisticados y de mayor alcance.[5]

Esta dinámica subraya una observación esencial: la escalada en la complejidad y frecuencia de los ciberataques está llevando al límite las operaciones cibernéticas tradicionales, impulsando la necesidad imperativa de incorporar la IA para potencializar y automatizar las capacidades defensivas, como es el caso de los CSIRTs, -ofrecen servicios de ciberseguridad para prevenir, detectar, mitigar y responder a incidentes cibernéticos-.[6] Las funciones de los CSIRTs han evolucionado más allá de la mera respuesta reactiva, abarcando un modelo integral de la gestión de la seguridad de la información y ciberseguridad que considera elementos técnicos, humanos, organizativos y económicos.

Paralelamente, las operaciones cibernéticas (CyberOps, por sus siglas en inglés) son actividades y procesos especializados dedicados a identificar, monitorear, proteger, analizar de manera proactiva ciberincidentes y ciberamenazas, como además el desarrollo de capacidades de ciberdefensa en el ciberespacio. Por lo cual la IA y los distintos tipos de familias y modelos de esta emergen como una tecnología clave para potenciar significativamente las capacidades en la detección, respuesta y mitigación de incidentes y amenazas en todo el ecosistema digital.

La IA hace referencia a los sistemas informáticos diseñados para emular las capacidades cognitivas humanas, como la percepción, el procesamiento del lenguaje, la toma de decisiones y el análisis de la información. Su rápido avance se debe principalmente a factores interrelacionados con el crecimiento exponencial de los macrodatos, la disponibilidad de big data, el desarrollo de algoritmos y técnicas de aprendizaje automático (ML, por sus siglas en inglés) más sofisticados y la creciente potencia de cálculo del hardware moderno.[7] [8]

Más aún, la capacidad de IA depende de cinco consideraciones principales: 1) entender la naturaleza del problema que se está resolviendo; es decir, entender si las predicciones o inferencias causales deben realizarse para resolver un problema; 2) considerar los tipos de datos necesarios para resolver el problema; 3) asegurar la disponibilidad de grandes volúmenes de datos de entrenamiento que permitirán que los algoritmos desarrollen sus capacidades de predicción antes del despliegue; 4) evaluación de la calidad de los datos que están integrados a través de las bases de datos pertinentes; y 5) protección y seguridad desde el diseño de los procesos, sistemas, modelos y algoritmos.[7]

Los avances recientes, en especial en la IA generativa (GenAI, por sus siglas en ingles), han dotado a los equipos de seguridad de herramientas para obtener conclusiones basadas en datos, generar informes de manera eficiente y recibir recomendaciones de mitigación. Esta capacidad de la IA permite una transición de lo reactivo a lo proactivo, capacitando a los sistemas para no

solo responder a incidentes conocidos, sino también poder predecir y neutralizar amenazas emergentes incluso antes de que puedan causar daño.

Además, la IA no es simplemente una herramienta para mejorar funciones de seguridad existentes, sino un catalizador para una redefinición fundamental de las ciberoperaciones. Su capacidad para automatizar la respuesta a incidentes y posibilitar "CyberOps unificadas", implica un cambio de paradigma que exige una reevaluación de los flujos de trabajo, los roles y estructuras organizativas dentro de los CSIRTs, para una adopción exitosa de la IA, se requiere una gestión del cambio cultural, organizacional y una visión estratégica clara sobre como las capacidades humanas y las de la IA se integrarán y redefinirán el panorama de seguridad.

Ante el incremento de incidentes cibernéticos de relevancia nacional en ALC, y la necesidad de los gobiernos de reaccionar ante ellos, a partir de 2003 los CSIRTs toman relevancia. Con el paso del tiempo llevó a que no solo se constituyeran centros nacionales sino también sectoriales.[6] En algunos casos, los CSIRTs se crearon por proyectos de ley, decretos o una Estrategia de Ciberseguridad Nacional.

Tal como ha ocurrido con otras grandes disrupciones tecnológicas, se prevé que la IA también conlleve una revolución económica a nivel mundial. En la actualidad, nos encontramos en el punto inicial de este camino, que brindará infinidad de oportunidades para aquellos países y regiones que sepan posicionarse o incluso lleguen a liderar esta disrupción. Las previsiones indican que la IA será uno de los principales motores de crecimiento de los próximos años. La IA se convertirá en la mayor oportunidad de negocio, y llegará a generar USD 16.500 trillones en el año 2030: USD 6.600 billones por mejoras en la productividad y USD 9.900 billones derivados del consumo adicional generado. Además, se estima que, en 2030, el 70% de las empresas a nivel mundial habrá adoptado algún tipo de IA. En este contexto, la información se convierte en un activo muy valioso, ya que constituye el elemento esencial para la creación de algoritmos de IA de valor.[9]

La implementación de IA en contextos cibernéticos trasciende las consideraciones puramente técnicas, demandando un enfoque holístico que integre principios éticos, legales y consideraciones de sostenibilidad. [10] [11] La naturaleza dual de la IA como herramienta defensiva y potencial vector de ataque exige una gobernanza que equilibre la innovación tecnológica con la responsabilidad operacional y la confianza institucional.

Finalmente, esta investigación examina desde un enfoque científico-técnico e ingenieril los fundamentos teóricos y prácticos para la gobernanza estratégica de la IA en las CyberOps.

1. Gobernanza Estratégica de la Inteligencia Artificial en las Ciberoperaciones

La gobernanza de la IA en las ciberoperaciones representa un paradigma emergente que integra aspectos normativos, estructuras organizacionales y procesos de toma de decisiones para gestionar la implementación responsable de sistemas autónomos e inteligentes de ciberseguridad. Lo cual surge una necesidad imperativa de establecer mecanismos de control y supervisión sobre tecnologías que operan en contextos de alta criticidad donde puede tener impactos significativos en la ciberresiliencia y la continuidad operativa.[1] [2] [3]

El concepto de gobernanza, entendido como el marco para lograr un desarrollo equilibrado y sostenible mediante la interacción del Estado, la sociedad civil y el mercado,[12] se traduce en la gobernanza de la IA[13] en las ciberoperaciones como el establecimiento de mecanismos institucionales, normativos y procedimentales sólidos que garanticen un uso responsable, ético y eficaz de la IA en entornos de seguridad digital.

En este contexto, la gobernanza implica primero establecer procesos transparentes para la selección, supervisión y regulación de los sistemas de IA utilizados en operaciones cibernéticas, que garanticen la rendición de cuentas al público y a las partes interesadas pertinentes; segundo garantizar la capacidad para gestionar los recursos tecnológicos con sensatez y formular, aplicar

y hacer cumplir políticas que alineen el despliegue de la IA con los objetivos sociales, los derechos humanos y los imperativos de seguridad; y tercero la adhesión mutua de los actores estatales, las entidades privadas y los ciudadanos a las normas legales y éticas[14] que rigen las interacciones impulsadas por la IA en el ciberespacio.

Así pues, la gobernanza de la IA no sólo debe garantizar la eficiencia operativa y la mitigación de riesgos, sino también defender los principios democráticos, la confianza institucional y el bien común en un mundo cada vez más digital e interconectado.

Los sistemas de gobernanza de la IA deben estar basados en un enfoque de gestión de riesgos que permita identificar de manera temprana posibles dificultades para el sistema y mitigar los daños no intencionales que pudieran haber resultado de un mal uso de la IA. Por lo anterior, se sugiere que los sistemas de gestión de riesgos estén basados en modelos de protección de la privacidad por defecto, a fin de que los datos personales usados en el sistema de IA estén protegidos en todo momento y que el usuario tenga el control sobre sus datos.[15]

Uno de los elementos más relevantes para el uso ético y responsable de la IA se relaciona con la identificación de las tareas que requieren supervisión, intervención o interacción humana, ya sea para corregir imprecisiones del sistema o generar una interacción adecuada entre usuarios y dicho sistema. Del mismo modo, y desde la posición de un enfoque de prevención de riesgos, el grado de involucramiento humano adecuado en los sistemas de IA debería potenciar la autonomía humana al tiempo que se minimizan posibles condicionamientos, engaños o subordinación injustificada a quienes utilizan esta tecnología.[15]

Los CSIRTs que hacen uso de sistemas de IA deberían llevar a cabo comunicaciones proactivas, claras y sistemáticas a las partes interesadas sobre las capacidades y limitaciones de los sistemas de IA. En este sentido, resulta importante manejar las expectativas sobre el alcance tecnológico.[15]

El relacionamiento con las partes interesadas se refiere de forma amplia a la implicación de los actores relevantes en todo el ciclo de vida de la IA, lo que incluye comunicaciones, colaboración, consultas, y rendición de cuentas con los titulares de la información, usuarios, clientes, inversionistas, socios, proveedores, entes reguladores y la sociedad civil en general.[15]

Los marcos de gobernanza contemporáneos, como COBIT 2019, ISO 27001 e ISO 42001 han evolucionado para abordar complejidades específicas de la integración de la IA en el dominio cibernético.[16] [17] Por tanto, la implementación efectiva requiere la armonización de múltiples dimensiones: transparencia algorítmica, rendición de cuentas automatizadas, evaluación de riesgos y mantenimiento de la supervisión humana.[3] [18]

Esta gobernanza se fundamenta en tres pilares conceptuales interconectados: la teoría de los agentes aplicado a sistema autónomos, donde los algoritmos actúan como agentes con capacidades decisorias delegadas; la teoría de riesgos sistémicos, que reconoce la naturaleza emergente y potencialmente de fallos en cascadas en sistemas IA; y la teoría de la responsabilidad compartida, que aborda la distribución de la responsabilidad en sistemas sociotécnico complejos. [1] [19] [20] Un régimen de gobernanza de IA adecuado es un desafío complejo, ya que requiere de una evaluación cuidadosa de riesgos y de impacto.

Dimensiones de la Gobernanza de la IA en las Ciberoperaciones

1. El nexo entre la gobernanza y la tecnología ya que se demostró una tensión fundamental entre el rápido avance de la IA y la evolución más lenta de los marcos normativos. Los participantes hicieron hincapié de manera sistemática en la necesidad de mecanismos de gobernanza adaptables que puedan evolucionar con las tecnologías emergentes y disruptivas, al tiempo que se mantiene la supervisión ética.

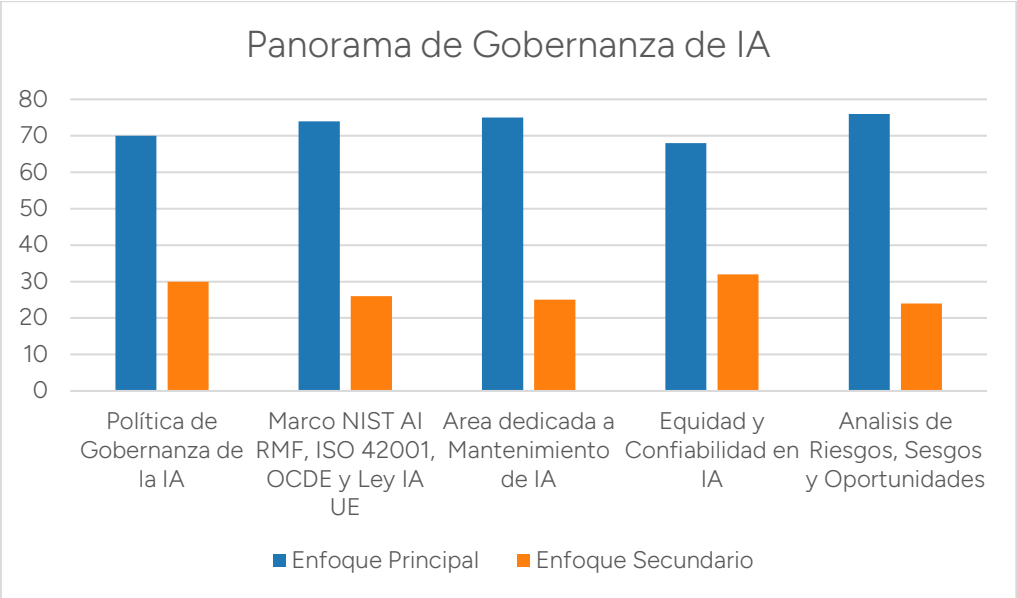
2. Hay una gran divergencia y convergencia en ALC, ya que, si bien los países muestran distintos niveles de madurez en la gobernanza de la IA, está surgiendo una convergencia en torno a enfoques de colaboracion y cooperacion en el fortalecimiento y desarrollo de capacidades a través de iniciativas impulsadas por la **OEA, CSIRT Américas, LAC4, ITU, BID, ISOC y LACNIC.**

3. En cuanto al paradigma de colaboración entre humanos e IA, en lugar de la automatización total, los expertos abogan por modelos que incorporen al ser humano en el proceso, aprovechando las capacidades de la IA y manteniendo al mismo tiempo la supervisión humana para las decisiones críticas o de alto riesgo.

4. Hay una tensión entre la sostenibilidad y seguridad que subraya la necesidad de una creciente conciencia de las implicaciones medioambientales de los sistemas IA en la ciberseguridad, donde se aboga por un llamado de atención y acción para equilibrar la potencia computacional con la eficiencia energética.

5. Imperativo de gobernanza multiactores, surgió un consenso en torno a la necesidad de involucrar al mundo académico, la industria, el gobierno y la sociedad civil en los esfuerzos de gobernanza para pasar del papel a acciones consolidadas, con trazabilidad y continuidad en el tiempo.

Gráfico 1. Panorama de los Esfuerzos Gobernanza de la IA



Nota: La categoría de enfoques "Principal" corresponde a la práctica más prevalente entre los CSIRTs (mayor porcentaje), mientras que "Secundario" representa la práctica menos común (menor porcentaje). Fuente: Elaboración propia.

El panorama de gobernanza de la IA en ALC muestra una paradoja convincente de concienciación unida a retos de implementación. El 70 % de los CSIRTs persiguen activamente la necidad de adopcion de políticas formales de gobernanza de la IA específicas para las operaciones ciberneticas, lo que indica una comprensión sofisticada de la necesidad de enfoques estructurados para la integración de la inteligencia artificial. Sin embargo, el 30 % restante se basa en directrices informales, lo que sugiere una adaptación pragmática a las limitaciones de recursos, al tiempo que se mantiene un cierto nivel de concienciación sobre la gobernanza. Esta distribución refleja los distintos niveles de madurez institucional en todo el ecosistema regional, donde los CSIRTs equilibran las aspiraciones de gobernanza con las realidades operativas y de sus capacidades.

Los patrones de selección de marcos revelan preferencias estratégicas que se alinean con orientaciones técnicas más que normativas. El 74 % de los CSIRTs ven la necesidad de implementar la gobernanza a través de marcos técnicos establecidos, como el NIST AI RMF, la ISO 42001 y las Directrices de la OCDE, en comparación con solo el 26 % que plantea tomar praticas de la Ley Europea de Inteligencia Artificial. Esta preferencia sugiere que los CSIRTs regionales dan prioridad a las normas técnicas aplicables frente a los marcos normativos que exigen un alto grado de cumplimiento, lo que podría reflejar los limitados mecanismos de aplicación de la normativa en la región o una elección estratégica consciente de dar prioridad a la eficacia técnica frente al cumplimiento legal y regulatorio. La preferencia por los marcos técnicos también indica que las redes de cooperación técnica internacional son más sólidas que los esfuerzos de armonización normativa.

Limitaciones de recursos y retos organizativos

La brecha más significativa entre la gobernanza y la implementación se da en el desarrollo de la infraestructura organizativa, donde el 75 % de los CSIRTs reconocen la necesidad de contar con áreas dedicadas al mantenimiento de las soluciones de IA, pero el 25 % no puede considerar su implementación debido a limitaciones presupuestarias y de talento humano. Esto representa un cuello de botella crítico en la adopción sostenible de la IA, lo que pone de relieve cómo las limitaciones de recursos se traducen directamente en deficiencias en la capacidad operativa. La alta tasa de reconocimiento demuestra una conciencia estratégica, mientras que las barreras de implementación revelan retos estructurales que requieren soluciones sistémicas más allá de los esfuerzos individuales de las organizaciones. Este patrón sugiere que la adopción sostenible de la IA en la región requiere modelos innovadores de intercambio de recursos e iniciativas de desarrollo de capacidades que aborden las limitaciones colectivas, en lugar de las individuales, de las organizaciones.

Las consideraciones de garantía de calidad y fiabilidad presentan otra dimensión de la brecha entre la gobernanza y la implementación. El 68 % de los CSIRTs están explorando opciones para garantizar la equidad y la fiabilidad en el desarrollo de soluciones de IA, mientras que el 32 % ha planteado ir elaborando planes de implementación específicos. Esta brecha entre el conocimiento y la acción indica que la comprensión de los requisitos de calidad de la IA supera las capacidades operativas actuales para implementar procesos sistemáticos de garantía de calidad. La fase de exploración sugiere un compromiso activo con los conceptos de calidad, pero la limitada planificación específica revela la complejidad de traducir la comprensión teórica en procedimientos operativos. Este patrón pone de relieve la necesidad de una orientación práctica que sirva de puente entre la conciencia conceptual y las prácticas de garantía de calidad aplicables.

Madurez de la evaluación de riesgos y enfoques sistemáticos

Los datos revelan una variabilidad preocupante en las prácticas de evaluación de riesgos en todo el ecosistema regional de CSIRT. El 76 % de los CSIRTs reconocen la importancia de un análisis de riesgos exhaustivo, una evaluación de sesgos y una evaluación de oportunidades de IA, lo que demuestra una conciencia sofisticada de los riesgos y oportunidades relacionados con la IA. Sin embargo, el 24 % solo realiza evaluaciones y análisis ad hoc, lo que indica niveles de madurez desiguales que podrían crear vulnerabilidades en las operaciones cibernéticas mejoradas por la IA. Esta disparidad sugiere que, mientras que las organizaciones líderes avanzan hacia enfoques sistemáticos de gestión de riesgos, una minoría significativa opera de forma reactiva, lo que podría crear brechas de seguridad regionales. La alta tasa de reconocimiento entre la mayoría indica una sólida comprensión fundamental, pero los enfoques ad hoc de casi una cuarta parte de las organizaciones ponen de relieve la necesidad de marcos de evaluación de riesgos estandarizados y adaptados a las diferentes capacidades organizativas.

Implicaciones estratégicas y gobernanza adaptativa

Estos hallazgos revelan una "paradoja de la preparación para la gobernanza" emergente, en la que una mayor conciencia de la gobernanza se correlaciona con mayores retos de implementación. Los CSIRTs con una comprensión sofisticada de los requisitos de gobernanza de la IA se enfrentan simultáneamente a las restricciones de recursos y las barreras de implementación más significativas. Este patrón contraintuitivo sugiere que una mayor conciencia de la gobernanza aumenta en realidad el reconocimiento de la complejidad y los requisitos de recursos para una implementación eficaz. En lugar de indicar un fracaso, este patrón demuestra una evaluación realista de las demandas de gobernanza y sugiere la necesidad de modelos de gobernanza adaptativos que puedan acomodar los distintos niveles de madurez organizativa, manteniendo al mismo tiempo la eficacia de la seguridad.

La preferencia por los marcos técnicos frente a los enfoques normativos, combinada con la prevalencia de directrices informales entre el 30 % de los CSIRTs, indica la aparición de ecosistemas de gobernanza híbridos. Estos modelos híbridos permiten a los CSIRTs adaptar los enfoques de gobernanza a sus limitaciones de recursos y contextos operativos, al tiempo que mantienen unos estándares mínimos de seguridad. Esta flexibilidad adaptativa representa una respuesta pragmática a las condiciones regionales y sugiere que los marcos de gobernanza eficaces deben adaptarse a la diversidad organizativa en lugar de imponer requisitos uniformes. El enfoque híbrido también permite una progresión gradual de la gobernanza informal a la formal a medida que se desarrolla la capacidad organizativa.

1.1 Marcos Éticos, Normativos y de Gobernanza para la IA en Contextos Cibernéticos

La transparencia y explicabilidad algorítmica en contextos de respuesta a ciberincidentes requiere mecanismos que permitan la trazabilidad de decisiones automatizadas sin comprometer la eficacia operacional.[21 22 23] Los sistemas de IA explicable (XAI, por sus siglas en inglés) deben proporcionar justificaciones comprensibles para las acciones recomendadas, especialmente en escenarios donde la intervención humana es crítica para la validación de respuestas automatizadas.

Por lo cual, la implementación práctica de transparencia exige el desarrollo de dashboards interpretativos donde se visualicen las cadenas de razonamiento de los algoritmos, incluyendo los factores de confianza asociados a cada recomendación.[24 25] Es aquí donde esta capacidad es esencial para mantener la supervisión humana efectiva y garantizar que los analistas de seguridad puedan evaluar críticamente las decisiones algorítmicas en tiempo real.

En ALC, los expertos destacan los siguientes requisitos para la implementación de la ética en los CSIRTs:

1. Transparencia y explicabilidad algorítmica (XAI).
2. Prevención de sesgos y equidad en las decisiones automatizadas.
3. Estructuras claras de rendición de cuentas para las acciones impulsadas por la IA.
4. Mecanismos regulares de auditoría y cumplimiento.

Además, la equidad y no discriminación en sistemas de detección mediante algoritmos de IA empleados en los CSIRTs deben incorporar controles contra sesgos sistemáticos que puedan resultar en tratamiento discriminatorio de diversos tipos de incidentes, organizaciones o sectores.[11 26] La equidad algorítmica demanda de la implementación de métricas de equidad o

justas -fairness- que evalúen regularmente la distribución de falsos positivos y negativos en diferentes categorías de amenazas y contextos organizacionales.

La auditoría continua de equidad debe incluir análisis de deriva -drift- algorítmico y evaluaciones de impacto diferencial, asegurando que las actualizaciones de modelos no introduzcan sesgos no intencionados que puedan afectar la calidad del servicio proporcionado a diferentes tipos de constituyentes.[27 28]

Ahora bien, en términos de privacidad y protección de datos es preciso considerar la integración y aplicación de principios de "privacidad por diseño y por defecto" en sistemas IA para los CSIRTs ya que requiere arquitecturas que minimicen la exposición de datos sensibles mientras se mantiene la eficacia analítica.[10 29] Esto incluye la implementación de técnicas de privacidad diferencial, aprendizaje federado y encriptación homomórfica que permitan el análisis colaborativo de amenazas sin comprometer la confidencialidad de los datos.

En ALC, los expertos destacan los siguientes retos de la ética en los CSIRTs:

1. Equilibrio entre la eficiencia y los requisitos de explicabilidad.
2. Gestión de las compensaciones entre privacidad y seguridad.
3. Abordar las lagunas en materia de responsabilidad en la toma de decisiones autónomas.
4. Prevenir resultados discriminatorios en las operaciones de seguridad.

Los mecanismos de anonimización y seudonimización deben ser robustos contra ciberataques de re-identificación, particularmente cuando se comparten indicadores de compromiso (IoCs, por sus siglas en inglés) e inteligencia de amenazas[29] entre diferentes organizaciones y sectores.[30]

Para muchos países de ALC, entre los riesgos específicos que presenta la inteligencia artificial y su desarrollo figuran la desinformación, la profundización de las desigualdades y la violación de derechos humanos. Sin embargo, la regulación tradicional y la autorregulación no son alternativas viables debido a la naturaleza de estas tecnologías y a los desafíos, limitaciones y circunstancias que los países enfrentan a la hora de regularlas.[31]

Se revela un consenso universal sobre la necesidad de actualizar los marcos normativos, y los participantes de múltiples contextos hacen hincapié en esta brecha crítica. El panorama normativo se caracteriza por un desajuste fundamental entre el ritmo del avance tecnológico y la evolución del marco jurídico, lo que plantea importantes retos de gobernanza.

- Los marcos jurídicos no logran seguir el ritmo de la evolución tecnológica de la IA, e incluso en el caso también de otras tecnologías de innovación emergentes y disruptivas.
- La disparidad entre las políticas de ciberseguridad y las estrategias de IA requiere una integración urgente.
- Necesidad de mecanismos de transparencia y auditabilidad algorítmica.
- Lagunas en la implementación política a pesar del desarrollo de documentos estratégicos.
- **Marcos de responsabilidad civil y penal** por el uso indebido de la IA, que van más allá de las directrices éticas para convertirse en mecanismos legales aplicables.

Además, los expertos señalan que ALC necesita un **enfoque de negociación de bloque regional** similar a la estrategia coordinada de la Unión Europea, ya que los países individuales carecen de la capacidad de negociación para influir eficazmente en los marcos de gobernanza de la IA a nivel mundial.

Además, la Organización para la Cooperación y el Desarrollo Económicos (OECD, por sus siglas en inglés), estableció los principios éticos y responsables para la IA, que proporcionan una orientación de alto nivel sobre lo que debe o no hacerse en el desarrollo y despliegue de sistemas de IA por parte de todos aquellos que tienen funciones activas en su ciclo de vida (incluyendo su diseño, desarrollo, despliegue, operación, mantenimiento, monitoreo etc.).[32] [33] No obstante, aún resulta incipiente la aplicación de recomendaciones específicas al respecto, con el propósito de adoptar una IA responsable en ALC.[34]

Las oportunidades que ofrece la IA para ALC también conllevan posibles riesgos que aún no se han materializado por completo, dado su estado de desarrollo actual. El tema de la ética de la IA es relativamente nuevo en el mundo e incipiente en la región. Si bien aún no existen avances significativos.[35] [36]

Principios Éticos de la OECD[37]

Crecimiento inclusivo, desarrollo sostenible y bienestar

Las partes interesadas deberán participar activamente en la gestión responsable de una IA que esté pensada para alcanzar resultados beneficiosos para las personas y el planeta. Con el uso adecuado de la IA se podrá promover el aumento de las capacidades humanas y de la creatividad, la inclusión de poblaciones minoritarias, la reducción de las desigualdades económicas y sociales, así como la protección de entornos naturales, estimulando para ello el crecimiento inclusivo, el desarrollo sostenible y el bienestar.

Valores centrados en el ser humano y la equidad

Las partes interesadas del ecosistema de IA deben respetar el estado de derecho, los derechos humanos y los valores democráticos a lo largo de todo su ciclo de vida. Entre estos últimos sobresalen la libertad, la dignidad y la autonomía, la privacidad y la protección de los datos, la no discriminación y la igualdad, la diversidad, la equidad, la justicia social y los derechos laborales internacionalmente reconocidos. Con este fin, los actores de la IA deben implementar

mecanismos y salvaguardias de protección de derechos como el de la autodeterminación de los individuos. Estos deben ajustarse al contexto y ser consistentes con el estado del arte.

