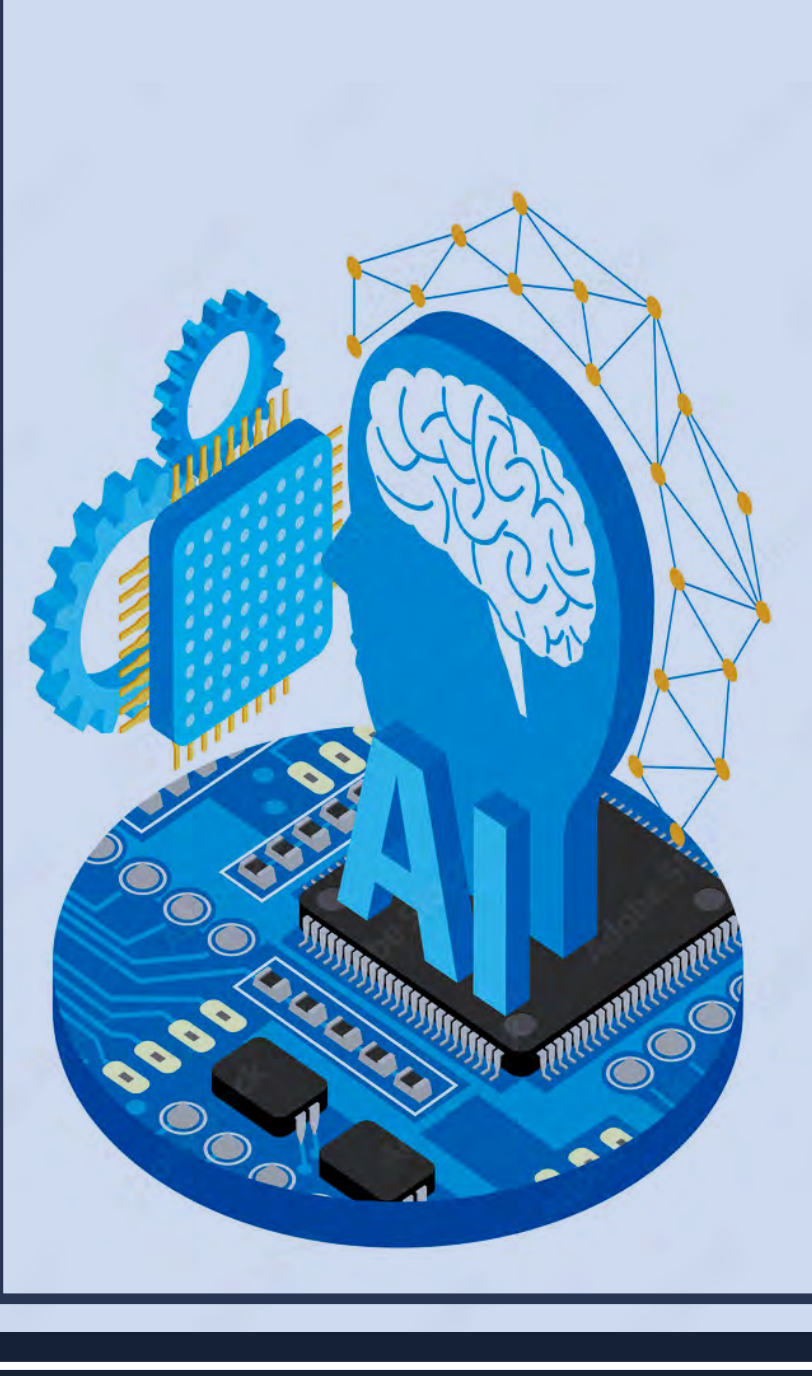
CCN-CERT

BP/30



Aproximación a la Inteligencia Artificial y la ciberseguridad

INFORME DE BUENAS PRÁCTICAS

OCTUBRE **2023**

ccn-cert centro criptológico nacional

CCN centro criptológico nacional

Edita:

Paseo de la Castellana 109, 28046 Madrid
© Centro Criptológico Nacional, 2023

Autores: Carlos Galán Cordero y Carlos M. Galán Pascual
Fecha de edición: octubre de 2023

LIMITACIÓN DE RESPONSABILIDAD

El presente documento se proporciona de acuerdo con los términos en él recogidos, rechazando expresamente cualquier tipo de garantía implícita que se pueda encontrar relacionada. En ningún caso, el Centro Criptológico Nacional puede ser considerado responsable del daño directo, indirecto, fortuito o extraordinario derivado de la utilización de la información y software que se indican incluso cuando se advierta de tal posibilidad.[1]

AVISO LEGAL

Quedan rigurosamente prohibidas, sin la autorización escrita del Centro Criptológico Nacional, bajo las sanciones establecidas en las leyes, la reproducción parcial o total de este documento por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares del mismo mediante alquiler o préstamo públicos.

Índice

Índice

Objeto del documento

El propósito del presente documento es, como su título aventura, realizar un recorrido de aproximación al área de trabajo de dos disciplinas, la Inteligencia Artificial y la Ciberseguridad, que, con orígenes claramente separados en el tiempo, han visto como sus superficies competenciales se han ido acercando a lo largo de los últimos años hasta lo que en la actualidad constituye una nueva actividad práctica, que aglutina el conocimiento y la experiencia previa de ambas: la **Inteligencia Artificial aplicada a la Ciberseguridad**; lo que podríamos bautizar con el nombre ***Artificial Intelligence CyberSecurity, AICS***.

Su carácter introductorio, que lo acerca más a un *survey* que a un tratado científico, persigue que este documento encuentre más fácilmente a los lectores a los que va dirigido: los profesionales o estudiosos de los sistemas de información, sus aplicaciones y sus retos, con especial proyección en los directivos de las organizaciones (públicas o privadas), en sus departamentos gerenciales, incluyendo las áreas técnicas y legales y, por supuesto, en los equipos de ciberseguridad.

Esperamos con este documento, facilitar una primera lectura —que podrá complementarse con otras más específicas como las que aquí referenciamos— y con aquellas con las que los venideros años den noticia de las nuevas y sin duda sorprendentes realidades que se vengan a materializar, de consuno, en la ciberseguridad y la inteligencia artificial.

DISCLAIMER:

En el presente texto, y para facilitar su comprensión, se identifican equipos, instrumentos o material comercial de distintas entidades. Dicha identificación no implica recomendación o aprobación por parte del Centro Criptológico Nacional, ni implica que los materiales o equipos identificados sean necesariamente los mejores disponibles para el propósito señalado en cada caso.

1. Introducción

1.1 Definición de inteligencia artificial (IA) y ciberseguridad

Aunque no existe un consenso que pueda ofrecer una definición universal, podemos decir que la **Inteligencia Artificial (IA)** es un subcampo de la informática que persigue desarrollar sistemas capaces de realizar tareas que, hasta ahora, requieren de la inteligencia humana; tareas que pueden incluir el **aprendizaje** (la adquisición de información y reglas para el uso de la información), el **razonamiento** (usando reglas para llegar a conclusiones aproximadas o definitivas) y la **auto-corrección**[1].

Por ejemplo, la actual Propuesta de Reglamento Europeo en materia de Inteligencia Artificial[2], en trílogos[3] al tiempo de redactarse estas líneas, señala que una definición de IA debe basarse en las principales características funcionales del software, y en particular en su capacidad para generar, en relación con un conjunto concreto de objetivos definidos por seres humanos, contenidos, predicciones, recomendaciones, decisiones

1 ENISA, en su documento ARTIFICIAL INTELLIGENCE AND CYBERSECURITY RESEARCH (Jun, 2023) señala al respecto: «No existe una definición común de la IA (European Commission. Joint Research Centre. AI watch: defining Artificial Intelligence: towards an operational definition and taxonomy of artificial intelligence. Publications Office, 2020. doi:10.2760/382730. (https://data.europa.eu/doi/10.2760/382730).

Aunque no existe una definición común, en las definiciones analizadas se observan algunos puntos en común (cf. CCI5) que pueden considerarse las principales características de la IA: (i) percepción del entorno, incluida la consideración de la complejidad del mundo real; (ii) procesamiento de la información (recogida e interpretación de entradas (en forma de datos); (iii) toma de decisiones (incluidos el razonamiento y el aprendizaje): emprender acciones, realizar tareas (incluidas la adaptación y la reacción a los cambios del entorno) con un cierto nivel de autonomía; (iv) consecución de objetivos específicos».

2 Propuesta de REGLAMENTO DEL PARLAMENTO EUROPEO Y DEL CONSEJO POR EL QUE SE ESTABLECEN NORMAS ARMONIZADAS EN MATERIA DE INTELIGENCIA ARTIFICIAL (LEY DE INTELIGENCIA ARTIFICIAL) Y SE MODIFICAN DETERMINADOS ACTOS LEGISLATIVOS DE LA UNIÓN. (Bruselas, 21.4.2021).

3 Los llamados trílogos son grupos informales que se crean para cada una de las propuestas legislativas y que están integrados por tres miembros: uno de la Comisión, otro del Parlamento y un tercero de la presidencia del Consejo (https://spanish-presidency.consilium.europa.eu/es/noticias/los-trilogos/)

u otra información de salida que influyan en el entorno con el que interactúa el sistema, ya sea en una dimensión física o digital; añadiendo que una definición de «sistema de IA» debe complementarse con una lista de las técnicas y estrategias concretas que se usan en su desarrollo.

Por su parte, el término **Ciberseguridad** se refiere, en esencia, a la práctica de proteger sistemas, redes y programas de ciberataques. Estos ciberataques suelen tener como objetivo acceder, alterar o destruir información valiosa o confidencial, extorsionar a los usuarios para obtener dinero o interrumpir procesos y servicios.

1.2 Breve historia de la IA en la ciberseguridad

La relación entre la IA y la ciberseguridad se ha consolidado con el paso de los años. Inicialmente, los sistemas de ciberseguridad se basaban principalmente en firmas y reglas predefinidas para detectar las amenazas. Sin embargo, con el aumento y la evolución de las ciberamenazas, la necesidad de sistemas más avanzados y adaptativos se ha hecho evidente.

En los años 90 surgieron los primeros intentos de utilizar **técnicas de aprendizaje automático para la detección de intrusiones**[4], pero no fue sino hasta la década de 2010, gracias a los avances en tecnologías de **aprendizaje profundo** y la disponibilidad de grandes conjuntos de datos, cuando la IA comenzó a jugar un papel significativo en materia de ciberseguridad, ofreciendo soluciones más eficientes y precisas para enfrentar unas amenazas en constante evolución.

4 Efectivamente, durante la década de los 90, hubo un creciente interés en el uso de técnicas de aprendizaje automático para la detección de intrusiones, reconociéndose que las técnicas tradicionales basadas en firmas no serían suficientes para detectar ataques nuevos o modificados, conocidos como ataques de día cero. En respuesta a esto, se exploraron técnicas basadas en el comportamiento y el aprendizaje automático, tales como: **IDES (Intrusion Detection Expert System)**: desarrollado a finales de los años 80 y principios de los 90 por SRI International, IDES fue uno de los primeros sistemas de detección de intrusiones basados en el comportamiento. Utilizó técnicas estadísticas para establecer un perfil de la actividad «normal» de un usuario o sistema y luego alertó sobre desviaciones significativas de este comportamiento. **Sistema ADAM**: En 1995, un sistema llamado ADAM (Automated Detection, Analysis, and Measurement) fue propuesto por Lee y Stolfo. Este sistema utilizaba algoritmos de agrupación para detectar actividad anómala en los sistemas de auditoría. **Neural Networks**: Durante los años 90, las redes neuronales también fueron exploradas como una herramienta para la detección de intrusiones. Por ejemplo, en 1998, Ghosh, Schwartzbard y Schatz propusieron usar redes neuronales para detectar comportamientos anómalos en conexiones de red. **Sistema LERAD**: A finales de los años 90, Barbara y Wu desarrollaron LERAD (Learning Rules for Anomaly Detection), que era una técnica basada en el aprendizaje automático para detectar actividades anómalas en conjuntos de datos de auditoría. **KDD Cup 1999 Dataset**: Quizás uno de los esfuerzos más influyentes en el área de detección de intrusiones basada en aprendizaje automático fue el KDD Cup de 1999. Proporcionó un conjunto de datos que contenía una variedad de intrusiones simuladas en un entorno de red. Este conjunto de datos ha sido ampliamente utilizado por la comunidad de investigación para evaluar y comparar diferentes métodos de detección de intrusiones.

PERIODOS	ACTIVIDADES
Inicios **DÉCADAS DE 1960 Y 1970**	Durante las primeras etapas de desarrollo de la informática, la idea de automatizar la seguridad no era una prioridad. Los sistemas informáticos no estaban ampliamente interconectados como en la actualidad, y el propio concepto de ciberseguridad estaba en su infancia. Las primeras aproximaciones a la IA durante este período estaban centradas en temas como el procesamiento del lenguaje natural y en los denominados sistemas expertos, pero no en ciberseguridad.
Nacimiento de la ciberseguridad **DÉCADA DE 1980**	Con el auge de la informática personal y el desarrollo de las primeras redes, surgieron las primeras ciberamenazas. Las herramientas de seguridad se basaban en firmas y patrones conocidos para detectar amenazas, lo que podría considerarse una forma primitiva de aprendizaje automático, aunque la IA como tal aún no se había integrado de manera significativa en la ciberseguridad.
Aproximaciones tempranas a la IA en ciberseguridad **DÉCADA DE 1990**	Los sistemas de detección de intrusiones (IDS) empezaron a incorporar técnicas básicas de aprendizaje automático para identificar patrones de tráfico anómalo. Comenzaron a surgir investigaciones y trabajos académicos explorando el uso de algoritmos de clasificación para mejorar la detección de malware y ataques.
Explosión del Big Data y avance de la IA **DÉCADA DE 2000**	Con la proliferación de Internet y la aparición de amenazas más avanzadas, los grandes conjuntos de datos (logs, tráfico de red, etc.) se convirtieron en una fuente crucial para la ciberseguridad. Las técnicas de IA empezaron a utilizarse para analizar estos grandes volúmenes de datos en busca de patrones sospechosos o comportamientos anómalos.
Aprendizaje profundo y ciberseguridad **DÉCADA DE 2010**	El auge del aprendizaje profundo (especialmente, las redes neuronales convolucionales y recurrentes) encontró aplicaciones en ciberseguridad, como la detección avanzada de malware basada en características y comportamientos en lugar de firmas. Se introdujeron sistemas de respuesta automática y orquestación[5], usando la IA para tomar decisiones en tiempo real frente a amenazas identificadas. Sin embargo, también surgió el concepto de «ataques adversarios»[6] contra modelos de IA, en los cuales los atacantes persiguen engañar o confundir a los modelos de aprendizaje automático.
Presente y futuro **DÉCADA DE 2020 EN ADELANTE**	En la actualidad, la IA es una herramienta esencial en la ciberseguridad, no solo para la detección y respuesta, sino también para la predicción de amenazas. A medida que las ciberamenazas se vuelven más sofisticadas, también lo hace la necesidad de disponer de soluciones de IA más avanzadas y robustas, entre ellas la IA generativa aplicada a ciberseguridad. Las preocupaciones sobre la ética, la privacidad y la responsabilidad en la IA aplicada a la ciberseguridad también están tomando protagonismo, y es probable que estas áreas vean un desarrollo significativo en los próximos años, como lo demuestra el Reglamento Europeo sobre IA que hemos comentado.

5 Security Orchestration, Automation, and Response (SOAR) systems.

6 Adversarial attacks.

1.3 Importancia y relevancia actual del tema

En la actualidad, la ciberseguridad no es solo una cuestión técnica, sino una preocupación global que afecta a instituciones públicas, empresas y personas. Con la digitalización de muchos servicios y la creación de infraestructuras críticas conectadas, la necesidad de asegurar estos sistemas es primordial; todo ello debido a la realidad que plantean las siguientes características de la **transformación digital** de la sociedad:

- **Crecimiento exponencial de datos:** podemos decir que vivimos en la era del *Big Data*. Cada día se generan petabytes de datos, y en este vasto océano de información detectar patrones maliciosos o comportamientos anómalos es una labor en extremo complicada. La IA, a través de algoritmos avanzados, es capaz de analizar grandes volúmenes de datos en tiempo real, identificando posibles amenazas que serían prácticamente imposibles de detectar por métodos manuales o tradicionales.
- **Amenazas en permanente evolución:** las ciberamenazas no son estáticas, sino que están en permanente evolución. Los agentes de las amenazas desarrollan constantemente nuevas técnicas y tácticas para eludir sistemas de seguridad. La IA permite adaptabilidad y aprendizaje continuo, lo que significa que puede «aprender» de las nuevas amenazas y adaptarse en consecuencia, proporcionando una capa adicional de protección.
- **Automatización y respuesta rápida:** ante un ciberataque la respuesta debe ser inmediata. La IA puede automatizar las acciones a tomar, tales como aislar un dispositivo comprometido o bloquear un acceso sospechoso, mucho más rápido de lo que lo haría un humano. Esto reduce el tiempo de exposición y potencialmente minimiza el daño.
- **Reconocimiento de patrones complejos:** la IA es excepcionalmente útil identificando patrones en grandes conjuntos de datos. En el contexto de la ciberseguridad, esto significa que puede identificar comportamientos maliciosos basándose en patrones sutiles que podrían pasar desapercibidos para los sistemas tradicionales.

En la actualidad, la ciberseguridad no es solo una cuestión técnica, sino una preocupación global que afecta a instituciones públicas, empresas y personas

- **Carencia de profesionales en ciberseguridad:** existe una creciente demanda de profesionales de ciberseguridad. La IA puede ayudar a llenar este vacío, asumiendo tareas que requieren análisis y respuesta en tiempo real y permitiendo que los expertos humanos se concentren en las tareas más estratégicas.
- **Coste económico y social:** Las brechas de seguridad pueden derivar en enormes pérdidas económicas, daño a la reputación y, en el caso de las infraestructuras críticas, incluso pueden poner en peligro vidas humanas. La IA aplicada a ciberseguridad no solo protege los activos y datos de una organización, sino que también puede desempeñar un papel crucial en la protección de la sociedad en su conjunto.
- **Desafíos éticos y regulatorios:** a medida que la IA se integra más profundamente en la ciberseguridad (y en la vida, en general), surgen nuevas cuestiones éticas y regulatorias. ¿Quién es responsable si una IA toma una decisión errónea? ¿Cómo garantizamos que la IA actúa de manera justa y no discriminatoria? Estas son preguntas cruciales que destacan la importancia de considerar la IA no solo desde una perspectiva técnica, sino también ética y social[7].