Transparencia y explicabilidad

Las partes interesadas del ecosistema de IA deberán comprometerse con la transparencia y la divulgación responsable de los sistemas relacionados. Deberán proporcionar información relevante que se ajuste al contexto y sea coherente con el estado del arte. Con lo anterior se busca: (i) fomentar una comprensión general de los sistemas de IA, (ii) procurar que las partes interesadas tomen plena conciencia de sus interacciones con los sistemas de IA, (iii) asegurarse de que los afectados por un sistema de IA entiendan el resultado, y (iv) permitir que las personas afectadas adversamente por un sistema de IA impugnen sus resultados basándose en información clara y fácil de entender sobre los factores y la lógica que sirvieron de base para la predicción, recomendación o decisión que se busca refutar.

Robustez, seguridad y protección

La robustez, la seguridad y la protección son elementos esenciales de todo sistema de IA por las siguientes razones:

- Los sistemas de IA deben ser robustos, seguros y protegidos durante todo su ciclo para que, en condiciones de uso normal, uso previsible, uso incorrecto u otras condiciones adversas, funcionen adecuadamente y no supongan un riesgo irrazonable para la seguridad.
- Para ello, se debe garantizar la trazabilidad permanente, incluso en relación con los conjuntos de datos, procesos y decisiones tomadas durante el ciclo de vida del sistema de IA. Así será posible analizar correctamente, y en consonancia con el estado del arte, sus resultados y respuestas a las preguntas que se les formulen.
- En función de sus labores, del contexto y de su capacidad de actuación, los actores de la IA deben aplicar continuamente un enfoque sistemático de gestión de riesgos en cada fase del ciclo de vida del sistema para abordarlos de la mejor manera, incluyendo los relativos a la privacidad, seguridad digital y sesgos.

Rendición de cuentas

Los actores de la IA deben ser responsables del buen funcionamiento de los sistemas de IA y del respeto por los principios antes mencionados, en función de sus deberes, del contexto y del estado del arte.

Por otra parte, ALC debe prepararse para aprovechar los beneficios de una IA confiable donde el ser humano esté en el centro de las decisiones, identificar los desafíos éticos y de privacidad que esta supone, y contar con mecanismos y estándares para el manejo y la mitigación de riesgos.[34] Desde el Banco Interamericano de Desarrollo (BID) lanzaron la iniciativa fAIr LAC que se apalanca en los sectores público y privado, con la misión de promover un desarrollo y aplicación responsables de la IA para mejorar la prestación de servicios sociales y atenuar la creciente desigualdad.

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.

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.

Uno de los retos éticos y de gobernanza más acuciantes en el desarrollo y la implementación de sistemas de IA radica en su falta de transparencia, a menudo denominada "efecto caja negra". Los modelos de aprendizaje profundo (DL, por sus siglas en inglés), debido a su complejidad inherente y al procesamiento de datos a gran escala, dificultan la comprensión de cómo se toman decisiones específicas, lo que suscita inquietudes en torno a la explicabilidad, la rendición de cuentas y la equidad.

Esta opacidad socava la confianza del público, especialmente en áreas críticas como la ciberseguridad, donde las decisiones tomadas por sistemas autónomos pueden tener consecuencias de gran alcance. El derecho a la explicación, que está surgiendo como un nuevo derecho digital, subraya la necesidad de que los algoritmos sean inteligibles y auditables, especialmente cuando afectan a derechos fundamentales. Por lo tanto, los mecanismos de transparencia y rendición de cuentas deben ser parte integral de cualquier marco ético de IA para garantizar que las tecnologías no solo sean eficaces, sino también justas y fiables.

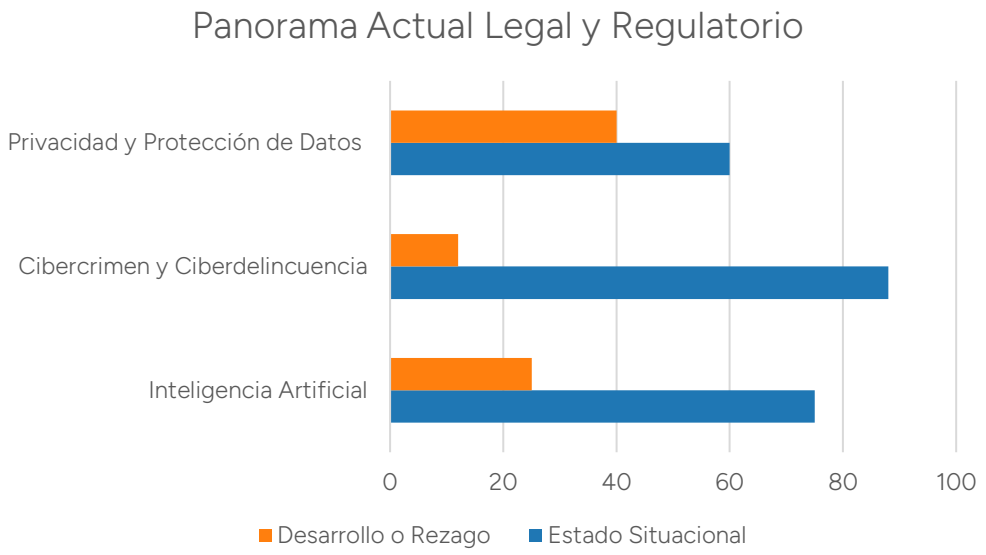
Además, los marcos de gobernanza deben abordar los riesgos asociados con la calidad de los datos, la privacidad y la discriminación algorítmica. Los sistemas de IA entrenados con conjuntos de datos sesgados o incompletos pueden replicar y amplificar las desigualdades sociales existentes, especialmente a través de procesos de toma de decisiones automatizados, como el reconocimiento facial o la elaboración de perfiles de amenazas.

Desde el punto de vista de la protección de datos, el uso generalizado de la información personal exige el cumplimiento estricto de normativas como el Reglamento General de Protección de Datos (RGPD) de la Unión Europea, que se ha convertido en un referente mundial. Es esencial establecer protocolos sólidos para la anonimización y la seguridad de los datos a fin de evitar el uso indebido o las brechas. Además, existe un riesgo crítico al diseñar proyectos de IA basados únicamente en capacidades tecnológicas en lugar de en objetivos sociales o políticos claramente definidos. Para garantizar un despliegue ético en contextos cibernéticos, las iniciativas de IA deben estar orientadas a la resolución de problemas, basadas en el interés público y guiadas por principios que salvaguarden los derechos, mitiguen los daños y promuevan el beneficio social.

Finalmente, en ALC la gestión de la privacidad sigue siendo un desafío persistente y se perfila cada vez más como una preocupación clave en la gobernanza de las tecnologías emergentes y disruptivas como la IA. Si bien algunos países de la región han avanzado en la regulación de la protección de datos, la capacidad institucional para hacer cumplir estos marcos y garantizar su cumplimiento suele ser insuficiente.

Esta fragmentación normativa, combinada con la escasa concienciación pública y las deficiencias en la infraestructura tecnológica, aumenta el riesgo de uso indebido de los datos personales y erosiona la confianza en los sistemas digitales.[38] A medida que los sistemas de IA se integran cada vez más en funciones críticas, incluidas las operaciones de ciberseguridad, la falta de normas de privacidad sólidas y armonizadas expone a las personas y las sociedades a vulnerabilidades que van desde la vigilancia hasta la discriminación. Para subsanar esta deficiencia es necesario no solo adoptar marcos jurídicos integrales, sino también fomentar la cooperación regional, invertir en capacidad técnica y promover culturas éticas de datos que den prioridad a la transparencia, el consentimiento y la rendición de cuentas en el uso de tecnologías basadas en la IA.

Gráfico 2 Panorama Actual Legal y Regulatorio



Nota: La categoría de "Estado Situacional" indica el nivel actual de implementación, adopción o enfoque predominante, mientras que "Desarrollo o Rezago" señala los elementos en fase de mejora, en proceso de desarrollo o con retrasos en su adopción. Fuente: Elaboración propia.

El entorno legislativo de la IA en ALC está experimentando un período de evolución dinámica caracterizado por un creciente reconocimiento de las implicaciones sociales y de seguridad de la IA. El 75 % de las jurisdicciones que supervisan los CSIRTs informan de que las iniciativas o planes normativos específicos sobre IA figuran en la agenda legislativa, lo que indica una intención proactiva de elaborar leyes específicas sobre IA en lugar de basarse en estatutos más generales. Por el contrario, el 25 % plantea ir regulando la IA a través de leyes generales cuyo alcance y capacidad de aplicación siguen siendo limitados. Esta división revela tanto el impulso hacia una gobernanza especializada en IA como la dependencia transitoria de los marcos jurídicos existentes, lo que subraya la necesidad de crear capacidad en materia de redacción legislativa y aplicación de la normativa para traducir las iniciativas de la agenda en leyes eficaces.

En el ámbito de la ciberdelincuencia, la base legislativa está establecida de manera más uniforme, ya que el 88 % de los países de ALC ya cuentan con legislación sobre ciberdelincuencia. Sin embargo, la aparición generalizada de nuevas amenazas digitales le esta demandando de algunas actualizaciónes exhaustiva de estas leyes para abarcar modalidades en evolución, como los ataques habilitados por la IA, el ransomware como servicio y la extorsión digital transfronteriza. Solo el 12 % de los países están actualmente desarrollando, revisando o actualizando sus leyes sobre ciberdelincuencia para adaptarlas a los retos tecnológicos actuales y previstos. Esta brecha entre la legislación existente y los esfuerzos de modernización activos pone de relieve la necesidad apremiante de acelerar los procesos de reforma, armonizar a nivel regional las definiciones y sanciones, e integrar disposiciones especializadas que aborden las ciberamenazas impulsadas por la IA.

La legislación sobre privacidad y protección de datos en la región refleja un perfil de madurez mixto en el contexto de las tecnologías emergentes. El 60 % de las jurisdicciones cuentan con leyes de protección de datos en vigor, pero estas leyes se consideran en general en algunos aspectos con capacidades insuficientes o limitadas cuando se enfrentan al procesamiento avanzado de datos impulsado por la IA, la elaboración de perfiles algorítmicos y el análisis de comportamiento a gran escala. Por el contrario, el 40 % de los países están trabajando en el proceso de redaccion, revision y actualizacion activa de sus marcos de privacidad para hacer frente a estos nuevos retos, lo que indica un avance progresivo hacia una gobernanza de datos sólida. La coexistencia de regímenes de protección de datos heredados junto con los esfuerzos legislativos en evolución subraya la doble necesidad de hacer cumplir los derechos de privacidad

actuales y, al mismo tiempo, acelerar la modernización de las garantías legales para preservar la autonomía individual y la soberanía de los datos en la era de la IA.

En los tres ámbitos legislativos de regulación de la IA, ciberdelincuencia y privacidad/protección de datos- la región muestra una clara trayectoria hacia marcos legales especializados y preparados para el futuro. La dependencia predominante de iniciativas programáticas y proyectos de reforma (75 % para la IA, 12 % para la actualización activa de la ciberdelincuencia y 40 % para la revisión de las leyes de privacidad), yuxtapuesta a la presencia generalizada de leyes básicas (25 %, 88 % y 60 %, respectivamente), ilustra una fase de transformación en la que las leyes existentes sirven como soluciones provisionales mientras se desarrollan regulaciones integrales y adaptadas a la tecnología. Esta transición legislativa se alinea con los objetivos más amplios de promover ecosistemas de ciberseguridad resilientes, responsables y adaptables, lo que requiere una colaboración regional coordinada, el desarrollo de la capacidad legislativa y la participación de las partes interesadas para garantizar que las nuevas leyes sean eficaces y aplicables ante el rápido avance de la IA y las amenazas cibernéticas.

Tabla 1 Modalidades y Enfoques de Prácticas

Privacidad	Prácticas	
	% Predominantes	% Complementarias
Ética por Diseño y Privacidad Diferencial	74	26
Controles y Medidas de Seguridad	58	42
Trazabilidad de decisiones algorítmicas	60	40
Transparencia de las decisiones de IA	47	53
Supervisión humana de sistemas IA	70	20

Nota: La categoría de práctica "Predominante" es la modalidad o enfoque que adopta la mayoría de los CSIRTs para ese indicador (el porcentaje más alto). Y en "Complementaria" es la modalidad o enfoque minoritario, secundaria o de apoyo, que adopta una minoría de CSIRTs en ese indicador (el porcentaje más bajo). Fuente: Elaboración propia.

Los datos sobre privacidad y prácticas de gobernanza de la IA revelan claras diferencias en la forma en que los CSIRT institucionalizan los controles claves. La ética por diseño y la privacidad diferencial son las medidas más arraigadas, ya que el 74 % de los equipos las consideran una práctica predominante y el 26 % las consideran medidas complementarias. Esta alta aceptación demuestra un amplio reconocimiento de que las técnicas de preservación de la privacidad, como la inyección de ruido y la minimización de datos, deben integrarse en los sistemas de IA desde el principio, en lugar de añadirse posteriormente.

La supervisión humana de los sistemas de IA también es una prioridad: el 70 % de los CSIRT exigen controles humanos predominantes en los puntos críticos de decisión, mientras que el 20 % aplica la supervisión de forma más selectiva como medida complementaria. Este fuerte énfasis en la responsabilidad refleja el entendimiento del sector de que las decisiones de IA totalmente autónomas pueden plantear riesgos inaceptables en las operaciones de ciberseguridad.

La trazabilidad de las decisiones algorítmicas ocupa el siguiente lugar, con un 60 % de las organizaciones que integran registros detallados, tarjetas de modelos o registros de rutas de decisión como práctica predominante, y un 40 % que considera la trazabilidad como una garantía complementaria. Estas proporciones sugieren que, si bien la mayoría aprecia el valor de la trazabilidad para el análisis posterior a los incidentes y el cumplimiento normativo, sigue existiendo la oportunidad de elevarla a la categoría de práctica estándar en todos los CSIRT.

Los controles y medidas de seguridad, como el cifrado de los artefactos de los modelos, las restricciones de acceso estrictas y los ciclos de vida de desarrollo seguros, están institucionalizados por el 58 % de los equipos como una práctica predominante, mientras que el

42 % los trata de forma complementaria. La considerable minoría que considera estos controles como medidas complementarias indica la necesidad de integrar más plenamente el refuerzo de la seguridad en los marcos de gobernanza de la IA.

Por último, la transparencia de las decisiones de IA está menos formalizada: solo el 47 % de los CSIRT adoptan la transparencia —exponer la lógica del modelo o las puntuaciones de confianza a las partes interesadas pertinentes— como práctica predominante, mientras que el 53 % la emplean de forma complementaria. Esto sugiere que, a pesar de los fuertes compromisos con la ética y la supervisión, la comunicación explícita de los resultados impulsados por la IA sigue siendo variable, posiblemente debido a la preocupación por revelar la lógica de detección a los adversarios o a la falta de herramientas de explicabilidad maduras.

En general, la tabla muestra que la ética por diseño 74 % y la supervisión humana 70 % son las prácticas de privacidad más sólidamente institucionalizadas, seguidas de la trazabilidad 60 % y los controles de seguridad 58 %. La transparencia se queda atrás con un 47 %, lo que pone de relieve la necesidad de elevar estas prácticas complementarias a requisitos básicos para reforzar la rendición de cuentas, el cumplimiento y la confianza en las operaciones cibernéticas impulsadas por la IA.

1.2 Modelo de Gobernanza Adaptativa para CSIRT Impulsados por IA

La evolución del panorama de amenazas cibernéticas y la incorporación de tecnologías de IA en los CSIRTs demanda un paradigma de gobernanza que trascienda los marcos tradicionales estáticos. El modelo de gobernanza adaptativa emerge como una respuesta estratégica a los desafíos dinámicos que enfrentan las organizaciones en la gestión de incidentes de seguridad cibernética.[39 40 41]

La gobernanza adaptativa reconoce la necesidad de flexibilidad ante la dinámica de las amenazas cibernéticas y la evolución tecnológica. Los CSIRTs impulsados por IA deben adoptar modelos de gobernanza que permitan la actualización continua de políticas, la integración de nuevas capacidades y la adaptación a cambios regulatorios. Esto implica estructuras organizativas ágiles, mecanismos de retroalimentación y aprendizaje organizacional, así como la participación activa de expertos multidisciplinarios.

> En ALC, CSIRT Americas es un claro ejemplo de gobernanza adaptativa en la práctica, con 52 CSIRT de 22 países que operan bajo protocolos TLP estandarizados para el intercambio de información. Esta red demuestra cómo la gobernanza adaptativa puede funcionar operativamente al tiempo que se mantienen los principios democráticos y los requisitos de supervisión humana.

Así mismo esta gobernanza se apoya en la monitorización de riesgos en tiempo real, la gestión automatizada de incidentes y la capacidad de respuesta proactiva. Los sistemas de IA deben ser auditables y configurables para ajustarse a los contextos operativos y regulatorios específicos de cada CSIRT, promoviendo la resiliencia y la sostenibilidad del ecosistema de ciberseguridad.

> Los expertos resaltan como innovacion clave la IA como un “sistema vivo” que se puede entrenar y adaptar a diario, y que requiere registros inmutables tanto para los modelos de IA como para el personal supervisor a fin de garantizar la rendición de cuentas.

El concepto de gobernanza dimensional propone un marco que rastrea dinámicamente cómo se distribuye la autoridad de decisión, la autonomía del proceso y la responsabilidad a través de las relaciones humano-IA. Este enfoque es especialmente relevante para los CSIRTs, donde la velocidad de respuesta a incidentes requiere un equilibrio cuidadoso entre la automatización y la supervisión humana.

Los modelos de gobernanza híbrida integran elementos de supervisión humana con capacidades de IA, creando estructuras que pueden adaptarse dinámicamente a diferentes escenarios de amenazas.[42] Estos modelos incorporan mecanismos de ML y el ser humano en el bucle (HITL, por

sus siglas en inglés), para asegurar que las decisiones automáticas mantengan la alineación con los objetivos organizacionales y los estándares éticos.

Esta gobernanza adaptativa para los CSIRTs impulsados por IA se fundamenta en la capacidad de ajuste dinámico en tiempo real ante la evolución constante de las amenazas y las capacidades tecnológicas. A diferencia de los marcos de gobernanza tradicionales que operan con estructuras rígidas y procesos predefinidos, este modelo incorpora mecanismos de retroalimentación continua y aprendizaje automático que permiten la reconfiguración automática de políticas, procedimientos y roles organizacionales.[40 43]

El concepto central radica en la resiliencia adaptativa, donde las políticas de gobernanza se ajustan mediante información de amenazas en tiempo real y evaluación post-incidente.[39] Esta aproximación reconoce que las amenazas cibernéticas modernas, especialmente las Amenazas Persistentes Avanzadas (APT, por sus siglas en inglés), revelan vulnerabilidades inherentes en las medidas de seguridad convencionales y subrayan la necesidad urgente de estrategias continuas, adaptativas y proactivas.[5]

Arquitectura del Modelo de Gobernanza Adaptativa

Componente de Inteligencia Aumentada

El núcleo del modelo se estructura sobre un sistema de inteligencia aumentada que integra algoritmos de machine learning con la experiencia humana especializada. Este componente procesa continuamente datos de múltiples fuentes: logs de seguridad, fuentes de inteligencia sobre amenazas, reportes de incidentes y métricas de rendimiento organizacional.[41 44] La arquitectura incorpora técnicas especializadas basadas en IA para extraer conocimiento de alta calidad sobre diversas entidades de amenazas y sus relaciones, construyendo y actualizando continuamente un grafo de conocimiento de amenazas.[45]

Marco de Toma de Decisiones Distribuida

La gobernanza adaptativa implementa un marco de decisiones distribuida que permite respuestas autónomas en diferentes niveles organizacionales. El sistema utiliza el paradigma de inteligencia artificial agentiva, donde sistemas autónomos ejecutan tareas con toma de decisiones autodirigida a través de detección de amenazas, orquestación de respuesta y aplicación de políticas.[2] Esta distribución de la capacidad decisoria reduce significativamente los tiempos de respuesta ante incidentes críticos mientras mantiene la coherencia estratégica organizacional.

Mecanismo de Retroalimentación Continua

El modelo integra un mecanismo de retroalimentación continua que captura lecciones aprendidas de cada incidente y las incorpora automáticamente en la base de conocimiento organizacional. Este proceso utiliza técnicas de procesamiento de lenguaje natural y análisis de patrones para identificar tendencias emergentes, gaps en las capacidades actuales y oportunidades de mejora en los procesos de respuesta.[42 46]

Componentes Operacionales del Modelo

Gestión Dinámica de Roles y Responsabilidades

La estructura organizacional del CSIRT se redefine mediante roles adaptativos que se ajustan según la naturaleza y severidad de los incidentes. El modelo incorpora algoritmos de asignación de recursos que consideran el expertise individual, la carga de trabajo actual y los requisitos específicos del incidente.[47] Esta flexibilidad organizacional permite una escalabilidad horizontal eficiente durante crisis cibernéticas masivas.

Políticas de Seguridad Adaptativas

Las políticas de seguridad operan como entidades dinámicas que se modifican automáticamente basándose en el análisis continuo del panorama de amenazas. El sistema implementa ML adaptativo para identificar patrones de ataque emergentes y ajustar proactivamente las reglas de detección, los umbrales de alerta y los procedimientos de respuesta.[48] Esta capacidad de autoajuste reduce significativamente la ventana de exposición ante amenazas zero-day.

Orquestación Inteligente de Respuesta

La respuesta a incidentes se orquesta mediante flujos de trabajo o workflows inteligentes que combinan automatización de tareas rutinarias con escalación inteligente para decisiones complejas. El sistema utiliza análisis predictivo para anticipar la evolución de incidentes y pre-posicionar recursos necesarios, optimizando así la eficiencia operacional y minimizando el impacto de los ataques.[41 46]

Mecanismos de Control y Supervisión

Marcos y Auditorias Continua

El modelo implementa un sistema de auditoria continua que documenta automáticamente todas las decisiones automatizadas, proporcionando trazabilidad completa para auditorías regulatorias y análisis post-incidente. Esta capacidad es crítica para mantener el cumplimiento normativo y la responsabilidad organizacional en entornos donde las decisiones automatizadas pueden tener consecuencias significativas.[2]

Supervisión Humana Estratégica

Aunque el modelo enfatiza la automatización, incorpora puntos de control humano estratégicos donde se requiere supervisión especializada. Estos puntos se identifican mediante análisis de riesgo automatizado y garantizan que las decisiones críticas mantengan el juicio humano apropiado, especialmente en situaciones con implicaciones legales, éticas o de negocio significativas.[1]

Evaluación y Métricas de Rendimiento

Indicadores de Rendimiento Adaptativos

El modelo utiliza métricas de rendimiento que evolucionan según las amenazas actuales y los objetivos organizacionales. Los indicadores clave de rendimiento (KPI, por sus siglas en inglés) se ajustan automáticamente para reflejar nuevas prioridades de seguridad, cambios en el perfil de riesgo organizacional y lecciones aprendidas de incidentes previos.[40] Esta adaptabilidad métrica asegura que la evaluación del rendimiento permanezca relevante y accionable. Estos indicadores deben abordar tanto la efectividad técnica como la conformidad ética y regulatoria, proporcionando métricas cuantificables para la evaluación continua del desempeño del sistema.

Análisis de Efectividad Longitudinal

El sistema implementa análisis longitudinal para evaluar la efectividad de las adaptaciones realizadas, identificando qué modificaciones han resultado en mejoras sostenibles y cuáles requieren refinamiento adicional. Esta capacidad de autoevaluación continua permite la optimización iterativa del modelo de gobernanza.

Consideraciones de Implementación

La implementación exitosa del modelo de gobernanza adaptativa requiere una transformación cultural que abrase la incertidumbre y la adaptación continua como elementos naturales del

entorno operacional. Las organizaciones deben desarrollar capacidades de gestión del cambio que permitan la evolución rápida de procesos sin comprometer la estabilidad operacional.

La integración con sistemas obsoleto o legacy representa un desafío técnico significativo que requiere arquitecturas de interfaz flexibles y estrategias de migración gradual. El modelo debe diseñarse para coexistir con sistemas existentes mientras proporciona un camino de evolución hacia capacidades completamente adaptativas.

El modelo de gobernanza adaptativa para CSIRT impulsados por IA representa una evolución paradigmática necesaria para enfrentar la complejidad creciente del paisaje de amenazas cibernéticas. Su implementación exitosa requiere un enfoque holístico que integre tecnología avanzada, rediseño organizacional y transformación cultural, posicionando a las organizaciones para responder efectivamente a las amenazas cibernéticas del futuro.

1.3 Gobernanza Responsable y Sostenible de Tecnologías Emergentes y Disruptivas en Entornos de la Alta Criticidad

La sostenibilidad en la gobernanza de la IA para entornos críticos abarca dimensiones ambientales, legales, económicas y sociales.[49] Los sistemas de IA consumen recursos computacionales significativos, lo que plantea desafíos de sostenibilidad ambiental que deben ser considerados desde la idea y en todo el ciclo de vida del sistema de IA para el CSIRT.[50]

Los principios de las tecnologías verdes o green tech tienen la necesidad de desarrollar sistemas de IA que minimicen el consumo energético mientras mantienen la efectividad operacional.[49] En el contexto de los CSIRTs, esto implica optimizar algoritmos de detección de amenazas para equilibrar precisión y eficiencia energética.

La gobernanza responsable requiere la implementación de mecanismos de transparencia algorítmica y explicabilidad.[51] Los sistemas de IA utilizados en respuesta a incidentes deben proporcionar justificaciones comprensibles para sus decisiones, especialmente cuando estas afectas operaciones críticas y de alta complejidad sistémica.

En ALC, los expertos plantean que todo este panorama trae consideraciones geopolíticas críticas, en particular el impacto de la rivalidad entre las grandes potencias en materia de IA en el desarrollo regional. Esto añade una dimensión de soberanía a la sostenibilidad, haciendo hincapié en la necesidad de la soberanía tecnológica y de datos, junto con las consideraciones medioambientales.

Los marcos de evaluación de impacto proporcionan metodologías sistemáticas para evaluar las consecuencias potenciales de los sistemas IA en entornos críticos. Estos marcos incluyen evaluaciones de riesgos de seguridad, privacidad,[52] sesgo algoritmo, ética, impacto social y los criterios medioambiente, social y de gobernanza (ESG, por sus siglas en inglés).

La arquitectura de gobernanza multinivel propuesta por investigaciones recientes establece tres niveles independientes de partes interesadas: gobiernos, organizaciones y ciudadanía.[53] En el contexto de las ciberoperaciones, esta estructura se traduce en:

Tabla 2 Gobernanza Multinivel

Nivel	Foco
Estratégico	Establece políticas nacionales de ciberseguridad que incorporan directrices específicas para el uso de IA en infraestructuras criticas.[54] [55] Los marcos regulatorios deben abordar la interoperabilidad entre sistemas IA de diferentes proveedores y la gestión de riesgos sistémicos. Hay que destacar el buen gobierno del dato y su calidad, además de la adecuada gestión de la privacidad.
Táctico	Implementa estándares técnicos y protocolos operacionales que garanticen la integración segura de sistemas IA en arquitecturas de seguridad.[56] [57] Esto

Nivel	Foco
	incluye la definición de métricas de rendimiento, umbrales de intervención humana y protocolos de escalamiento.
Operacional	Gestiona la ejecución diría de las operaciones automatizadas, manteniendo capacidades de monitoreo continuo, análisis del comportamiento anómalo y respuesta adaptativa a amenazas emergentes.[58] [59]

Un gran desafío es la fragmentación regulatoria entre jurisdicciones, la velocidad de la evolución tecnológica versus la adaptabilidad normativa, y la necesidad de equilibrar la autonomía operacional con el control humano.[18] Lo cual subraya que una gobernanza efectiva requiere el desarrollo de capacidades en la dinámica organizacional especializada y la implementación de formación continua para el talento técnico.[3]

La sostenibilidad de la gobernanza depende de la capacidad de adaptarse a las amenazas emergentes y tecnologías disruptivas.[57] Esto requiere la implementación de mecanismos de retroalimentación continua, evaluación periodo de la efectividad de las políticas y actualización proactiva de los marcos normativos.

La eficiencia energética y la huella de carbono es esencial para la sostenibilidad ambiental de los sistemas de IA en los CSIRTs ya que requiere la optimización del consumo energético tanto en gases de entrenamiento como de inferencias.[50] [60] [61] La implementación de modelos eficientes energéticamente, incluyendo técnicas de compresión, incluyendo pruning, quantization, y destilación del conocimiento, permiten la reducción significativa de la huella de carbono computacional de modelos de IA sin degradación substancial del rendimiento,[61] [62] al tiempo de no comprometer la eficiencia operacional. La implementación de cuantización dinámica reduce la precisión del peso y las activaciones de los modelos durante inferencia, logrando aceleraciones de 2-4x con pérdida mínima de precisión y exactitud.

La aplicación de búsqueda de arquitectura neuronal (NAS, por sus siglas en inglés) automatizada permite el diseño de arquitecturas específicamente optimizadas para tareas de ciberseguridad bajo restricciones de los recursos y capacidades computacionales.[63] [64] Estos modelos optimizados demuestran particular valor en implementaciones Edge donde recursos computacionales son limitados.

La adopción estratégica de la computación verde -green computing- o tecnologías verdes -green tech- incluye la utilización de energías renovables para centros de datos, optimización de algoritmos para reducir la sobrecarga computacional -computacional overhead-, y la implementación de la programación inteligente -schedulling inteligence- que aproveche periodos de menor demanda energética.[61] [62] [65] La medición continua de las emisiones de carbono asociadas con operaciones de IA debe integrarse en métricas de performance organizacional.

La sostenibilidad social en la implementación de IA en CSIRT requiere inversiones continuas en desarrollo de capacidades humanas y programas de mejoras y fortalecimiento de competencias que permitan a los profesionales de ciberseguridad adaptarse a entornos de trabajo colaborativo humano-IA.[21] [66] [67] Esto incluye la creación de programas de formación continua en interpretabilidad de IA, toma de decisiones éticas, y supervisión de sistemas autónomos.

Las consideraciones de equidad social deben abordar el impacto distributivo de la automatización en el empleo del sector ciberseguridad,[68] promoviendo transiciones laborales justas y oportunidades de reentrenamiento para profesionales cuyas funciones puedan ser afectadas por la autonomía y desempeño total de modelos o sistemas de IA en algunos procesos de las ciberoperaciones.

La sostenibilidad económica requiere modelos de evaluación de retorno de la inversión (ROI, por sus siglas en inglés) que consideren tanto beneficios cuantificables (reducción de tiempo de respuesta, mejora en tasas de detección) como beneficios intangibles (mejora en reputación organizacional, cumplimiento regulatorio).[69] Los marcos de evaluación deben incluir análisis de

costo-beneficio a largo plazo que consideren costos de mantenimiento, actualización y obsolescencia tecnológica.

La implementación escalonada de capacidades de IA permite la optimización gradual de inversiones y la demostración de valor antes de comprometerse con despliegues de mayor escala.[70][71] Esta aproximación facilita el ajuste continuo de estrategias de implementación basado en resultados empíricos y feedback operacional.

Es preciso disponer de mecanismos de supervisión y control como el establecimiento de comités de ética de IA específicos para contextos de ciberseguridad requiere la inclusión de expertos multidisciplinario, incluyendo especialistas en ética tecnológica, profesionales de ciberseguridad, expertos legales, y representantes de las partes interesadas afectadas.[19][72][73] Estos comités deben desarrollar directrices específicas para dilemas éticos comunes en respuesta a incidentes, incluyendo solución de compromiso "trade-offs" entre velocidad de respuesta y precisión, y decisiones sobre intercambio de inteligencia de amenazas.

Los sistemas de auditoría para IA en los CSIRTs deben incorporar métricas específicas que evalúen tanto performance técnica como adherencia a principios éticos.[74][75] Esto incluye evaluaciones regulares de equidad o justicia -fairness- algorítmica, robustez contra ataques adversarios, y la efectividad de mecanismos de supervisión humana.

La implementación de sistemas de monitoreo continuo debe incluir alertas automáticas para desviaciones significativas del rendimiento, deriva del modelo -drift- en distribuciones de datos, y violaciones de umbrales éticos y técnicos predefinidos.[76][77] Esta capacidad de monitoreo en tiempo real es esencial para mantener la confianza en sistemas de IA operando en entornos críticos.

Green AI y Carbon-Aware Computing

La implementación de estrategias de IA verde o green AI incluye programación consciente del carbono -carbon aware schedulling- que optimiza el tiempo de entrenamiento de los modelos para aprovechar períodos de mayor disponibilidad de energía renovable.[60][66][68] Los sistemas de monitoreo de huella de carbono integrados proporcionan métricas en tiempo real de emisiones asociadas con operaciones de IA.

La adopción de técnicas de aprendizaje por transferencia y aprendizaje con pocos ejemplos reduce significativamente los requisitos de potencia computacional necesarios para adaptar los modelos a nuevos dominios o tipos de amenazas.[62][78] Este enfoque es especialmente valioso para organizaciones con recursos computacionales limitados que desean implementar capacidades avanzadas de inteligencia artificial.

Arquitecturas Sostenibles de Data Center

Infraestructura como Código y Automatización

La implementación de infraestructura como código (IaC, por sus siglas en inglés) para entornos de IA permite la optimización continua de la asignación de recursos basada en patrones y flujos de carga de trabajo con métricas de eficiencia.[79][80] Los sistemas inteligentes de autoescalado ajustan dinámicamente los recursos computacionales en función de los patrones de demanda y los objetivos de optimización, que incluyen tanto el rendimiento como la eficiencia energética.

La orquestación automatizada de las cargas de trabajo de entrenamiento e inferencia de IA mediante herramientas como Kubernetes, Apache Airflow y MLflow facilita la implementación de políticas de sostenibilidad que priorizan el uso eficiente de los recursos energéticos.[81][82]

Optimización de Refrigeración e Integración de Energías Renovables

Los sistemas de optimización de la refrigeración utilizan la inteligencia artificial para predecir las cargas térmicas y optimizar los sistemas de climatización en tiempo real, logrando reducciones del 10-15 % en el consumo energético.[83] [84] La integración con fuentes de energía renovables requiere un equilibrio inteligente de la carga que coordine las cargas de trabajo de IA con los patrones de disponibilidad de la energía solar y eólica.

2. Desarrollo Tecnológico de la IA en Entornos de Respuesta Cibernética

En ALC hay que destacar[85] que, para garantizar el desarrollo eficaz y seguro de soluciones de IA, es esencial adoptar metodologías ágiles que promuevan la innovación iterativa, lo que permite la creación de pruebas de concepto, prototipos y productos mínimos viables. Estas etapas sirven como entornos experimentales que fomentan el aprendizaje, la retroalimentación y la mejora continua.

Paralelamente, los CSIRTs deben establecer marcos estructurales que apoyen el desarrollo y la adopción de la IA, al tiempo que evalúan y garantizan que los equipos posean las habilidades y capacidades necesarias. Un componente crítico de este proceso es reconocer la importancia central de los datos e implementar prácticas sólidas de gobernanza de datos. Esto incluye identificar los conjuntos de datos necesarios, las fuentes de datos disponibles y definir la arquitectura y el flujo de información dentro del sistema.

En ALC, los expertos destacan los siguientes retos en las capacidades tecnicas en los CSIRTs:

1. Detección de amenazas en tiempo real e identificación de anomalías.
2. Análisis predictivo para la anticipación de amenazas emergentes.
3. Respuesta automatizada a incidentes y remediación.
4. Integración con la infraestructura de seguridad existente y los sistemas heredados.

Además, desde la fase de diseño, es crucial evaluar las herramientas tecnológicas y la infraestructura de datos necesarias, en particular en lo que respecta a las capacidades de almacenamiento y procesamiento, para apoyar la implementación de soluciones de IA. La selección del modelo debe basarse en criterios objetivos alineados con la naturaleza del problema, el tipo y la calidad de los datos, los recursos computacionales, los objetivos de rendimiento y la necesidad de explicabilidad. Por último, las consideraciones éticas, la protección de la privacidad y los requisitos de seguridad deben integrarse desde las primeras etapas del diseño para garantizar una implementación responsable, fiable y sostenible de la IA.

El desarrollo tecnológico de sistemas de inteligencia artificial para entornos de respuesta cibernética ha experimentado una evolución significativa hacia arquitecturas híbridas que combinan aprendizaje automático, procesamiento de lenguaje natural y análisis predictivo.[44] Estas arquitecturas integran múltiples componentes tecnológicos: sistemas de detección de anomalías basados en redes neuronales convolucionales (CNN, por sus siglas en inglés), redes neuronales recurrentes (RNN, por sus siglas en inglés) para análisis temporal de patrones de ataque, y algoritmos de detección de anomalías basados en grafos (GBAD, por sus siglas en inglés) para identificación de comportamientos maliciosos.[86]

En ALC, los expertos destacan los siguientes retos en el desarrollo de la IA en los CSIRTs:

1. Escasez de talento en IA especializada para aplicaciones de ciberseguridad.

2. Altos requisitos de consumo computacional y energético.
3. Calidad y estandarización de los datos entre organizaciones.
4. Complejidad de la integración de los sistemas heredados.

La implementación de plataformas de inteligencia de amenazas automatizadas representa un avance crítico en la capacidad de respuesta proactiva.[44] Estas plataformas emplean algoritmos de machine learning para el procesamiento automatizado de IoC, la correlación de amenazas múltiples vectores y la generación de inteligencia accionable en tiempo real. Los sistemas contemporáneos incorporan capacidades de procesamiento de big data para el análisis de volúmenes masivos de logs de seguridad, tráfico de red y telemetría de endpoints.[86]

Estas tecnologías facilitan el procesamiento de datos en tiempo real y los protocolos de respuesta ágil, esenciales para adelantarse a los atacantes, que a menudo explotan la automatización y la velocidad como parte de sus tácticas.

Una tendencia crítica es el cambio de una ciberseguridad reactiva a una proactiva, aprovechando las capacidades predictivas impulsadas por la inteligencia artificial para prever y prevenir posibles incidentes. Mediante el análisis de grandes cantidades de datos históricos y actuales, los modelos de aprendizaje automático pueden identificar patrones emergentes, anticipar posibles vulnerabilidades y priorizar las respuestas en función del riesgo.

Esta orientación predictiva, combinada con la automatización inteligente, permite a las organizaciones optimizar la asignación de recursos, reducir el tiempo de permanencia de las amenazas y maximizar la rentabilidad de las operaciones de seguridad. Sin embargo, este progreso también plantea nuevos retos, como la deriva de los modelos, la manipulación adversaria y el problema de la explicabilidad de la "caja negra", lo que requiere un perfeccionamiento técnico continuo y marcos de validación robustos.

Otra necesidad tecnológica imperiosa es la sostenibilidad y la resiliencia en el diseño de soluciones de IA. Los modelos de aprendizaje automático de alto rendimiento pueden imponer elevados costes computacionales y medioambientales, lo que hace que la eficiencia y la gestión del ciclo de vida sean una prioridad.

Las prácticas de ingeniería sostenible requieren algoritmos ligeros, canales de datos optimizados y arquitecturas modulares que admitan la escalabilidad, la interoperabilidad y la reutilización. La gestión del ciclo de vida también debe incluir una formación continua sobre conjuntos de datos diversos y representativos para mitigar la degradación del rendimiento y permitir la adaptación a nuevas tipologías de amenazas sin necesidad de un rediseño completo.

Finalmente, el desarrollo tecnológico en este contexto debe integrar estrechamente la resiliencia como principio básico de diseño. Esto incluye redundancias, mecanismos de respaldo y "degradación gradual", lo que garantiza que, incluso en casos de fallo parcial del sistema, las funciones críticas del CSIRT persistan. La confluencia de la robustez (resiliencia a los ataques y errores), la adaptabilidad (reconfiguración en respuesta a las novedades) y el autoaprendizaje (mejora a partir de la retroalimentación humana y los resultados operativos) marca el futuro del desarrollo tecnológico para la respuesta de ciberseguridad impulsada por la IA.

Desarrollo Seguro de la IA

El ciclo de vida de desarrollo seguro de la IA es un marco fundamental para garantizar que los sistemas y modelos de IA se diseñen, desarrollen, implementen y mantengan de manera que se minimicen los riesgos y se refuerce la confianza. En el contexto de las operaciones de ciberseguridad, especialmente dentro de los CSIRTs, el desarrollo seguro de la IA debe incorporar los principios tradicionales de seguridad del software, como la programación segura, la autenticación y el control de acceso, al tiempo que aborda los riesgos específicos de la IA.

Estos incluyen el envenenamiento de datos, las entradas adversas, la inversión de modelos y la propagación de sesgos. Un sistema de IA robusto comienza con prácticas de diseño seguro, incluido el modelado de amenazas en una etapa temprana que identifica posibles vectores de ataque contra los procesos de IA, como la ingesta de datos, el entrenamiento de modelos y la inferencia. Esta etapa requiere la colaboración interfuncional entre científicos de datos, ingenieros de software y profesionales de la ciberseguridad para anticipar las vulnerabilidades específicas del sistema.

Algo central del desarrollo seguro de la IA es la garantía de la integridad y la privacidad de los datos a lo largo de todo el ciclo de vida de la IA. Dado que los sistemas de IA se basan intrínsecamente en los datos, es fundamental garantizar la procedencia, la calidad y la protección de los conjuntos de datos de entrenamiento. Esto es especialmente importante para los CSIRTs que utilizan la IA para detectar anomalías o predecir ataques, ya que los datos comprometidos pueden inducir a error a los modelos y mermar la precisión de la detección.

Técnicas como la privacidad diferencial, el aprendizaje federado y el cálculo multipartito seguro (SMPC, por sus siglas en inglés) pueden proporcionar salvaguardias contra la fuga de datos y los ataques de inferencia de modelos. Además, la validación y las pruebas continuas con ejemplos adversarios son esenciales para reforzar la solidez de los modelos y evitar la explotación mediante entradas cuidadosamente elaboradas que puedan eludir la detección.

Para seguir apoyando la implementación segura de la IA en los CSIRTs, la resiliencia frente al aprendizaje automático adversario (AML, por sus siglas en inglés) debe ser una prioridad. Los modelos deben someterse a pruebas de estrés no solo en cuanto a rendimiento, sino también en cuanto a seguridad en condiciones adversas. Esto incluye la implementación de mecanismos para el entrenamiento adversario, la explicabilidad de los modelos y los activadores a prueba de fallos que permiten a los analistas humanos intervenir cuando el comportamiento del modelo se desvía de las normas esperadas.

Los CSIRTs, que a menudo se encargan de tomar decisiones críticas bajo complejidad e incertidumbre, deben confiar en sistemas de IA que proporcionen tanto precisión como interpretabilidad. Esto requiere principios de transparencia desde el diseño que incorporen la trazabilidad y la auditabilidad en los resultados de la IA, lo que permite a los equipos de seguridad validar las decisiones automatizadas, comprender las anomalías y justificar las medidas de respuesta de acuerdo con las normas legales y éticas.

La integración de la IA en las operaciones del CSIRT puede mejorar significativamente las capacidades de detección, clasificación y respuesta. Las herramientas basadas en IA pueden analizar grandes cantidades de datos de telemetría para identificar patrones que sugieran amenazas emergentes, automatizando las etapas iniciales de clasificación y priorización de incidentes. Por ejemplo, el procesamiento del lenguaje natural (NLP, por sus siglas en inglés) se puede utilizar para analizar fuentes de inteligencia sobre amenazas o informes de incidentes, mientras que los algoritmos de aprendizaje automático pueden correlacionar los IoC entre sistemas en tiempo real.

Además, la IA puede apoyar la toma de decisiones durante los incidentes proporcionando evaluaciones probabilísticas de posibles vías de ataque, recomendando medidas de contención y simulando escenarios de impacto. Sin embargo, una implementación exitosa requiere no solo una alineación técnica, sino también una preparación organizativa, que incluye la formación del personal del CSIRT para interpretar los resultados de la IA, gestionar los falsos positivos y proporcionar retroalimentación para mejorar continuamente los modelos de IA.

Tabla 3 Enfoques de las Prácticas

Enfoques	Prácticas	
	% Predominante	% Complementaria
Calidad de Datos	74	26
Desafios en Integraciones con IA	56	44
Revision Periodica de Algoritmos y Modelos	71	29
Tecnicas de Estimacion de Datos	52	48
Monitoreo de Rendimiento	77	23
Prueba de concepto	61	39

Nota: La categoría de práctica "Predominante" es la modalidad o enfoque que adopta la mayoría de los CSIRTs para ese indicador (el porcentaje más alto), como un componente central de sus operaciones. Y en "Complementaria" es la modalidad o enfoque minoritario, secundaria o de apoyo, que adopta una minoría de CSIRTs en ese indicador (el porcentaje más bajo), se refiere a su uso de manera esporádica o de apoyo según las necesidades. Fuente: Elaboración propia.

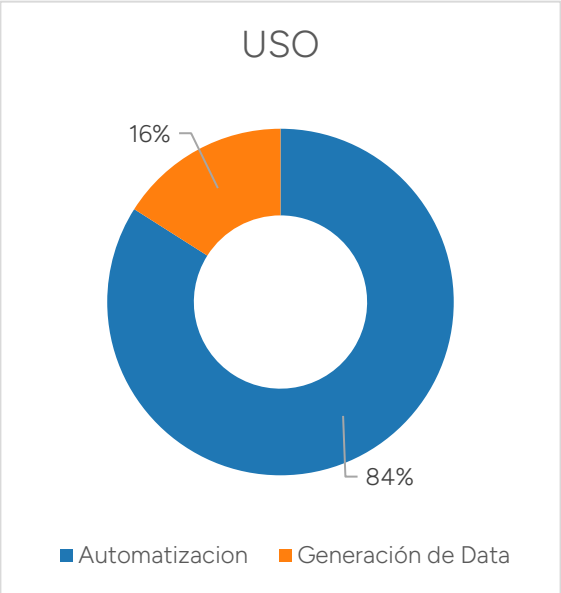
La tabla destaca un patrón claro en la forma en que los CSIRTs consideran la integran y gestion de la IA en las CyberOps, revelando enfoques predominantes y complementarios en seis ámbitos clave. En primer lugar, la calidad de los datos destaca como el área más prioritaria de forma universal, ya que el 74% de los equipos la consideran una práctica predominante; solo el 26% la utiliza de forma complementaria, lo que subraya que los datos de alta fidelidad se reconocen ampliamente como fundamentales para cualquier iniciativa de IA.

Del mismo modo, el monitoreo del rendimiento -transparencia y confianza- es una actividad fundamental para el 77% de los CSIRTs, lo que refleja la necesidad crítica de realizar un seguimiento continuo de la eficacia de la IA y protegerse contra la deriva de los modelos, mientras que solo el 23% la gestiona de forma más puntual. La revisión periódica de los algoritmos y modelos también está institucionalizada de forma predominante por el 71% de los equipos, lo que indica la madurez de los procesos de gobernanza para garantizar que los sistemas de IA se mantengan alineados con los cambiantes panoramas de amenazas; el 29% restante revisa los modelos solo como medida complementaria.

Por el contrario, las técnicas de estimación de datos, utilizadas para inferir información faltante o incierta, se distribuyen de manera más uniforme, ya que el 52% las considera una práctica principal y el 48% una práctica complementaria, lo que sugiere que casi la mitad de los equipos carecen de capacidades de estimación sólidas o las utilizan solo cuando es necesario. Los retos de la integración de la IA (por ejemplo, la interoperabilidad técnica, la resistencia cultural o las limitaciones de recursos) son abordados como un objetivo predominante por el 56% de los equipos, pero siguen siendo una preocupación secundaria para el 44%, lo que indica que una minoría sustancial puede no dar la prioridad necesaria a la superación de las barreras de integración.

Por último, el 61% de los CSIRTs llevan a cabo pruebas de concepto (PoC, por sus siglas en ingles) como iniciativas formales y predominantes, mientras que el 39% utiliza las PoC de forma más moderada o informal, lo que apunta a la variabilidad en el rigor con el que los CSIRTs prueban las soluciones de IA antes de su plena implementación. En general, esta distribución de prácticas revela que, si bien las actividades fundamentales como la calidad de los datos, la supervisión del rendimiento y la revisión de modelos están ampliamente integradas en las operaciones de los CSIRTs, áreas como las técnicas de estimación, los retos de integración y el rigor de las PoC muestran una mayor heterogeneidad, lo que pone de relieve las oportunidades de estandarizar las mejores prácticas y reforzar el ciclo de vida de la adopción de la IA en toda la región.

Gráfico 3. Perspectiva y Foco de Uso



Fuente: Elaboración propia.

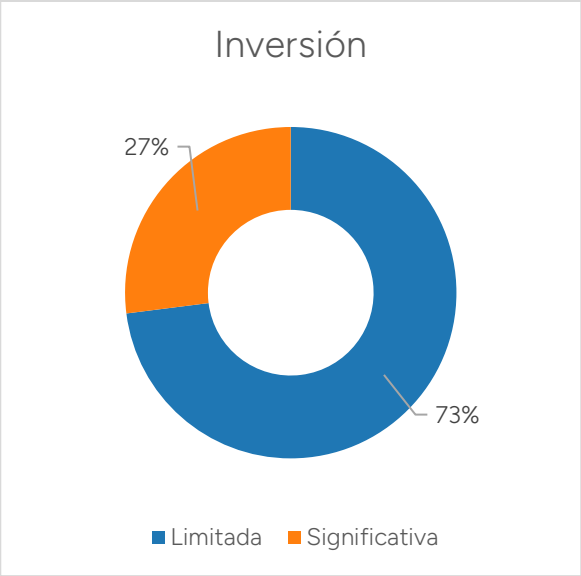
El 84% de los CSIRTs de todo el mundo han integrado la IA principalmente para la automatización de tareas, aprovechando los algoritmos de aprendizaje automático para optimizar funciones rutinarias como la clasificación de alertas, la correlación de inteligencia sobre amenazas y los flujos de trabajo de gestión de incidentes. Esta adopción generalizada subraya la eficacia demostrada de la IA para reducir la carga de trabajo humana y acelerar los tiempos de respuesta.

Por el contrario, el 16% de los equipos se centra en procesos que generan datos para su posterior análisis y explotación, utilizando la IA para enriquecer los conjuntos de datos mediante la búsqueda automatizada de amenazas, la detección de anomalías y los procesos de análisis de comportamiento.

Este grupo más reducido refleja un cambio estratégico emergente hacia el aprovechamiento de la IA no solo para la eficiencia operativa, sino también para mejorar la calidad y el volumen de la inteligencia procesable.

La disparidad entre las aplicaciones de IA centradas en la automatización 84% y las basadas en el análisis y generación de datos 16% pone de relieve un modelo de madurez de dos niveles: la mayoría de los CSIRTs dan prioridad a las ganancias de eficiencia inmediatas, mientras que una minoría invierte en la creación de infraestructuras de datos basadas en la IA que pueden sustentar capacidades avanzadas de ciberseguridad predictivas y prescriptivas.

Gráfico 4 Gestión de Inversión



Fuente: Elaboración propia.

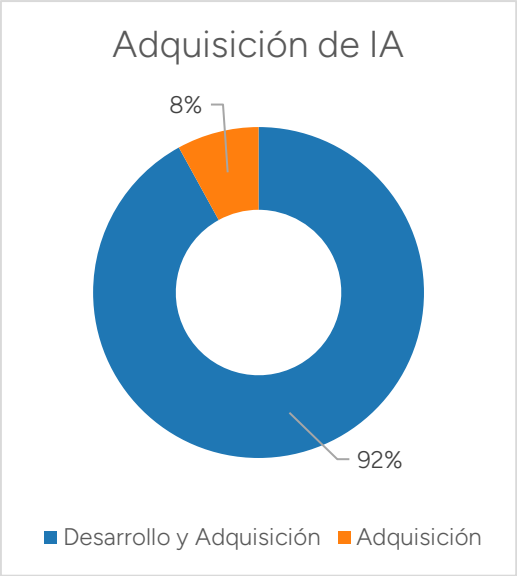
El 73% de los CSIRTs tienen la intención de destinar una parte limitada de sus presupuestos (entre el 0% y el 5%) a la implementación de soluciones de IA, lo que refleja un enfoque cauteloso y gradual, condicionado por la dependencia de las asignaciones presupuestarias estatales y las restricciones fiscales que estas conllevan.

Por el contrario, el 27 % de los equipos planea destinar una parte sustancial de sus presupuestos (10-20%) al despliegue de IA, lo que indica una priorización estratégica de las capacidades de IA y una voluntad de invertir más en tecnologías avanzadas de ciberdefensa, a pesar de los retos que plantea la obtención de mayores fondos públicos.

Esta distribución sugiere un panorama de inversión bifurcado en el que la mayoría de los CSIRTs adoptan posturas conservadoras en materia de financiación, impulsadas por los ciclos presupuestarios gubernamentales, posiblemente debido a la incertidumbre presupuestaria o a la

aversión al riesgo, mientras que una minoría significativa persigue estrategias de inversión agresivas destinadas a ampliar rápidamente las operaciones de seguridad basadas en la IA.

Gráfico 5 Gestión del Desarrollo y Adquisición de Soluciones



Fuente: Elaboración propia.

El 92% de los CSIRTs emplean un enfoque equilibrado que combina el desarrollo interno de capacidades de IA con la adquisición de soluciones comerciales o de código abierto, aprovechando la experiencia interna para personalizar las herramientas y cubriendo las carencias con ofertas de terceros.