La IA aplicada a ciberseguridad no solo protege los activos y datos de una organización, sino que también puede desempeñar un papel crucial en la protección de la sociedad en su conjunto

Todo ello sin olvidar que, como señala ENISA[8], existen múltiples **actores y agentes de la amenaza** que ya están usando técnicas de IA para desarrollar sus acciones, entre ellos:

- Los **ciberdelincuentes**, cuya motivación principal es el lucro y tenderán a utilizar la IA como herramienta para realizar ataques, pero también para explotar las vulnerabilidades de los sistemas de IA existentes. Por ejemplo, podrían intentar atacar chatbots con IA para sustraer información de tarjetas de crédito u otros datos. También podrían lanzar un ataque de ransomware contra sistemas basados en IA que estén siendo utilizados para la gestión de la cadena de suministro y el almacenamiento.

7 Para más información al respecto, véase «La certificación como mecanismo de control de la inteligencia artificial en Europa» (C. Galán. Instituto Español de Estudios Estratégicos. 2019). https://www.ieee.es/Galerias/fichero/docs_opinion/2019/DIEEEO46_2019CARGAL-InteligenciaArtificial.pdf

8 ENISA- AI Cybersecurity Challenges (2021).

- Las **personas con acceso a información privilegiada**, incluidos los empleados y contratistas que tienen acceso a las redes de una organización, pueden desarrollar acciones dañinas, ser tanto malintencionadas como involuntarias. Los intrusos malintencionados podrían, por ejemplo, tratar de sustraer o sabotear el conjunto de datos de entrenamiento utilizado por los sistemas de IA de la empresa. Por el contrario, otras personas, inadvertidamente, podrían corromper accidentalmente dicho conjunto de datos.
- Los **Estados-nación** o **agentes patrocinados por Estados** que, además de desarrollar formas de aprovechar los sistemas de IA para atacar a otros países (incluidas industrias e infraestructuras críticas), así como de utilizar los sistemas de IA para defender sus propias redes, buscarán activamente vulnerabilidades en los sistemas de IA que puedan explotar. Esto puede ser un medio para causar daño a otro país o para recabar información.
- Los **terroristas**, que persiguen causar daños físicos o incluso la pérdida de vidas humanas, por ejemplo, ciberatacando vehículos sin conductor para utilizarlos como arma.
- Los **hacktivistas**, que en su mayoría tienden a estar motivados ideológicamente, también pueden tratar de piratear los sistemas de IA para demostrar que es algo que se puede hacer. Cada vez hay más grupos preocupados por los peligros potenciales de la IA, y no es inconcebible que puedan piratear un sistema de IA para obtener publicidad.
- También hay **actores no sofisticados**, como los scripts kiddies, que pueden tener motivaciones criminales o ideológicas. Por lo general, se trata de personas no cualificadas que utilizan scripts o programas pre-escritos para atacar sistemas, ya que carecen de los conocimientos necesarios para codificarlos ellos mismos.
- Además de estos tradicionales actores de amenazas, cada vez parece más necesario incluir también a los **competidores** como actores de tales amenazas, ya que algunas empresas podrían estar empezando a evidenciar su intención de atacar a sus rivales para ganar cuota de mercado.

Todo ello configura un panorama de amenazas amplio y extraordinariamente delicado[9].

Los intrusos malintencionados podrían, por ejemplo, tratar de sustraer o sabotear el conjunto de datos de entrenamiento utilizado por los sistemas de IA de la empresa

9 ENISA (2021), *op. cit*.

AL THREAT TAXONOMY

Nefarious activity/abuse (NAA): «Acciones intencionadas dirigidas a sistemas, infraestructuras y redes TIC mediante actos dañinos con el objetivo de sustraer, alterar o destruir un objetivo específico».

Eavesdropping/Interception/Hijacking (EIH): «Acciones destinadas a escuchar, interrumpir o hacerse con el control de una comunicación de terceros sin consentimiento».

Physical Attacks (PA): «Acciones cuyo objetivo es destruir, exponer, alterar, inutilizar, sustraer u obtener acceso no autorizado a activos físicos como infraestructuras, hardware o interconexiones».

Unintentional Damages (UD): Acciones no intencionadas que causan «destrucción, daño o lesión de bienes o personas y provocan un fallo o una reducción de la utilidad».

Failures/Malfunctions (FM): «Funcionamiento parcial o totalmente insuficiente de un activo (hardware o software)».

Outages (OUT): «Interrupciones inesperadas del servicio o disminución de la calidad por debajo de un nivel requerido».

Disasters (DIS): «Accidente repentino o catástrofe natural que causa grandes daños o pérdidas de vidas humanas».

Legal (LEG): «Acciones legales de terceros (contratantes o no), con el fin de prohibir acciones o compensar pérdidas sobre la base de la legislación aplicable».

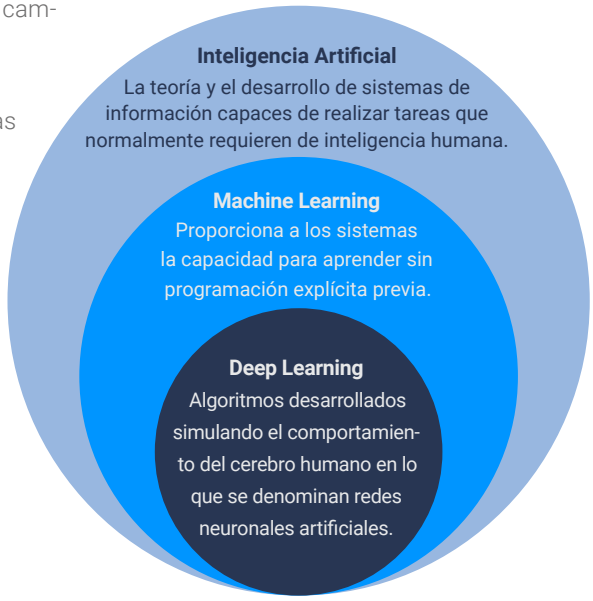
2. Fundamentos de la Inteligencia Artificial

La inteligencia artificial (IA) es un amplio campo de estudio que engloba diversas técnicas y tecnologías. Desde los primeros días de la informática hasta la actualidad, la IA ha pasado de ser un concepto teórico a una herramienta práctica que tiene aplicaciones en innumerables dominios, incluida la ciberseguridad. En este contexto de ciberseguridad, la IA actúa como un **multiplicador de fuerzas**, ofreciendo capacidades avanzadas que van más allá de lo que es posible con métodos tradicionales.

Para comprender cómo la IA beneficia a la ciberseguridad, es esencial familiarizarse con las técnicas y tecnologías específicas que se están aplicando. Estas técnicas abarcan desde el aprendizaje automático y sus subdominios hasta la lógica difusa, pasando por redes neuronales o la más reciente IA generativa. Cada una de estas técnicas posee sus propias características, ventajas, desafíos y aplicaciones dentro de la ciberseguridad, y, conjuntamente, componen un arsenal que las organizaciones pueden utilizar para defenderse contra las crecientes y cambiantes amenazas cibernéticas.

En esta sección, exploraremos algunas de las principales técnicas y tecnologías de IA que se están utilizando en ciberseguridad, proporcionando una base sólida para comprender cómo la IA está revolucionando la forma en que protegemos nuestros sistemas y datos.

Desde el punto de vista descriptivo, conviene situar los diferentes **modelos de IA** dentro de un contexto que permita una mejor comprensión de sus técnicas y características. La figura siguiente desarrolla esta idea.



2.1 Aprendizaje automático (*Machine Learning, ML*)

El **aprendizaje automático** es un método de análisis de datos que automatiza la construcción de modelos analíticos. En lugar de ser programadas explícitamente para realizar una tarea, las máquinas se «entrenan» usando grandes conjuntos de datos y ejecutando algoritmos que les dan la capacidad de *aprender* a realizar la tarea.

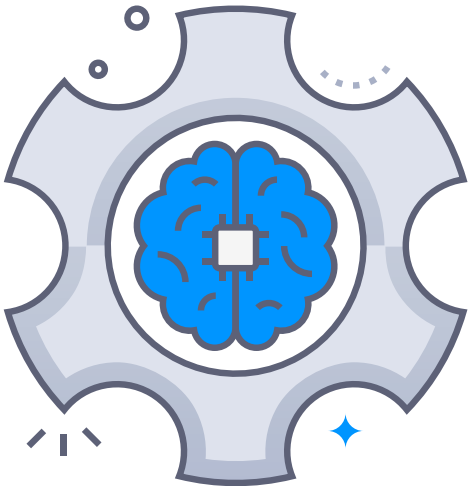
Podemos clasificar las técnicas de aprendizaje automático en los siguientes **tipos**:

Aprendizaje supervisado	Es la técnica más común. En ella, el modelo se entrena usando un conjunto de datos etiquetado, lo que significa que cada ejemplo en el conjunto de datos se acompaña de la «respuesta correcta». Una vez entrenado, el modelo puede empezar a hacer predicciones o tomar decisiones sin intervención humana. Algunos ejemplos de aplicación son la clasificación de correos electrónicos como «spam» o «no spam» o la predicción de precios de viviendas basada en características como el tamaño y la ubicación.
Aprendizaje no supervisado	En este caso, el modelo se entrena con un conjunto de datos sin etiquetas, y su objetivo es descubrir estructuras ocultas en los datos. Las técnicas comunes incluyen la agrupación y la reducción de dimensionalidad. Un ejemplo podría ser segmentar a los clientes en diferentes grupos, en base a sus comportamientos de compra.
Aprendizaje por refuerzo	Es un tipo de aprendizaje donde un agente aprende cómo comportarse en un determinado entorno llevando a cabo ciertas acciones y recibiendo como respuesta recompensas o penalizaciones. Suele usarse en robótica, juegos y navegación.

Dentro de estos tipos se han desarrollado técnicas específicas, tales como los **árboles de decisión** (*decisión trees*), las **máquinas de vectores de soporte** (*support vector machines*), el **clasificador de Bayes** (*Naive Bayes'classifier*), la denominada **K-means clusterig,** el **Hidden Markov Model** o los **algoritmos genéticos**, que examinaremos sumariamente más adelante.

En materia de ciberseguridad, el aprendizaje automático (ML) puede resultar útil en aplicaciones tales como la **detección de amenazas** (puesto que puede analizar grandes volúmenes de datos para identificar patrones de comportamiento anómalo o sospechoso, permitiendo una detección de amenazas más rápida y precisa), el **análisis de código** (puesto que al entrenar modelos de IA sobre conjuntos de datos de malware, el ML puede ayudar a identificar y clasificar nuevas variantes, incluso si no se han observado previamente), el **phishing y la detección de fraudes** (haciendo que los modelos de ML analicen las características de sitios web y correos electrónicos, para determinar si son maliciosos o legítimos).

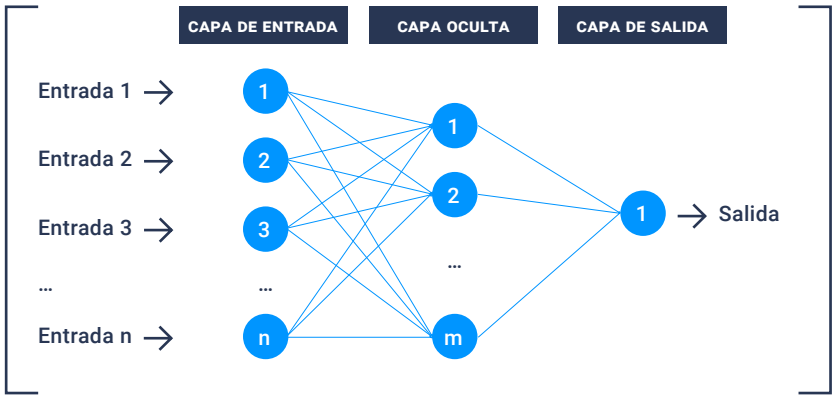
No obstante, construir sistemas de ciberseguridad sustentados en ML tiene exigencias y plantea ciertos desafíos, tales como **disponer de datos de calidad**, puesto que el aprendizaje automático es tan bueno como los datos con los que se entrena, de modo que si los datos están sesgados o son de baja calidad, los modelos resultantes también serán deficientes; o el llamado **overfitting**, puesto que un modelo puede ser demasiado complejo y «memorizar» el conjunto de datos de entrenamiento, en lugar de generalizar sobre datos nuevos y no vistos, o los llamados **ataques adversarios,** mediante los cuales un atacante puede intentar engañar a un modelo de ML presentando datos diseñados específicamente para confundirlo y generar decisiones erróneas.



2.2 Aprendizaje profundo (*Deep Learning, DL*)

El denominado **aprendizaje profundo** es una técnica de aprendizaje automático que utiliza **redes neuronales**[10] con tres o más capas.

Las **redes neuronales artificiales (RNA)** son modelos computacionales inspirados en el funcionamiento de las neuronas en el cerebro humano. Están compuestas por unidades o nodos llamados «neuronas» que se organizan en capas: entrada, oculta(s) y salida. Cada conexión entre neuronas tiene un peso asociado, que se ajusta durante el proceso de entrenamiento.



Estas redes son capaces de aprender patrones y representaciones de datos en niveles de complejidad cada vez mayores, lo que les permite realizar tareas que se consideraban demasiado complejas para los algoritmos de aprendizaje automático tradicionales.

Podemos dividir estas tecnologías en los siguientes **tipos**:

10 En la actualidad existen varios conjuntos de datos de entrenamiento. Los más usados son: KDD Cup99; DEFCON; CTU-13; Spam Base; SMS Spam Collection; CICIDS2017; CICAndMal2017; Android Validation; IoT-23 data set; cada uno de ellos con especiales características para abordar problemáticas específicas.

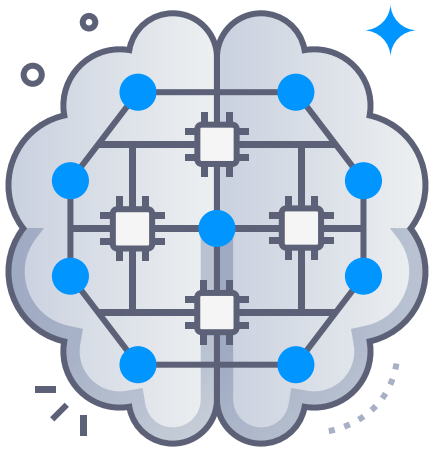
Redes Neuronales Convolucionales (CNNs)	Especialmente útiles para tareas relacionadas con imágenes y video, puesto que pueden identificar y extraer características de imágenes de manera eficiente.
Redes Neuronales Recurrentes (RNNs)	Son especialmente eficaces para trabajar con secuencias de datos, tales como series temporales o texto, debido a su capacidad para «recordar» información previa de la secuencia.
Redes Neuronales de Memoria a Largo Corto Plazo (LSTM)	Se trata de una variante de las RNNs, diseñada para abordar el problema del gradiente desvaneciente[11] y retener información a largo plazo. Al igual que las RNNs, se utilizan para análisis de secuencias, aunque con mayor precisión en secuencias más extensas.
Redes Generativas Adversarias (GANs)	Se trata de un tipo de modelo que utiliza dos redes (una generativa y una discriminativa) que trabajan conjuntamente para generar datos que parezcan auténticos.

En relación con la ciberseguridad, este tipo de técnicas de DL pueden usarse para:

- **Detección de código dañino:** puesto que las redes neuronales pueden entrenarse para identificar software malicioso basado en patrones y características extraídos de archivos. Por ejemplo, una CNN podría analizar el contenido de un archivo binario y determinar si tiene o no características de malware.
- **Análisis de tráfico de red:** las RNNs, dada su aptitud para analizar secuencias, pueden ser útiles para inspeccionar el tráfico de red en busca de patrones anómalos o maliciosos.
- **Detección de phishing:** las CNNs pueden entrenarse para analizar el contenido visual de sitios web y determinar si imitan o replican sitios legítimos, con el propósito de engañar a los usuarios.

11 El problema del gradiente desvaneciente (vanishing gradient) es un obstáculo que surge en el entrenamiento de redes neuronales artificiales tradicionales, en especial las redes neuronales recurrentes (RNNs). Se refiere a la tendencia de los gradientes a disminuir exponencialmente a medida que se retropropagan a través de las capas y a través del tiempo en las RNNs. Cuando los gradientes se acercan a cero, implica que las ponderaciones (o pesos) de las neuronales no se actualizan de manera efectiva durante el proceso de entrenamiento, lo que lleva a un entrenamiento ineficiente o estancado. Las redes LSTM (Long Short-Term Memory) fueron diseñadas específicamente para abordar este problema, así como el problema relacionado del gradiente explotante, donde los gradientes pueden crecer exponencialmente. Las LSTMs logran esto a través de su estructura celular, que incluye puertas de entrada, olvido y salida. Estas puertas, combinadas con una celda de estado, permiten que las LSTMs retengan o descarten información a lo largo de secuencias largas, asegurando que el gradiente se mantenga y se propague adecuadamente a través de la red sin desvanecerse ni explotar. Este diseño especial permite a las LSTMs aprender dependencias a largo plazo en los datos, lo que las hace particularmente útiles para tareas como la traducción automática, el procesamiento del lenguaje natural, la predicción de series temporales y más, donde es crucial recordar información de partes anteriores de una secuencia.

- **Generación de muestras de malware para pruebas:** las GANs pueden usarse para crear muestras de malware. Así, el componente generativo crearía muestras mientras que el componente discriminativo evaluaría su autenticidad, lo que puede ayudar a mejorar la robustez de determinadas herramientas.
- **Análisis de comportamiento:** las redes neuronales pueden aprender patrones de comportamiento de los usuarios o de los sistemas y detectar desviaciones de tales patrones, lo que podría indicar una actividad maliciosa o un compromiso.



No obstante, igual que sucedía en el caso anterior, la utilización de este tipo de técnicas comporta considerar ciertas cuestiones, tales como la **necesidad de disponer de grandes conjuntos de datos**, puesto que el DL requiere frecuentemente grandes cantidades de datos etiquetados para el entrenamiento; o el **tiempo de entrenamiento**, puesto que entrenar modelos de aprendizaje profundo puede ser computacionalmente intensivo; o la **interpretabilidad**, puesto que, a diferencia de otros algoritmos, las redes neuronales profundas actúan frecuentemente como «cajas negras», lo que significa que sus decisiones no son fácilmente interpretables por los humanos.