Por el contrario, el 8% depende exclusivamente de soluciones de IA adquiridas, que pueden ofrecer una rápida implementación, pero que introducen limitaciones como la dependencia de un proveedor, una menor personalización a las amenazas locales, posibles retos de cumplimiento con los requisitos de soberanía de los datos y la dependencia de ciclos de soporte externos que pueden no ajustarse a las necesidades cambiantes de respuesta a incidentes.

Esta dependencia de productos comerciales puede dificultar una integración más profunda con los flujos de trabajo existentes y ralentizar la adaptación cuando se necesitan ajustes a medida.

2.1 Tecnologías emergentes habilitadoras: algoritmos, automatización y capacidades predictivas

Los modelos de IA se utilizan cada vez más para anticipar y predecir eventos futuros, basándose en modelos estadísticos, diseño algorítmico y ejecución automatizada de tareas específicas. En esencia, la IA se basa principalmente en algoritmos de ML, sistemas que se adaptan en función de los datos de entrada y mejoran su rendimiento con el tiempo a través de la experiencia.

Es fundamental distinguir entre el ML y el DL. Mientras que los modelos de ML aprenden de los datos y mejoran con la retroalimentación, los algoritmos de DL se componen de múltiples capas conectadas de forma no lineal, formando lo que se conoce como redes neuronales artificiales. En estas arquitecturas, cada unidad de la capa inferior recibe datos externos, como píxeles de imagen para sistemas de reconocimiento facial, y pasa esta información procesada a unidades seleccionadas de la siguiente capa. Estas unidades, a su vez, integran los datos utilizando reglas matemáticas y envían el resultado a través de capas sucesivas.

A medida que los datos pasan por estas múltiples etapas de procesamiento, se someten a una compleja cadena de transformaciones estadísticas, lo que finalmente da lugar a un resultado probabilístico. Dado que estas capas están interconectadas de forma no lineal, el proceso de toma de decisiones en los modelos de ML se vuelve cada vez más opaco. Este fenómeno, a menudo denominado "efecto caja negra", significa que incluso los expertos pueden ser incapaces de interpretar o explicar completamente cómo se toman determinadas decisiones de IA.

A pesar de esta opacidad, la arquitectura multicapa permite a los sistemas de IA no solo seguir las instrucciones programadas, sino también adaptarse a los cambios en su entorno. Esta capacidad de adaptación es una característica distintiva de la IA: puede aprender de forma autónoma a partir de nuevos datos y generar resultados específicos para cada contexto con un control humano limitado o nulo.

En ALC, los expertos resaltan que necesiario para los CSIRTs desarrollar honeypots dinámicas y entornos sandbox adaptativos que permitan demostrar aplicaciones prácticas de la IA más allá de los sistemas de detección tradicionales.

Una de las aplicaciones más transformadoras de esta adaptabilidad es la IA generativa, un ámbito emergente dentro de la inteligencia artificial que está revolucionando el uso de la tecnología. A diferencia de los modelos tradicionales de IA que clasifican, analizan o procesan datos existentes, la IA generativa tiene como objetivo crear datos completamente nuevos que se asemejen a los originales y que a menudo son indistinguibles del contenido generado por humanos. Estos modelos utilizan técnicas de aprendizaje avanzadas, redes neuronales y arquitecturas de última generación para capturar y reproducir patrones complejos en conjuntos de datos masivos. Aunque todavía está en evolución, la IA generativa es muy prometedora como asistente de conocimiento y creador de contenido en múltiples industrias.

Estos avances algorítmicos y computacionales son especialmente potentes en el ámbito de la ciberseguridad, donde la IA permite pasar de una defensa reactiva a una anticipación proactiva de los riesgos. La base tecnológica de los sistemas de ciberseguridad basados en IA se sustenta en sofisticados algoritmos de aprendizaje automático que permiten la detección avanzada de amenazas, el reconocimiento de patrones y el análisis predictivo, capacidades que son fundamentales para las operaciones modernas de los CSIRTs.

Los sistemas de ciberseguridad actuales utilizan una amplia gama de técnicas de inteligencia artificial, entre las que se incluyen:

- Redes neuronales de aprendizaje profundo para el análisis de patrones complejos;
- Métodos de conjunto como Random Forest y XGBoost para la clasificación de malware;
- Enfoques de aprendizaje no supervisado, como bosques de aislamiento y autocodificadores, para la detección de anomalías.

Estos métodos permiten a las plataformas de ciberseguridad procesar grandes cantidades de datos en tiempo real, detectar comportamientos maliciosos sutiles y adaptarse a las amenazas en constante evolución. A diferencia de los sistemas basados en reglas, las defensas impulsadas por IA pueden aprender dinámicamente del entorno de amenazas, lo que mejora significativamente la precisión de la detección y reduce los falsos positivos.

Una de las características más impactantes de la IA en la ciberseguridad es su capacidad para ofrecer análisis predictivos. Esta capacidad marca un cambio de paradigma al pasar de simplemente reaccionar ante los incidentes a anticiparlos. Al aprovechar:

- Registros históricos de incidentes,
- Fuentes de inteligencia sobre amenazas,
- Análisis del comportamiento y patrones de las actividades en el ciberespacio,

Los modelos de IA pueden pronosticar los vectores de ataque probables, estimar la probabilidad de explotación de vulnerabilidades y predecir el impacto potencial de futuras amenazas. Esta previsión permite a los CSIRTs asignar recursos de manera eficaz, implementar contramedidas preventivas y diseñar posturas de defensa estratégicas que sean tanto proactivas como adaptativas, mejorando así la ciberresiliencia.

Los sistemas de ciberseguridad basados en IA también se benefician de la automatización inteligente, que es esencial para gestionar la velocidad y el volumen de las amenazas cibernéticas actuales. La automatización permite:

- Búsqueda automatizada de amenazas,
- Clasificación y escalados autónomos de incidentes, y
- Coordinación de respuestas en tiempo real, a menudo sin intervención humana.

Estas capacidades son cruciales para defenderse de amenazas automatizadas y de alta velocidad, como las APTs, en las que los ataques a velocidad de máquina pueden desbordar los sistemas de defensa tradicionales impulsados por humanos. Al aprovechar la automatización impulsada por la IA, las organizaciones mantienen operaciones de seguridad eficaces incluso bajo la presión de adversarios sofisticados.

Sin embargo, para aprovechar plenamente el potencial de la IA en la ciberseguridad, las organizaciones deben adoptar ecosistemas tecnológicos integrados. Estos requieren:

- Mecanismos de coordinación sofisticados que coordinen diversos algoritmos, fuentes de datos y sistemas de respuesta;
- Arquitecturas de sistemas capaces de procesar con baja latencia y tomar decisiones en tiempo real;
- Interacción fluida entre los procesos automatizados y los analistas humanos.

Esta integración no solo aumenta la precisión de la detección y reduce el tiempo de respuesta, sino que también proporciona la flexibilidad y la escalabilidad necesarias para hacer frente a los retos de ciberseguridad en constante evolución en entornos operativos muy dinámicos.

La convergencia de algoritmos avanzados, automatización y capacidades predictivas en la IA marca una era transformadora para la ciberseguridad. Estas tecnologías no son complementarias, sino que son facilitadores fundamentales de la ciberdefensa de próxima generación. A medida que las organizaciones se enfrentan a una complejidad digital cada vez mayor, a la presión regulatoria y a la sofisticación de las amenazas, invertir en enfoques basados en la IA se convierte no solo en una ventaja estratégica, sino en un imperativo de ciberseguridad.

A pesar de la creciente integración de la IA en las operaciones de ciberseguridad, "los CSIRTs siguen enfrentándose a importantes retos y carencias en cuanto a capacidades". Entre ellos se incluyen la visibilidad en tiempo real limitada en infraestructuras distribuidas e híbridas, las dificultades para correlacionar la inteligencia sobre amenazas procedente de múltiples fuentes y las restricciones de recursos que impiden la supervisión continua de las amenazas y la respuesta rápida ante incidentes.

Además, la sofisticación de las amenazas cibernéticas, como el malware polimórfico, los exploits de día cero y los ataques a la cadena de suministro, a menudo supera los ciclos de aprendizaje y las capacidades de toma de decisiones de los modelos tradicionales de IA. La falta de transparencia (problemas de "caja negra"), en los sistemas de aprendizaje profundo complica aún más la confianza, la responsabilidad y la explicabilidad en entornos de respuesta a incidentes de alto riesgo. Estas limitaciones exponen a los CSIRTs a la fatiga operativa, a posturas de seguridad reactivas y a dificultades para lograr una ciberresiliencia sistémica.

De cara al futuro, la inteligencia artificial cuántica (QAI, por sus siglas en inglés) ofrece una oportunidad transformadora para abordar muchas de estas limitaciones sistémicas. Al combinar las capacidades de reconocimiento de patrones y aprendizaje de la IA con la potencia de procesamiento exponencial de la "computación cuántica", la QAI podría permitir el procesamiento en tiempo real de datos de amenazas multidimensionales y masivos que superan las capacidades clásicas.

Esto incluye un análisis criptográfico mejorado, una detección más rápida de anomalías en el tráfico cifrado y la capacidad de modelar interdependencias complejas de riesgos cibernéticos con mayor precisión. Para los CSIRTs, la QAI tiene el potencial de reducir drásticamente la latencia de respuesta, apoyar el desarrollo de estrategias de defensa predictivas a gran escala y superar los actuales cuellos de botella computacionales en el modelado de amenazas y la toma de decisiones, lo que marcará el comienzo de una nueva era de "operaciones de ciberdefensa resilientes a las tecnologías cuántica y mejoradas por la IA".

Sistemas de IA Distribuida y Edge Computing

La implementación de arquitecturas de IA distribuida en entornos CSIRT permite el procesamiento de inteligencia de amenazas en tiempo real mientras minimiza la latencia y dependencia de conectividad centralizada.[87 88 89] Estas arquitecturas emplean computación Edge para realizar análisis inicial de anomalías en puntos de recolección de datos, aplicando técnicas de aprendizaje federado que permiten el entrenamiento colaborativo de modelos sin centralizar datos sensibles.

La integración de marcos de orquestación como Kubernetes facilita el despliegue escalable de microservicios de IA, permitiendo la especialización funcional de diferentes componentes del sistema "detección de malware, análisis de tráfico de red, correlación de eventos", mientras mantiene capacidades de interoperabilidad y gestión unificada de recursos.[25 90 91]

Las arquitecturas híbridas de nube y Edge proporcionan flexibilidad para equilibrar los requisitos de latencia con las necesidades de potencia computacional, lo que permite el procesamiento local para decisiones urgentes, al tiempo que se aprovechan los recursos de la nube para tareas de análisis más complejas y que requieren muchos recursos.[81 92]

Modelos de Aprendizaje Automático Especializados

Deep Learning para Detección de Amenazas

Los modelos de DL especializados en ciberseguridad incorporan arquitecturas como las CNN para el análisis de patrones en el tráfico de red, las RNN y los transformadores para el análisis secuencial de registros y eventos, y las redes generativas adversarias (GAN por sus siglas en inglés), para generar datos de entrenamiento sintéticos.[87 93]

La implementación de mecanismos de atención en los modelos de transformadores permite el análisis contextual de secuencias de eventos de seguridad, identificando patrones complejos que pueden indicar actividades maliciosas coordinadas a través de múltiples vectores de ataque.[22 44] Estos modelos son especialmente eficaces para detectar APT que utilizan sofisticadas técnicas de evasión.

Métodos de Conjunto y Meta-aprendizaje

Los métodos de conjunto combinan múltiples algoritmos especializados para mejorar la solidez y la precisión en la detección de amenazas, utilizando técnicas como votación, bagging y boosting para agregar predicciones de diferentes modelos base.[61 94] La implementación del meta aprendizaje permite una rápida adaptación a nuevos tipos de amenazas con un mínimo de datos de entrenamiento adicionales.

La aplicación de marcos AutoML facilita la optimización automática de hiperparámetros y la selección de arquitectura, lo que reduce el tiempo necesario para el desarrollo y la implementación de nuevos modelos, al tiempo que mejora el rendimiento.[94 95] Esta automatización es particularmente valiosa en entornos CSIRT donde la velocidad de adaptación a nuevas amenazas es crítica.

Tecnologías de Procesamiento de Datos en Tiempo Real

Procesamiento de Flujos y Procesamiento de Eventos Complejos

La implementación de sistemas de procesamiento de flujos utilizando marcos como Apache Kafka, Apache Storm y Apache Flink permite el análisis en tiempo real de grandes volúmenes de datos de seguridad.[65 96] Estos sistemas deben gestionar un rendimiento de millones de eventos por segundo, manteniendo al mismo tiempo latencias inferiores a un segundo para la detección crítica.

El procesamiento de eventos complejos (CEP, por sus siglas en inglés), permite identificar patrones de eventos distribuidos temporalmente, correlacionando actividades en múltiples fuentes de datos para detectar campañas de ataques coordinados.[97 98] La integración de CEP con capacidades de aprendizaje automático permite el perfeccionamiento continuo de las reglas de correlación basadas en la información obtenida de incidentes confirmados.

Data Lakes y Arquitectura Lambda

Las arquitecturas de data lakes proporcionan un almacenamiento escalable y flexible para datos estructurados y no estructurados procedentes de múltiples fuentes, incluyendo registros del sistema, flujos de red, fuentes de inteligencia sobre amenazas y datos de detección y respuesta de los endpoint.[82 84] La implementación de la arquitectura Lambda permite el procesamiento paralelo por lotes y en flujo, optimizando tanto el análisis histórico como la respuesta en tiempo real.

La aplicación de técnicas específicas de gobernanza de datos para entornos de ciberseguridad incluye la clasificación automatizada de datos, políticas de retención diferenciadas según el tipo de amenaza y los requisitos normativos, y mecanismos de linaje de datos que facilitan las auditorías y las investigaciones forenses.[99 100]

Sistemas de Respuesta Automatizada e Inteligencia Aumentada

Los sistemas de respuesta automatizada han evolucionado hasta convertirse en modelos de orquestación, automatización y respuesta de seguridad (SOAR, por sus siglas en inglés) que integran capacidades de IA para la toma de decisiones autónoma. Respuesta guiada de Microsoft Copilot para la seguridad (CGR, por sus siglas en inglés) es un claro ejemplo de esta evolución, ya que proporciona arquitecturas de aprendizaje automático a escala industrial que guían a los analistas de seguridad a través de tres tareas críticas: investigación con contexto histórico, clasificación para determinar los falsos positivos y recomendaciones de respuesta automatizada.[25]

La integración de los modelos de lenguaje grandes (LLM, por sus siglas en inglés) en los sistemas de respuesta ha demostrado capacidades significativas para automatizar los procesos de planificación y revisión de la respuesta a incidentes.[42] Estos sistemas proporcionan interfaces conversacionales que permiten a los equipos de seguridad consultar bases de conocimiento complejas, generar informes automatizados y recibir recomendaciones contextualizadas para la mitigación de amenazas.[22]

La implementación del aprendizaje por refuerzo permite la optimización continua de las estrategias de respuesta basadas en los resultados de incidentes pasados. La coordinación de las acciones de respuesta entre múltiples herramientas de seguridad (cortafuegos, detección y respuesta en los puntos finales, sistemas de gestión de identidades) requiere de interfaz de programación de aplicaciones (API, por sus siglas en inglés) estandarizadas y mecanismos de gestión de errores que garanticen la ejecución coherente de las acciones de corrección.[98 101]

Tecnologías de Detección y Análisis Predictivo

Las tecnologías de detección contemporáneas emplean enfoques multimodales que combinan el análisis del comportamiento, la correlación de eventos y la predicción de amenazas.[42] Se han desarrollado sistemas de XAI para abordar el problema de la "caja negra" en los sistemas de detección automatizados, proporcionando justificaciones comprensibles para las alertas generadas y facilitando la confianza entre los analistas de seguridad.

HuntGPT representa una innovación significativa en la integración de detección de anomalías basada en ML con capacidades de explicabilidad a través de LLMs.[22] Este sistema combina algoritmos de detección automatizada con interfaces conversacionales que permiten a los equipos de respuesta interrogar el sistema sobre la lógica detrás de las alertas[102] y obtener recomendaciones específicas para la investigación y mitigación.

Estrategias de Respuesta Adaptativa

Los sistemas de respuesta adaptativa utilizan algoritmos de -multi-armed bandit- y - contextual bandits- para optimizar la selección de estrategias de respuesta en condiciones de incertidumbre.[1 20] Estos algoritmos equilibran la exploración de nuevas estrategias de respuesta con la explotación de enfoques que han demostrado su eficacia en escenarios similares.

La implementación de capacidades de aprendizaje continuo permite el perfeccionamiento de las estrategias de respuesta basándose en los comentarios de los analistas de seguridad y los resultados de los incidentes.[21 103] Esta adaptabilidad es crucial para la eficacia frente a las técnicas de ataque en constante evolución y los vectores de amenazas emergentes.

Análisis de Comportamiento y Detección de Anomalías

La implementación de análisis de comportamiento avanzados emplea algoritmos de aprendizaje no supervisado para establecer bases de referencia de comportamiento normal en el tráfico de red, la actividad de los usuarios y las operaciones del sistema.[44 93 104] Los modelos de detección de anomalías identifican desviaciones significativas que pueden indicar actividad maliciosa, utilizando técnicas como bosques aislados, máquina de vectores de soporte (SVM, por sus siglas en inglés) de una clase y autocodificadores -autoencoders-.

La aplicación del modelado de secuencias permite detectar patrones de ataque que se desarrollan durante largos periodos de tiempo, identificando la progresión del ataque en la cadena de ataque -cyber kill chain- y las actividades de movimiento lateral que los sistemas tradicionales basados en firmas pueden pasar por alto.[24 42 105] Estos sistemas demuestran una eficacia especial contra las amenazas internas y las APTs.

Fusión Automatizada de Inteligencia sobre Amenazas

Los sistemas automatizados de inteligencia sobre amenazas fusionan y correlacionan información procedente de múltiples fuentes, "fuentes comerciales de amenazas, inteligencia de código abierto, telemetría interna e intercambio entre pares" para generar evaluaciones de amenazas contextualizadas.[89 106 107] La implementación del procesamiento del lenguaje natural permite la extracción de IoC de informes de amenazas no estructurados y avisos de seguridad.

La aplicación de grafos de conocimiento para representar las relaciones entre los actores de amenazas, las técnicas, la infraestructura y los objetivos facilita las operaciones de consulta complejas y el razonamiento sobre los entornos de amenazas.[24 91] Esta representación estructurada permite realizar análisis sofisticados, como la predicción de la progresión de los ataques y la evaluación de las capacidades de los actores maliciosos.

Arquitecturas de Colaboración Humano-IA

El marco A2C (automatizado, aumentado y colaborativo) establece un modelo de varias etapas para la toma de decisiones colaborativa en equipos formados por humanos e inteligencia artificial.[103] Este marco admite tres modos de funcionamiento diferentes: automatizado para respuestas rutinarias y de bajo riesgo; aumentado para situaciones que requieren análisis humano con apoyo de IA; y colaborativo para escenarios complejos que exigen la exploración conjunta de soluciones.

La implementación de "cyber ranges" con asistentes de IA integrados, como CyberAlly, ha demostrado mejoras significativas en la eficiencia y eficacia de los equipos Blue Team durante los eventos de respuesta a incidentes.[24] Estos sistemas combinan grafos de conocimiento con capacidades de inteligencia artificial para proporcionar soporte contextual en tiempo real durante la investigación y respuesta ante amenazas.

Innovaciones en Tecnologías de Forense Digital

Las tecnologías forenses digitales han evolucionado hasta convertirse en modelos de forense digital como servicio (DFaaS, por sus siglas en inglés) que proporcionan capacidades escalables y automatizadas para la investigación de incidentes.[26] Estos sistemas implementan marcos de cinco fases basados en las normas ISO/IEC 27037 y NIST SP 800-101, lo que proporciona una trazabilidad sistemática de los procedimientos forenses y capacidades mejoradas para preservar la integridad de las pruebas digitales.

La integración de tecnologías blockchain para el intercambio seguro de información sobre ciberseguridad entre los CSIRTs representa una innovación significativa en la preservación de la confidencialidad, integridad y disponibilidad de los datos sobre amenazas.[27] Estas implementaciones utilizan Hyperledger Composer y el Sistema de Archivos Interplanetario (IPFS, por sus siglas en inglés) para crear redes descentralizadas de intercambio de información sobre amenazas.

Tabla 4 Ciberoperaciones Impulsadas por IA

CyberOps impulsadas por IA	%Porcentaje
Detección de anomalías y amenazas avanzadas (ej. día cero, malware polimórfico)	25
Predicción de amenazas y análisis predictivo	25
Búsqueda avanzada de amenazas (Threat Hunting)	19
Seguridad de punto final (ej. detección de comportamientos sospechosos)	13
Respuesta automatizada a incidentes (ej. bloqueo de acceso, aislamiento de dispositivos)	6
Prevención de phishing y fraude	6
SecOps unificadas (integración de SIEM, NDR, XDR)	6

Fuente: Elaboración propia.

La distribución de las capacidades impulsadas por IA en las funciones de CyberOps revela un panorama de adopción diverso pero desigual. A la vanguardia, la detección de anomalías y amenazas avanzadas y la predicción de amenazas/análisis predictivo representan cada uno el 25% de los casos de uso de IA. Esta paridad indica que los CSIRTs dan la misma importancia estratégica a la identificación de amenazas novedosas en tiempo real y a la anticipación de riesgos emergentes. La mitad combinada de las inversiones en IA en funciones de detección y predicción subraya una postura de defensa proactiva, que aprovecha el aprendizaje automático

para descubrir ataques sofisticados, como exploits de día cero y malware polimórfico, y pronosticar campañas de amenazas basadas en indicadores históricos y de comportamiento.

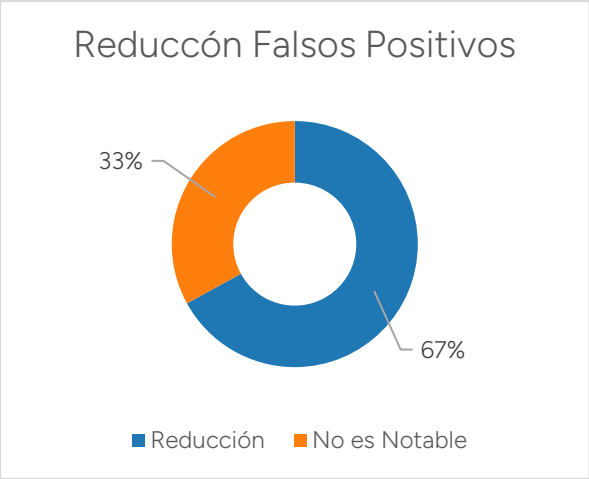
Tras estas funciones principales, la búsqueda avanzada de amenazas representa el 19% de las aplicaciones de IA. Esta cuota sustancial refleja la creciente confianza de los CSIRTs en el uso de la IA para examinar grandes volúmenes de datos de telemetría, descubrir actividades sigilosas de los adversarios y apoyar a los analistas humanos en investigaciones en profundidad. Aunque ligeramente inferior a la detección y la predicción, la asignación de casi una quinta parte demuestra que muchos equipos consideran la IA como un habilitador esencial para ampliar la búsqueda de amenazas dirigida por humanos a gran escala.

La categoría intermedia de seguridad de los puntos finales 13% destaca el papel de la IA en la supervisión de los comportamientos a nivel de los dispositivos. Al centrarse en la ejecución de procesos sospechosos, las conexiones de red anómalas y las modificaciones de archivos, la IA mejora las plataformas tradicionales de protección de los puntos finales para detectar el malware evasivo y las amenazas internas. El hecho de que la seguridad de los puntos finales represente una parte menor, pero significativa, del uso de la IA subraya que, aunque son críticas, las aplicaciones centradas en los dispositivos suelen depender de soluciones integradas en lugar del desarrollo de IA a medida.

En el extremo inferior del espectro de adopción se encuentran la respuesta automatizada a incidentes, la prevención del phishing y el fraude y la integración unificada de SecOps, cada una de las cuales representa el 6% del uso de la IA. Las funciones de respuesta automatizada, como el bloqueo dinámico del acceso o el aislamiento de dispositivos, ilustran los esfuerzos incipientes por cerrar el ciclo desde la detección hasta la mitigación sin intervención manual. Del mismo modo, la prevención del phishing y el fraude impulsada por la IA sigue siendo una capacidad emergente, probablemente obstaculizada por la complejidad del procesamiento del lenguaje natural y la necesidad de datos de entrenamiento de alta fidelidad. La integración unificada de SecOps (SIEM, NDR, XDR) también se sitúa en este nivel básico, lo que sugiere que la orquestación de todo el espectro de la telemetría de seguridad diversa a través de la IA se encuentra todavía en fases tempranas de adopción.

En general, los datos muestran una inversión prioritaria en funciones de detección y predicción, con un interés creciente por la búsqueda y la supervisión de terminales habilitadas por la IA. Los porcentajes relativamente más bajos en respuesta automatizada, prevención del phishing e integración unificada ponen de relieve las oportunidades de crecimiento, donde una mayor inversión, la mejora de los canales de datos y una mayor interoperabilidad entre herramientas podrían impulsar una adopción más amplia de la IA en todo el ciclo de vida de la respuesta a incidentes.

Gráfico 6 Perspectiva Operativa



La integración de la IA en las operaciones cibernéticas ha dado lugar a una reducción moderada de los falsos positivos en el 67% de los CSIRTs, lo que refleja la eficacia de la IA para perfeccionar los algoritmos de detección de amenazas y minimizar la fatiga por alertas.

Esta mejora permite a los analistas centrarse más en los incidentes reales, lo que reduce el esfuerzo innecesario y acelera los flujos de trabajo de investigación.

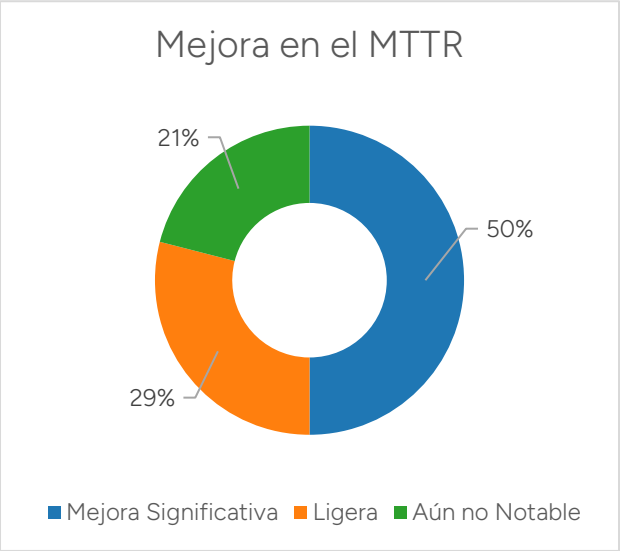
Fuente: Elaboración propia.