2.3 Algoritmos de clasificación

Los algoritmos de clasificación son una rama del aprendizaje automático supervisado. Su objetivo principal es tomar una entrada (o instancia) y asignarla a una de las clases predefinidas. En el contexto de la ciberseguridad, estas clases podrían ser, por ejemplo, «malicioso» o «benigno».

Así pues, un algoritmo de clasificación persigue aprender, a partir de un conjunto de datos de entrenamiento, cómo categorizar entradas no vistas en una o más categorías o clases.

Podemos clasificar los algoritmos de clasificación en los siguientes **tipos**, señalando sus aplicaciones en materia de ciberseguridad.

TIPO	CARACTERÍSTICAS	APLICACIÓN EN CIBERSEGURIDAD
Regresión logística	Es un método estadístico para analizar conjuntos de datos en los que hay una o más variables independientes que determinan un resultado. El resultado se mide con una variable dicotómica (sí/no, 1/0, verdadero/falso).	**Determinar si una actividad es maliciosa o no basándose en varias características.**
Máquinas de vectores de soporte (SVM)	Estos algoritmos buscan encontrar el hiperplano que mejor divide un conjunto de datos en clases.	**Clasificación de correos electrónicos como spam o no spam, detección de malware basada en características.**
Árboles de decisión y bosques aleatorios	Los árboles de decisión dividen el conjunto de datos en subconjuntos basados en el valor de las características de entrada. Los bosques aleatorios son una colección de árboles de decisión que participan conjuntamente para ofrecer una predicción final.	**Detección de intrusiones basadas en características como el tipo de protocolo, dirección IP, duración, etc.**
Redes neuronales	Ya estudiadas con anterioridad. Se trata de estructuras inspiradas en el cerebro humano que pueden aprender patrones complejos.	**Detección de malware, análisis de comportamiento de usuarios, detección de anomalías, entre otros.**
K-Vecinos más cercanos (K-NN)	Se trata de un algoritmo que clasifica una entrada en base a cómo sus k vecinos más cercanos están clasificados.	**Detección de actividad maliciosa basada en su similitud con comportamientos conocidos.**
Naive Bayes	Se basa en el teorema de Bayes y asume independencia entre las características. Es especialmente útil cuando la dimensión de los datos es alta.	**Clasificación de correos electrónicos como spam o legítimos, análisis de texto para identificar comunicaciones maliciosas.**

Como siempre, la aplicación práctica de estas técnicas requiere tomar en consideración ciertas cuestiones, tales como la **calidad de los datos** (puesto que, como hemos señalado antes, un modelo es tan bueno como los datos con los que se entrena), el **desbalance de clases** (puesto que en muchos escenarios relativos a ciberseguridad, como la detección de malware, la mayoría de las instancias pueden ser benignas, y solo un pequeño porcentaje maliciosas, lo que puede conducir a un sesgo del modelo si no se maneja adecuadamente), o la **interpretabilidad** (puesto que es importante entender las razones detrás de las decisiones de un modelo, especialmente en un contexto crítico como la ciberseguridad, o la **adaptabilidad** (puesto que si los agentes de las amenazas están evolucionando constantemente sus tácticas, técnicas y procedimientos, es esencial que los modelos de clasificación puedan adaptarse rápidamente a nuevas amenazas).

2.4 IA generativa

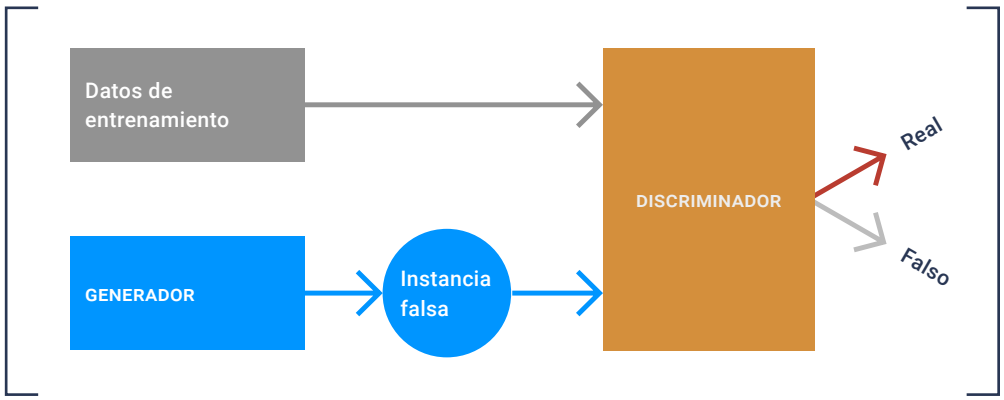
La **inteligencia artificial (IA) generativa** se refiere a un subconjunto de técnicas de aprendizaje automático que tienen como objetivo generar datos nuevos que son similares, pero no idénticos, a los datos con los que fueron entrenadas. A diferencia de las técnicas de aprendizaje automático discriminativo, que aprenden a diferenciar entre diferentes tipos de datos (por ejemplo, clasificar correos electrónicos como «spam» o «no spam»), las técnicas generativas buscan producir datos que se asemejen a los datos de entrada.

Uno de los enfoques más populares y efectivos dentro de la IA generativa son las **Redes Generativas Adversarias (GANs)**, que ya hemos mencionado. Estas redes se basan en dos modelos que trabajan conjuntamente:

1. Un elemento **Generador,** cuyo objetivo es generar datos. Inicialmente, produce datos aleatoriamente, pero con el tiempo y la retroalimentación del discriminador, mejora su capacidad de generar datos que se parezcan a los datos reales.
2. Un elemento **Discriminador**, que examina los datos y trata de distinguir entre datos reales y datos generados por el elemento generador. Proporciona retroalimentación al generador sobre qué tan bien (o mal) lo está haciendo.

Así, el elemento generador intenta producir datos falsos cada vez más convincentes, mientras que el elemento discriminador mejora constantemente su capacidad para detectar estos datos falsos. Con suficiente entrenamiento, el generador puede llegar a producir datos que son casi indistinguibles de los datos reales para humanos y máquinas.

La inteligencia artificial (IA) generativa se refiere a un subconjunto de técnicas de aprendizaje automático que tienen como objetivo generar datos nuevos que son similares, pero no idénticos, a los datos con los que fueron entrenadas

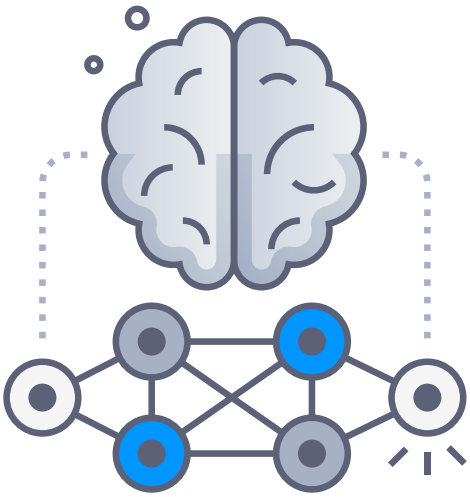


Las aplicaciones de la IA generativa son múltiples: Tratamiento de Imágenes (creación de imágenes artísticas, mejora de la resolución de imágenes, generación de imágenes de objetos o escenas que no existen en la realidad, etc.); **Audio** (creación de música, efectos sonoros o voces sintéticas); **Texto** (generación de textos coherentes, historias, poesía, etc.); **Video** (creación de vídeos sintéticos o modificación de vídeos existentes, como los llamados deepfakes); **Datos sintéticos** (generación de conjuntos de datos para entrenamiento cuando no se dispone de datos reales o son insuficientes); **Diseño y modelado** (generación de diseños para productos, arquitectura o modelado 3D, etc.)

En materia de ciberseguridad, la IA generativa también encuentra varias aplicaciones, tales como las siguientes:

- **Creación de muestras de malware:** las GANs pueden entrenarse para generar variantes de malware que evadan la detección de soluciones de seguridad tradicionales. Aunque esto puede parecer peligroso, los investigadores de seguridad pueden usar esta técnica en entornos controlados para mejorar la robustez de los sistemas de detección.
- **Fortalecimiento de sistemas de detección:** al generar malware o tráfico de red malicioso, los equipos de seguridad pueden usar estas muestras para entrenar y mejorar sus sistemas de detección. En esencia, se enfrenta a la IA contra sí misma para mejorar la detección.

- **Simulación de tráfico de red:** la IA generativa puede simular tráfico de red normal o tráfico de ataque para probar la robustez de una red o sistema. Esto es especialmente útil en la formación de defensores cibernéticos y en la prueba de sistemas de seguridad.
- **Creación de dominios falsos:** en el ámbito de la protección contra amenazas de día cero, las GANs pueden ser usadas para generar dominios falsos que se parezcan a dominios maliciosos reales. Esto ayuda a los sistemas de seguridad a predecir y bloquear dominios que podrían ser utilizados en futuros ataques.
- **Ataques adversarios:** como mencionamos anteriormente, los ataques adversarios involucran la introducción de pequeñas perturbaciones en los datos para engañar a los modelos de aprendizaje automático. Las GANs se pueden utilizar para generar estas perturbaciones de manera eficiente, lo que puede ayudar a los defensores a entender y mitigar estos ataques.
- **Phishing y generación de contenidos maliciosos:** las GANs pueden entrenarse para generar correos electrónicos o páginas web que imiten a las legítimas, haciéndolas herramientas potencialmente útiles para ataques de phishing. Sin embargo, también se pueden usar en la defensa, generando ejemplos de phishing para entrenar sistemas de detección.



3. Aplicaciones de la IA en Ciberseguridad

Como es sabido, la ciberseguridad es una pugna interminable entre atacantes y defensores. Mientras que los atacantes buscan nuevas vulnerabilidades y formas de comprometer sistemas, los defensores persiguen anticiparse a ellos (prevención), detectarlos (detección) y responder a estos ataques (respuesta).

La IA, con su capacidad para procesar grandes cantidades de datos a extraordinaria velocidad y aprender de ellos, ofrece significativas soluciones para los desafíos (actuales y emergentes) de la ciberseguridad.

Tradicionalmente, algunas de las aplicaciones fundamentales de la IA en ciberseguridad son las mostradas en el cuadro siguiente:

Detección y respuesta ante amenazas	Los sistemas basados en IA pueden analizar patrones en el tráfico de red o en los comportamientos de los usuarios para identificar anomalías o actividades sospechosas. Una vez detectadas, la IA puede actuar rápidamente, a menudo más rápido que un equipo humano, para mitigar o neutralizar la amenaza.
Análisis predictivo	La IA puede usar datos históricos para predecir futuras amenazas o vulnerabilidades, permitiendo a las organizaciones prepararse y protegerse proactivamente[12].
Autenticación y gestión de identidad	La IA puede emplear biometría avanzada, comportamiento del usuario y otros factores para autenticar individuos con alta precisión, reduciendo el riesgo de accesos no autorizados.
Protección contra phishing	Al analizar el contenido, las imágenes y los patrones de textos o documentos (por ejemplo, correos electrónicos), la IA puede identificar intentos de phishing con gran precisión, protegiendo a los usuarios de posibles estafas.
Optimización de la configuración de seguridad	La IA puede evaluar configuraciones y políticas de seguridad para identificar posibles debilidades y proponer mejoras.

12 Para más información de lo que se ha dado del análisis anticipativo, en lo que se denomina Modelo del Diamante, puede encontrarse en la Guía CCN-STIC 425 Ciclo de Inteligencia y Análisis de Intrusiones. https://www.ccn-cert.cni.es/series-ccn-stic/guias-de-acceso-publico-ccn-stic/1093-ccn-stic-425-ciclo-de-inteligencia-y-analisis-de-intrusiones/file.html

A medida que nos adentramos en este apartado, exploraremos en detalle cómo la IA se integra en estos y otros dominios de la ciberseguridad, su potencial y, por supuesto, las consideraciones éticas y de privacidad asociadas a su uso.

3.1 Detección de amenazas y análisis de comportamiento

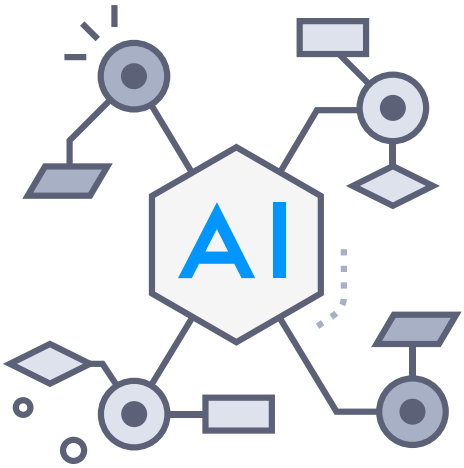
La detección de amenazas y el análisis de comportamiento son esenciales para identificar y responder a ciberataques en tiempo real. Con la implementación de la IA en estos campos, la ciberseguridad ha experimentado una mejora notable en la eficiencia y precisión de la detección.

La cantidad de datos que las organizaciones (públicas o privadas) procesan diariamente es inmensa. La detección manual de amenazas en tales volúmenes es prácticamente imposible. Por su parte, los ciberataques modernos emplean frecuentemente tácticas sigilosas, como el movimiento lateral y la persistencia de perfil bajo, dificultando su detección con los métodos tradicionales.

Así pues, en lugar de basarse solo en firmas conocidas de malware, la IA se centra en **patrones de comportamiento anómalo**. Esto permite detectar amenazas previamente desconocidas o variantes de malware que hubieren sido ligeramente modificadas. Analizando el comportamiento del usuario y del sistema, la IA puede identificar actividades inusuales, como el acceso a archivos en horas extrañas o la transferencia inusual de grandes cantidades de datos.

La **detección de amenazas basada en el comportamiento** ha experimentado un rápido crecimiento en popularidad y adopción, y han ido apareciendo distintas herramientas y sistemas, tanto comerciales como de código abierto, especializadas en este enfoque. Se enumeran seguidamente algunas de las herramientas más conocidas:

- **Darktrace**[13]**:** Darktrace utiliza aprendizaje automático y algoritmos de IA para detectar, responder y mitigar amenazas en tiempo real basándose en patrones de comportamiento anómalo. La herramienta es conocida por su «*Enterprise Immune System*», que aprende y establece lo que puede entenderse como una «situación de normalidad» en la red y luego identifica desviaciones de esta norma.
- **Vectra**[14]**:** Vectra ofrece detección de amenazas en tiempo real utilizando IA. Centra su enfoque en detectar comportamientos maliciosos dentro del tráfico de red y proporciona una visión detallada de la cadena de ataque en curso, permitiendo a los equipos de seguridad responder rápidamente.
- **CrowdStrike Falcon**[15]**:** CrowdStrike es firma muy conocida por sus soluciones de protección de endpoints. Su plataforma Falcon utiliza técnicas basadas en comportamiento para detectar y prevenir amenazas que otros sistemas basados en firmas podrían obviar.
- **Cylance**[16]**:** CylancePROTECT es una solución de protección de endpoints que utiliza modelos de IA para identificar y bloquear malware basándose en sus características y comportamientos, en lugar de firmas conocidas.
- **Gurucul**[17]**:** Ofrece soluciones de análisis de comportamiento de usuarios y entidades (UEBA) que utilizan algoritmos de aprendizaje automático para detectar amenazas internas, fraudes y accesos no autorizados.
- **Wazuh**[18]**:** Se trata de una plataforma de código abierto para la detección de amenazas, gestión de vulnerabilidades y monitorización de la integridad. Utiliza reglas y decodificadores para analizar eventos de seguridad y detectar comportamientos anómalos.
- **Snort**[19]**:** Aunque es más conocido como un sistema de detección y prevención de intrusiones (IDPS), Snort ha evolucionado para incorporar capacidades basadas en comportamiento. La comunidad Snort desarrolla y comparte nuevas reglas que pueden detectar comportamientos anómalos.



13 https://es.darktrace.com/
14 www.vectra.ai
15 www.crowdstrike.com
16 www.cylance.com
17 www.gurucul.com
18 www.wazuh.com
19 www.snort.org

- **ELK Stack (Elasticsearch, Logstash, Kibana)**[20]**:** Aunque ELK en sí misma no es una herramienta de detección basada en comportamiento, se puede configurar con complementos y reglas específicas para realizar análisis de comportamiento de logs y eventos.

Por su parte, los sistemas de IA que operan bajo el modelo de **Aprendizaje Automático para el Análisis de Comportamiento** se entrenan utilizando grandes conjuntos de datos de comportamientos, tanto legítimos como maliciosos. A través del aprendizaje supervisado, la IA puede aprender a clasificar y detectar actividad anómala. Así, con el tiempo y a medida que se procesan más datos, estos sistemas pueden mejorar su precisión mediante aprendizaje no supervisado y aprendizaje por refuerzo.

Los sistemas de IA que operan bajo el modelo de Aprendizaje Automático para el Análisis de Comportamiento se entrenan utilizando grandes conjuntos de datos de comportamientos, tanto legítimos como maliciosos

Muchas herramientas modernas de ciberseguridad han incorporado el aprendizaje automático (ML) en sus capacidades para mejorar la detección y respuesta ante amenazas. Estas herramientas usan ML para aprender y adaptarse a nuevas amenazas al estudiar patrones y comportamientos en los datos. Además de las herramientas señaladas, se enumeran seguidamente algunos de las soluciones más conocidas:

- **Endgame**[21]**:** esta plataforma utiliza ML para la protección de endpoints, detección de amenazas y respuesta. Su capacidad de ML se centra en detectar técnicas y tácticas de ataque sin depender únicamente de firmas.
- **PatternEx**[22]**:** es una solución de análisis de comportamiento de usuarios y entidades (UEBA) que utiliza aprendizaje automático. Analiza grandes volúmenes de datos para identificar patrones que sugieren actividades maliciosas.
- **SentinelOne**[23]**:** es una solución de protección de endpoints que utiliza aprendizaje automático para detectar, clasificar y responder a comportamientos maliciosos y anómalos.
- **Kaspersky Machine Learning for Anomaly Detection (MLAD)**[24]**:** diseñado para sistemas industriales, MLAD de Kaspersky utiliza aprendizaje automático para detectar desviaciones en la operación de máquinas industriales.

20 www.elastic.co

21 (Adquirido por Elastic): https://www.elastic.co

22 https://www.patternex.com

23 https://www.sentinelone.com

24 https://www.kaspersky.com

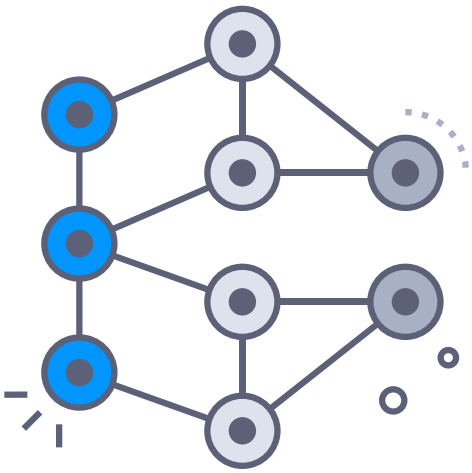
- **Splunk**[25]**:** si bien Splunk es principalmente una herramienta de análisis de datos y SIEM (gestión de eventos e información de seguridad), tiene capacidades que permiten a los usuarios implementar modelos de aprendizaje automático para identificar patrones y anomalías en grandes volúmenes de datos.

Por su parte, las **redes neuronales,** especialmente, las profundas (Deep Learning), se han mostrado eficaces para detectar patrones en grandes conjuntos de datos, pudiéndose utilizar para identificar malware en archivos basándose en sus características, detectar ataques DDoS basándose en patrones de tráfico o identificar intentos de phishing a través del análisis de texto y contenido.

Además de las firmas comerciales antes reseñadas, que también están trabajando en redes neuronales, se enumeran seguidamente alguna de las herramientas más conocidas que hacen uso de estas técnicas:

- **Deep Instinct**[26]**:** esta empresa utiliza redes neuronales de aprendizaje profundo para prevenir y detectar malware en tiempo real, ofreciendo soluciones tanto para endpoints como para dispositivos móviles.
- **SparkCognition**[27]**:** esta empresa ofrece DeepArmor, una solución que utiliza redes neuronales para ofrecer protección contra amenazas en tiempo real.
- **NVIDIA**[28]**:** Aunque no es una herramienta de seguridad per se, NVIDIA ofrece plataformas y bibliotecas como CUDA y cuDNN que aceleran la computación en redes neuronales.

Por otro lado, son múltiples las herramientas que, trayendo causa de conceptos más tradicionales, han sabido incorporar mecanismos de IA a sus tecnologías, al objeto de hacerlas más eficaces.



25 https://www.splunk.com

26 https://www.deepinstinct.com

27 https://www.sparkcognition.com

28 https://www.nvidia.com

Algunas de tales **aplicaciones prácticas tradicionales** son las siguientes:

Sistemas de detección y prevención de intrusiones (IDPS)	Utilizando la IA, estos sistemas pueden detectar y bloquear tráfico malicioso en tiempo real con mayor precisión. Un sistema de detección y prevención de intrusiones (IDPS) es fundamental para detectar y responder a actividades maliciosas en una red o sistema. La integración de la inteligencia artificial (IA) en estos sistemas ha mejorado significativamente su capacidad para identificar y reaccionar a amenazas en tiempo real. Algunos ejemplos son: • **Darktrace** (https://www.darktrace.com)**:** como hemos mencionado antes, Darktrace es conocido por su enfoque basado en IA para la detección y prevención de amenazas. Su tecnología de «Enterprise Immune System» utiliza aprendizaje automático para detectar comportamientos anómalos en tiempo real. • **Vectra** (https://www.vectra.ai)**:** Vectra Cognito utiliza IA para detectar y priorizar automáticamente comportamientos anómalos en tiempo real para descubrir ataques activos y amenazas internas. • **Cisco Stealthwatch** (https://www.cisco.com)**:** aunque no se trata de un IDPS en el sentido tradicional, Stealthwatch utiliza aprendizaje automático para detectar comportamientos anómalos en la red y se integra con otras soluciones de Cisco para proporcionar capacidades de prevención. • **Lastline** (https://www.lastline.com)**:** ofrece soluciones que utilizan técnicas de IA, como aprendizaje automático, para detectar y responder a amenazas avanzadas, evasivas y de día cero. • **Awake Security** (https://www.awakesecurity.com)**:** su plataforma utiliza IA para analizar el tráfico de red y detectar amenazas. Puede identificar comportamientos maliciosos y de riesgo sin depender de firmas o conocimientos previos. • **Fortinet** (https://www.fortinet.com)**:** aunque Fortinet ofrece una variedad de soluciones de seguridad, su FortiGate con funcionalidad IDPS incorporada también ha incluido IA para mejorar la detección de amenazas.

Herramientas de análisis forense	La IA puede acelerar las investigaciones después de un incidente de seguridad, identificando rápidamente los indicadores de compromiso y trazando el camino de un atacante. El análisis forense digital, especialmente cuando se aplica a incidentes de seguridad (véase la disciplina DEFIR: Digital Forensics and Incident Response), puede generar grandes cantidades de datos para investigar. La inteligencia artificial y, en concreto, el aprendizaje automático (ML) están desarrollando un papel significativo en esta área, ayudando a identificar patrones, realizar análisis más rápidos y obtener insights más precisos. Se enumeran algunas de las herramientas más conocidas que integran IA en sus capacidades de análisis forense: • **Autopsy** (https://www.sleuthkit.org/autopsy/): aunque es principalmente una herramienta de análisis forense digital, tiene módulos y plugins que pueden aprovechar capacidades basadas en IA para analizar datos y buscar patrones específicos. • **Cellebrite** (https://www.cellebrite.com): conocida por sus soluciones de análisis forense de dispositivos móviles, Cellebrite utiliza IA para ayudar en la identificación y categorización de datos relevantes en dispositivos móviles. • **Brainspace** (https://www.brainspace.com): esta es una plataforma de análisis y visualización que utiliza aprendizaje automático para asistir en investigaciones, revisiones de documentos y análisis de datos. Es utilizado en investigaciones legales, pero también puede ser aplicado en análisis forense digital. • **Cyber Triage** (https://www.cybertriage.com): Asociado con Autopsy, esta herramienta utiliza técnicas de IA para la evaluación rápida de sistemas comprometidos, buscando evidencia de actividad maliciosa. • **Endgame** (parte de Elastic; https://www.elastic.co): su plataforma proporciona capacidades de respuesta a incidentes y amenazas, y utiliza técnicas de ML para analizar datos y detectar actividades maliciosas. • **ReversingLabs** (https://www.reversinglabs.com): ofrece soluciones para el análisis de archivos y artefactos maliciosos con capacidades basadas en IA para identificar, clasificar y desglosar amenazas. Estas herramientas, combinadas con la experiencia humana, pueden proporcionar análisis forenses más rápidos y precisos, lo que es crucial en respuesta a incidentes e investigaciones.

Sistemas de respuesta automatizada	Una vez que se detecta una amenaza, la IA puede iniciar acciones predefinidas para contener o mitigar el ataque, como aislar un sistema comprometido o bloquear una dirección IP sospechosa. La respuesta automatizada, a menudo combinada con la detección de amenazas, es un componente crucial en la seguridad moderna. Al utilizar inteligencia artificial (IA), estos sistemas pueden tomar decisiones en tiempo real para contener, mitigar o neutralizar amenazas sin intervención humana inmediata. Además de las firmas comerciales ya citadas, se enumeran seguidamente algunas de las herramientas y soluciones que integran IA para ofrecer capacidades de respuesta automatizada, más conocidas: • **Darktrace Antigena** (https://www.darktrace.com): Antigena es una extensión del sistema de detección basado en IA de Darktrace, que tiene la capacidad de tomar acciones automáticas en respuesta a amenazas detectadas, como bloquear conexiones o poner en cuarentena dispositivos. • **Palo Alto Networks - Cortex XDR** (https://www.paloaltonetworks.com): esta plataforma detecta amenazas y automatiza la respuesta. Utiliza técnicas de aprendizaje automático para identificar amenazas y puede realizar acciones como bloquear procesos maliciosos o actualizar reglas de firewall automáticamente. • **FireEye Helix** (https://www.fireeye.com): es una plataforma de seguridad que utiliza IA para detectar amenazas y automatizar respuestas. Puede integrarse con una variedad de herramientas y sistemas para ejecutar acciones de respuesta. • **IBM Resilient** (https://www.ibm.com): es una plataforma de respuesta a incidentes que, combinada con Watson, el sistema de IA de IBM, puede ofrecer recomendaciones y automatizar acciones en respuesta a incidentes de seguridad. • **Fortinet FortiResponder** (https://www.fortinet.com): es una solución de respuesta a incidentes que se integra con otros productos de Fortinet para ofrecer capacidades automatizadas basadas en reglas. Aunque la respuesta se basa principalmente en reglas definidas, la detección y los insights pueden ser potenciados por técnicas de IA. Como parece lógico suponer, la automatización debe utilizarse con cuidado. Una configuración incorrecta o la falta de supervisión adecuada pueden llevar a respuestas no deseadas que afecten negativamente las operaciones. La IA y la automatización deben ser vistas como herramientas que complementan, pero no reemplazan, a los expertos en seguridad humana.

La utilización de estas técnicas también plantea desafíos. Aunque la IA puede mejorar la precisión, todavía existe el riesgo de **falsos positivos**, que en alto número puede llevar a la fatiga del equipo de seguridad y derivar en posibles omisiones.

Por su parte, y al igual que los sistemas basados en firmas, los atacantes están desarrollando técnicas para **evadir la detección basada en la IA**, como el llamado *envenenamiento de datos* o la manipulación de modelos.

Efectivamente, la evasión de sistemas basados en IA es una táctica empleada por los agentes de las amenazas para evitar ser detectados por sistemas de seguridad que utilizan técnicas de inteligencia artificial o aprendizaje automático. Estos métodos se basan en la comprensión y explotación de las debilidades inherentes o los sesgos de los modelos de aprendizaje automático. Algunas herramientas y técnicas son diseñadas específicamente para este propósito, mientras que otras se han adaptado para poder evadir la IA.

Se muestran seguidamente algunos de los conceptos o herramientas y recursos que han sido utilizados o estudiados en relación con la **evasión de la IA**:

ADVERSARIAL MACHINE LEARNING	Se trata más bien de una categoría que de una herramienta específica. Los ataques adversarios buscan introducir pequeñas perturbaciones en los datos de entrada para engañar a los modelos de aprendizaje automático.
CLEVERHANS[29]	Es una biblioteca de software que proporciona herramientas para probar la robustez de los modelos de aprendizaje automático frente a ataques adversarios.
DEEP-PWNING[30]	Es una herramienta de evaluación de seguridad para asistir en el análisis de sistemas que utilizan aprendizaje profundo. Puede ser utilizada para evaluar la resistencia de modelos frente a modificaciones adversarias.
GAN (REDES GENERATIVAS ADVERSARIAS)	Aunque no son herramientas de evasión per se, las GANs pueden ser usadas para generar datos que engañen a sistemas de IA. Como hemos señalado con anterioridad, estas redes se componen de dos elementos: un generador que crea imágenes y un discriminador que intenta distinguir entre imágenes reales y generadas.
FGSM (FAST GRADIENT SIGN METHOD)	Es una técnica de ataque adversario que introduce perturbaciones en los datos de entrada para confundir al modelo de aprendizaje automático.

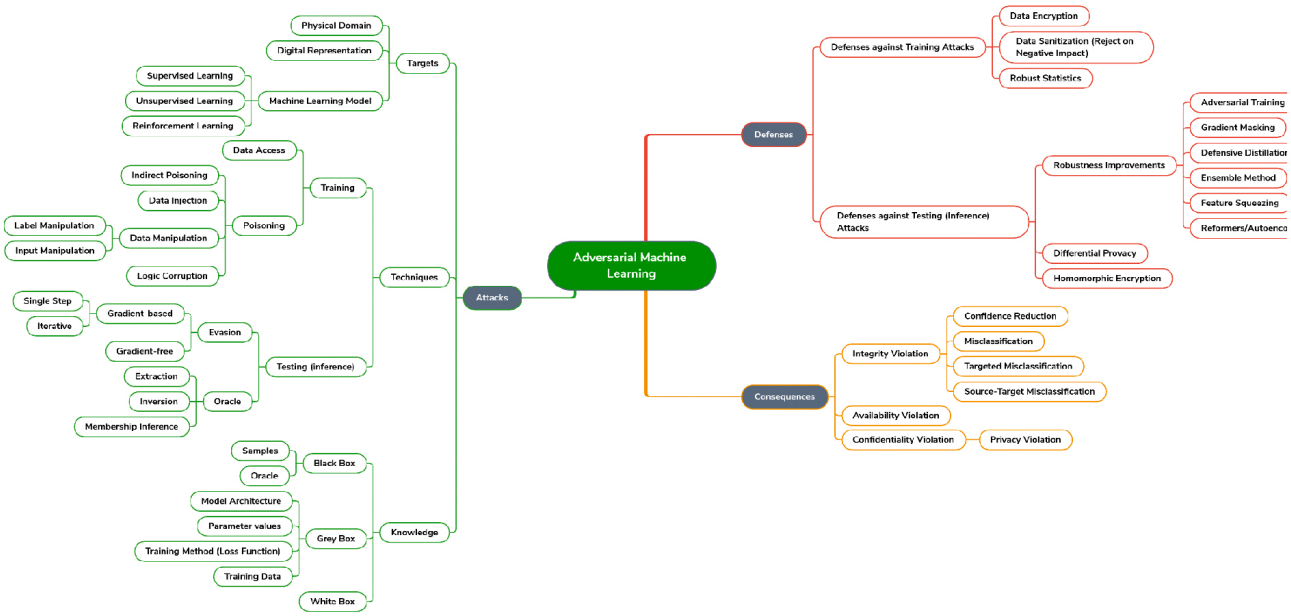
Estos retos de seguridad contemplan, como hemos visto, el enorme potencial de la manipulación adversaria de los datos de entrenamiento y la explotación adversaria de las sensibilidades del modelo, al objeto de perturbar el rendimiento de la clasificación y la regresión de ML.

29 https://github.com/tensorflow/cleverhans

30 https://github.com/cchio/deep-pwning

Así, se entiende por AML (*Adversarial Machine Learning*) el diseño de algoritmos de ML que puedan resistir los retos de seguridad, el estudio de las capacidades de los atacantes y la comprensión de las consecuencias de los ataques. Puesto que los ataques son lanzados por adversarios con intenciones dañinas, la seguridad del ML debe referirse a las defensas destinadas a prevenir o mitigar las consecuencias de dichos ataques. Aunque los componentes de ML también pueden verse afectados negativamente por diversos factores no intencionados, como fallos de diseño o sesgos en los datos, estos factores no son ataques adversarios intencionados, y no entran dentro del ámbito de la seguridad abordado por la literatura sobre el AML.

Por su indudable interés, reproducimos la **Taxonomía de Ataques, Defensas y Consecuencias en AML**, del National Institute Of Standards and Technology (NIST)[31].



Finalmente, parece claro que la adopción de la IA para la detección de amenazas y el análisis de comportamiento está en vías de transformar la capacidad de las organizaciones para defenderse contra ciberataques. Sin embargo, al igual que con cualquier herramienta, es esencial utilizarla en conjunto con otras técnicas y enfoques de ciberseguridad para garantizar una defensa integral.

31 NIST - Draft NISTIR 8269 - A Taxonomy and Terminology of Adversarial Machine Learning (2019).

3.2 Respuesta automática y orquestación

La denominada **respuesta automática y orquestación** (*Security Orchestration, Automation and Response - SOAR*), se refiere a la capacidad de un sistema de seguridad para detectar automáticamente una amenaza o vulnerabilidad y responder a ella sin intervención humana, a menudo coordinando múltiples sistemas y herramientas en el proceso.

Este modelo, por tanto, exige el concurso de tres componentes: la **orquestación,** entendida como la coordinación y gestión integrada de herramientas y sistemas de seguridad, permitiendo que trabajen juntos de manera armónica[32]; la **automatización,** entendida como la capacidad de realizar tareas específicas sin intervención humana[33]; y la **respuesta,** referida a las acciones tomadas en respuesta a un evento de seguridad, y que pueden ser automáticas (por ejemplo, bloquear una IP) o podrían requerir intervención humana (como investigar una posible intrusión).

La utilización de herramientas configuradas bajo el modelo SOAR es particularmente útil, puesto que, dada la velocidad de propagación de las amenazas, la capacidad de responder automáticamente a las mismas puede ser crucial para minimizar el daño.

Además, la automatización permite a los equipos de seguridad centrarse en tareas de mayor valor o en amenazas más sofisticadas, dejando tareas repetitivas o rutinarias a las soluciones automatizadas, eliminando el factor humano del tratamiento de estas últimas y, en consecuencia, reduciendo el riesgo de errores o inconsistencias.

Finalmente, los sistemas SOAR pueden manejar un gran volumen de alertas y eventos, muchos de los cuales serían abrumadores para un equipo humano.

La denominada respuesta automática y orquestación (*Security Orchestration, Automation and Response - SOAR*), se refiere a la capacidad de un sistema de seguridad para detectar automáticamente una amenaza o vulnerabilidad y responder a ella sin intervención humana, a menudo coordinando múltiples sistemas y herramientas en el proceso

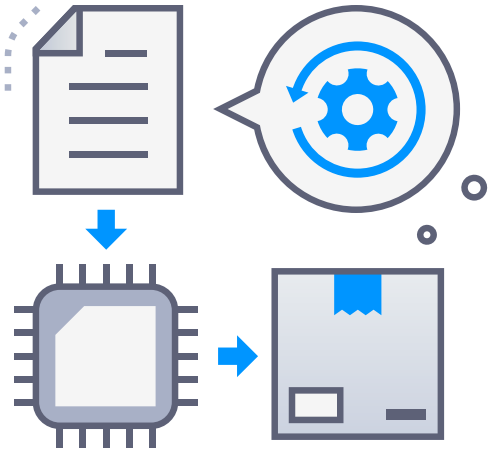
32 Por ejemplo, si un sistema detecta un código dañino (malware), la orquestación podría asegurarse de que la información sea compartida con todas las herramientas pertinentes para el subsiguiente análisis y respuesta.

33 Que podrían contemplar bloquear una IP maliciosa, deshabilitar cuentas de usuario comprometidas o incluso parchear un software vulnerable.