Por el contrario, el 33% de los equipos no ha observado cambios notables en las tasas de falsos positivos, lo que indica posibles deficiencias en el entrenamiento de los modelos, problemas de calidad de los datos o configuraciones de umbrales desajustadas que requieren esfuerzos de optimización específicos.

Gráfico 7 Perspectiva de Mejora en Capacidad del MTTR

La influencia de la IA en el tiempo medio de respuesta (MTTR, por sus siglas en inglés) es igualmente pronunciada. Más de la mitad de los CSIRTs 50,4% informan de una reducción significativa del MTTR de más del cincuenta por ciento, lo que ejemplifica la capacidad de la IA para automatizar la clasificación inicial, priorizar las alertas críticas y recomendar medidas de respuesta.

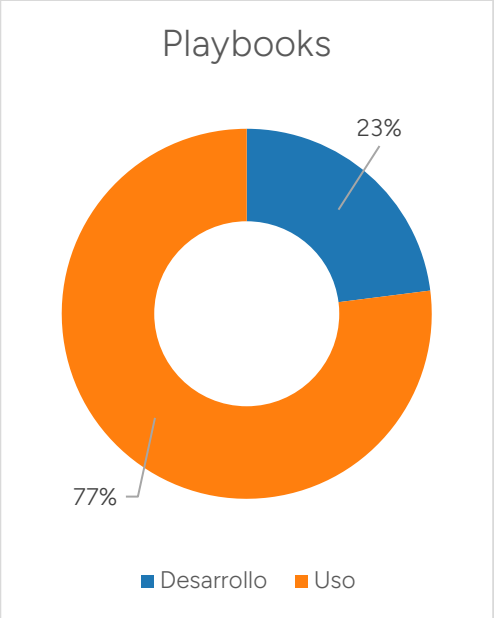
Otro 2 % experimenta una ligera mejora del MTTR de hasta el cincuenta por ciento, lo que demuestra que las mejoras incrementales de la IA pueden seguir produciendo ganancias significativas en términos de eficiencia.



Fuente: Elaboración propia.

Sin embargo, el 21% no observa ningún cambio notable en el MTTR, lo que sugiere que los beneficios de la IA pueden verse limitados por los retos de integración, los flujos de trabajo de toma de decisiones o la confianza de los analistas en las recomendaciones basadas en la IA.

Gráfico 8 Playbooks en las CyberOps IA



Fuente: Elaboración propia.

La prevalencia de playbooks automatizados basados en IA en la respuesta a incidentes subraya aún más el impacto operativo de la IA. El 78,9% de los CSIRTs emplean actualmente manuales automatizados, incorporando la lógica de decisión basada en IA en procedimientos de respuesta estandarizados y garantizando acciones coherentes y rápidas durante los incidentes de seguridad.

El 21,1% restante se encuentra en la fase de desarrollo de guías automatizadas, lo que indica que se están realizando esfuerzos activos para codificar las mejores prácticas y pasar de flujos de trabajo manuales a flujos de trabajo aumentados por IA. Este patrón de adopción pone de relieve una curva de maduración regional en la que la mayoría de los equipos han pasado de la experimentación al uso operativo, mientras que un subconjunto continúa pilotando y perfeccionando la automatización de las guías.

2.2 Sostenibilidad e Innovación en el Ciclo de Vida del desarrollo del Desarrollo de Soluciones de IA

Los Objetivos de Desarrollo Sostenible (ODS) establecidos por las Naciones Unidas en 2015 apuntan a garantizar una mayor prosperidad, al tiempo que buscan minimizar los impactos adversos del cambio climático sobre los recursos naturales existentes. Sin embargo, se espera que las tecnologías de producción y el comportamiento de consumo actuales generen resultados positivos solamente hasta cierto punto más allá del cual utilizar el capital natural acarrea consecuencias negativas para el crecimiento y el bienestar común.[108]

Grazzi, Sasso y Kemp definen la innovación verde (también conocida como eco-innovación, innovación medioambiental o innovación sostenible) se define como los bienes y servicios, procesos, métodos de marketing, estructuras organizativas y arreglos institucionales nuevos o significativamente mejorados que, con o sin intención, conducen a mejoras ambientales en comparación con las alternativas relevantes.[109]

En ALC, sostenibilidad medioambiental surge como un tema destacado, lo que refleja la creciente conciencia sobre el impacto del consumo energético de la IA. Esto representa una consideración novedosa en la implementación de la IA en la ciberseguridad en los CSIRTs:

1. Impacto medioambiental de los requisitos computacionales.
2. Sostenibilidad económica de las inversiones a largo plazo en IA.
3. Sostenibilidad social a través del acceso inclusivo y las consideraciones sobre la brecha digital.
4. Sostenibilidad de la innovación a través del desarrollo de capacidades locales.

En otras palabras, la innovación verde incluye la creación y comercialización de nuevas tecnologías que son más benignas para el medio ambiente, así como la difusión y adopción de tecnologías más ecológicas por parte de las empresas. Dichas acciones cumplirán un rol crucial en el logro de las metas climáticas de los países de nuestra región y también crearán sendas de crecimiento y generación de nuevos empleos significativos.[110]

Aunque la innovación y la eficiencia son objetivos esenciales del avance tecnológico, debe darse prioridad al acceso equitativo. El desarrollo tecnológico no debe servir únicamente a objetivos funcionales o económicos, sino que debe promover activamente valores sociales fundamentales como la integridad, la tolerancia y la diversidad. Es fundamental evitar reforzar los modelos de consumo excesivo que contribuyen a la degradación del medio ambiente, incluidos los residuos, la contaminación y el cambio climático.[34]

Desde una perspectiva socio-organizativa, la implantación generalizada de la tecnología no debe conducir a una bifurcación en la calidad del servicio, en la que las soluciones impersonales y automatizadas se dirijan a las poblaciones desfavorecidas, mientras que los servicios individualizados y centrados en el ser humano se reserven a quienes disponen de mayores medios económicos. Por el contrario, la tecnología debe aprovecharse para reducir las desigualdades, no para profundizarlas.

El desarrollo de soluciones de IA sostenibles para aplicaciones de ciberseguridad requiere enfoques integrales de gestión del ciclo de vida que optimicen la utilización de los recursos, minimicen el impacto medioambiental y garanticen la mantenibilidad del sistema a largo plazo, al tiempo que proporcionan capacidades de seguridad eficaces.

Las prácticas de desarrollo sostenible de la IA abarcan el diseño de algoritmos eficientes desde el punto de vista energético, arquitecturas computacionales optimizadas y principios de economía circular que dan prioridad a la reutilización de los recursos y a la longevidad del sistema frente a las ganancias de rendimiento a corto plazo. Estos enfoques reconocen que la intensidad computacional de los sistemas de IA puede generar importantes costes medioambientales, lo que requiere metodologías de desarrollo que equilibren la eficacia de la seguridad con los objetivos de sostenibilidad mediante técnicas inteligentes de gestión y optimización de recursos.

En ALC, los expertos hacen hincapié en que, si bien la IA puede optimizar los recursos, los **altos requisitos de consumo de energía** crean retos de sostenibilidad que requieren **enfoques equilibrados.**

La innovación en el desarrollo del ciclo de vida de las soluciones de IA para aplicaciones de ciberseguridad se centra en la creación de sistemas adaptables y autocorrectivos que puedan evolucionar y optimizar su rendimiento a lo largo del tiempo sin necesidad de sustituir completamente el sistema ni realizar una reconfiguración manual exhaustiva.

Los marcos de desarrollo de IA modernos incorporan mecanismos de aprendizaje continuo, procedimientos de actualización automática de modelos y parámetros de autoajuste que permiten a los sistemas adaptarse a los cambiantes entornos de amenazas y requisitos operativos. Este enfoque de la innovación hace hincapié en el desarrollo de soluciones de IA resilientes y duraderas que puedan mantener su eficacia durante largos periodos operativos, minimizando al mismo tiempo la necesidad de ciclos de rediseño o sustitución que requieren muchos recursos.

La innovación sostenible en soluciones de ciberseguridad basadas en IA también requiere el desarrollo de arquitecturas de sistemas modulares e interoperables que permitan la reutilización de componentes, la extensibilidad del sistema y la mejora gradual de las capacidades, en lugar de la sustitución total del sistema.

Este enfoque arquitectónico permite a los CSIRTs invertir en una infraestructura de IA que puede evolucionar y adaptarse con el tiempo, reduciendo la necesidad de costosas renovaciones completas del sistema y manteniendo al mismo tiempo capacidades de seguridad de vanguardia. El paradigma de desarrollo resultante da prioridad a la creación de soluciones de IA flexibles y adaptables que puedan crecer y evolucionar con las necesidades que van demandando los CSIRTs y los avances tecnológicos, maximizando así el valor a largo plazo y la sostenibilidad de las inversiones en IA en aplicaciones de ciberseguridad.

2.3 Desarrollo Centrado en la Ciberresiliencia: IA como Herramienta de Anticipación y Mitigación de Amenazas.

Teniendo en cuenta que el ciberespacio es un ámbito de oportunidades en términos de progreso tecnológico, conectividad, integración y conexión global a Internet. Pero también es terreno de amenazas y riesgos. Los ciberataques pueden dañar a las organizaciones e infligirles importantes daños económicos y de imagen. Para que una organización esté preparada para defenderse de las amenazas cibernéticas, debe dominar una gran cantidad de especializaciones: tecnológicas, organizativas y de procesos.[111]

Los enfoques de desarrollo centrados en la resiliencia para los sistemas de ciberseguridad basados en IA dan prioridad a la creación de mecanismos de defensa robustos y adaptables que puedan anticipar, resistir y recuperarse de ciberataques sofisticados, al tiempo que mantienen la continuidad operativa durante todo el ciclo de vida del incidente.

En ALC, los expertos destacan que el cambio **de los ataques de interrupción del servicio a los ataques de manipulación social y control del discurso**, especialmente relevantes durante los **procesos electorales**. Esto representa un cambio fundamental en el panorama

de las amenazas que requiere nuevas capacidades de respuesta de la IA, lo que supone replantear las capàcidades de los CSIRTs.

Este paradigma de desarrollo hace hincapié en la creación de sistemas de IA que presenten capacidades de autorreparación, mecanismos de respuesta adaptables y características de degradación gradual que permitan el funcionamiento continuo incluso en condiciones adversas. Los sistemas de IA centrados en la resiliencia incorporan múltiples capas de redundancia, mecanismos de conmutación por error y modos operativos alternativos que garantizan que las funciones de seguridad críticas sigan estando disponibles incluso cuando los sistemas primarios se ven comprometidos o son objeto de un ataque.

El desarrollo de sistemas de anticipación de amenazas basados en IA requiere sofisticadas capacidades de modelización predictiva que puedan identificar patrones de amenazas emergentes, pronosticar la progresión de los ataques y estimar posibles escenarios de impacto en múltiples vectores de amenaza simultáneamente.

Por lo cual, estos sistemas aprovechan técnicas avanzadas de aprendizaje automático, como el análisis de patrones temporales, la modelización del comportamiento y la simulación de escenarios, para proporcionar capacidades de alerta temprana que permitan adoptar medidas de defensa proactivas. Los sistemas de anticipación de amenazas también deben incorporar mecanismos de cuantificación de la incertidumbre que puedan evaluar la fiabilidad de las predicciones y proporcionar intervalos de confianza para las evaluaciones de amenazas, lo que permite a los equipos de seguridad tomar decisiones informadas basadas en inteligencia de amenazas probabilística en lugar de determinista.

La dependencia de la infraestructura crítica de la IA crea nuevas categorías de vulnerabilidad que los enfoques tradicionales de ciberseguridad no pueden abordar.

Los sistemas de mitigación de amenazas basados en la inteligencia artificial representan un componente crítico del desarrollo centrado en la resiliencia, ya que proporcionan capacidades de respuesta automatizadas que pueden ejecutar contramedidas a la velocidad de una máquina, al tiempo que se coordinan con los operadores humanos para garantizar una supervisión y un control adecuados.

Estos sistemas incorporan algoritmos inteligentes de toma de decisiones que pueden evaluar la gravedad de las amenazas, valorar las posibles opciones de respuesta y ejecutar estrategias de mitigación óptimas basadas en la inteligencia sobre amenazas en tiempo real y la tolerancia al riesgo de la organización. El desarrollo de sistemas de mitigación eficaces requiere una comprensión sofisticada de los mecanismos de propagación de las amenazas, las interdependencias del sistema y los modos de fallo en cascada para garantizar que las respuestas automatizadas mejoren, en lugar de comprometer, la resiliencia general del sistema.

La integración de capacidades de anticipación y mitigación en sistemas de IA centrados en la resiliencia requiere el desarrollo de bucles de retroalimentación y mecanismos de aprendizaje integrales que permitan la mejora continua de las capacidades de detección, predicción y respuesta ante amenazas basadas en la experiencia operativa.

Finalmente, sistemas de aprendizaje deben incorporar tanto el aprendizaje supervisado a partir de datos de incidentes etiquetados como el aprendizaje por refuerzo a partir de los resultados operativos para optimizar el rendimiento del sistema a lo largo del tiempo. El enfoque de desarrollo también debe abordar el reto del aprendizaje adversario, en el que los actores maliciosos pueden intentar manipular los sistemas de IA mediante entradas cuidadosamente elaboradas o el envenenamiento de los datos de entrenamiento, lo que requiere defensas robustas contra adversarios adaptables que pueden atacar específicamente los sistemas de seguridad basados en IA.

3. Adopción Responsable y Adaptativa de la IA en los CSIRTs

La adopción responsable y adaptativa de la IA en los CSIRTs implica más que una simple integración técnica; representa una transformación organizativa que afecta a la cultura, las competencias, el liderazgo y la confianza compartida tanto en las personas como en las máquinas. Un factor crítico para el éxito es fomentar una cultura abierta a la experimentación, la transparencia y el aprendizaje continuo. Los equipos CSIRTs que adoptan el cambio de forma proactiva, invierten en formación y fomentan la colaboración interfuncional están mejor preparados para integrar la IA como una parte fundamental de su arsenal defensivo, en lugar de como una imposición externa disruptiva.

En ALC, los expertos señalan factores de exitos que son clave en la adopción de la IA en los CSIRTs:

1. Implementacion gradual comenzando con casos de uso de bajo riesgo.
2. Programas integrales de gestión del cambio, formación y desarrollo de capacidades y competencias.
3. Transformación cultural que aborde las preocupaciones de los analistas sobre la pérdida de puestos de trabajo.
4. Demostración clara del valor de la IA a través de resultados medibles.

El desarrollo de competencias es fundamental. Además del dominio técnico de la IA y los métodos de ciencia de datos, el personal de los CSIRTs debe cultivar nuevas habilidades: comprender las limitaciones de los sistemas de IA, supervisar los sesgos y los fallos, y dominar los protocolos de colaboración entre humanos y máquinas. El liderazgo debe comunicar una visión clara para la adopción de la IA, establecer políticas para su uso ético, garantizar estructuras de responsabilidad e incentivar el comportamiento responsable. Esto incluye refinar las definiciones de roles, aclarando cuándo deben intervenir los humanos, cómo deben interpretarse los resultados de la IA y qué cadena de mando debe seguirse durante incidentes complejos.

La adopción adaptativa también se basa en mecanismos sólidos de evaluación de riesgos y gobernanza adaptados a entornos dinámicos y de alto impacto. Las organizaciones deben evaluar sistemáticamente tanto las promesas como los peligros de la IA, desde la fiabilidad y la seguridad hasta las posibles interrupciones operativas y las faltas éticas.

La toma de decisiones debe basarse en pruebas, respaldarse en una planificación exhaustiva de escenarios y adaptarse a los distintos grados de automatización y complejidad en las diferentes etapas de la respuesta a incidentes. Las estrategias de mitigación de riesgos deben incorporar protocolos de respaldo, supervisión constante y bucles de retroalimentación iterativos para garantizar que la adopción de la IA mejore, en lugar de poner en peligro, los resultados de las misiones de seguridad.

Además, la resiliencia y la adaptabilidad de los CSIRTs habilitados para la IA exigen un aprendizaje organizativo continuo. La retroalimentación del rendimiento en el mundo real, incluidos los incidentes de fallo de la IA, los falsos positivos o los ataques adversarios, debe codificarse y utilizarse para actualizar las políticas, volver a entrenar los modelos e informar las futuras evaluaciones de riesgos.

Los CSIRTs deben considerar la IA no como una herramienta estática, sino como un sistema vivo que evoluciona en colaboración con analistas humanos y actores maliciosos por igual. Las organizaciones que tengan éxito serán aquellas que combinen la innovación técnica con una gobernanza disciplinada, un diseño centrado en el ser humano y un compromiso inquebrantable con la mejora en medio del cambio y la incertidumbre.

Por último, el Consejo de Inteligencia Artificial de la OCDE insta basar la adopción de IA en los principios[37] en aras de promover el desarrollo sostenible y el bienestar humano debe ser un objetivo fundamental en el diseño, el desarrollo y la implementación de la IA. Esto implica respetar el estado de derecho, los derechos humanos y los valores democráticos, incluidos la equidad, la inclusión y el derecho a la privacidad. Los sistemas de IA deben desarrollarse con un fuerte énfasis en la transparencia, la explicabilidad, la solidez, la seguridad y la protección. Además, es fundamental garantizar la rendición de cuentas a lo largo de todo el ciclo de vida de la IA.

Para lograr una IA fiable y beneficiosa, es esencial:

- Establecer políticas nacionales coherentes y promover la cooperación internacional con el fin de crear una IA fiable.
- Aumentar la inversión en investigación y desarrollo de IA para fomentar la innovación y abordar las necesidades de la sociedad.
- Fomentar el desarrollo de un ecosistema inclusivo que promueva la adopción responsable y equitativa de la IA.
- Crear un entorno político y de gobernanza propicio que sea interoperable, con visión de futuro y alineado con las normas internacionales.
- Fortalecer las capacidades humanas y preparar a las sociedades para la transformación del mercado laboral que traerá consigo la IA.
- Reforzar la colaboración internacional para garantizar valores compartidos, normas comunes y confianza mutua en el desarrollo y el uso de la IA.

3.1 Factores Críticos de Éxito para la Adopción de la IA en CSIRT: Cultura, Competencias y Liderazgo

La adopción exitosa de tecnologías de IA en los equipos de respuesta a incidentes de seguridad informática depende fundamentalmente del fomento de culturas organizativas que adopten la innovación, el aprendizaje continuo y los enfoques adaptativos para la resolución de problemas, al tiempo que se mantienen los rigurosos estándares operativos esenciales para unas operaciones de ciberseguridad eficaces.[112]

La cultura organizativa actúa como un facilitador o una barrera fundamental para la adopción de la IA, y las investigaciones indican que las culturas orientadas a la innovación son significativamente más propensas a integrar con éxito las tecnologías de IA que aquellas que se resisten al cambio tecnológico. Por lo tanto, los CSIRTs deben dar prioridad a las iniciativas de transformación cultural que promuevan la experimentación con tecnologías emergentes, la aceptación de riesgos calculados y los enfoques colaborativos para la resolución de problemas que aprovechen tanto la experiencia humana como las capacidades de la IA para lograr resultados de seguridad superiores.

Las estrategias de liderazgo desempeñan un papel fundamental en el impulso de la adopción exitosa de la IA dentro de los CSIRTs, lo que requiere líderes que puedan articular visiones claras para la integración de la IA, proporcionar los recursos necesarios para la adopción de la tecnología y fomentar entornos que apoyen tanto la innovación técnica como la excelencia operativa.

El liderazgo eficaz en los CSIRTs habilitados para la IA debe equilibrar los enfoques transformadores que inspiran el cambio organizativo con consideraciones prácticas relacionadas con la continuidad operativa, la gestión de riesgos y el cumplimiento normativo. Los líderes

también deben demostrar su compromiso con las prácticas responsables de IA mediante el establecimiento de directrices éticas, la garantía de la transparencia en los procesos de toma de decisiones de IA y el mantenimiento de mecanismos de supervisión humana adecuados durante la implementación y el funcionamiento del sistema de IA.

El desarrollo de competencias y el fortalecimiento de capacidades es esencial para la adopción de la IA en los CSIRTs abarca tanto habilidades técnicas relacionadas con el funcionamiento y el mantenimiento de los sistemas de IA como capacidades complementarias, entre las que se incluyen el análisis de datos, la integración de sistemas y las técnicas de colaboración entre humanos e IA.

Este desarrollo técnicas debe abordar múltiples ámbitos, entre los que se incluyen los algoritmos de aprendizaje automático, el preprocesamiento de datos y la ingeniería de características, la validación y prueba de modelos, y la supervisión y el mantenimiento de sistemas de IA. Además, el personal de los CSIRTs debe desarrollar una comprensión sofisticada de las limitaciones de los sistemas de IA, los posibles modos de fallo y los mecanismos de supervisión humana adecuados para garantizar la integración eficaz de las capacidades de IA en los procedimientos operativos y los procesos de toma de decisiones existentes.

Las consideraciones sobre los factores humanos en la adopción de la IA van más allá del desarrollo de competencias técnicas y abarcan capacidades organizativas más amplias, como la gestión del cambio, la comunicación con las partes interesadas y los enfoques colaborativos para la resolución de problemas que aprovechan tanto la intuición humana como las capacidades analíticas de la IA.

> Hay una crisis de talento, la cual revela que revela que **los CSIRTs nacionales operan con solo 7-8 empleados,** lo que proporciona una prueba concreta de la crisis de recursos humanos que limita fundamentalmente las capacidades de integración de la IA en toda ALC.

Los recursos financieros desempeñan un rol fundamental en las operaciones del CSIRT, por lo que una financiación adecuada es esencial no solo para adquirir herramientas e infraestructura avanzadas de IA, como computación de alto rendimiento, almacenamiento de datos y plataformas de análisis, sino también para mantener el desarrollo, la integración y el mantenimiento continuos.[6]

Además, la inversión financiera permite la contratación y retención de talento especializado, apoya las iniciativas de mejora de las competencias y permite la investigación y la experimentación continuas, fundamentales para adaptar los modelos de IA a los cambiantes escenarios de amenazas.

Una financiación insuficiente o inconsistente puede dar lugar a una implementación fragmentada, a una reducción de la eficacia operativa y a una dependencia excesiva de tecnologías obsoletas, lo que, en última instancia, socava la capacidad del CSIRT para responder de forma proactiva e inteligente a las amenazas cibernéticas. Por lo tanto, alinear la planificación financiera con los objetivos estratégicos de IA es un factor fundamental para la resiliencia y la innovación en las operaciones de ciberdefensa.

Muy presente es el fenómeno de la fuga de cerebros es un reto importante que debe tenerse en cuenta en el contexto de la adopción de la IA en los entornos CSIRT. A medida que la demanda de profesionales cualificados en inteligencia artificial y ciberseguridad sigue creciendo a nivel mundial, muchos expertos altamente capacitados abandonan los CSIRTs nacionales o del sector público en busca de oportunidades mejor remuneradas en el sector privado, en organizaciones internacionales o en países más avanzados tecnológicamente. Esta migración de talento puede debilitar gravemente la eficacia operativa y la capacidad estratégica de los CSIRTs, especialmente en regiones con recursos limitados o ecosistemas digitales poco desarrollados.

La implementación de la IA intensifica este desafío, ya que requiere una combinación poco común de conocimientos especializados en ciencia de datos, aprendizaje automático, ciberseguridad y respuesta a incidentes. La competencia por este talento interdisciplinario es feroz y, sin los incentivos adecuados, ya sean económicos, profesionales o intelectuales, muchas personas cualificadas pueden buscar entornos que ofrezcan más innovación, estabilidad o prestigio. Como resultado, los CSIRTs pueden experimentar una escasez de personal capaz de desarrollar, gestionar y auditar sistemas de IA, lo que puede retrasar o comprometer el despliegue de capacidades críticas impulsadas por la IA.

Además, la fuga de cerebros puede socavar la memoria institucional y la continuidad, especialmente cuando el personal con experiencia se marcha sin un plan sostenible de sucesión o transferencia de conocimientos. Esto no solo ralentiza el progreso, sino que también aumenta la dependencia de proveedores o consultores externos, lo que puede generar riesgos adicionales relacionados con la dependencia de proveedores, la soberanía de los datos y la pérdida de autonomía estratégica. Para contrarrestar la fuga de cerebros, las organizaciones deben invertir no solo en una remuneración competitiva, sino también en la creación de una cultura atractiva impulsada por la misión, ofreciendo vías de desarrollo profesional y fomentando la colaboración internacional que permita a los expertos locales crecer sin abandonar sus funciones institucionales.

En última instancia, retener el talento es tan importante como adquirirlo. En el contexto de la adopción de la IA por parte de los CSIRTs, abordar las causas fundamentales de la fuga de cerebros debe considerarse una prioridad estratégica, integrada en una planificación más amplia del desarrollo de la fuerza laboral y la resiliencia organizativa.

La cooperación, la colaboración, la coordinación y la gobernanza eficaces son pilares fundamentales para la adopción exitosa de la IA en las operaciones de los CSIRTs. La complejidad de las amenazas cibernéticas y la rápida evolución de las tecnologías de IA exigen un enfoque multilateral que rompa los silos entre los equipos técnicos, la dirección, los responsables políticos, el mundo académico y los socios industriales. Los mecanismos de colaboración, como las iniciativas de investigación conjunta, las plataformas de intercambio de información, los programas de formación intersectoriales y las asociaciones público-privadas, mejoran la conciencia situacional colectiva y aceleran la innovación, al tiempo que fomentan la confianza y la interoperabilidad.

Al mismo tiempo, se necesitan marcos de coordinación sólidos para armonizar los esfuerzos de integración de la IA en todos los niveles de la organización, garantizando la coherencia en la gestión de riesgos, la gobernanza de datos y los protocolos de respuesta a incidentes. La gobernanza estratégica proporciona la estructura general que guía estos esfuerzos, estableciendo funciones claras, responsabilidades, límites éticos y el cumplimiento de las normas reglamentarias. En conjunto, estos elementos crean un ecosistema resiliente en el que los CSIRTs pueden adoptar las tecnologías de IA de forma responsable, eficiente y con una visión estratégica compartida.

Para que la adopción de la IA sea satisfactoria, es necesario abordar la posible resistencia al cambio tecnológico, las preocupaciones sobre la pérdida de puestos de trabajo o la modificación de funciones, y la necesidad de aprendizaje y adaptación continuos a medida que evolucionan las tecnologías de IA.[113] Por lo tanto, se deben implementar estrategias integrales de gestión del cambio que aborden tanto las dimensiones técnicas como humanas de la adopción de la IA, asegurándose de que el personal comprenda la propuesta de valor de la integración de la IA y mantenga la confianza en su importancia continua dentro de los entornos operativos mejorados por la IA.