Pese a sus ventajas, el uso de herramientas SOAR plantea igualmente desafíos: por ejemplo, una respuesta automática mal configurada puede causar más problemas de los que resuelve, como bloquear tráfico legítimo o ignorar amenazas verdaderas; o la sobredependencia de tales sistemas sin revisión o supervisión humana, que puede conllevar la falta de detección de amenazas más sofisticadas o complejas.

Las herramientas SOAR suelen utilizarse en los siguientes **escenarios**:

- **Respuesta a Incidentes:** si un sistema detecta un comportamiento anómalo, como un aumento inusual del tráfico hacia un destino específico, puede bloquear automáticamente esa actividad y alertar al equipo de seguridad.
- **Integración de Herramientas:** Al integrar múltiples herramientas (como sistemas de detección de intrusiones, firewalls y soluciones de endpoint), la orquestación permite una respuesta más holística y coordinada a las amenazas.
- **Automatización de Workflows:** Por ejemplo, ante la detección de un software vulnerable, un sistema SOAR podría iniciar automáticamente un proceso de parcheo o actualización.



Se relacionan seguidamente algunas de las herramientas SOAR más conocidas:

- **Splunk Phantom** (https://www.splunk.com/en_us/software/splunk-security-orchestration.html)
- **Siemplify** (https://www.siemplify.co/)
- **Palo Alto Networks - Cortex XSOAR** (https://www.paloaltonetworks.com/cortex/xsoar)
- **IBM Resilient** (https://www.ibm.com/security/incident-response/resilient-soar-platform)
- **CyberSponse** (https://www.cybersponse.com/)

3.3 Predicción de amenazas

El concepto de «predicción de amenazas» es una evolución de la tradicional detección de amenazas y representa un cambio en el enfoque de la ciberseguridad, desde una postura reactiva a una proactiva. Por tanto, la predicción de amenazas se refiere al proceso de anticipar y detectar posibles ciberataques o vulnerabilidades antes de que se materialicen, utilizando para ello análisis avanzados y, por la parte que ahora nos interesa, técnicas de inteligencia artificial para identificar patrones y señales que sugieren un ataque inminente o el surgimiento de nuevas vulnerabilidades.

El concepto de «predicción de amenazas» es una evolución de la tradicional detección de amenazas y representa un cambio en el enfoque de la ciberseguridad, desde una postura reactiva a una proactiva

Las ventajas de este modelo son evidentes: parece claro que, al predecir una amenaza antes de que ocurra, las organizaciones tienen tiempo para prepararse, fortalecer sus defensas y reducir el impacto potencial del ataque. Además, la predicción permite a las organizaciones centrar sus esfuerzos y recursos en aquellas amenazas que son más probables que ocurran, en lugar de dispersarlos en una amplia gama de posibles escenarios y, finalmente, al anticipar amenazas y actuar sobre ellas de manera proactiva, las organizaciones pueden mejorar su postura general de seguridad.

Esencialmente, los **métodos de predicción** usados por las herramientas de este modelo son los tres siguientes:

Análisis de tendencias	Al analizar tendencias pasadas, las organizaciones pueden prever tipos de amenazas que podrían surgir en el futuro.
Inteligencia de amenazas	Implica recopilar y analizar información sobre amenazas existentes y emergentes de diversas fuentes, como feeds de inteligencia, informes de investigadores y datos de eventos de seguridad.
Modelos predictivos	Utiliza algoritmos y modelos matemáticos, a menudo potenciados por IA y aprendizaje automático, para analizar grandes conjuntos de datos e identificar patrones que sugieran una amenaza inminente.

Como en los modelos anteriores, este concepto también presenta significativos desafíos.

Efectivamente, la predicción de amenazas, especialmente cuando se basa en modelos predictivos, puede llevar a **falsos positivos**, lo que puede desviar recursos y atención de otras áreas críticas. Por otro lado, la creación y mantenimiento de modelos predictivos, especialmente aquellos que utilizan técnicas avanzadas de IA, pueden ser tareas **complejas** y requerir personal especializado. Finalmente, como siempre, la precisión de las predicciones depende en gran medida de la **calidad, relevancia y actualización de los datos de entrada**.

La predicción de amenazas, especialmente cuando se basa en modelos predictivos, puede llevar a falsos positivos, lo que puede desviar recursos y atención de otras áreas críticas

Las aplicaciones de este modelo en ciberseguridad se vienen centrando en la **predicción de código dañino** (basándose en las características de los malwares conocidos, se pueden prever nuevas variantes o evoluciones del malware), la **predicción de ataques de phishing** (al analizar patrones en campañas de phishing previas, se puede anticipar futuros ataques o identificar dominios sospechosos) y en la **predicción de ataques DDoS** (al observar patrones de tráfico y otras señales, es posible anticipar un ataque DDoS antes de que ocurra).

Se muestran seguidamente algunos de los ejemplos más conocidos de herramientas que han utilizado este modelo, en sus distintas variantes, desde el mero análisis estadístico tradicional hasta el aprendizaje automático y la inteligencia artificial, para anticipar amenazas antes de que ocurran. Algunas de ellas ya se han mencionado con anterioridad.

- **Darktrace Antigena** (https://www.darktrace.com/en/products/antigena/)
- **Recorded Future** (https://www.recordedfuture.com/)
- **Palo Alto Networks – AutoFocus** (https://www.paloaltonetworks.com/cortex/autofocus)
- **CyberInt** (https://www.cyberint.com/)
- **Lookout** (https://www.lookout.com/)
- **SparkCognition DeepArmor** (https://www.sparkcognition.com/deeparmor-endpoint-protection/)
- **CrowdStrike Falcon** (https://www.crowdstrike.com/es-es/products/falcon-platform/)
- **CylancePROTECT** (https://www.blackberry.com/us/en/products/blackberry-protect)
- **Kenna Security Platform** (https://www.kennasecurity.com/platform/)

3.4 Identificación y autenticación biométrica

La identificación y autenticación biométrica se refiere al uso de características únicas, físicas o de comportamiento, de un individuo, para verificar o confirmar su identidad. Estas características pueden incluir, pero no se limitan a, huellas o impresiones dactilares, reconocimiento facial, reconocimiento de voz, patrones de iris, entre otras.

Podemos distinguir los siguientes **tipos de biometría**:

Biometría Física	Se basa en características físicas del cuerpo, tales como: • **Huellas (impresiones) Dactilares:** Las crestas y valles únicos en las yemas de los dedos. • **Reconocimiento Facial:** La estructura y características del rostro. • **Reconocimiento de Iris:** Los patrones únicos en el iris del ojo. • **Geometría de la Mano:** La forma y tamaño de la mano.
Biometría de Comportamiento	Se basa en las acciones realizadas por el individuo, tales como: • **Dinámica de Tecleo:** La forma en que un individuo pulsa las teclas de un teclado. • **Reconocimiento de Voz:** Las características únicas de la voz de una persona. • **Patrón de paseo:** La manera en que una persona camina.

La utilización de métodos biométricos plantea una serie de ventajas, tales como la **unicidad** (las características biométricas son únicas para cada individuo, lo que reduce la probabilidad de duplicidad o suplantación), la **conveniencia** (los usuarios no necesitan recordar contraseñas o códigos PIN) y la **dificultad de su falsificación** (puesto que es complicado replicar o falsificar datos biométricos, sobre todo si lo comparamos con contraseñas).

Sin embargo, el uso de mecanismos biométricos plantea ciertos **desafíos y limitaciones**, por ejemplo:

- **Errores de Reconocimiento:** ningún sistema biométrico es 100% preciso. Puede haber falsos positivos (reconociendo a alguien que no es el usuario) o falsos negativos (no reconociendo al usuario legítimo).
- **Preocupaciones de Privacidad:** la recolección, almacenamiento y uso de datos biométricos plantea preocupaciones sobre la privacidad, el consentimiento y la conformidad legal.
- **Irrevocabilidad:** a diferencia de las contraseñas, que se pueden cambiar, las características biométricas son permanentes. Si los datos biométricos se ven comprometidos, no pueden ser reemplazados o alterados.
- **Coste:** la implementación de sistemas biométricos puede requerir hardware y software especializado, lo que puede suponer un coste adicional.

Pese a todo, el uso de la biometría ha encontrado distintas **aplicaciones en la ciberseguridad**, tales como el **acceso lógico seguro** (muchos dispositivos y aplicaciones ofrecen opciones de autenticación biométrica como una capa adicional de seguridad), **transacciones en línea** (la autenticación biométrica puede utilizarse en transacciones bancarias en línea y pagos móviles para verificar la identidad del usuario), o el **control de acceso físico** (los sistemas biométricos pueden utilizarse para controlar el acceso a edificios, salas u otras áreas restringidas, etc.).

Pese a todo, el uso de la biometría ha encontrado distintas aplicaciones en la ciberseguridad, tales como el acceso lógico seguro, transacciones en línea, o el control de acceso físico

Como decimos, la identificación y autenticación biométrica se han popularizado en muchos dispositivos debido a su capacidad para proporcionar una capa de seguridad adicional. Se muestran seguidamente algunos de los ejemplos más conocidos de herramientas y sistemas que utilizan la biometría:

- **Apple Face ID y Touch ID:** Face ID permite el desbloqueo del iPhone, iPad y algunos Macs mediante el reconocimiento facial, mientras que Touch ID utiliza la huella dactilar (Face ID y Touch ID)
- **Windows Hello:** es una función a partir de Windows 10 que permite a los usuarios acceder a sus dispositivos utilizando el reconocimiento facial o de huellas dactilares (https://www.microsoft.com/es-es/windows/windows-hello)

- **Samsung Pass:** es una herramienta de autenticación biométrica que permite a los usuarios de dispositivos Samsung desbloquear sus smartphones y acceder a aplicaciones y sitios web mediante el reconocimiento de iris, facial o huellas dactilares (https://www.samsung.com/global/galaxy/apps/samsung-pass/)
- **BioID:** es una plataforma de autenticación facial basada en la nube que puede ser integrada en diversas aplicaciones para ofrecer autenticación biométrica (https://www.bioid.com/)
- **AuthenTrend:** ofrece soluciones de autenticación basadas en huellas dactilares para diferentes aplicaciones, desde unidades USB hasta soluciones empresariales (https://www.authentrend.com/)
- **Nuance VocalPassword:** es una solución de reconocimiento de voz que verifica la identidad del usuario basándose en las características únicas de su voz (https://www.nuance.com/omni-channel-customer-engagement/security/vocalpassword.html)

Estos son solo algunos ejemplos tradicionales de las múltiples soluciones de autenticación biométrica disponibles en el mercado. Es esencial tener en cuenta que, al considerar cualquier solución biométrica, es crucial evaluar la seguridad, la privacidad y la facilidad de uso para garantizar que cumple con los requisitos específicos de la organización, del usuario o exigencias regulatorias[34].

Además de lo anterior, las soluciones de identificación y autenticación biométrica han comenzado a integrar **capacidades avanzadas de IA**, especialmente en áreas como el reconocimiento facial y el análisis de comportamiento, para mejorar la precisión y reducir los falsos positivos. Se relacionan seguidamente algunos de los ejemplos más conocidos:

- **Trueface:** utiliza IA para ofrecer soluciones de reconocimiento facial. Sus algoritmos aprenden y mejoran con el tiempo, lo que incrementa la precisión en la identificación (https://www.trueface.ai/)
- **Kairos:** es una plataforma basada en la nube que utiliza IA para analizar rostros en videos y fotos, ofreciendo soluciones de reconocimiento facial (https://www.kairos.com/)
- **BehavioSec:** esta plataforma utiliza la IA para analizar patrones de comportamiento en tiempo real, como la dinámica del tecleo y la forma en que se maneja el ratón, para autenticar usuarios (https://www.behaviosec.com/)

34 Como sería el caso del cumplimiento del Esquema Nacional de Seguridad (Real Decreto 311/2022, de 3 de mayo) para las entidades de su ámbito de aplicación.

- **ID R&D:** utiliza la IA en sus soluciones biométricas de voz y de comportamiento para ofrecer una autenticación más segura y eficiente (https://www.idrnd.net/)
- **Deepware Scanner:** es un escáner de huellas dactilares basado en redes neuronales profundas. Utiliza IA para analizar y verificar huellas dactilares con alta precisión.
- **Affectiva:** aunque se centra principalmente en la interpretación emocional a través del análisis facial, Affectiva utiliza la IA para el análisis en tiempo real de las expresiones faciales, lo que tiene aplicaciones potenciales en áreas de autenticación basada en comportamientos o respuestas emocionales (https://www.affectiva.com/)

3.5 Análisis de vulnerabilidades y pentesting automatizado

Como es sabido, el **análisis de vulnerabilidades** es un proceso sistemático para evaluar, identificar y clasificar las vulnerabilidades de seguridad de los sistemas de información. Estas vulnerabilidades pueden obedecer a errores en el software, configuraciones inadecuadas, fallos en el hardware, o incluso prácticas de gestión de la seguridad deficientes.

Este proceso de análisis suele contemplar la **identificación** (las herramientas escanean sistemas, redes y aplicaciones en busca de vulnerabilidades conocidas), la **clasificación** (una vez detectadas, las vulnerabilidades se clasifican según su gravedad y riesgo), la **remediación** (se proponen soluciones para mitigar o solucionar las vulnerabilidades detectadas) y la **verificación** (después de la remediación, se realiza una nueva verificación para confirmar que las vulnerabilidades se hayan solucionado adecuadamente).

Por su parte, las **pruebas de penetración**, comúnmente conocidas como **pentesting**, son simulacros de ataques a un sistema con el objetivo de descubrir vulnerabilidades antes de que los atacantes reales lo hagan. A diferencia del análisis de vulnerabilidades, que normalmente utiliza escaneos automatizados para identificar vulnerabilidades conocidas, el pentesting a menudo involucra a expertos que intentan activamente explotar vulnerabilidades y penetrar en los sistemas, simulando las tácticas, técnicas y procedimientos (TTP) de los adversarios reales.

El proceso, generalmente, contempla el **reconocimiento** (recopilación de información sobre el objetivo), el **escaneo** (identificación de posibles puntos de entrada), la **penetración** (explotación de vulnerabilidades), el **mantenimiento del acceso** (simulación de movimientos de un atacante después de obtener acceso) y el **análisis** (conteniendo el informe de los hallazgos y recomendaciones para fortificar el sistema).

Como es lógico suponer, la **Inteligencia Artificial** también se ha incorporado al análisis de vulnerabilidades y a las pruebas de penetración, con los siguientes procedimientos:

Automatización mejorada	Con la IA, las herramientas pueden escanear redes y sistemas más rápidamente y con mayor precisión, identificando vulnerabilidades que las herramientas tradicionales podrían pasar por alto.
Aprendizaje continuo	Las herramientas basadas en IA pueden aprender de cada escaneo, adaptándose a las nuevas vulnerabilidades y técnicas de ataque.
Simulación avanzada	En el *pentesting*, la IA puede simular comportamientos de atacantes más complejos, probando sistemas contra amenazas emergentes y avanzadas.
Priorización de riesgos	La IA puede ayudar a priorizar vulnerabilidades basándose en el contexto y los datos históricos, permitiendo que los equipos de seguridad se concentren en las amenazas más inminentes o dañinas.
Integración y correlación	Las soluciones basadas en IA pueden correlacionar datos de múltiples fuentes, ofreciendo una visión más holística de la postura de seguridad de una organización.

Herramientas como **Tenable.io**, **Qualys Cloud Platform** o **Checkmarx** ya están utilizando capacidades de IA para mejorar sus escaneos y análisis. Además, plataformas de pentesting como **Cobalt** están incorporando IA para automatizar y mejorar partes del proceso.

La integración de la IA en estas áreas es prometedora, pero es esencial recordar que, por ahora, la combinación de expertos humanos con estas herramientas avanzadas proporciona el enfoque más robusto y completo hacia la ciberseguridad.

3.6 Defensa contra adversarios automatizados

A medida que la tecnología avanza, no solo los defensores mejoran sus herramientas, sino también los atacantes. Los **adversarios automatizados** son aquellos programas, bots y scripts diseñados para llevar a cabo ataques sin intervención humana directa. Estos ataques pueden variar desde simples ataques de fuerza bruta hasta modelos más sofisticados que pueden adaptarse y cambiar tácticas sobre la marcha.

La tabla siguiente muestra una **tipología** de los adversarios automatizados más comunes y sus **características** generales.

TIPOS		CARACTERÍSTICAS
Bots y Scrapers	Pueden ser usados para muchas tareas, como el scraping de sitios web, pero también pueden ser empleados para ataques, como intentos de inicio de sesión o la explotación de vulnerabilidades en un sitio web.	**1. Velocidad:** Pueden lanzar ataques a una velocidad que es prácticamente imposible para un humano. **2. Adaptabilidad:** Algunos sistemas automatizados avanzados pueden cambiar tácticas si detectan que un enfoque concreto no está funcionando. **3. Escala:** Son capaces de dirigirse a miles o incluso millones de objetivos simultáneamente. **4. Persistencia:** Pueden continuar sus ataques durante largos periodos sin cansancio o distracción.
Gusanos) worms	Son programas maliciosos que se replican automáticamente para propagarse a otras computadoras, a menudo explotando vulnerabilidades en el software.	
Bots de DDoS	Parte de las redes botnet se utilizan para lanzar ataques DDoS coordinados, inundando objetivos con tráfico para derribar servicios o infraestructuras.	
Sistemas de phishing automatizados	Estos pueden generar rápidamente sitios web fraudulentos o enviar correos electrónicos masivos con enlaces maliciosos.	

La defensa con Inteligencia Artificial

Para protegerse contra estos adversarios automatizados, la defensa también debe ser ágil, adaptable y, en muchos casos, también automatizada. Aquí es donde entra la inteligencia artificial. Veamos los escenarios más habituales:

1. **Detección de anomalías:** la IA puede analizar grandes conjuntos de datos para detectar patrones anómalos que pueden indicar un ataque automatizado.

La detección de anomalías es una de las aplicaciones más comunes de la inteligencia artificial en ciberseguridad. La idea es identificar patrones de comportamiento «normales» y luego detectar desviaciones o «anomalías» de esos patrones, lo que podría indicar actividad maliciosa o no autorizada. Se enumeran algunas de las herramientas más conocidas al respecto que emplean IA para la detección de anomalías, alguna de las cuales ya ha sido mencionada con anterioridad:

- **Darktrace** (https://www.darktrace.com/)
- **Splunk User Behavior Analytics (UBA)** (https://www.splunk.com/en_us/software/user-behavior-analytics.html)
- **Vectra Cognito** (https://www.vectra.ai/products)
- **Gurucul Risk Analytics** (https://gurucul.com/products/risk-analytics)
- **Exabeam Advanced Analytics** (https://www.exabeam.com/product/advanced-analytics/)

Como sucede con todas las herramientas de seguridad, es esencial mantenerlas actualizadas y utilizarlas como parte de una estrategia de seguridad más amplia.

2. **Identificación de bots:** a través del análisis de comportamiento, la IA puede identificar y bloquear bots basándose en sus patrones de interacción.

La identificación de bots es esencial, especialmente en el contexto del tráfico web, publicidad digital y redes sociales, donde los bots pueden incrementar artificialmente las métricas, desviar tráfico o difundir información errónea o desinformación. Varias soluciones utilizan inteligencia artificial y aprendizaje automático para identificar y bloquear el tráfico de bots en tiempo real. Se relacionan seguidamente algunas de las herramientas más conocidas:

- **Imperva Bot Management** (anteriormente Distil Networks) (https://www.imperva.com/products/bot-management/)
- **Akamai Bot Manager** (https://www.akamai.com/us/en/products/security/bot-manager.jsp)
- **Cloudflare Bot Management** (https://www.cloudflare.com/bots/)
- **DataDome** (https://www.datadome.co/)
- **Reblaze** (https://www.reblaze.com/)
- **Cofense Triage** (https://cofense.com/product-services/triage/)

Estas herramientas ofrecen protección en tiempo real contra el tráfico de bots, permitiendo a las organizaciones proteger sus activos en línea y garantizar que sus métricas y análisis sean precisos. Es esencial que las organizaciones elijan una solución que se ajuste a sus necesidades específicas y que esté en consonancia con su infraestructura y objetivos.

3. **Aprendizaje continuo:** a medida que los adversarios automatizados evolucionan, las soluciones basadas en IA pueden aprender de los ataques y adaptarse.

En el contexto de la ciberseguridad, el «aprendizaje continuo» (también conocido como «aprendizaje en línea» o «en tiempo real») se refiere a la capacidad de un sistema para adaptarse continuamente a las amenazas cambiantes, en tiempo real.

Se relacionan seguidamente algunas de las herramientas y sistemas más conocidos que utilizan aprendizaje continuo y técnicas de inteligencia artificial para la protección contra adversarios automatizados (algunas ya las hemos mencionado con anterioridad):

- **Darktrace Antigena** (https://www.darktrace.com/en/products/darktrace-antigena/)
- **CylancePROTECT** (https://www.blackberry.com/us/en/products/blackberry-protect)
- **SentinelOne Singularity Platform** (https://www.sentinelone.com/)
- **Endgame** (https://www.elastic.co/security)
- **CrowdStrike Falcon** (https://www.crowdstrike.com/products/falcon-platform/)

La ventaja clave de este tipo de herramientas está en su capacidad para adaptarse y aprender de las amenazas en tiempo real, lo que les permite mantenerse un paso adelante de los adversarios, incluso cuando estos cambian sus tácticas.

4. **Respuesta rápida:** ante la detección de un ataque, la IA puede tomar medidas inmediatas para mitigarlo, ya sea bloqueando el tráfico, cerrando procesos o alertando a los equipos de seguridad.

La respuesta rápida contra adversarios automatizados es esencial, ya que estos actores pueden escalar ataques o evolucionar rápidamente. Las soluciones basadas en IA tienen la capacidad de responder en tiempo real a amenazas identificadas, y algunas, incluso, pueden tomar acciones autónomas para mitigar o neutralizar la amenaza.

Se relacionan seguidamente algunas herramientas más conocidas que utilizan IA para proporcionar una respuesta rápida contra adversarios automatizados (alguna de ellas ya ha sido comentada previamente):

- **Darktrace Antigena** (https://www.darktrace.com/en/products/darktrace-antigena/)
- **Cisco Threat Response** (https://www.cisco.com/c/en/us/products/security/threat-response.html)
- **Palo Alto Networks Cortex XSOAR** (anteriormente Demisto) (https://www.paloaltonetworks.com/cortex/soar)
- **FireEye Helix** (https://www.fireeye.com/helix.html)
- **Symantec Endpoint Protection (SEP) Adaptive Threat Protection (ATP)** (https://www.broadcom.com/products/cyber-security/endpoint/endpoint-protection)
- **Netscout Arbor DDoS Protection** (https://www.netscout.com/solutions/ddos-protection)

Estas soluciones utilizan técnicas de inteligencia artificial para analizar y responder a eventos de seguridad en tiempo real. Además, muchas de ellas ofrecen la capacidad de integrarse con otras soluciones de seguridad, lo que permite a las organizaciones construir un enfoque de defensa en profundidad y responder rápidamente a amenazas desde múltiples vectores.

5. **Simulación de adversarios:** usar IA para simular ataques en entornos controlados (red teaming) ayuda a identificar debilidades y preparar mejor las defensas.

La simulación de adversarios, también conocida como «red teaming», ha adoptado la inteligencia artificial para mejorar la simulación y para probar de manera más efectiva las defensas en diferentes escenarios. Algunas de las herramientas más conocidas son:

- **Endgame Red Team Tools** (ahora parte de Elastic) (https://www.elastic.co/what-is/endpoint-security)
- **MITRE Caldera** (https://github.com/mitre/caldera)

Estas soluciones ayudan a las organizaciones a comprender sus vulnerabilidades y a mejorar sus posturas de defensa al simular ataques realistas. Sin embargo, es esencial recordar que las simulaciones de adversarios son solo una parte de una estrategia de ciberseguridad completa. La formación continua, la actualización de sistemas y software, y una vigilancia constante son cruciales para una defensa efectiva.

La defensa contra adversarios automatizados es una carrera en constante evolución. Con la capacidad de los atacantes para automatizar y adaptar sus ataques, las defensas tradicionales, basadas únicamente en firmas o reglas estáticas, pueden quedar rápidamente desactualizadas. Integrar la inteligencia artificial en la defensa proporciona la agilidad y adaptabilidad necesarias para mantenerse un paso por delante de estos adversarios avanzados.

Con la capacidad de los atacantes para automatizar y adaptar sus ataques, las defensas tradicionales, basadas únicamente en firmas o reglas estáticas, pueden quedar rápidamente desactualizadas

3.7 IA generativa y Ciberseguridad

La **inteligencia artificial (IA) generativa** se ha convertido en una herramienta valiosa en una variedad de campos, desde la creación de arte hasta la síntesis de datos. En el contexto de la ciberseguridad, la IA generativa puede ser tanto una solución como una amenaza potencial, veamos cómo:

BENEFICIOS DE LA IA GENERATIVA EN CIBERSEGURIDAD:	
Generación de datos sintéticos	La IA generativa puede utilizarse para crear conjuntos de datos sintéticos que simulan tráfico de red o comportamientos de usuarios, sin comprometer datos reales. Estos datos pueden ser utilizados para entrenar sistemas de detección de intrusos, sin violar la privacidad del usuario.
Simulación de ataques	A través de redes generativas adversarias (GANs), es posible simular cómo actuaría un atacante, permitiendo a las organizaciones probar la robustez de sus sistemas y realizar mejoras antes de que ocurran incidentes reales.
Creación de escenarios de pruebas	La IA generativa puede ayudar a crear escenarios realistas de pruebas de penetración, mejorando las prácticas tradicionales que a menudo se basan en escenarios predefinidos y menos dinámicos.
Reforzamiento del aprendizaje	La IA generativa, especialmente las GANs, puede ser útil en el reforzamiento del aprendizaje, donde un agente (la red generativa) y un adversario (la red discriminativa) trabajan en conjunto. Esta técnica puede ser utilizada para enseñar a sistemas de ciberseguridad cómo mejorar su detección y respuesta a amenazas en tiempo real.

AMENAZAS Y DESAFÍOS DE LA IA GENERATIVA EN CIBERSEGURIDAD:	
Vulnerabilidades durante y después del entrenamiento del modelo	Dado que los modelos generativos de IA se entrenan con datos que se recopilan de todo tipo de fuentes —y no siempre de forma transparente—, se desconoce exactamente qué datos quedan expuestos a esta superficie de ataque adicional. Combinado con el hecho de que estas herramientas de IA generativa a veces almacenan datos durante largos períodos de tiempo y no siempre cuentan con las mejores reglas de seguridad y salvaguardas, es muy posible que los actores de amenazas accedan y manipulen los datos de entrenamiento en cualquier etapa de dicho proceso de entrenamiento.
Violación de la privacidad de los datos personales	No existe una estructura que regule qué tipo de datos introducen los usuarios en los modelos generativos. Esto significa que los usuarios corporativos —y, en realidad, cualquier otra persona— pueden utilizar datos sensibles o personales sin cumplir la normativa ni obtener permiso de la fuente. Una vez más, con la forma en que se entrenan estos modelos y se almacenan los datos, la información de identificación personal puede llegar fácilmente a las manos equivocadas y conducir a situaciones indeseables.
Exposición de la propiedad intelectual	Se han dado casos de empresas que han expuesto involuntariamente datos de propiedad de la empresa a modelos generativos de forma perjudicial. Esta exposición se produce con mayor frecuencia cuando los empleados cargan en el sistema elementos u obras sujetas a propiedad intelectual o industrial, claves API y otra información confidencial.
Jailbreaks y soluciones de ciberseguridad	Muchos foros en línea ofrecen «jailbreaks», o formas secretas para que los usuarios enseñen a los modelos generativos a trabajar en contra de sus reglas establecidas. Por ejemplo, ChatGPT consiguió recientemente engañar a un humano para que resolviera un CAPTCHA en su nombre[35]. La capacidad de utilizar herramientas de IA generativa para generar contenidos de formas tan diferentes y similares a las humanas ha permitido sofisticados esquemas de phishing y malware que son más difíciles de detectar que las aproximaciones tradicionales.
Creación de malware y ataques	Las técnicas generativas pueden ser utilizadas por actores maliciosos para generar variantes de malware que pueden eludir sistemas de detección tradicionales.
Phishing y engaño	Las herramientas de IA generativa pueden utilizarse para crear sitios web falsos, correos electrónicos o comunicaciones que imitan a las legítimas, lo que sin duda aumenta la efectividad de los ataques de phishing.

35 https://cdn.openai.com/papers/gpt-4.pdf

Manipulación y falsificación de datos	Las GANs y otras técnicas pueden usarse para crear registros de logs falsos o manipular datos, lo que puede hacer que los ataques sean indetectables o desviar la atención de los equipos de seguridad.
Deepfakes en ciberseguridad	La habilidad de crear deepfakes (vídeos o audios falsos que parecen reales) puede ser explotada en ataques dirigidos para engañar a empleados o ejecutivos para que realicen acciones que comprometan la seguridad.
Mitigación y adaptación	La clave para abordar las amenazas asociadas con la IA generativa en ciberseguridad radica en la adaptación y actualización constante de las herramientas y técnicas de defensa, es decir: · **Monitorización continua** de las últimas investigaciones y tendencias en IA generativa. · **Entrenamiento regular** de los equipos de ciberseguridad en las capacidades y amenazas asociadas con la IA generativa. · **Adopción de sistemas** de IA que puedan adaptarse y aprender de técnicas generativas, manteniéndose un paso por delante de las amenazas.

Puesto que el uso de IA generativa presenta, como hemos visto, significativos desafíos, no parece que esté de más seleccionar algunos **Consejos y Buenas Prácticas en materia de Ciberseguridad para el uso de la IA generativa**[36], a saber:

- **Leer atentamente las políticas de seguridad de los proveedores de IA generativa**

  Tras las protestas iniciales por la falta de transparencia de ciertos proveedores de IA generativa en el entrenamiento de sus modelos, muchos de los principales proveedores han empezado a ofrecer una amplia documentación explicando cómo funcionan sus herramientas y en qué se basan los acuerdos con los usuarios.

  Para saber qué ocurre con los datos de entrada, hay que consultar las políticas de los proveedores sobre eliminación de datos y plazos, así como qué información utilizan para entrenar sus modelos. También es una buena práctica buscar en su documentación menciones a la trazabilidad, el historial de registros, la anonimización y otras funciones que pueda necesitar para cumplir requisitos normativos específicos.

  Y lo que es especialmente importante: hay que buscar cualquier mención a las opciones de inclusión y exclusión voluntarias y cómo elegir que los datos se utilicen o almacenen.

36 Fuente: Hiter, S. Generative AI and Cybersecurity. eWeek (Junio, 2023)

- **No introducir datos sensibles cuando se utilicen modelos generativos**

  La mejor manera de proteger los datos más sensibles es mantenerlos fuera de los modelos generativos, especialmente de aquellos con los que no se está familiarizado.

  A menudo es difícil saber qué datos pueden utilizarse o se utilizarán para entrenar futuras iteraciones de un modelo generativo, por no mencionar cuántos de los datos corporativos se almacenarán en los registros de datos del proveedor y durante cuánto tiempo.

  En lugar de confiar ciegamente en los protocolos de seguridad que estos proveedores puedan tener o no, es mejor crear copias sintéticas de los datos o evitar por completo el uso de estas herramientas cuando se trabaja con datos sensibles.

- **Mantener actualizados los modelos generativos de IA**

  Los modelos generativos reciben actualizaciones periódicas y, a veces, esas actualizaciones incluyen correcciones de errores y otras optimizaciones de seguridad. Es necesario estar atento a las oportunidades de actualizar las herramientas para que sigan siendo eficaces.

- **Formar a los empleados sobre el uso adecuado**

  Es sabido que las herramientas de IA generativa son fáciles de usar y, por tanto, de utilizar indebidamente. Es importante que los empleados sepan qué tipo de datos pueden utilizar como entradas, qué partes de su flujo de trabajo pueden beneficiarse de las herramientas de IA generativa y cuáles son las expectativas de cumplimiento normativo, todo ello, además del cumplimiento de las obligaciones normativas generales de la organización respecto del uso de los medios electrónicos.

- **Utilizar herramientas de gobierno y seguridad de datos**

  Las herramientas de gobierno y seguridad de los datos pueden proteger toda su superficie de ataque, incluidas las herramientas de IA generativa de terceros que pueda estar utilizando.

  Hay que considerar la posibilidad de invertir en herramientas de prevención de pérdida de datos, inteligencia de amenazas, plataforma de protección de aplicaciones nativas de la nube (CNAPP) y/o detección y respuesta ampliadas (XDR).

La mejor manera de proteger los datos más sensibles es mantenerlos fuera de los modelos generativos, especialmente de aquellos con los que no se está familiarizado

Se muestran seguidamente algunos ejemplos de **herramientas y soluciones de seguridad que utilizan IA generativa.**

	Google Cloud Security AI Workbench	Este nuevo desarrollo de Google se basa en Vertex AI de Google Cloud y está impulsado por Sec-PaLM. Google Cloud Security AI Workbench está diseñado para soportar inteligencia avanzada de amenazas y seguridad, detección de malware, análisis de comportamiento y gestión de vulnerabilidades. https://cloud.google.com/blog/products/identity-security/rsa-google-cloud-security-ai-workbench-generative-ai
	Microsoft Security Copilot	Microsoft Security Copilot es una de las soluciones de seguridad más específicas del arsenal de productos de IA generativa de Microsoft. Trabaja para optimizar la respuesta a incidentes, la detección de amenazas y la generación de informes de seguridad para los usuarios, e integra perspectivas e información de herramientas como Microsoft Sentinel, Microsoft Defender y Microsoft Intune. https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot
	CrowdStrike Charlotte AI	Esta herramienta de CrowdStrike permite a los usuarios gestionar la ciberseguridad a través del lenguaje natural en la plataforma Falcon. Al igual que muchas de estas herramientas de IA de ciberseguridad emergentes, Charlotte AI está diseñada para complementar a los equipos de seguridad existentes y reducir el impacto de las carencias de competencias. Charlotte AI se utiliza generalmente para apoyar la detección de amenazas y los esfuerzos de corrección. https://www.crowdstrike.com/press-releases/crowdstrike-introduces-charlotte-ai-to-deliver-generative-ai-powered-cybersecurity/
	Cisco Security Cloud	Cisco está añadiendo capacidades de IA generativa a la Nube de Seguridad y a sus carteras de Colaboración y Seguridad. Las nuevas funciones están diseñadas para hacer más fácil -incluso conversacional- la gestión de políticas y la respuesta ante amenazas. https://investor.cisco.com/news/news-details/2023/Cisco-Unveils-Next-Gen-Solutions-that-Empower-Security-and-Productivity-with-Generative-AI/default.aspx

AIRGAP	**Airgap Networks ThreatGPT**	Basado en GPT-3 y bases de datos gráficas, ThreatGPT es una solución de Airgap Networks que ayuda a las empresas a analizar de forma más eficaz y holística las amenazas a la seguridad en entornos de tecnología operativa (OT) y sistemas legacy. https://airgap.io/embargo-until-tbd/
	SentinelOne	La entidad actualizó recientemente (y restringida) su plataforma de captura de amenazas con funciones de IA generativa. Está diseñada para escalar las operaciones de seguridad y detección de amenazas, basándose en redes neuronales integradas y un amplio modelo de lenguaje que proporcionan información mejor y más próxima al tiempo real sobre posibles amenazas y soluciones. https://www.sentinelone.com/press/sentinelone-unveils-revolutionary-ai-platform-for-cybersecurity/
	Synthesis Humans	Synthesis Humans es una de las muchas herramientas generativas que ofrece Synthesis AI. Esta solución está diseñada para entrenar sistemas biométricos de control de acceso de una forma más ágil. En combinación con Synthesis Scenarios, esta herramienta puede utilizarse para respaldar la seguridad de las instalaciones, así como la ciberseguridad. https://synthesis.ai/synthesis-humans/
	SecurityScorecard	SecurityScorecard ha lanzado una plataforma de calificación de la seguridad basada parcialmente en GPT-4 de OpenAI. Con esta solución, los equipos de seguridad pueden formular preguntas abiertas y en lenguaje sencillo sobre la seguridad de su red y de terceros proveedores, y recibir respuestas proactivas y orientación para la gestión de riesgos. https://securityscorecard.com/company/press/securityscorecard-launches-first-and-only-security-ratings-platform-with-openais-gpt-4-search-system-providing-customers-with-faster-security-insights/
M	**MOSTLY AI**	MOSTLY AI es una herramienta de generación de datos sintéticos diseñada específicamente para generar datos anónimos que cumplan diversos requisitos de seguridad y conformidad. Debido a su marcado enfoque en la seguridad y el cumplimiento normativo, se utiliza con frecuencia en sectores regulados como la banca y los seguros. https://mostly.ai/

4. Escenarios de estudio

Los escenarios de estudio no solo nos brindan una comprensión tangible de cómo la inteligencia artificial (IA) se está utilizando en el mundo real para combatir las ciberamenazas, sino que también revelan las fortalezas y debilidades inherentes a estos enfoques.