3.2 Evaluación de Riesgos y Toma de Decisiones en la Adopción de IA Ciber defensiva

La adopción de tecnologías de IA en las ciberoperaciones requiere metodologías exhaustivas de evaluación de riesgos que puedan evaluar tanto los beneficios potenciales como los riesgos inherentes asociados al despliegue de sistemas de IA en entornos de seguridad críticos.

Los marcos de evaluación de riesgos para la adopción de la IA deben abordar múltiples categorías de riesgos potenciales, incluidos los riesgos técnicos relacionados con la fiabilidad y el rendimiento del sistema, los riesgos operativos relacionados con la integración con los procesos y procedimientos existentes, y los riesgos estratégicos asociados a la dependencia de los sistemas de IA para funciones de seguridad críticas.

> Los expertos plantean que el sesgo algorítmico refleja la realidad social en los datos, lo que requiere soluciones sistemáticas a nivel de datos en lugar de correcciones a posteriori. El sistema solicita programación para evitar conductas discriminatorias, lo que proporciona enfoques técnicos de mitigación.

Estas evaluaciones también deben tener en cuenta la naturaleza dinámica tanto de las amenazas cibernéticas como de las tecnologías de IA, incorporando mecanismos para la supervisión y reevaluación continuas de los riesgos a medida que los sistemas evolucionan y cambian los panoramas de amenazas.

Los marcos de toma de decisiones para la adopción de la IA en contextos de ciberseguridad deben equilibrar los importantes beneficios potenciales de las tecnologías de IA, incluidas las capacidades mejoradas de detección de amenazas y los tiempos de respuesta optimizados, con los riesgos e incertidumbres inherentes asociados a las tecnologías emergentes.[114]

Una toma de decisiones eficaz requiere una comprensión sofisticada de las capacidades y limitaciones de los sistemas de IA, un análisis exhaustivo de la preparación de la organización para la adopción de la IA y una consideración cuidadosa de los requisitos de recursos y los retos de implementación.[114] [115] [116] Los responsables de la toma de decisiones también deben considerar las implicaciones competitivas de la adopción de la IA, reconociendo que las organizaciones que no adopten tecnologías de IA pueden enfrentarse a desventajas en su capacidad para hacer frente a ciberamenazas sofisticadas habilitadas por la IA.

La integración de las tecnologías de IA en las operaciones de ciberseguridad existentes requiere una cuidadosa consideración de la compatibilidad del sistema, la calidad y disponibilidad de los datos, y los requisitos de interacción entre humanos y máquinas que pueden afectar significativamente al éxito de la implementación.

Los retos de integración técnica incluyen garantizar la compatibilidad entre los sistemas de IA y la infraestructura heredada, abordar los requisitos de formato y calidad de los datos para un funcionamiento eficaz de la IA, y establecer interfaces adecuadas entre los sistemas de IA y los operadores humanos. Las organizaciones también deben considerar las implicaciones de la adopción de la IA para los procedimientos operativos existentes, las funciones y responsabilidades del personal y las estructuras organizativas, asegurándose de que la integración de la IA mejore, en lugar de perturbar, las operaciones de ciberseguridad eficaces.[116]

Las estrategias de mitigación de riesgos para la adopción de la IA en entornos de ciberseguridad deben abordar tanto los riesgos inmediatos de implementación como las consideraciones operativas a largo plazo, incluyendo el mantenimiento del sistema, la supervisión del rendimiento y la adaptación a los requisitos cambiantes.

Una mitigación eficaz de los riesgos requiere el establecimiento de procedimientos sólidos de prueba y validación para los sistemas de IA, la implementación de mecanismos integrales de supervisión y alerta para el rendimiento de los sistemas de IA y el mantenimiento de

procedimientos de respaldo que garanticen la continuidad operativa en caso de fallos del sistema de IA. Los CSIRTs también deben desarrollar estrategias para gestionar la naturaleza cambiante de las tecnologías de innovación emergentes y disruptivas como es el caso de la IA, incluidos procedimientos para la actualización de los sistemas, la optimización del rendimiento y la adaptación a los cambiantes escenarios de amenazas y requisitos operativos.

3.3 Modelos Adaptativos y Resilientes de Operaciones Basados en IA para Entornos CSIRTs

El desarrollo de modelos operativos adaptativos para entornos CSIRT habilitados para IA requiere arquitecturas de sistemas sofisticadas que puedan ajustar dinámicamente sus parámetros operativos, estrategias de respuesta y asignación de recursos basándose en información sobre amenazas en tiempo real y retroalimentación operativa.

En ALC, los expertos señalan que los modelos de adaptativos y resilientes basada en el ecosistema IA surge como un marco novedoso, que va más allá de la resiliencia de los sistemas individuales de IA para abarcar enfoques organizativos y regionales integrales. Esto representa una evolución significativa con respecto al pensamiento tradicional en seguridad y ciberseguridad.

Los sistemas adaptativos deben incorporar múltiples modos operativos que puedan seleccionarse de forma automática o manual en función de la gravedad de la amenaza, la complejidad del incidente y la disponibilidad de recursos, garantizando un rendimiento óptimo en diversos escenarios operativos.[117] Estos modelos operativos también deben abordar el reto de mantener la coherencia y la previsibilidad en el comportamiento del sistema, al tiempo que permiten una adaptación dinámica a las condiciones cambiantes, lo que requiere un sofisticado equilibrio entre la estabilidad y la flexibilidad en el diseño y la implementación del sistema.

Los modelos operativos resilientes para los sistemas CSIRT basados en IA deben incorporar mecanismos integrales de redundancia, conmutación por error y recuperación que garanticen el funcionamiento continuo incluso en condiciones adversas, como fallos del sistema, ciberataques dirigidos a los propios sistemas de IA y limitaciones de recursos que puedan restringir las capacidades del sistema.

La resiliencia requiere enfoques multicapa que aborden los posibles modos de fallo en diferentes niveles del sistema, desde fallos de algoritmos individuales hasta compromisos completos del sistema, garantizando que las funciones de seguridad críticas sigan estando disponibles en diversos escenarios de fallo.[110] Estos modelos también deben incorporar capacidades de degradación gradual que permitan a los sistemas mantener las funciones esenciales incluso cuando funcionan a capacidad reducida, proporcionando continuidad a las operaciones de seguridad críticas en condiciones difíciles.

Se identifican las siguientes capacidades de resiliencia de la aplicacion de la IA en los CSIRTs en ALC:

1. Detección proactiva de amenazas con adaptación continua.
2. Estrategias de defensa multicapa con integración de IA.
3. Capacidades de autorreparación y autorremediación.
4. Cooperación regional para la respuesta colectiva a las amenazas.

La integración de mecanismos de supervisión y control humanos en modelos operativos de IA adaptables y resilientes representa un requisito fundamental para mantener una responsabilidad y un control adecuados sobre las operaciones de seguridad impulsadas por la IA.

Los marcos de colaboración entre humanos e IA deben establecer una delimitación clara de las responsabilidades entre los sistemas automatizados y los operadores humanos, garantizando que las decisiones críticas sigan estando sujetas a una supervisión humana adecuada, al tiempo que se permite a los sistemas de IA proporcionar capacidades analíticas mejoradas y tiempos de respuesta rápidos.[118] [119] Estos marcos también deben abordar el reto de mantener la conciencia situacional y la capacidad de toma de decisiones humanas en entornos en los que los sistemas de IA gestionan una proporción cada vez mayor de tareas operativas rutinarias.

En ALC, los expertos proyectan que la la IA debe ser considerada un **"multiplicador de capacidad"** para el procesamiento de la información y la aceleración del análisis, en lugar de un sustituto del ser humano, lo que permite **al personal del CSIRT centrarse en el análisis profundo** mientras la IA se encarga del procesamiento de grandes volúmenes.

La mejora continua y la capacidad de aprendizaje son componentes esenciales de los modelos operativos adaptables y resilientes, ya que permiten a los sistemas de IA evolucionar y optimizar su rendimiento basándose en la experiencia operativa y en los cambios en el panorama de amenazas. Los mecanismos de aprendizaje deben incorporar tanto la adaptación automatizada basada en métricas de rendimiento del sistema como la incorporación estructurada de la retroalimentación humana y la experiencia en el ámbito para garantizar que los sistemas de IA se desarrollen en consonancia con los objetivos organizativos y los requisitos operativos.

Se identifican las siguientes capacidades de adaptabilidad para la aplicacion de la IA en los CSIRTs en ALC:

1. Aprendizaje continuo a partir de los entornos de amenazas en evolución.
2. Capacidades de ajuste dinámico de políticas y normas.
3. Respuestas de seguridad sensibles al contexto.
4. Integración de inteligencia sobre amenazas en tiempo real.

El modelo operativo también debe abordar el reto de gestionar la evolución del sistema manteniendo la estabilidad y la previsibilidad en las operaciones de seguridad críticas, lo que requiere enfoques sofisticados para la mejora incremental y la validación de las modificaciones del sistema.

En ALC, los expertos proponen la **Infraestructura pública digital (DPA) a nivel regional** para soluciones de IA compartidas, **distribuyendo la carga de innovación entre los países** para la reducción dinámica de costes y la mejora de la capacidad.

Finalmente, el marco presentado en esta investigación demuestra que la integración exitosa de las tecnologías de IA en las operaciones del CSIRT requiere una atención coordinada a los factores de gobernanza, desarrollo tecnológico y adopción organizativa. Las pruebas indican que los modelos de respuesta de seguridad sostenibles, resilientes, responsables y adaptables surgen a través de la alineación estratégica de marcos éticos, soluciones técnicas innovadoras y enfoques de implementación centrados en las personas.

Conclusiones

Esta investigación subraya que los CSIRT de ALC se encuentran en un momento crucial en la adopción de la ciberseguridad basada en la IA. Si bien existe una gran conciencia sobre la gobernanza y la intención estratégica, la implementación sigue siendo desigual, lo que revela una importante brecha entre la ambición y la madurez operativa. La heterogeneidad en cuanto a capacidades, recursos y conocimientos técnicos pone de relieve la necesidad de enfoques de gobernanza que se adapten a las realidades organizativas y sean capaces de apoyar tanto a equipos con muchos recursos como a equipos con recursos limitados. Los mecanismos de cooperación complementarios a nivel regional, como las plataformas compartidas, los programas de formación conjuntos y los conjuntos de herramientas colaborativas, se perfilan como habilitadores esenciales para salvar estas brechas y fomentar un ecosistema de ciberdefensa más cohesionado.

Una idea clave es la prevalencia de un enfoque pragmático basado en normas técnicas en lugar del estricto cumplimiento normativo. Esta orientación refleja tanto una necesidad como una oportunidad: los equipos pueden aplicar inmediatamente prácticas viables, como la supervisión continua de modelos, el control de calidad y la evaluación iterativa de los sistemas de IA. Sin embargo, este pragmatismo también apunta al papel fundamental de la legislación y los marcos formales en la institucionalización de la rendición de cuentas, la garantía de la equidad y la provisión de salvaguardias aplicables. La convergencia de la orientación técnica práctica con una regulación coherente será decisiva para mantener las capacidades impulsadas por la IA a largo plazo.

El estudio hace hincapié además en que la arquitectura organizativa y el capital humano son fundamentales para integrar la IA como una capacidad continua y no como un proyecto puntual. El reconocimiento de la necesidad de funciones de mantenimiento dedicadas a la IA - especialistas en MLOps, ingenieros de seguridad de IA y responsables de riesgos de modelos- señala un paradigma emergente en el que la IA se integra en los procesos operativos básicos. La formalización de los procesos de MLOps y la gestión del riesgo de los modelos, junto con el desarrollo profesional y el reciclaje de habilidades, es crucial para traducir la conciencia conceptual de equidad, fiabilidad y explicabilidad en prácticas institucionalizadas. Sin inversión tanto en tecnología como en capital humano, los CSIRT corren el riesgo de limitar la IA a iniciativas piloto en lugar de lograr un impacto transformador.

Las pruebas cuantitativas y cualitativas confirman que la IA mejora la detección de amenazas y la respuesta a incidentes, lo que valida la inversión en estas capacidades. Sin embargo, los avances son desiguales: las barreras de integración, la confianza limitada en los sistemas automatizados y la calidad inconsistente de los datos generan variabilidad entre los equipos. Esto subraya la importancia de las métricas de rendimiento estandarizadas, los protocolos de evaluación estructurados y la evaluación comparativa entre organizaciones. Al establecer bases de referencia y marcos de medición comunes, los CSIRT pueden garantizar la mejora continua, aumentar la confianza en los resultados de la IA y acelerar la transferencia de conocimientos en todo el ecosistema regional.

Otra dimensión tiene que ver con las prácticas de desarrollo y adquisición estratégicas. Los equipos que adoptan un enfoque equilibrado que combina el desarrollo interno con soluciones externas se benefician de la personalización y la rápida implementación. Sin embargo, la dependencia exclusiva de proveedores externos introduce riesgos como la dependencia de un proveedor, la desalineación con el panorama local de amenazas y posibles lagunas en la retención de capacidades a largo plazo. Por lo tanto, los marcos de adquisición e implementación deben incorporar requisitos de explicabilidad, soberanía de los datos y retención de habilidades profesionales, salvaguardando la agilidad y la resiliencia de la organización a lo largo del tiempo.

Por último, el entorno normativo presenta tanto retos como oportunidades. Las normativas transitorias sobre IA, las leyes fundamentales sobre ciberdelincuencia y las leyes heredadas de protección de datos constituyen un punto de partida, pero requieren una armonización con la

gobernanza técnica para permitir la innovación. La armonización de las normas relacionadas con la responsabilidad algorítmica, la procedencia de los datos y la aplicación transfronteriza será fundamental para catalizar la ciberdefensa mejorada por la IA en la región. Dicha armonización también ofrece la oportunidad de fortalecer la colaboración entre los CSIRTs, armonizar las prácticas operativas y reducir las redundancias, lo que en última instancia mejora la defensa colectiva.

En conjunto, estas conclusiones apuntan a una estrategia multidimensional para promover los CSIRTs habilitados por la IA en ALC. Para lograr una transformación sistémica se necesitará:

1. Modelos de gobernanza adaptables alineados con las capacidades organizativas.
2. Una inversión sostenida en capital humano e infraestructura técnica.
3. Una evaluación y medición del rendimiento estandarizadas en todos los equipos.
4. Una armonización normativa pragmática que respalde la rendición de cuentas y la innovación.
5. Mecanismos regionales de cooperación para poner en común conocimientos especializados, compartir recursos y reducir la fragmentación operativa.
6. Prácticas sostenibles que minimicen el impacto medioambiental y mantengan la eficacia operativa.

Finalmente, si se implementan de manera colectiva, estas estrategias pueden transformar las iniciativas dispersas de IA en un ecosistema de ciberdefensa coherente, resiliente y responsable. Los CSIRTs de la región no solo mejorarán su eficacia operativa, sino que también servirán de modelo para integrar las tecnologías emergentes en las operaciones públicas de ALC para avanzar desde una respuesta reactiva a los incidentes hacia una postura de ciberseguridad proactiva, predictiva y adaptativa, capaz de abordar amenazas complejas y, al mismo tiempo, mantener la rendición de cuentas, la equidad y la responsabilidad medioambiental.

Recomendaciones

La eficacia futura de los CSIRTs depende de intervenciones estratégicas, estructuradas y escalables en materia de gobernanza, tecnología, capital humano y armonización normativa. Si bien muchos equipos han adoptado políticas formales y marcos técnicos, las deficiencias en materia de infraestructura, talento, integración y gestión del ciclo de vida limitan su capacidad para aprovechar plenamente la IA. Las recomendaciones tienen por objeto proporcionar una hoja de ruta completa para consolidar los logros, acelerar la adopción y mejorar la resiliencia, garantizando que la IA se implemente de forma responsable, segura y sostenible.

1. Adoptar una arquitectura y modelo de gobernanza por niveles. Los CSIRT deben implementar un modelo de gobernanza flexible y multinivel que se ajuste a la capacidad de la organización y garantice al mismo tiempo las salvaguardias esenciales en todos los equipos:

    - Nivel básico con los controles mínimos para la calidad de los datos, la gestión del acceso, el registro y la supervisión humana, lo que permite a los equipos más pequeños o emergentes implementar salvaguardias fundamentales sin sobrecargar los recursos.
    - Nivel intermedio con la gestión formalizada del riesgo del modelo, supervisión continua de los sesgos y las desviaciones, y revisiones periódicas estructuradas para garantizar la equidad, la fiabilidad y la trazabilidad.
    - Nivel avanzado con respuesta automatizada controlada, procesos reproducibles y ejercicios basados en simulaciones con equipos dedicados, lo que permite capacidades operativas de alta seguridad.

Este enfoque favorece una madurez progresiva, garantizando que las prácticas de gobernanza fundamentales se apliquen de forma coherente, al tiempo que proporciona una hoja de ruta para aumentar la sofisticación y la automatización.

2. Institucionalizar la garantía del ciclo de vida del modelo. La gestión del ciclo de vida es fundamental para el funcionamiento fiable de la IA. Los equipos deben adoptar las mejores prácticas de MLOps, entre las que se incluyen:

    - Control de versiones y seguimiento del linaje de los modelos y conjuntos de datos.
    - Implementaciones en segundo plano y lanzamientos canarios para probar los cambios de forma segura antes de su despliegue completo en producción.
    - Listas de verificación de evaluación estandarizadas, rutinas de calibración continua y supervisión de sesgos.
    - Las iniciativas ad hoc o piloto de equidad y fiabilidad deben formalizarse en procedimientos operativos.

La incorporación de prácticas de ética por diseño y de preservación de la privacidad, como la minimización de datos, la inyección de ruido y el manejo seguro, en el ciclo de vida garantiza que los sistemas de IA sean responsables desde el principio. La supervisión humana en los puntos críticos de decisión debe seguir siendo fundamental para evitar errores operativos y mantener la confianza.

3. Reforzar las prácticas de trazabilidad, seguridad y transparencia. Las operaciones de IA fiables requieren visibilidad y control de extremo a extremo. Las medidas recomendadas incluyen:

    - Registro detallado de las decisiones algorítmicas, las tarjetas de modelos y las rutas de decisión para respaldar el análisis posterior al incidente, la auditoría y el cumplimiento normativo.
    - Cifrado de los artefactos del modelo, controles de acceso estrictos y ciclos de vida de desarrollo seguros para evitar el uso no autorizado y las violaciones de datos.
    - Medidas proactivas para mejorar la transparencia, incluyendo la exposición de las puntuaciones de confianza y la justificación de las decisiones a las partes interesadas, al tiempo que se equilibra la necesidad de proteger la lógica de detección de los adversarios.

Elevar la trazabilidad, la seguridad y la transparencia de prácticas complementarias a prácticas estándar reforzará la responsabilidad, el cumplimiento y la confianza de las partes interesadas.

4. Abordar los cuellos de botella de la integración con arquitecturas de referencia. La integración tecnológica fragmentada sigue siendo un obstáculo importante. Los CSIRT deben:
    - Desarrollar planos independientes del proveedor para la interoperabilidad entre las plataformas SIEM, EDR, NDR y SOAR.
    - Implementar patrones de integración de alto impacto, como microservicios de enriquecimiento, actualizaciones automatizadas de tickets y armonización de puntuaciones de confianza, para mejorar la fidelidad de la detección sin necesidad de sustituir todo el sistema.
    - Utilizar adaptadores modulares e interfaces abiertas para evitar la dependencia de un proveedor y facilitar la adopción incremental de nuevas herramientas y modelos de IA.

Estas medidas acelerarán la unificación de SecOps y sentarán las bases para una automatización fiable y una toma de decisiones asistida por IA.

5. Implementar servicios compartidos regionales para mitigar las limitaciones de recursos. Las limitaciones presupuestarias y de talento pueden aliviarse mediante marcos cooperativos, entre los que se incluyen:

- Plataformas centralizadas de inteligencia sobre amenazas y repositorios de modelos compartidos.
- Programas rotativos para funciones especializadas, como ingenieros de MLOps, responsables de riesgos de modelos y expertos en privacidad/cumplimiento.
- Centros regionales de excelencia que proporcionan formación, evaluación y apoyo consultivo.

Los servicios compartidos reducen los costes operativos, elevan los estándares mínimos de capacidad y fomentan el acceso equitativo a las capacidades mejoradas por la IA en toda la región.

6. Dirigir las inversiones hacia las etapas menos desarrolladas del ciclo de vida de los incidentes. Las intervenciones impulsadas por la IA deben dar prioridad a las acciones de bajo riesgo y a los proyectos piloto controlados:

    - Aislamiento temporal, bloqueo restringido con confirmación humana y protocolos de escalado.
    - Expansión de la automatización en torno a los manuales operativos existentes, garantizando transiciones seguras desde la detección hasta la respuesta.
    - Supervisión continua del rendimiento de los proyectos piloto para guiar la ampliación gradual de las operaciones autónomas.

Este enfoque por etapas garantiza la seguridad operativa al tiempo que aumenta progresivamente el papel de la IA en la gestión de incidentes.

7. Modernizar los controles legales y de adquisición. La adopción sostenible de la IA requiere habilitar marcos normativos y de adquisición:

    - Alinear la legislación sobre IA, las leyes de privacidad y los estatutos sobre ciberdelitos con las normas de gobernanza operativa.
    - Incluir requisitos de documentación de modelos, explicabilidad, pruebas de sesgo, portabilidad de datos y desmantelamiento en las cláusulas de adquisición.
    - Establecer mecanismos para hacer cumplir la responsabilidad, reducir el riesgo e incentivar la innovación segura.

La modernización legal y de adquisiciones garantiza que la adopción de la IA sea responsable, auditable y esté alineada con los objetivos de interés público.

8. Desarrollar una cantera de talentos alineada con la IA en CyberOps. El capital humano es el eje central de los CSIRT habilitados para la IA. Las recomendaciones incluyen:
    - Establecer funciones especializadas (ingenieros de seguridad de IA, especialistas en MLOps, responsables de riesgos de modelos, expertos en privacidad/cumplimiento).
    - Crear programas de microcredenciales e iniciativas de formación cruzada para poner en práctica competencias de equidad, fiabilidad y supervisión.
    - Fomentar la colaboración entre los científicos de datos y los gestores de incidentes para cerrar las brechas de traducción entre los resultados de la IA y los procedimientos operativos aplicables.

Una estrategia estructurada de desarrollo del talento garantiza la sostenibilidad a largo plazo, la resiliencia y la excelencia operativa en la ciberdefensa impulsada por la IA.

Estas recomendaciones proporcionan una hoja de ruta integral para que los CSIRTs de ALC avancen en sus capacidades basadas en la IA de una manera responsable, segura y sostenible. Al combinar una gobernanza por niveles, una gestión sólida del ciclo de vida, una trazabilidad y transparencia reforzadas, arquitecturas de integración modulares, servicios regionales compartidos, inversiones específicas, modernización legal y de adquisiciones, y desarrollo

estratégico del talento, las organizaciones pueden salvar las brechas actuales y transformar las iniciativas experimentales en operaciones fiables y escalables.

En conjunto, estas medidas permitirán a los CSIRTs mejorar la resiliencia operativa, fomentar la colaboración regional y garantizar que la IA contribuya no solo a una respuesta más rápida y eficaz ante las amenazas, sino también a prácticas de ciberseguridad responsables, transparentes y alineadas con la ética en toda la región.

Perspectivas de Futuro

La trayectoria de los CSIRTs mejorados con IA en ALC estará determinada por la interacción entre la adopción de tecnología, la madurez de la gobernanza, el desarrollo del talento, la armonización normativa y la colaboración regional. Si bien los avances iniciales han demostrado un valor operativo cuantificable, es necesario abordar los retos sistémicos —entre ellos, la integración desigual, la escasa experiencia especializada y la adopción variable de salvaguardias de privacidad y gobernanza— para garantizar un impacto sostenido. El desarrollo futuro debe guiarse por estrategias escalonadas y basadas en pruebas que equilibren la innovación con la rendición de cuentas, la seguridad operativa y la resiliencia.

Corto plazo (12-24 meses): Consolidación y estandarización

A corto plazo, los CSIRTs se centrarán en consolidar las capacidades de IA existentes e institucionalizar las prácticas de gobernanza esenciales. Las iniciativas clave incluyen:

- Incorporar principios de ética desde el diseño y técnicas de preservación de la privacidad en todas las nuevas implementaciones de IA, garantizando que la minimización de datos, la inyección de ruido y el manejo seguro se integren desde el principio.
- Establecer mecanismos de supervisión humana sólidos en los puntos de decisión críticos para mantener la rendición de cuentas y reducir el riesgo operativo.
- Ampliar el uso de controles estandarizados de trazabilidad y seguridad, como el registro de modelos, los ciclos de vida de desarrollo seguro y la gestión de accesos, para crear bases operativas coherentes.
- La integración gradual de mejoras modulares en el flujo de trabajo en las plataformas SIEM, EDR, NDR y SOAR, lo que permite mejoras pequeñas, pero de gran impacto en la detección, el enriquecimiento y la coordinación de incidentes sin necesidad de realizar revisiones completas de la plataforma.
- Programas piloto iniciales para la automatización controlada en escenarios de bajo riesgo, aprovechando los manuales de estrategias existentes y los umbrales de confirmación humana para salvar la brecha entre la detección y la respuesta.

Esta fase hace hincapié en la escalabilidad consciente del riesgo, garantizando que las prácticas técnicas y de gobernanza fundamentales se conviertan en rutina antes de ampliar el alcance de la automatización o la autonomía.

Medio plazo (24-36 meses): Toma de decisiones y refuerzo preventivo

A medida que los CSIRTs maduran, el enfoque operativo pasará de la gestión reactiva de incidentes a la toma de decisiones proactiva y las medidas preventivas de ciberseguridad. Entre los avances previstos se incluyen:

- Despliegue de modelos predictivos para asignar los recursos de forma más eficaz, anticipar posibles amenazas y optimizar las estrategias de respuesta a incidentes.
- Expansión controlada de acciones de respuesta semiautónomas, con umbrales claramente definidos, supervisión humana en el bucle y mecanismos de retroceso para mantener la seguridad.

- Fortalecimiento de las prácticas de trazabilidad y transparencia, incluyendo registros detallados, tarjetas de modelos, rutas de decisión y puntuaciones de confianza, lo que permite la revisión posterior al incidente, la rendición de cuentas y el cumplimiento normativo.
- Armonización de las normas jurídicas y técnicas: se actualizarán la legislación sobre IA, las leyes sobre ciberdelincuencia y las normas de privacidad para codificar la supervisión, la documentación, la minimización de datos y la responsabilidad de los modelos. Esta armonización reducirá la aversión al riesgo institucional y fomentará una adopción más amplia de las herramientas de IA.
- Desarrollo de servicios regionales compartidos y redes de conocimiento colaborativas, incluidos repositorios de modelos compartidos, laboratorios de evaluación y plataformas de formación, para acelerar la difusión de las mejores prácticas y reducir las desigualdades operativas.