A lo largo de los años, la integración de la IA en la ciberseguridad ha conducido a significativos éxitos, así como a lecciones aprendidas de incidentes donde las soluciones basadas en IA no pudieron detectar o prevenir ataques. Estos casos prácticos ilustran cómo las organizaciones, públicas o privadas, grandes o pequeñas, están utilizando la IA para proteger sus activos digitales.

En este epígrafe, exploraremos algunos ejemplos donde las soluciones de IA han detectado, prevenido o mitigado con éxito ciberataques, destacando cómo la tecnología ha podido ayudar a superar las capacidades tradicionales.

Los escenarios de estudio no solo nos brindan una comprensión tangible de cómo la inteligencia artificial (IA) se está utilizando en el mundo real para combatir las ciberamenazas, sino que también revelan las fortalezas y debilidades inherentes a estos enfoques

4.1 Sistemas modernos de detección y respuesta ante amenazas

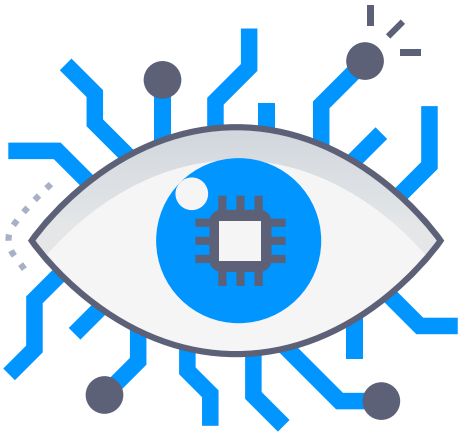
La detección y respuesta a amenazas ha sido un pilar en el mundo de la ciberseguridad durante años. Sin embargo, actualmente, con la adopción masiva de tecnologías basadas en inteligencia artificial (IA), estos sistemas han evolucionado considerablemente.

Efectivamente, estos sistemas modernos, comúnmente conocidos como **soluciones gestionadas de detección y respuesta ante amenazas (MDR) o sistemas de detección y respuesta de endpoints (EDR)**, a menudo incorporan capacidades de IA para mejorar la eficacia de sus operaciones.

Las características esenciales de estos sistemas actuales son la **automatización avanzada**, puesto que utilizan IA para reconocer patrones y comportamientos maliciosos en grandes conjuntos de datos en tiempo real, lo que permite una respuesta más rápida a las amenazas; el **aprendizaje continuo**, dado que se adaptan y evolucionan con el tiempo, aprendiendo de nuevos tipos de ataques y adaptándose a nuevos patrones de amenazas; la **integración y orquestación**, dada la capacidad de integrarse con otras herramientas y sistemas para proporcionar una respuesta cohesiva y orquestada a las amenazas.

Seguidamente se exponen sumariamente algunos **escenarios de aplicación de éxito**.

Estos sistemas modernos, comúnmente conocidos como soluciones gestionadas de detección y respuesta ante amenazas (MDR) o sistemas de detección y respuesta de endpoints (EDR), a menudo incorporan capacidades de IA para mejorar la eficacia de sus operaciones



Detección de ataques de día cero	Utilización de un sistema MDR basado en IA para identificar y prevenir un ataque de día cero antes de que se convierta en una amenaza generalizada. • **Contexto:** los ataques de día cero hacen referencia a vulnerabilidades no conocidas en el software, que los actores maliciosos explotan antes de que los desarrolladores tengan la oportunidad de crear y distribuir un parche. Dada su naturaleza, estos ataques son difíciles de prevenir con sistemas tradicionales de ciberseguridad. • **Situación:** las organizaciones pueden implementar sistemas MDR avanzado basados en IA, lo que propiciará la detección de actividad anómala en, por ejemplo, un software muy utilizado en la organización, pero que aún no hubiera sido reportado como vulnerable. • **Acción:** el sistema de IA, utilizando análisis de comportamiento, habría identificado patrones de acceso y modificación de datos que no se alineaban con los patrones normales de uso. En lugar de depender de firmas de malware conocidas, se habría centrado en el comportamiento inusual. Todo esto habría facilitado que la organización se pusiera en estado de alerta y pudiera aislar el software afectado, evitando una posible brecha de seguridad a gran escala. • **Resultado:** gracias a la detección temprana, no solo se podría proteger la información de la organización, sino que también se podría alertar al desarrollador del software y a la comunidad de seguridad, permitiendo una respuesta rápida para proteger a otros usuarios.
Respuesta a ransomware automatizado	El caso en el que un sistema EDR detecta y mitiga un intento de ransomware en segundos, salvando a una organización de una interrupción significativa. • **Contexto:** el ransomware, un tipo de malware que cifra los datos del usuario y exige un rescate para descifrarlos, ha evolucionado en complejidad a lo largo de los años. Los ataques de ransomware pueden paralizar organizaciones enteras, con costos significativos en términos de tiempo de inactividad, pérdida de datos, pérdidas económicas y reputacionales. • **Situación:** una organización es atacada por una variante desconocida de ransomware. En cuestión de segundos, el código dañino habría comenzado a cifrar archivos en varios sistemas. • **Acción:** el sistema EDR basado en IA, que habría sido instalado en la organización, habría detectado el comportamiento anómalo: el acceso rápido y masivo a archivos seguido de modificaciones consistentes con el cifrado. El EDR automáticamente podría haber aislado los sistemas afectados, pudiendo revertir los cambios realizados por el ransomware en muy poco tiempo. • **Resultado:** el ataque podría haberse contenido rápidamente, y la organización habría evitado pérdidas significativas y tiempo de inactividad. Además de ello, podrían haberse recabado datos valiosos sobre el ransomware, lo que puede ayudar a fortalecer las defensas no solo de la organización atacada sino también de otras organizaciones al compartir indicadores de compromiso.

No obstante, lo anterior, la utilización de estas técnicas, como hemos visto en los epígrafes precedentes, también plantea **desafíos y lecciones aprendidas**; a saber:

Desafíos:

1. **Detección de falsos positivos:** los sistemas basados en IA, especialmente cuando se entrenan o se configuran por primera vez, pueden generar alertas sobre actividades que, aunque inusuales, no son necesariamente maliciosas. Esto puede desencadenar respuestas innecesarias y desviar recursos.
2. **Adaptabilidad de los adversarios:** los actores maliciosos no son estáticos; evolucionan y cambian sus tácticas, técnicas y procedimientos (TTP) para eludir los sistemas de seguridad. Esto significa que lo que funciona hoy para detectar un ataque podría no ser eficaz mañana.
3. **Integración con la infraestructura existente:** no todas las organizaciones tienen la capacidad de implementar sistemas de ciberseguridad de última generación desde cero. A menudo, deben integrar nuevas soluciones con sistemas legacy, lo que puede presentar desafíos de compatibilidad y eficiencia.
4. **Dificultad en la interpretación:** las decisiones tomadas por modelos de IA avanzada pueden ser, a menudo, «cajas negras»; es decir, difíciles de interpretar o entender para los humanos, lo que puede generar desconfianza o confusión entre los equipos de seguridad.

Lecciones aprendidas:

1. **Necesidad de entrenamiento constante:** al igual que un antivirus necesita actualizaciones regulares de firmas, los sistemas de IA requieren entrenamiento continuo con datos recientes para mantener su eficacia.
2. **Importancia de la retroalimentación humana:** es esencial que los analistas de ciberseguridad proporcionen retroalimentación al sistema sobre la precisión de las alertas. Esto ayuda a ajustar y mejorar el modelo a lo largo del tiempo.
3. **Defensa en profundidad:** en materia de ciberseguridad, no se debe depender únicamente de un sistema basado en IA. Es importante contar con múltiples capas de defensa y no descuidar las prácticas básicas de higiene de seguridad.
4. **Colaboración y compartición de inteligencia:** en el mundo interconectado de hoy, compartir indicadores de compromiso, tácticas y otras formas de inteligencia sobre amenazas puede ayudar a otras organizaciones a prepararse y defenderse contra amenazas emergentes.
5. **Adopción gradual:** es prudente implementar y evaluar los sistemas basados en IA en entornos controlados o pilotos (sandbox) antes de acometer un despliegue completo. Esto permite identificar y abordar posibles problemas en un entorno más controlado.

Estos desafíos y lecciones subrayan la complejidad del panorama actual de la IA aplicada a ciberseguridad y la necesidad de enfoques innovadores, pero también reflexivos y holísticos, para abordar las amenazas.

Efectivamente, abordar el despliegue de herramientas de este tipo exige una planificación escrupulosa que debe considerar los siguientes elementos y fases:

Adopción y adaptación

La adopción y adaptación de nuevas tecnologías, en particular en el ámbito de la ciberseguridad, requiere un enfoque cuidadoso. Nos centraremos en la adopción y adaptación de sistemas basados en IA para la detección y respuesta ante amenazas:

Fase I: evaluación previa

- **Necesidades y carencias actuales:** es fundamental identificar primero las áreas en las que la organización enfrenta desafíos de ciberseguridad. Estos pueden incluir puntos ciegos en la detección, tiempo de respuesta lento o incluso un alto volumen de falsos positivos.
- **Requisitos de integración:** ¿cómo se integrará el nuevo sistema en la infraestructura tecnológica existente? Se deben considerar aspectos técnicos, pero también de procesos y equipos humanos.

Fase II: Selección de la solución

- **Personalización vs. soluciones genéricas:** algunas organizaciones pueden optar por sistemas personalizados, ajustados a sus necesidades específicas, mientras que otras pueden encontrar adecuadas las soluciones genéricas disponibles en el mercado.
- **Pruebas piloto:** antes de adoptar una solución en toda la organización, es conveniente realizar pruebas en un entorno limitado para evaluar su eficacia y garantizar que se integre bien con los sistemas existentes.

Fase III: Implementación y ajuste

- **Formación del personal:** es fundamental que el personal encargado de la ciberseguridad comprenda cómo funciona el nuevo sistema, cómo interpretar sus resultados y cómo actuar sobre ellos.
- **Retroalimentación inicial y ajuste:** los primeros meses de implementación son críticos para recopilar feedback. Esta retroalimentación se utiliza para ajustar el sistema, reduciendo los falsos positivos y mejorando la detección de amenazas legítimas.

Fase IV: Evaluación continua y adaptación

- **Evaluación del rendimiento:** a medida que el paisaje de amenazas cambia, es esencial evaluar regularmente cómo está funcionando el sistema y si está cumpliendo con las expectativas.
- **Adaptación a nuevas amenazas:** la inteligencia artificial, especialmente en el aprendizaje automático, puede requerir nuevos datos o reajustes para abordar amenazas emergentes. Las soluciones de ciberseguridad deben ser dinámicas y adaptarse a la evolución del paisaje de amenazas.

Fase V: Revisión y mejora

- **Incorporación de nuevas características o capacidades:** a medida que la tecnología avanza, pueden surgir nuevas características o capacidades que sea conveniente incorporar en el sistema existente.
- **Iteración basada en feedback:** las lecciones aprendidas durante la operación del sistema deben ser la base para mejoras continuas, garantizando así que la solución sigue siendo relevante y efectiva frente a las amenazas emergentes.

Corolario

La adopción y adaptación de sistemas basados en IA para ciberseguridad no es un proceso estático, sino que requiere un compromiso continuo, evaluación y ajuste para garantizar que la organización se mantiene protegida contra unas amenazas en constante evolución.

Evolución de las amenazas en respuesta a los modernos sistemas:

La evolución de las amenazas en respuesta a los modernos sistemas de defensa es un fenómeno complejo y dinámico. A medida que se implementan nuevas soluciones tecnológicas, los ciberdelincuentes también se adaptan, desarrollando tácticas y técnicas más avanzadas. Esto lleva a un ciclo continuo de adaptación y evolución entre defensores y atacantes.

1. **Evasión de la detección basada en IA:**
   - **Ataques polimórficos:** estos ataques cambian automáticamente su apariencia/*code signature* para evitar la detección. Esto puede hacerse a través de cambios en el código malicioso o la ofuscación de su comportamiento.
   - **Técnicas de *adversarial machine learning*:** como hemos visto con anterioridad, se refiere a estrategias diseñadas específicamente para confundir modelos de IA, como la introducción de pequeñas perturbaciones en los datos que pueden llevar a la clasificación errónea de contenido malicioso como benigno.
2. **Aprovechamiento de la automatización:**
   - **Ataques a gran escala y de rápida propagación:** los sistemas automatizados pueden lanzar ataques a una escala y velocidad que sería imposible para un humano, como la rápida propagación de ransomware o worms.
   - **Ataques de saturación:** estos intentan abrumar las capacidades de detección y respuesta de un sistema, enviando una avalancha de tráfico malicioso o solicitudes, como en el caso de los ataques DDoS.
3. **Ataques más dirigidos:**
   - **Spear phishing y ataques dirigidos:** en lugar de ataques masivos, los ciberdelincuentes pueden usar la información recolectada para dirigirse específicamente a individuos u organizaciones, a menudo usando tácticas de ingeniería social altamente personalizadas.

- **ATA, APT (Advanced Targeted Attacks, Advanced Persistent Threats):** estos ataques (muchos de ellos patrocinados por Estados) son altamente sofisticados y pueden involucrar una multiplicidad de tácticas y técnicas para evadir la detección y alcanzar su objetivo.

4. **Explotación de tecnologías emergentes**
   - **IoT (Internet of Things) y Edge Computing:** la proliferación de dispositivos conectados presenta nuevas oportunidades para los agentes de las amenazas, especialmente porque muchos de estos dispositivos carecen de medidas de seguridad adecuadas.
   - **Ataques en entornos cloud:** a medida que más organizaciones mueven sus operaciones y datos a la nube, los ciberdelincuentes buscan explotar las vulnerabilidades de las configuraciones y los servicios basados en la nube.
5. **Contramedidas y contrainteligencia:**
   - **Descubrimiento de defensas:** herramientas y tácticas que persiguen descubrir las defensas de un objetivo, identificar sus debilidades para adaptar el ataque posterior en consecuencia.
   - **Ataques de desinformación:** lo que puede involucrar la creación y difusión de información falsa para desviar la atención de las defensas reales o para desacreditar las alertas de seguridad legítimas.

Insistimos en que la continua evolución de las amenazas en respuesta a los avances en ciberseguridad subraya la importancia de la adaptación y la innovación constante en el campo de la defensa cibernética. Las organizaciones deben adoptar un enfoque proactivo, anticipando las tácticas emergentes de los atacantes y ajustando sus defensas en consecuencia.

En resumen, mientras que los sistemas modernos de detección y respuesta basados en IA están ofreciendo capacidades sin precedentes en la lucha contra las ciberamenazas, también presentan nuevos desafíos y requerimientos de adaptación tanto para las herramientas como para los profesionales que las utilizan.

4.2 Implementaciones exitosas de la IA en la ciberseguridad

Las implementaciones exitosas de IA en la ciberseguridad constituyen valiosos **escenarios de estudio** para comprender cómo la tecnología puede reforzar la postura de seguridad de una organización. Estos ejemplos también ofrecen lecciones sobre cómo integrar de manera efectiva la IA en las infraestructuras existentes y cómo superar los desafíos más habituales.

ESCENARIO EJEMPLO		LECCIÓN
Detección de amenazas avanzadas	Las organizaciones pueden detectar actividades sospechosas dentro de su red que otros sistemas habían pasado por alto. Utilizando algoritmos de aprendizaje automático, la solución podría analizar patrones de tráfico y detectar anomalías que indiquen un compromiso de datos.	**El aprendizaje automático puede ser particularmente eficaz para detectar amenazas desconocidas o de día cero al observar desviaciones del comportamiento normal.**
Respuesta automatizada a incidentes	Una organización que preste servicios de comercio o tramitación electrónica puede implementar sistemas basados en IA que, al detectar un aumento en el tráfico web que podría dar evidencia de un ataque DDoS, automáticamente redistribuyera y filtrara el tráfico, minimizando el impacto en sus operaciones.	**Una respuesta rápida y automatizada puede mitigar el daño de un ataque en tiempo real, especialmente cuando se trata de amenazas de este mismo tipo.**
Autenticación biométrica	Una entidad puede implementar un sistema de reconocimiento facial para sus aplicaciones móviles, proporcionando una capa adicional de seguridad. La IA no solo podría analizar características faciales, sino también patrones de comportamiento, como la forma en que un usuario sostiene su dispositivo.	**La IA puede añadir capas de autenticación multifactorial basadas en características intrínsecas y comportamientos del usuario.**
Simulación de adversarios	Una organización podría utilizar IA para simular ataques en su propia red, lo que les permitiría identificar vulnerabilidades y fortalecer su postura de seguridad antes de que ocurra un ataque real.	**Las simulaciones basadas en IA pueden ayudar a las organizaciones a prepararse para amenazas reales, identificando puntos débiles en su infraestructura.**

Análisis forense	Después de un ataque, una organización podría utilizar herramientas de IA para analizar rápidamente registros y logs, identificando cómo los atacantes pudieron penetrar en el sistema, qué datos comprometieron y cómo se movieron dentro de la red. La IA puede acelerar significativamente el proceso de análisis forense, permitiendo una recuperación más rápida y proporcionando información vital para evitar futuros incidentes.

Mientras que cada organización enfrentará desafíos únicos, estas implementaciones ofrecen evidencia tangible de los beneficios y ventajas que la IA puede proporcionar en la lucha contra las amenazas cibernéticas

Estos escenarios muestran la variedad de formas en que la IA se puede integrar con éxito en el panorama de la ciberseguridad. Mientras que cada organización enfrentará desafíos únicos, estas implementaciones ofrecen evidencia tangible de los beneficios y ventajas que la IA puede proporcionar en la lucha contra las amenazas cibernéticas.