Este periodo se define por la transición de la detección a la inteligencia estratégica, en la que la IA informa sobre la asignación de recursos y la planificación operativa, al tiempo que garantiza la gobernanza y el cumplimiento.

Largo plazo (36-60 meses): Automatización de alta seguridad y colaboración regional

A largo plazo, la resiliencia transformacional dependerá de la profunda integración de la IA, la cooperación regional y las prácticas operativas de alta seguridad. Entre los avances previstos se incluyen:

- Aprendizaje federado y uso compartido de modelos que preservan la privacidad, lo que permite a los CSIRTs aprovechar conjuntos de datos más amplios sin comprometer la soberanía de los datos nacionales u organizativos.
- Despliegue de automatización de alta seguridad, respaldada por pruebas adversarias, evaluación causal y garantía del ciclo de vida, lo que permite acciones de respuesta más complejas al tiempo que se mantiene la seguridad operativa.
- Expansión de centros regionales de excelencia para coordinar el desarrollo, la evaluación y la formación de modelos, garantizando un acceso equitativo a conocimientos especializados y promoviendo la interoperabilidad transfronteriza.
- Integración avanzada de mecanismos de supervisión continua, evaluación del rendimiento en tiempo real y gobernanza adaptativa, que proporcionan una supervisión dinámica de los sistemas de IA, la equidad y la fiabilidad.
- Establecimiento de marcos de resiliencia en todo el sistema, en los que los conocimientos basados en la IA, la automatización y la experiencia humana se fusionan en estrategias de defensa regionales coordinadas.

Los factores clave en esta etapa incluyen la financiación multilateral de servicios compartidos, ejercicios cibernéticos coordinados y marcos normativos y de contratación armonizados. Entre los riesgos se encuentran la fragmentación de las normas, la desconfianza del público en la IA, la austeridad presupuestaria prolongada y la escasez de talento, todo lo cual debe mitigarse mediante una gobernanza colaborativa y sólidos programas de desarrollo de capacidades.

La evolución de los CSIRTs habilitados por la IA en ALC es gradual, se basa en la evidencia y depende en gran medida de la gobernanza, el talento y la colaboración regional. Los esfuerzos a corto plazo consolidarán las prácticas existentes y establecerán salvaguardias fundamentales. Las iniciativas a medio plazo ampliarán el papel de la IA en la toma de decisiones y las operaciones preventivas, con el apoyo de una legislación armonizada y servicios compartidos. El progreso a largo plazo dependerá de la cooperación regional, la automatización de alta seguridad y los marcos de gobernanza integrados que, en conjunto, permitan un ecosistema de ciberseguridad sostenible, resiliente y responsable basado en la IA.

Llamado de Acción

Los CSIRTs de ALC han superado la fase experimental y han demostrado mejoras operativas cuantificables gracias a la detección de amenazas mejorada con IA, la reducción de falsos positivos y los playbooks de procedimientos automatizados. Sin embargo, siguen existiendo retos: gobernanza desigual, obstáculos a la integración, escasez de talento e inversión insuficiente.

La fortaleza institucional, y no solo las herramientas, es fundamental. Las medidas inmediatas incluyen formalizar la ética por diseño, la supervisión humana, la trazabilidad y los controles de seguridad; establecer un seguimiento continuo y revisiones periódicas de los modelos; y ampliar los manuales automatizados en escenarios controlados. Las prácticas de transparencia deben elevarse a requisitos estándar, respaldando la rendición de cuentas, el cumplimiento y la confianza de las partes interesadas.

Los servicios compartidos regionales y las interfaces abiertas pueden aliviar las limitaciones de recursos, manteniendo un enfoque equilibrado de desarrollo y adquisición, al tiempo que se reduce la dependencia de los proveedores. Los responsables políticos y los financiadores deben apoyar ecosistemas de IA fiables mediante infraestructuras compartidas, marcos jurídicos actualizados y regímenes de contratación pública que impongan la explicabilidad, la seguridad y la soberanía de los datos. El mundo académico y la sociedad civil deben proporcionar una evaluación independiente y una supervisión sociocrítica para garantizar una adopción equitativa y mitigar los sesgos sistémicos.

La resiliencia adaptativa es posible. Mediante la inversión colectiva en capas de gobernanza, garantía del ciclo de vida, desarrollo del talento humano y servicios regionales compartidos, los CSIRT pueden transformar los proyectos piloto en operaciones de ciberdefensa basadas en IA duraderas, responsables y escalables. Las acciones a corto plazo se centran en la estandarización y la supervisión, los objetivos a medio plazo en la automatización controlada y las integraciones interoperables, y las prioridades a largo plazo en las redes de servicios regionales y los marcos jurídicos armonizados. La ejecución disciplinada y la gobernanza colaborativa permitirán a la región construir un ecosistema de ciberseguridad sostenible, de alto rendimiento y responsable impulsado por la IA.

Reflexiones Finales

Gobernanza

- La gobernanza de la IA en las CyberOps está pasando de ser un ejercicio de cumplimiento o nivel de madurez para los CSIRTs y pasa a convertirse en un imperativo estratégico. Los enfoques tradicionales —privacidad y protección de datos (por ejemplo, el RGPD o GDPR) y marcos éticos (por ejemplo, Ethics by Design)— ya no son suficientes por sí solos. En su lugar, los expertos en ALC hacen hincapié en la integración de metodologías proactivas de gestión de riesgos, como el Marco de Gestión de Riesgos de IA del NIST, con capacidades de inteligencia que garanticen una alineación dinámica con los requisitos legales en constante evolución, como las diversas iniciativas que en ALC se están desarrollando de IA.

- La gobernanza también debe ir más allá de la estrategia e integrarse en mecanismos técnicos de aplicación: aplicación de políticas, registros de auditoría, controles criptográficos, controles de acceso basados en roles, identidades, agentes y modelos estructurados de supervisión humana *(Human-In-The-Loop, Human-On-The-Loop y Human-In-Command).* Estos controles son vitales para garantizar la responsabilidad, la explicabilidad y la transparencia, incluso en sistemas de IA parcial o totalmente autónomos.

- Por último, la ampliación de la gobernanza a toda la cadena de suministro —mediante contratos seguros, la evaluación de modelos por terceros y la supervisión continua— refleja el creciente reconocimiento de que la exposición de los proveedores representa una vulnerabilidad crítica en la implementación de la IA.

Desarrollo

- Las Cyber Ops habilitadas por IA deben ir más allá de la ingeniería tradicional en silos hacia un enfoque integrado de DevSecOps y MLOps, reforzado por principios de diseño ético y una gobernanza de datos rigurosa. En este paradigma, la explicabilidad, la reproducibilidad y la verificación formal no son complementos opcionales, sino componentes nativos del ciclo de vida del desarrollo.

- Los expertos en ALC destacan el valor de las arquitecturas seguras y modulares basadas en entornos contenedorizados, que garantizan la escalabilidad, la resiliencia y la facilidad de mantenimiento. Prácticas que antes se consideraban experimentales, como las métricas de equidad automatizadas, las pruebas de robustez adversaria y el aprendizaje federado, están emergiendo ahora como requisitos fundamentales del diseño de la IA, especialmente beneficiosos para los CSIRTs y la ciberdefensa crítica.

- Además, la adopción de artefactos de datos con control de versiones, registros de modelos inmutables y herramientas de desarrollo colaborativo introduce un rigor del que a menudo carecen las operaciones cibernéticas tradicionales. En última instancia, estos hallazgos confirman que el desarrollo seguro de la IA es inseparable de la responsabilidad organizativa, lo que exige procesos de desarrollo que admitan la validación en tiempo real y la garantía ética continua.

Adopción

- La adopción de la IA en las CyberOps es un proceso de madurez progresivo, no un evento binario. Esta progresión debe regirse por estrategias de implementación conscientes del riesgo que equilibren la innovación con la seguridad.

- Se destaca que la adopción exitosa va más allá de la preparación técnica: requiere alineación cultural, políticas adaptativas y desarrollo estructurado del talento y las capacidades humanas. Los habilitadores técnicos, como los enclaves seguros, el cifrado homomórfico y los elementos de hardware de confianza, permiten que la IA funcione en entornos sensibles sin comprometer la confidencialidad o la integridad.

- A nivel operativo, las plataformas de observabilidad y supervisión, los mecanismos de reversión sensibles a las desviaciones y los marcos contractuales para los sistemas de IA de terceros actúan como habilitadores de la confianza en entornos sistémicos altamente complejos. Igualmente importante es que la adopción está condicionada por la preparación humana: la formación basada en simulaciones, los ejercicios de equipos rojos/azules y la planificación de escenarios ayudan a los analistas a comprender el comportamiento de la IA, sus limitaciones y los riesgos de uso indebido.

- Por lo tanto, la integración de la IA debe producirse no solo en herramientas, sistemas o modelos, sino también en el tejido organizativo de los CSIRTs mediante formación, pruebas iterativas e integración alineada con la gobernanza.

Seguridad

- Las Cyber Ops mejoradas con IA son multifacéticas y exigen una estrategia de defensa por capas adaptada a las amenazas específicas de la IA. A diferencia de los sistemas tradicionales, la IA introduce nuevas vulnerabilidades, como el envenenamiento de datos, las entradas adversarias y la inversión de modelos, que obligan a replantearse los controles de seguridad.

- Los expertos en ALC, hacen hincapié en la necesidad de protecciones integrales, que incluyan defensas adversarias certificadas, privacidad diferencial, limitación de consultas y procedencia basada en blockchain tanto para los modelos como para los conjuntos de datos. La extensión de los principios de confianza cero a los flujos de trabajo de IA, tratando cada interacción, flujo de datos y punto de decisión como potencialmente comprometido, representa un cambio conceptual en la forma de establecer y mantener la confianza.

- Medidas adicionales, como una lista de materiales de inteligencia artificial (AIBOM), cálculos multipartitos seguros y entornos de ejecución confiables (TEE), subrayan la convergencia de la ciberseguridad y la garantía de la IA en una nueva disciplina. La adopción de esta postura de seguridad integral garantiza que la IA no cree nuevas superficies de ataque, sino que refuerce la resiliencia y la integridad de los sistemas cibernéticos.

Resiliencia

- La resiliencia de los sistemas de IA en las Cyber Ops debe integrarse desde el diseño, y no añadirse tras la implementación. La resiliencia va más allá del mantenimiento del tiempo de actividad; requiere precisión continua, conciencia contextual y confianza en condiciones de incertidumbre, complejidad y riesgo.

- Los mecanismos técnicos, como las arquitecturas modulares basadas en conjuntos, el cambio dinámico de modelos y los protocolos de respaldo, proporcionan redundancia. Al mismo tiempo, las pruebas de estrés y la ingeniería del caos exponen las debilidades antes de que los adversarios las aprovechen. La detección de desviaciones y el reentrenamiento automatizado añaden propiedades de autorreparación, lo que permite a los sistemas adaptarse a los cambios ambientales y adversos.

- Las estrategias avanzadas de tolerancia a fallos, como el consenso bizantino y los mecanismos de reversión, representan un cambio de paradigma en el que los agentes de IA se convierten en participantes activos en la continuidad de la ciberdefensa.

- En última instancia, la resiliencia evoluciona de una medida reactiva a una capacidad sistémica, lo que permite a los CSIRTs resistir, absorber y recuperarse de las interrupciones sin comprometer la integridad operativa.

Adaptabilidad

- La adaptabilidad es esencial para que los sistemas de IA sigan siendo eficaces en entornos volátiles y adversos. Se está produciendo un cambio claro: se está pasando de modelos estáticos a sistemas de aprendizaje dinámicos respaldados por una supervisión continua, la detección de desviaciones estadísticas y el reentrenamiento en tiempo real.

- Entre los habilitadores técnicos se incluyen los algoritmos de aprendizaje en línea, la búsqueda de arquitectura neuronal y el aprendizaje por refuerzo, que permiten a los sistemas de IA perfeccionar de forma autónoma sus estrategias en respuesta a las amenazas en constante evolución. Enfoques complementarios como el aprendizaje federado y el aprendizaje por transferencia apoyan la adaptación distribuida sin necesidad de una consolidación centralizada de los datos, lo que preserva la privacidad y acelera el aprendizaje.

- A nivel de infraestructura, las pipeline de datos en streaming (por ejemplo, Apache Flink, Kafka) combinadas con la observabilidad en tiempo real establecen bucles de retroalimentación que transforman cada incidente cibernético en una oportunidad de aprendizaje.

- Esta adaptabilidad contextual se está convirtiendo en un habilitador crítico para los CSIRTs, ya que les permite mantener la superioridad operativa frente a adversarios ágiles y sofisticados.

Responsabilidad

- La responsabilidad en las operaciones de los CSIRTs habilitadas por IA ha pasado de ser una preocupación ética abstracta a convertirse en una necesidad técnica y procedimental. La explicabilidad, la equidad y la responsabilidad deben integrarse en todo el ciclo de vida del sistema. Las técnicas de IA explicable (XAI), como SHAP, LIME y el razonamiento contrafactual, no son opcionales, sino obligatorias para validar los resultados en entornos de alto riesgo.

- Esto requiere evaluaciones continuas de equidad, modelos éticos formales que utilicen especificaciones basadas en la lógica y mecanismos sólidos de gobernanza de datos alineados con las normativas de privacidad y protección de datos para garantizar resultados legales, transparentes y no discriminatorios. Al mismo tiempo, las pistas de auditoría inmutables y el seguimiento del linaje de los modelos permiten la investigación forense y la verificación del cumplimiento.

- La responsabilidad también se extiende a la interacción entre humanos y IA. Es esencial contar con flujos de trabajo claramente definidos con puntos de intervención humana, protocolos a prueba de fallos y estructuras de supervisión multidisciplinares para evitar el uso indebido o la automatización incontrolada.

- De este modo, la responsabilidad se convierte en un marco tangible para la confianza operativa y la legitimidad organizativa, lo que garantiza que la integración de la IA en la ciberdefensa refuerce, en lugar de socavar, la resiliencia y la gobernanza.

Sostenibilidad

- Los resultados confirman que la sostenibilidad de la IA para los CSIRTs debe abarcar tanto la eficiencia medioambiental como la responsabilidad social. Desde una perspectiva medioambiental, la eficiencia energética, la optimización de los recursos y la gestión del ciclo de vida son fundamentales.

- El impacto ecológico del entrenamiento y la inferencia de la IA puede reducirse mediante la compresión de modelos (por ejemplo, poda, cuantificación), la programación dinámica de la carga de trabajo y las prácticas de tecnología verde. Del mismo modo, la IA de vanguardia y el aprendizaje federado reducen la dependencia de la infraestructura centralizada en la nube y minimizan las transferencias de datos innecesarias, mientras que la orquestación inteligente garantiza un uso eficiente de los recursos computacionales.

- Del mismo modo, la sostenibilidad social es vital. La adopción de modelos de código abierto, plataformas compartidas de inteligencia sobre amenazas e iniciativas de educación inclusiva ayuda a democratizar el acceso a capacidades avanzadas de IA en diversas comunidades de seguridad. Los programas de desarrollo de capacidades y las estrategias de retención de talento garantizan que los conocimientos especializados no solo se desarrollen, sino que también se mantengan dentro de los CSIRTs, evitando la fuga de cerebros y reforzando la resiliencia de las organizaciones.

- La gestión del ciclo de vida, que incluye el seguimiento de la huella de carbono, las evaluaciones del ciclo de vida del hardware y la reutilización de modelos,

garantiza además que la escalabilidad no se logre a expensas de la viabilidad ecológica u operativa.

- En última instancia, la adopción sostenible de la IA permite a los CSIRTs desarrollar capacidades de ciberdefensa a largo plazo, éticas y escalables, respaldadas tanto por la eficiencia tecnológica como por el desarrollo del capital humano.

Apéndice: **Contexto de Investigación**

Descripción del Problema

Las innovaciones tras los avances de tecnologías emergentes y disruptivas en América Latina y el Caribe está sentando precedentes. Tal es el caso de algunos países de ALC con la puesta en marcha de sus Estrategias Nacionales de Inteligencia Artificial, donde los CSIRTs están a la vanguardia de la identificación, detección, gestión y mitigación de las ciberamenazas. No obstante, el volumen, la velocidad, la sofisticación y la complejidad creciente de los ciberataques, especialmente en infraestructuras críticas y sectores de alto riesgo, han superado las capacidades de las herramientas tradicionales y los análisis por los humanos.

Hay que destacar que los CSIRTs de la región desde el 2020 al 2024 ha detectado más de 900 millones de ciberataques, incluyendo la gestión de ciberincidentes en las infraestructuras de servicios esenciales. Por tanto, la IA proporciona un potencial significativo para mejorar las operaciones de los CSIRTs, desde la detección de amenazas y el triaje hasta el apoyo a la toma de decisiones y el análisis predictivo, contribuyendo al desarrollo de capacidades de anticipación y prospectiva en la región mediante la colaboración y cooperación desde CSIRT-Américas. Sin embargo, la integración de IA conlleva una nueva serie de retos y desafíos:

1. Falta de mecanismos claros de gobernanza y responsabilidad para la IA en entornos operativos.
2. Riesgos de falsos positivos o negativos, sesgo algorítmico y desviación del modelo.
3. Interoperabilidad y fiabilidad de las herramientas basadas en IA.
4. Comprensión limitada de cómo alinear el uso de la IA con los marcos éticos, legales y organizativos.

Esta investigación reviste importancia a las comunidades de ciberseguridad en la región ya que subraya la necesidad de una adopción sostenible, segura de la IA en las CyberOps. La ausencia de marcos claros para el desarrollo despliegue y gobernanza de la IA en entornos CSIRT es una brecha crítica que esta investigación busca abordar.

Objetivos de Investigación

Objetivo General

Analizar cómo la gobernanza, el desarrollo y la adopción responsable de la inteligencia artificial en las ciberoperaciones pueden contribuir a la transformación de los CSIRTs hacia modelos de respuesta cibernética más resilientes, sostenibles y adaptativos en entornos de creciente complejidad.

Objetivos Específicos

1. Examinar los marcos normativos y éticos aplicables a la gobernanza de la IA en contextos de ciberseguridad crítica.
2. Identificar buenas prácticas en el desarrollo de soluciones de IA enfocadas en la sostenibilidad, robustez y alineación con los principios de seguridad cibernética.
3. Evaluar los factores organizacionales, técnicos y humanos que inciden en la adopción efectiva de IA en los CSIRTs.
4. Definir un modelo conceptual de operación ciberdefensiva basado en IA que promueva la resiliencia adaptativa y la respuesta responsable ante incidentes cibernéticos.

Apéndice: **Metodología**

La metodología consistió en principio por una revisión de la literatura dimensionado tanto en publicaciones de rigor científico e informes técnico-industriales en aras de conocer, comprender y dimensionar el estado del arte de las tendencias pasadas, actuales y futuras. También, la investigación se enmarcó en fundamentos de enfoques positivistas, el pragmatismo, el realismo crítico, el sociocrítico y el constructivismo.

Estos enfoques garantizaron que la investigación sea holística, adaptable y con rigor científico; lo cual equilibró las pruebas empíricas, los modelos teóricos y la aplicabilidad en el mundo real.

En este estudio se aplicó un diseño de investigación multimétodo, integrando enfoques cualitativos y cuantitativos, dado el contexto de complejidad de las ciberoperaciones y su relación con la IA, se adopta un diseño mixto que combina elementos descriptivos, correlacional, explicativo y exploratorio, en aras de garantizar un análisis exhaustivo de gobernanza, el desarrollo y la adopción responsable de la IA en los CSIRTs.

Se construyo una base de datos donde se tomaron cuatro CSIRTs y veinte analistas de CyberOps de ALC como unidad primaria de análisis, mediante un instrumento diseñado. Con una muestra de selección aleatoria de profesionales y expertos en ciberseguridad, analistas de operaciones cibernéticas, riesgos, transformación digital, gobernanza, protección de datos, inteligencia artificial y policy maker como parte de las entrevistas semiestructuradas al pool de expertos.

Esta población proporcionó una visión clara de las prácticas actuales, las estrategias en el contexto de las tecnologías de innovación emergente y disruptivas en el ámbito de la ciberseguridad. Los participantes se seleccionaron en función de su rol desempeñado, experiencia y participación tanto en las ciberoperaciones como la incidencia en la región sobre estas tecnologías, lo que garantizará la pertinencia y la profundidad de la recopilación de datos para los objetivos de la investigación.

Para recopilar los datos, se diseñaron dos instrumentos uno destinado al CSIRT para conocer y profundizar en la gobernanza, el desarrollo y la adopción de la IA en las CyberOps, teniendo en cuenta la resiliencia, sostenibilidad y adaptación; correlacionado con eso se aplicó uno para CSIRT-Américas en aras de tener una perspectiva regional de los 22 países miembros. Así mismo las entrevistas semiestructuradas al pool de expertos.

En particular, en el análisis de datos se enfocó en matizar la correlación de los aspectos vinculados con la ciberseguridad y la IA en el contexto de la investigación lo que permitió apoyar la triangulación de datos entre fuentes y para validar las conclusiones de las entrevistas aplicando el método Delphi con estudios de casos y el análisis observacional logrando el examen empírico de las practicas gobernanza.

De igual manera, para algunas entrevistas se consideraron los organismos de colaboración y cooperación regional como el Latin America and Caribbean Cyber Competence Centre (LAC4, por sus siglas en inglés), la Red de los Equipos de Respuesta ante Incidentes Cibernéticos (CSIRT- Américas) de la Organización de los Estados Americanos (OEA) y el Banco Interamericano de Desarrollo (BID) a través de la Red de Excelencia en Ciberseguridad de Latinoamérica y el Caribe (Red Ciberlac), el Center for Public Intelligence (CPI), la Sociedad Dominicana de Ciencia de Datos (BigDataDO), Internet Society (ISOC), la Red Iberoamericana de Protección de Datos (RIPD) y el LACNIC CSIRT.

Por otra parte, la estrategia de estadística estuvo diseñada para analizar rigurosamente los datos recopilados en consonancia con los objetivos de investigación, por lo cual se esbozan el uso de combinaciones de técnicas estadística, incluidas estadística descriptiva e inferencial, análisis de correlación y regresión y enfoques multivariantes para explorar las relaciones y extraer conclusiones significativas de los datos. Además, los datos cualitativos se analizaron

temáticamente para identificar patrones; es por esto por lo que este enfoque garantiza una exploración exhaustiva y basada en datos de las preguntas de investigación.

Para el análisis de datos cuantitativos se empleó SPSS, complementado con scripts en R y Python, integrando además algoritmos de inteligencia artificial predictiva para potenciar la modelización y proyección de escenarios. Este abordaje se vinculó con el análisis cualitativo, estableciendo correlaciones cruzadas que permitieron una interpretación más robusta, rigurosa y sistémica de los resultados.

Además, se aplicaron los métodos inductivo, deductivo, analítico y de variables concomitantes en la investigación porque su combinación permite un enfoque integral: el método inductivo facilitó la observación y generalización de hechos concretos; el deductivo permitió contrastar hipótesis con teorías existentes; el analítico descompuso el problema para entender sus componentes, y el de variables concomitantes ayudó a identificar relaciones causales entre fenómenos. Esta integración fortaleció la validez y profundidad del análisis.

Como parte del desarrollo metodológico, se solicitó carta de credenciales formales al LACNIC, para desarrollar la investigación y poder presentarla a las partes interesadas y así acceder a información técnica especializada, bases de datos no públicas y entrevistas con expertos, documentación interna controlada y varios estudios de casos, esenciales para la validación empírica y el análisis crítico de la investigación.

En todo momento se siguieron las buenas prácticas de ética en la investigación científica y estándares internacionales de gobierno del dato, garantizando la integridad, calidad, seguridad, privacidad, confidencialidad, disponibilidad, autenticidad, trazabilidad y fiabilidad de los datos. El ciclo de vida del dato "desde su generación hasta su interpretación" se gestionará conforme a los principios (FAIR, por sus siglas en inglés) de localización, accesibilidad, interoperabilidad y reutilización datos científicos. A su vez, se tuvo en cuenta la Protección y Privacidad de Datos Personales.

Finalmente, la metodología propuesta garantizó un análisis riguroso, sistémico y aplicable al entorno real. Por lo que la combinación de modelado, validación empírica, y las sinergias de colaboración y cooperación con actores claves permitió obtener resultados sólidos, relevantes y alineados con los objetivos del proyecto de investigación.

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. K. Tallam, "Transforming Cyber Defense: Harnessing Agentic and Frontier AI for Proactive, Ethical Threat Intelligence," Feb. 2025.
2. Sri Keerthi Suggu, "Agentic AI Workflows in Cybersecurity: Opportunities, Challenges, and Governance via the MCP Model," *Journal of Information Systems Engineering and Management*, vol. 10, no. 52s, pp. 612–624, Jun. 2025, doi: 10.52783/jisem.v10i52s.10767.
3. L. Potter, "Naming is framing: How cybersecurity's language problems are repeating in AI governance," Apr. 2025.
4. E. Vergara Cobos, *Cybersecurity Economics for Emerging Markets*. Washington, DC: World Bank, 2024. doi: 10.1596/978-1-4648-2120-2.
5. World Economic Forum and Accenture, "Global Cybersecurity Outlook 2025," Geneva, Jan. 2025. [Online]. Available: https://reports.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2025.pdf
6. Organización de los Estados Americanos, "Guía Práctica para CSIRTs: Un modelo de negocio sustentable," Jul. 2023. Accessed: Jul. 12, 2025. [Online]. Available: https://www.oas.org/ext/DesktopModules/MVC/OASDnnModules/Views/Item/Download.aspx?type=1&id=663&lang=2
7. Banco Interamericano de Desarrollo, "Exponential Disruption in the Digital Economy," Washington, D. C., Apr. 2018. doi: 10.18235/0001068.
8. Executive Office of the President and National Science and Technology Council, "Preparing for the future of artificial intelligence," Oct. 2016. Accessed: Jul. 12, 2025. [Online]. Available: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
9. E. Iglesias Rodriguez, A. García Zaballos, P. Puig Gabarró, and I. Benzaquén, *Inteligencia artificial: Gran oportunidad del siglo XXI: Documento de reflexión y propuesta de actuación*. Inter-American Development Bank, 2020. doi: 10.18235/0003037.
10. P. A. Zubaedah, R. Harliyanto, S. M. T. Situmeang, D. S. Siagian, and E. Septaria, "The Legal Implications of Data Privacy Laws, Cybersecurity Regulations, and AI Ethics in a Digital Society," *The Journal of Academic Science*, vol. 1, no. 2, Jul. 2024, doi: 10.59613/29qypw51.
11. F. H. O. Kolo, "Responsible AI for Cybersecurity: Assessing the Barriers, Biases and Governance Gaps in Implementation with E-commerce Systems," *Journal of Engineering Research and Reports*, vol. 27, no. 5, pp. 510–532, May 2025, doi: 10.9734/jerr/2025/v27i51520.
12. K. Aart, P. Zoido-Lobatón, and D. Kaufman, "La gobernabilidad es fundamental Del análisis a la acción," *Finanzas & Desarrollo*, 2000, Accessed: Jul. 12, 2025. [Online]. Available: https://www.imf.org/external/pubs/ft/fandd/spa/2000/06/pdf/kauf.pdf
13. Y. Martínez Mancilla and M. Vega, "Gobernanza participativa de la inteligencia artificial," Washington, D. C., Sep. 2020. doi: 10.18235/0002719.
14. J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3518482.
15. C. S. Rosales Torres, C. Buenadicha Sánchez, and T. Narita, "Ethical Assessment of AI for Actors within the Entrepreneurial Ecosystem: Application Guide," Washington, D. C., May 2021. doi: 10.18235/0003269.
16. T. R. McIntosh *et al.*, "From COBIT to ISO 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models," *Comput Secur*, vol. 144, p. 103964, Sep. 2024, doi: 10.1016/j.cose.2024.103964.
17. J. D. Azabache Santos, N. A. Ángeles Piedra, and A. C. Mendoza de los Santos, "IMPACTO DE LA INTEGRACIÓN DEL GOBIERNO DE TI EN LA ADOPCIÓN DE LA INTELIGENCIA ARTIFICIAL," *INVESTIGACION & DESARROLLO*, vol. 23, no. 2, Jan. 2024, doi: 10.23881/idupbo.023.2-9e.
18. D. D. Ordoñez-Iturralde and C. X. Proaño-Piedra, "Gobernanza y desafíos éticos en la adopción empresarial de la Inteligencia Artificial," *Innovación Empresarial*, vol. 4, no. 2, pp. 73–82, Aug. 2024, doi: 10.37711/rcie.2024.4.2.633.
19. A. Herrera-Poyatos, J. Del Ser, M. L. de Prado, F.-Y. Wang, E. Herrera-Viedma, and F. Herrera, "Responsible Artificial Intelligence Systems: A Roadmap to Society's Trust through Trustworthy AI, Auditability, Accountability, and Governance," Feb. 2025.
20. Y. Ge and Q. Zhu, "The Game-Theoretic Symbiosis of Trust and AI in Networked Systems," Nov. 2024.
21. I. H. Sarker, H. Janicke, N. Mohammad, P. Watters, and S. Nepal, "AI Potentiality and Awareness: A Position Paper from the Perspective of Human-AI Teaming in Cybersecurity," Sep. 2023.
22. T. Ali and P. Kostakos, "HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs)," Sep. 2023.
23. C. Claire Adaji, A. Adewale Bello, C. Emmanuela Ukatu, N. Okika, O. Kabiru Agboola, and C. Godwin Amomo, "AI-Powered Cybersecurity Governance: The Role of Business Analysts in Ethical AI Deployment," *Int J Innov Sci Res Technol*, pp. 1384–1396, Mar. 2025, doi: 10.38124/ijisrt/25mar924.
24. M. Kim *et al.*, "CyberAlly: Leveraging LLMs and Knowledge Graphs to Empower Cyber Defenders," Apr. 2025.
25. S. Freitas, J. Kalajdjieski, A. Gharib, and R. McCann, "AI-Driven Guided Response for Security Operation Centers with Microsoft Copilot for Security," Nov. 2024.
26. U. Nedunuri, "Factors Influencing the Adoption of Responsible AI," in *2023 10th International Conference on Electrical and Electronics Engineering (ICEEE)*, IEEE, May 2023, pp. 214–218. doi: 10.1109/ICEEE59925.2023.00046.
27. V. R. T. P. V. Dr. Khaja Mohinuddeen J. and D. N. R. Dr. K. M. J. , Dr. Biswo Ranjan Mishra, "Ethical Implications of Artificial Intelligence in Business Decision-making: A Framework for Responsible AI Adoption," *Journal of Informatics Education and Research*, vol. 4, no. 1, Feb. 2024, doi: 10.52783/jier.v4i1.597.
28. D. Bruschi and N. Diomede, "A framework for assessing AI ethics with applications to cybersecurity," *AI and Ethics*, vol. 3, no. 1, pp. 65–72, Feb. 2023, doi: 10.1007/s43681-022-00162-8.
29. F. Hariyanto and K. Ramli, "Design and Analysis of Cybersecurity Information Sharing Mechanism Between Computer Security Incident Response Teams (CSIRT) in Indonesia on Blockchain Technology Through Hyperledger Composer and Interplanetary File System (IPFS)," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1390–1402, Jul. 2024, doi: 10.57152/malcom.v4i4.1466.
30. C. Novelli, F. Casolari, P. Hacker, G. Spedicato, and L. Floridi, "Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity," *Computer Law & Security Review*, vol. 55, p. 106066, Nov. 2024, doi: 10.1016/j.clsr.2024.106066.
31. E. Ulate Brenes *et al.*, "La importancia de establecer un marco orientador de política pública para el uso responsable y ético de la inteligencia artificial y su aplicación en Costa Rica," Washington, D. C., Nov. 2021. doi: 10.18235/0003773.
32. Organización para la Cooperación y el Desarrollo Económicos (OECD), *Artificial Intelligence in Society*. OECD, 2019. doi: 10.1787/eedfee77-en.
33. IEEE Standards Association, "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," *IEEE*, no. 1, 2019, Accessed: Jul. 12, 2025. [Online]. Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e.pdf
34. M. Cabrol, C. Pombo, R. Sánchez Ávalos, and N. González Alarcón, "fAIr LAC: Responsible and Widespread Adoption of Artificial Intelligence in Latin America and the Caribbean," Washington, D. C., Jan. 2020. doi: 10.18235/0002169.
35. F. Scrollini, "Automatizar Con Cautela: Datos e Inteligencia Artificial en AmericaLatina," 2018, Accessed: Jul. 12, 2025. [Online]. Available: https://idatosabiertos. org/automatizarcon-cautela-datos-e-inteligencia-artificial-en-america-latina

36. C. M. Del Pozo, C. Gómez Mont, A. V. Martín del Campo Alcocer, and C. Martínez Pinto, "Artificial Intelligence for Social Good in Latin America and the Caribbean: The Regional Landscape and 12 Country Snapshots," Washington, D. C., May 2020. doi: 10.18235/0002393.
37. OECD, "Recommendation of the Council on Artificial Intelligence," no. OECD/LEGAL/0449, 2019, Accessed: Jul. 12, 2025. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
38. L. Latorre, L. De Leo, and E. Rego, "Tech Report: Digital Ethics," Washington, D. C., Apr. 2025. doi: 10.18235/0013487.
39. Oluwatoyin Funmilayo Ayodele and Adesola Oluwatosin Adelaja, "Advancing Cybersecurity Governance: Adaptive Resilience and Strategic Third-Party Risk Management in Financial Services," *World Journal of Advanced Research and Reviews*, vol. 24, no. 2, pp. 293–302, Nov. 2024, doi: 10.30574/wjarr.2024.24.2.3312.
40. H. M. Melaku, "A Dynamic and Adaptive Cybersecurity Governance Framework," *Journal of Cybersecurity and Privacy*, vol. 3, no. 3, pp. 327–350, Jun. 2023, doi: 10.3390/jcp3030017.
41. R. G. Maezo and A. E. Rey, "Boosted CSIRT with AI powered open source framework," in *2023 JNIC Cybersecurity Conference (JNIC)*, IEEE, Jun. 2023, pp. 1–8. doi: 10.23919/JNIC58574.2023.10205787.
42. S. Hays and J. White, "Employing LLMs for Incident Response Planning and Review," Mar. 2024.
43. I. Brass and J. H. Sowell, "Adaptive governance for the Internet of Things: Coping with emerging security risks," *Regul Gov*, vol. 15, no. 4, pp. 1092–1110, Oct. 2021, doi: 10.1111/rego.12343.
44. L. Alevizos and M. Dekker, "Towards an AI-Enhanced Cyber Threat Intelligence Processing Pipeline," *Electronics (Basel)*, vol. 13, no. 11, p. 2021, May 2024, doi: 10.3390/electronics13112021.
45. P. Gao, X. Liu, E. Choi, S. Ma, X. Yang, and D. Song, "ThreatKG: An AI-Powered System for Automated Open-Source Cyber Threat Intelligence Gathering and Management," Oct. 2024.
46. M. A. M. Farzaan, M. C. Ghanem, A. El-Hajjar, and D. N. Ratnayake, "AI-Enabled System for Efficient and Effective Cyber Incident Detection and Response in Cloud Environments," Jan. 2025.
47. R. Pandey, S. Anjimoon, V. Asha, A. Singla, I. Khan, and Z. A. H. Abed, "Developing Robust Cybersecurity Policies and Governance Frameworks in Response to Evolving Legal and Regulatory Landscapes," in *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/OTCON60325.2024.10687438.
48. T. Becher and S. Torka, "Exploring AI-Enabled Cybersecurity Frameworks: Deep-Learning Techniques, GPU Support, and Future Enhancements," Dec. 2024.
49. R. Raman, D. Pattnaik, H. H. Lathabai, C. Kumar, K. Govindan, and P. Nedungadi, "Green and sustainable AI research: an integrated thematic and topic modeling analysis," *J Big Data*, vol. 11, no. 1, p. 55, Apr. 2024, doi: 10.1186/s40537-024-00920-x.
50. United Nations Environment Programme, "Artificial Intelligence (AI) end-to-end: The Environmental Impact of the Full AI Lifecycle Needs to be Comprehensively," 2024, Accessed: Aug. 04, 2025. [Online]. Available: https://wedocs.unep.org/20.500.11822/46288.
51. K. Kalodanis, P. Rizomiliotis, G. Feretzakis, D. Kalles, V. Verykios, and D. Anagnostopoulos, "Enhancing Transparency in Large Language Models to Meet EU AI Act Requirements," in *Proceedings of the 28th Pan-Hellenic Conference on Progress in Computing and Informatics*, New York, NY, USA: ACM, Dec. 2024, pp. 281–286. doi: 10.1145/3716554.3716597.
52. T. M. Kolade, N. T. Aideyan, S. M. Oyekunle, O. S. Ogungbemi, D. L. Dapo-Oyewole, and O. O. Olaniyi, "Artificial Intelligence and Information Governance: Strengthening Global Security, through Compliance Frameworks, and Data Security," *Asian Journal of Research in Computer Science*, vol. 17, no. 12, pp. 36–57, Dec. 2024, doi: 10.9734/ajrcos/2024/v17i12528.
53. H. Choung, P. David, and J. S. Seberger, "A multilevel framework for AI governance," Jul. 2023.
54. A. E. Fernández-Osorio, L. F. Villalba-García, and E. F. Velandia-Pardo, "Gobernanza policéntrica, big data e inteligencia artificial: herramientas para la seguridad ciudadana en Colombia," *Revista Criminalidad*, vol. 66, no. 3, pp. 11–25, Dec. 2024, doi: 10.47741/17943108.658.
55. F. Filgueira, "Desafíos de gobernanza de inteligencia artificial en América Latina. Infraestructura, descolonización y nueva dependencia," *Revista del CLAD Reforma y Democracia*, no. 87, pp. 44–70, Dec. 2023, doi: 10.69733/clad.ryd.n87.a3.
56. A. Abisoye and J. I. Akerele, "A High-Impact Data-Driven Decision-Making Model for Integrating Cutting-Edge Cybersecurity Strategies into Public Policy, Governance, and Organizational Frameworks," *International Journal of Multidisciplinary Research and Growth Evaluation*, vol. 2, no. 1, pp. 623–637, 2021, doi: 10.54660/.IJMRGE.2021.2.1.623-637.
57. Y. Kumar Jain, CA. Dhaarna Singh Rathore, A. Johrawanshi, M. Gupta, D. K. Choudhary, and A. Pandey, "Cybersecurity Frameworks: A Roadmap for Business Resilience," in *2024 International Conference on Cybernation and Computation (CYBERCOM)*, IEEE, Nov. 2024, pp. 102–108. doi: 10.1109/CYBERCOM63683.2024.10803234.
58. A. Alnaffar, "Cybersecurity Resilience Awareness in the Era of AI," *International Journal of Science and Research (IJSR)*, vol. 13, no. 3, pp. 244–245, Mar. 2024, doi: 10.21275/SR24305145928.
59. R. Pandey, S. Anjimoon, V. Asha, A. Singla, I. Khan, and Z. A. H. Abed, "Developing Robust Cybersecurity Policies and Governance Frameworks in Response to Evolving Legal and Regulatory Landscapes," in *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/OTCON60325.2024.10687438.
60. S. Chakraborty, "Towards A Comprehensive Assessment of AI's Environmental Impact," May 2024.
61. N. F. Rozam, H. Jati, and E. Marpanaji, "Comparative Study of Machine Learning Algorithms for DDoS Attack Detection in SDN Networks: A Carbon Emission Analysis with Hyperparameter Optimization Using Bayesian Optimization," *Int J Adv Sci Eng Inf Technol*, vol. 15, no. 2, pp. 530–537, Apr. 2025, doi: 10.18517/ijaseit.15.2.20767.
62. F. Rohde *et al.*, "Broadening the perspective for sustainable AI: Comprehensive sustainability criteria and indicators for AI systems," Nov. 2023.
63. L. Cruz, X. F. Gutierrez, and S. Martínez-Fernández, "Innovating for Tomorrow: The Convergence of SE and Green AI," Jun. 2024.
64. C. Clemm, L. Stobbe, K. Wimalawarne, and J. Druschke, "Towards Green AI: Current status and future research," May 2024.
65. Hazem Salim Abdullah and Salim Abdullah Saleh, "Securing Smart Cities - The Impact of Machine Learning Technology," *Jurnal Komputer, Informasi dan Teknologi*, vol. 5, no. 1, p. 19, Jun. 2025, doi: 10.53697/jkomitek.v5i1.2748.
66. P. Li, Y. Liu, J. Yang, and S. Ren, "Towards Socially and Environmentally Responsible AI," Apr. 2024.
67. J. Dacon, "Are You Worthy of My Trust?: A Socioethical Perspective on the Impacts of Trustworthy AI Systems on the Environment and Human Society," Sep. 2023.
68. M. Hajiesmaili, S. Ren, R. K. Sitaraman, and A. Wierman, "Towards Environmentally Equitable AI," Dec. 2024.
69. A. O. da C. Mazzinghy *et al.*, "Assessment of the Benefits of the ISO/IEC 42001 AI Management System: Insights from Selected Brazilian Logistics Experts: An Empirical Study," *Standards*, vol. 5, no. 2, p. 10, Mar. 2025, doi: 10.3390/standards5020010.
70. A. Reuel *et al.*, "Responsible AI in the Global Context: Maturity Model and Survey," Oct. 2024.
71. R. Alghazzawi, "AI adoption and organizational readiness: boosting accounting efficiency in Jordan," *Journal of Financial Reporting and Accounting*, Dec. 2024, doi: 10.1108/JFRA-08-2024-0570.
72. M. Mäntymäki, M. Minkkinen, T. Birkstedt, and M. Viljanen, "Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance," Jan. 2023.
73. E. Hadley, A. Blatecky, and M. Comfort, "Investigating Algorithm Review Boards for Organizational Responsible Artificial Intelligence Governance," Jan. 2024.

74. L. Freeman, A. Rahman, and F. A. Batarseh, "Enabling Artificial Intelligence Adoption through Assurance," *Soc Sci*, vol. 10, no. 9, p. 322, Aug. 2021, doi: 10.3390/socsci10090322.
75. K. Lam, B. Lange, B. Blili-Hamelin, J. Davidovic, S. Brown, and A. Hasan, "A Framework for Assurance Audits of Algorithmic Systems," May 2024, doi: 10.1145/3630106.3658957.
76. M. Rodriguez, R. A. Popa, F. Flynn, L. Liang, A. Dafoe, and A. Wang, "A Framework for Evaluating Emerging Cyberattack Capabilities of AI," Apr. 2025.
77. K. Madhavan, A. Yazdinejad, F. Zarrinkalam, and A. Dehghantanha, "Quantifying Security Vulnerabilities: A Metric-Driven Security Analysis of Gaps in Current AI Standards," Jul. 2025.
78. N. van Uffelen, L. Lauwaert, M. Coeckelbergh, and O. Kudina, "Towards an Environmental Ethics of Artificial Intelligence," Dec. 2024.
79. W. Y. Leong, "Advancing Smart Energy Networks Through ESG-Driven Engineering: Sustainable Innovations and Practices," *ASM Science Journal*, vol. 20, no. 1, pp. 1–9, Jun. 2025, doi: 10.32802/asmscj.2025.2007.
80. P. Kumar and Professor. V.N.Gupta, "Values That Industry 4.0 Technologies Create For Economic, Environmental, and Social Sustainability," *International Journal of Business and Management Invention*, vol. 13, no. 9, pp. 137–142, Sep. 2024, doi: 10.35629/8028-1309137142.
81. A. Baicoianu, D. Pescaru, D. Sasu, C. Husar, R. Luca, and L. P. Larsen, "Integrated Digital Platform for Marine Energy Management," in *2024 IEEE International Conference on Electrical Systems for Aircraft, Railway, Ship Propulsion and Road Vehicles & International Transportation Electrification Conference (ESARS-ITEC)*, IEEE, Nov. 2024, pp. 1–5. doi: 10.1109/ESARS-ITEC60450.2024.10819756.
82. P. Sharma, A. Kumari, V. Gupta, V. Jindal, and S. Gupta, "Smart Grid Solutions," in *2024 1st International Conference on Sustainable Computing and Integrated Communication in Changing Landscape of AI (ICSCAI)*, IEEE, Jul. 2024, pp. 1–5. doi: 10.1109/ICSCAI61790.2024.10866782.
83. W. Y. Leong, "Advancing Smart Energy Networks Through ESG-Driven Engineering: Sustainable Innovations and Practices," *ASM Science Journal*, vol. 20, no. 1, pp. 1–9, Jun. 2025, doi: 10.32802/asmscj.2025.2007.
84. A. J. M. Rani, A. Srinivasan, S. A. Shiney, B. Kalpana, S. Subramaniam, and V. S. Pandi, "Artificial Intelligence - Enabled Smart Grids: Enhancing Efficiency and Sustainability," in *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, Nov. 2023, pp. 175–180. doi: 10.1109/ICECA58529.2023.10395590.
85. G. Cruz, A. Riobó, M. Pfeifer, and D. Duarte, *IA desde los cimientos: desafíos y oportunidades en el contexto de América Latina y el Caribe*. Inter-American Development Bank, 2024. doi: 10.18235/0013275.
86. Ramya V. J., "Data-Driven Framework for Cloud Storage Security Optimization: Leveraging Predictive Analytics and Machine Learning to Enhance Threat Detection and Incident Response," *Journal of Electrical Systems*, vol. 20, no. 10s, pp. 6646–6653, Jul. 2024, doi: 10.52783/jes.6721.
87. M. R. Naeem *et al.*, "Harnessing AI and analytics to enhance cybersecurity and privacy for collective intelligence systems," *PeerJ Comput Sci*, vol. 10, p. e2264, Sep. 2024, doi: 10.7717/peerj-cs.2264.
88. A. Abisoye and J. I. Akerele, "A Practical Framework for Advancing Cybersecurity, Artificial Intelligence and Technological Ecosystems to Support Regional Economic Development and Innovation," *International Journal of Multidisciplinary Research and Growth Evaluation*, vol. 3, no. 1, pp. 700–713, 2022, doi: 10.54660/.IJMRGE.2022.3.1.700-713.
89. M. Omar, "Integrative Approaches in Cybersecurity and AI," Aug. 2024.
90. H. U. Khan, R. A. Khan, H. S. Alwageed, A. O. Almagrabi, S. Ayouni, and M. Maddeh, "AI-driven cybersecurity framework for software development based on the ANN-ISM paradigm," *Sci Rep*, vol. 15, no. 1, p. 13423, Apr. 2025, doi: 10.1038/s41598-025-97204-y.
91. V. Mayoral-Vilches *et al.*, "CAI: An Open, Bug Bounty-Ready Cybersecurity AI," Apr. 2025.
92. S. Khan, "Cloud Infrastructure and Real-Time Data Analytics in Travel: Advancing Sustainability and Cybersecurity," *Int J Innov Res Sci Eng Technol*, vol. 13, no. 09, Sep. 2024, doi: 10.15680/IJIRSET.2024.1309091.
93. Sheetal Temara, "Harnessing the power of artificial intelligence to enhance next-generation cybersecurity," *World Journal of Advanced Research and Reviews*, vol. 23, no. 2, pp. 797–811, Aug. 2024, doi: 10.30574/wjarr.2024.23.2.2428.
94. Naveen Vemuri, Naresh Thaneeru, and Venkata Manoj Tatikonda, "Adaptive generative AI for dynamic cybersecurity threat detection in enterprises," *International Journal of Science and Research Archive*, vol. 11, no. 1, pp. 2259–2265, Feb. 2024, doi: 10.30574/ijsra.2024.11.1.0313.
95. E. S. V. -, Mr. K. A. -, K. C. -, and A. P. -, "Enhancing Cybersecurity Through AI-Driven Threat Detection: A Transfer Learning Approach," *International Journal For Multidisciplinary Research*, vol. 6, no. 3, May 2024, doi: 10.36948/ijfmr.2024.v06i03.18022.
96. W. Fuertes, F. Reyes, P. Valladares, F. Tapia, T. Toulkeridis, and E. Pérez, "An Integral Model to Provide Reactive and Proactive Services in an Academic CSIRT Based on Business Intelligence," *Systems*, vol. 5, no. 4, p. 52, Nov. 2017, doi: 10.3390/systems5040052.
97. Z. Abdelouahab and F. A. Pestana Júnior, "Information Sharing between CSIRT and IDS," in *Novel Algorithms and Techniques In Telecommunications, Automation and Industrial Electronics*, Dordrecht: Springer Netherlands, 2008, pp. 561–565. doi: 10.1007/978-1-4020-8737-0_101.
98. S. Bhatt, P. K. Manadhata, and L. Zomlot, "The Operational Role of Security Information and Event Management Systems," *IEEE Secur Priv*, vol. 12, no. 5, pp. 35–41, Sep. 2014, doi: 10.1109/MSP.2014.103.
99. J. Nasirov and L. Cheng, "Enhancing Cybersecurity in Climate Technology Transfer: The Interaction of Environmental Law with Digital Security," *International Journal of Digital Law and Governance*, vol. 1, no. 1, pp. 187–206, Apr. 2024, doi: 10.1515/ijdlg-2024-0009.
100. S. Ramos and J. Ellul, "Blockchain for Artificial Intelligence (AI): enhancing compliance with the EU AI Act through distributed ledger technology. A cybersecurity perspective," *International Cybersecurity Law Review*, vol. 5, no. 1, pp. 1–20, Mar. 2024, doi: 10.1365/s43439-023-00107-9.
101. T. Sander and J. Hailpern, "UX Aspects of Threat Information Sharing Platforms," in *Proceedings of the 2nd ACM Workshop on Information Sharing and Collaborative Security*, New York, NY, USA: ACM, Oct. 2015, pp. 51–59. doi: 10.1145/2808128.2808136.
102. M. Lescano-Delgado, "Avances en el uso de inteligencia artificial para la mejora del control y la detección de fraudes en organizaciones," *Revista Científica de Sistemas e Informática*, vol. 3, no. 1, p. e494, Jan. 2023, doi: 10.51252/rcsi.v3i1.494.
103. S. Tariq, M. B. Chhetri, S. Nepal, and C. Paris, "A2C: A Modular Multi-stage Collaborative Decision Framework for Human-AI Teams," Jan. 2024.
104. M. Schmitt and P. Koutroumpis, "Cyber Shadows: Neutralizing Security Threats with AI and Targeted Policy Measures," Jan. 2025, doi: 10.1109/TAI.2025.3527398.
105. 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.
106. L. Ofusori, T. Bokaba, and S. Mhlongo, "Artificial Intelligence in Cybersecurity: A Comprehensive Review and Future Direction," *Applied Artificial Intelligence*, vol. 38, no. 1, Dec. 2024, doi: 10.1080/08839514.2024.2439609.
107. I. H. Sarker, "AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions," Jan. 25, 2021. doi: 10.20944/preprints202101.0457.v1.
108. M. Grazzi, S. Sasso, and R. Kemp, "A Conceptual Framework to Measure Green Innovation in Latin America and the Caribbean," Washington, D. C., Dec. 2019. doi: 10.18235/0002082.

109. T. Könnöla, P. del Río González, J. Carrillo-Hermosilla, and F. J. Díaz López, "Innovación verde en América Latina y el Caribe: marco conceptual," Washington, D. C., May 2023. doi: 10.18235/0004890.
110. O. Eigner *et al.*, "Towards Resilient Artificial Intelligence: Survey and Research Issues," Sep. 2021, doi: 10.1109/CSR51186.2021.9527986.
111. Dirección Nacional de Ciberseguridad de Israel, "Preguntas de orientación para formuladores de políticas cibernéticas: mejores prácticas en ciberseguridad," Washington, D. C., Jul. 2022. doi: 10.18235/0004371.
112. E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," Jan. 2023. doi: 10.6028/NIST.AI.100-1.
113. C. Nott, "Organizational Adaptation to Generative AI in Cybersecurity: A Systematic Review," May 2025.
114. M. Dekker and L. Alevizos, "A Threat-Intelligence Driven Methodology to Incorporate Uncertainty in Cyber Risk Analysis and Enhance Decision Making," Jun. 2023, doi: 10.1002/spy2.333.
115. S. U. Lee *et al.*, "Responsible AI Question Bank: A Comprehensive Tool for AI Risk Assessment," Jan. 2025.
116. J. M. Camacho, A. Couce-Vieira, D. Arroyo, and D. R. Insua, "A Cybersecurity Risk Analysis Framework for Systems with Artificial Intelligence Components," Jan. 2024.
117. V. Moskalenko, V. Kharchenko, and S. Semenov, "Model and Method for Providing Resilience to Resource-Constrained AI-System," *Sensors*, vol. 24, no. 18, p. 5951, Sep. 2024, doi: 10.3390/s24185951.
118. Sudheer Kotilingala, "Human + AI for Enhanced Security: Creating Resilient Systems through Collaboration," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 2, pp. 3079–3089, Apr. 2025, doi: 10.32628/CSEIT25112780.
119. Kumrashan Indranil Iyer, "Cognitive Trust Architecture for Mitigating Agentic AI Threats: Adaptive Reasoning and Resilient Cyber Defense," *Journal of Information Systems Engineering and Management*, vol. 10, no. 47s, pp. 1011–1019, May 2025, doi: 10.52783/jisem.v10i47s.9423.