Se muestran seguidamente enlaces a algunos casos de éxito concretos relacionados con la implementación de IA en la ciberseguridad. Conviene mencionar, no obstante, que, debido a la naturaleza confidencial de muchos incidentes de ciberseguridad, algunas empresas podrían no divulgar detalles específicos de los incidentes o de cómo se resolvieron.

1. **Detección de amenazas avanzadas**:
   - **Empresa:** Darktrace
   - **Detalles**: Darktrace utiliza aprendizaje automático y algoritmos basados en IA para detectar, responder y mitigar amenazas cibernéticas en tiempo real. Uno de sus casos de éxito involucró a una empresa de energía en la que se detectó un compromiso en una de sus estaciones de trabajo, que estaba siendo utilizada para escanear la red interna.
   - **Resultado**: Según la entidad, la actividad fue identificada y detenida rápidamente, evitando un potencial compromiso a mayor escala.
   - **URL de referencia**: Darktrace Casos de éxito

2. **Respuesta automatizada a incidentes**:
   - **Empresa**: Cloudflare
   - **Detalles**: Cloudflare ofrece soluciones para proteger sitios web contra todo tipo de amenazas, incluidos ataques DDoS. En una

ocasión, protegieron a uno de sus clientes de un ataque DDoS que excedía los 400 Gbps.

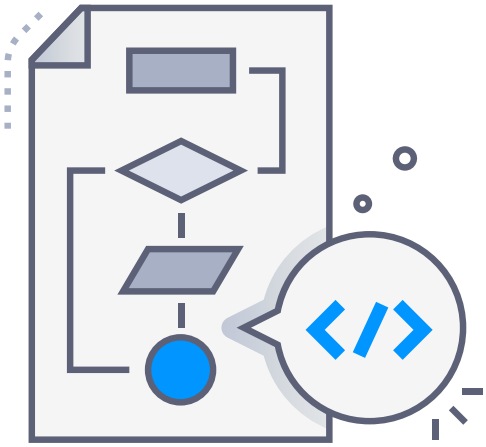
- **Resultado**: El tráfico malicioso fue filtrado con éxito, y el sitio web del cliente permaneció en línea sin interrupciones.
- **URL de referencia**: Cloudflare Blog

3. **Autenticación biométrica**:
   - **Empresa**: HSBC
   - **Detalles**: La entidad financiera HSBC implementó tecnología de reconocimiento de voz para verificar la identidad de sus clientes cuando se ponen en contacto con la entidad. Según la entidad, la identificación de la solución se basa en más de 100 características únicas de la voz de una persona.
   - **Resultado**: Se redujo el tiempo de autenticación y se mejoró la experiencia del cliente, a la vez que se añadió una capa adicional de seguridad.
   - **URL de referencia**: HSBC Voice ID

4. **Simulación de adversarios**:
   - **Empresa**: Cymulate
   - **Detalles**: Cymulate es una plataforma que permite a las organizaciones simular ataques en sus propias redes. Un cliente, una empresa de seguros, utilizó Cymulate para identificar y mitigar vulnerabilidades antes de ser explotadas.
   - **Resultado**: La empresa pudo fortalecer su postura de seguridad y estaba mejor preparada para enfrentar amenazas reales.
   - **URL de referencia**: Cymulate

Estos casos de éxito proporcionan una visión de cómo la inteligencia artificial y el aprendizaje automático están siendo utilizados en situaciones del mundo real para mejorar la ciberseguridad. Es crucial, sin embargo, investigar en profundidad sobre cada uno de estos casos para obtener detalles específicos y comprender completamente su impacto y alcance.

Es crucial investigar en profundidad sobre cada uno de estos casos para obtener detalles específicos y comprender completamente su impacto y alcance



4.3 Fallos y lecciones aprendidas

Analizar los fallos y extraer lecciones aprendidas es esencial para entender el panorama completo de cualquier tecnología o aplicación. En el contexto de la inteligencia artificial aplicada a ciberseguridad, si bien ha habido muchos éxitos, también ha habido desafíos y errores que han servido como puntos de aprendizaje cruciales para la industria. Se muestran algunos ejemplos.

TIPOLOGÍA Y DESCRIPCIÓN		LECCIONES APRENDIDAS
Confianza excesiva en soluciones automatizadas	En algunas ocasiones, las organizaciones han confiado demasiado en sus sistemas de IA para la detección de amenazas, asumiendo que la IA detectaría todas las amenazas posibles. Sin embargo, ningún sistema es infalible.	**Es esencial tener un equilibrio entre las soluciones de IA y la supervisión humana. La experiencia y el juicio humano son esenciales en el mundo de la ciberseguridad.**
Ataques adversarios contra modelos de IA	Han surgido ataques que buscan engañar o confundir modelos de aprendizaje automático. Por ejemplo, se pueden modificar ligeramente las muestras de malware para que sean indetectables por sistemas basados en IA.	**Es crucial actualizar y entrenar constantemente los modelos de IA con datos recientes y relevantes. Además, se deben aplicar técnicas de defensa específicas contra ataques adversarios.**
Falsos positivos	En algunas implementaciones, los sistemas de IA han generado una cantidad significativa de falsos positivos, lo que puede llevar a la sobrecarga de los equipos de seguridad y a la posibilidad de perder amenazas reales en medio del ruido generado.	**Es esencial ajustar y optimizar constantemente los modelos y algoritmos de IA para reducir la cantidad de falsos positivos y mejorar la precisión.**
Dependencia de datos de calidad	La eficacia de la IA depende de la calidad de los datos con los que se entrena. Si un sistema de IA se entrena con datos inadecuados o sesgados, sus predicciones o detecciones pueden ser incorrectas o ineficaces.	**Es fundamental garantizar la calidad, la diversidad y la representatividad de los datos utilizados para entrenar sistemas de IA.**
Coste de implementación	La adopción e implementación de soluciones de IA puede ser costosa, no solo en términos económicos, sino también en cuanto a tiempo y recursos. Algunas organizaciones han podido subestimar estos costes y han tenido dificultades en la fase de implementación.	**Es fundamental realizar un análisis detallado de costes y beneficios antes de implementar soluciones de IA en ciberseguridad.**

Estos fallos y lecciones aprendidas resaltan la importancia de adoptar un enfoque equilibrado y cuidadoso al implementar la IA en la ciberseguridad. Mientras que la IA ofrece herramientas y capacidades poderosas, sigue siendo esencial tener en cuenta sus limitaciones y desafíos.

5. Desafíos y limitaciones de la IA en Ciberseguridad

Como hemos dicho, la inteligencia artificial está impactando directamente en el campo de la ciberseguridad, ofreciendo soluciones innovadoras para la detección y prevención de amenazas, el análisis de comportamiento y la respuesta automática a incidentes. Sin embargo, como toda tecnología emergente, la IA no está exenta de desafíos y limitaciones. A pesar de su potencial transformador, las expectativas hacia la IA deben ser equilibradas con un entendimiento claro de sus restricciones.

Estos desafíos no solo abarcan aspectos técnicos, como la calidad del entrenamiento de datos o la interpretación de resultados, sino también dilemas éticos y preocupaciones sobre la privacidad. Además, a medida que los ciberdelincuentes se adaptan y evolucionan, surgen nuevos obstáculos para los sistemas basados en IA, desde ataques adversarios hasta la manipulación de modelos.

En este epígrafe, exploraremos en detalle los desafíos inherentes al uso de la IA en ciberseguridad, las limitaciones actuales de esta tecnología y las áreas en las que, a pesar de los avances, la intervención y el juicio humano siguen siendo insustituibles. Al hacerlo, buscamos ofrecer una perspectiva equilibrada y realista que permita a las organizaciones aprovechar al máximo las ventajas de la IA mientras se mantienen alerta ante sus posibles limitaciones.

A pesar de su potencial transformador, las expectativas hacia la IA deben ser equilibradas con un entendimiento claro de sus restricciones

5.1 Ataques adversarios contra modelos de IA

Los ataques adversarios contra modelos de IA han aflorado como una preocupación crítica en el campo de la ciberseguridad. Como hemos señalado en el presente trabajo, estos ataques se diseñan para engañar o confundir a los modelos de aprendizaje automático, lo que podría conducir a decisiones erróneas o malintencionadas por parte de tales sistemas.

Efectivamente, un ataque adversario implica la introducción de pequeñas perturbaciones en los datos de entrada, diseñadas para ser casi imperceptibles para el ser humano, pero que pueden llevar al modelo a hacer predicciones incorrectas. Estas perturbaciones son cuidadosamente calculadas para maximizar el error en la predicción del modelo.

Los ataques adversarios pueden ser de dos **tipos:**

ATAQUES DE CAJA BLANCA	**En este escenario, el atacante tiene un conocimiento completo del modelo, incluida su arquitectura y los parámetros. Esto le permite diseñar perturbaciones que son especialmente efectivas contra el modelo específico.**
ATAQUES DE CAJA NEGRA	**En este caso, el atacante no tiene acceso directo al modelo y a sus parámetros, pero puede tener acceso a sus predicciones. Aunque este escenario es más desafiante para el atacante, todavía es posible generar perturbaciones adversarias efectivas.**

Estos ataques adversarios presentan varias **implicaciones negativas para la ciberseguridad.**

Por ejemplo, si se trata de **detección de malware**, si un sistema de IA se utiliza para detectar malware, un atacante podría diseñar malware que, una vez alterado, pase desapercibido por el modelo. En el caso de **sistemas de autenticación**, si un sistema basado en IA se encarga de la autenticación, por ejemplo, mediante reconocimiento facial, un ataque adversario podría permitir el acceso no autorizado a un intruso. Por último, en el caso de **análisis de tráfico de red** los atacantes pueden manipular características específicas de dicho tráfico para evadir la detección de un sistema basado en IA.

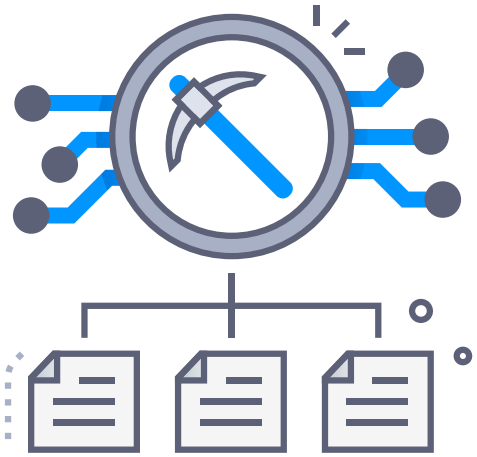
Frente a ello pueden desarrollarse determinadas **contramedidas**, entre ellas:

- **Entrenamiento adversario:** esta técnica implica entrenar el modelo con ejemplos adversarios, lo que puede aumentar su robustez frente a este tipo de ataques.
- **Detección de perturbaciones:** algunos métodos persiguen detectar directamente las perturbaciones adversarias en lugar de intentar hacer predicciones precisas en su presencia.
- **Regularización y técnicas de defensa:** Se trata de técnicas diseñadas para hacer que los modelos sean intrínsecamente más resistentes a los ataques adversarios, ajustando su comportamiento durante el entrenamiento.

Los ataques adversarios contra modelos de IA son una manifestación de una verdad fundamental en ciberseguridad: cualquier sistema, por avanzado que sea, tiene vulnerabilidades. La meta a alcanzar sería mantenerse un paso por delante de los atacantes, adaptándose y evolucionando constantemente en respuesta a nuevas amenazas.

Tanto los atacantes como los defensores hacen uso de **herramientas avanzadas**, y muchas de estas herramientas integran capacidades de IA. Se muestran seguidamente algunas de las más conocidas, tanto para ataque como para defensa.

Los ataques adversarios contra modelos de IA son una manifestación de una verdad fundamental en ciberseguridad: cualquier sistema, por avanzado que sea, tiene vulnerabilidades



Herramientas OFENSIVAS, que pueden ser usadas por los atacantes:	**DeepExploit:** es una herramienta automatizada de pentesting que utiliza aprendizaje profundo. Es capaz de aprender de los resultados de las pruebas de penetración anteriores y adaptar sus técnicas en consecuencia. https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit
	Snallygaster: herramienta que busca archivos expuestos en servidores web, utilizando técnicas de IA para identificar posibles vectores de ataque. https://github.com/hannob/snallygaster
	GPT-2: aunque originalmente no se diseñó como herramienta de ataque, esta tecnología de lenguaje natural desarrollada por OpenAI puede ser utilizada para generar contenido falso, como correos electrónicos de phishing. https://github.com/openai/gpt-2
Herramientas DEFENSIVAS:	**TensorFlow Privacy:** se trata de una biblioteca que ayuda a los desarrolladores a entrenar modelos de aprendizaje automático con privacidad diferencial, lo que puede ayudar a proteger los datos de entrenamiento. https://github.com/tensorflow/privacy
	IBM's Adversarial Robustness Toolbox (ART): es una biblioteca de Python que proporciona herramientas para mejorar la robustez de los modelos de aprendizaje automático y profundizar contra ataques adversarios. (https://github.com/Trusted-AI/adversarial-robustness-toolbox)
	DeepArmor: es una solución de ciberseguridad que utiliza técnicas de aprendizaje profundo para detectar y prevenir malware en tiempo real. https://www.sparkcognition.com/deeparmor-endpoint-security/
	CylancePROTECT: se trata de un software de endpoint que utiliza modelos de IA para predecir y prevenir ejecuciones de malware y scripts avanzados. https://www.cylance.com/cylanceprotect

Estas herramientas son solo un pequeño subconjunto de las opciones disponibles en el mercado.

5.2 Dependencia excesiva de soluciones automatizadas

La **dependencia excesiva de soluciones automatizadas** en ciberseguridad, y en particular de las basadas en Inteligencia Artificial (IA) y Aprendizaje Automático (ML), tiene implicaciones significativas y riesgos asociados, a saber:

1. **Carencia de Interpretabilidad:**
   La IA, especialmente el aprendizaje profundo, puede funcionar como una «caja negra». Esto significa que, aunque un modelo pueda predecir o clasificar con alta precisión, a menudo es difícil entender cómo llegó a una decisión particular. Esto plantea preocupaciones en la ciberseguridad, donde la trazabilidad y la comprensión de las decisiones tomadas son fundamentales para evaluar la eficacia y la confiabilidad del sistema, además de que podría constituir una no conformidad legal, si al sistema en cuestión estuviera sometido a una regulación específica, como prescribe el Reglamento Europeo de Inteligencia Artificial, para sistemas de riesgo.

2. **Falsa sensación de seguridad:**
   El despliegue de soluciones de IA puede llevar a las organizaciones a creer que están totalmente protegidas contra amenazas. Sin embargo, ningún sistema es infalible. Si las organizaciones confían únicamente en soluciones automatizadas, pueden pasar por alto áreas críticas de vulnerabilidad o no estar preparadas para responder cuando estas soluciones fallan o son eludidas.

3. **Evolución de amenazas:**
   Los atacantes están constantemente adaptando y evolucionando sus métodos para eludir los sistemas de defensa. Si las soluciones de IA no se actualizan y adaptan continuamente al cambiante panorama de amenazas, pueden volverse obsoletas rápidamente.

La dependencia excesiva de soluciones automatizadas en ciberseguridad, tiene implicaciones significativas y riesgos asociados

4. **Ataques específicos contra la IA:**
   Los atacantes son cada vez más conscientes de cómo funcionan los sistemas basados en IA y están desarrollando técnicas específicas, como ataques adversarios, para engañar o eludir estos sistemas. Una dependencia excesiva de soluciones de IA sin la debida diligencia puede exponer a las organizaciones a estos ataques especializados.

5. **Fallos en la automatización:**
   Los sistemas de IA son tan buenos como los datos con los que han sido entrenados. En el contexto de la ciberseguridad, esto significa que, si un sistema ha sido entrenado con datos no representativos o sesgados, puede hacer predicciones incorrectas o no detectar ciertas amenazas.

6. **Desplazamiento del juicio humano:**
   A pesar de los avances en IA, el juicio y la experiencia humanos siguen siendo cruciales en ciberseguridad. El equipo de ciberseguridad tiene una comprensión intuitiva y contextual de los sistemas y redes que administran, lo que es extraordinariamente importante para identificar y responder a amenazas que podrían pasar desapercibidas para un sistema automatizado.

7. **Coste de mantenimiento y actualización:**
   Aunque la automatización puede parecer rentable a corto plazo, mantener y actualizar sistemas de IA para asegurar que sigan siendo efectivos frente a las amenazas emergentes puede requerir inversiones significativas en tiempo y recursos.

Así pues, como **conclusión** de todo ello, podemos decir que mientras que la IA y la automatización pueden ofrecer capacidades revolucionarias en el campo de la ciberseguridad, es esencial abordar estos sistemas con un enfoque equilibrado, y deben ser vistos como una herramienta en un arsenal más amplio de defensa, complementando, no reemplazando, otros métodos y técnicas tradicionales.

La combinación de la experiencia humana con las capacidades de la IA y la conformidad con las normas legales que resulten de aplicación es la mejor defensa contra las ciberamenazas en evolución.

Los atacantes son cada vez más conscientes de cómo funcionan los sistemas basados en IA y están desarrollando técnicas específicas, como ataques adversarios, para engañar o eludir estos sistemas

5.3 Falsos positivos y falsos negativos

Los **falsos positivos** y **falsos negativos** son un desafío crucial en cualquier sistema de detección o clasificación, y su prevalencia en sistemas basados en Inteligencia Artificial (IA) o Aprendizaje Automático (ML) puede tener graves consecuencias en el ámbito de la ciberseguridad.

Se dice que aparece un **Falso Positivo (FP)** cuando el sistema identifica erróneamente una actividad benigna como maliciosa. En términos de seguridad, podría tratarse de un software legítimo identificado erróneamente como malware.

Por su parte, se dice que aparece un **Falso Negativo (FN)** cuando el sistema falla en detectar una actividad maliciosa, clasificándola erróneamente como benigna. Por ejemplo, un malware real que no fuera detectado por el sistema de seguridad.

Ambos tipos de falsos, positivos y negativos, tienen importantes implicaciones para la ciberseguridad, a saber:

IMPLICACIONES DE LOS FALSOS POSITIVOS	IMPLICACIONES DE LOS FALSOS NEGATIVOS
Interrupciones innecesarias: Los falsos positivos pueden llevar a bloquear o detener aplicaciones y procesos legítimos, causando interrupciones en las operaciones normales del negocio.	**Brechas de seguridad no detectadas:** Un falso negativo permite que las amenazas reales eludan las defensas, lo que puede derivar en violaciones de datos, compromiso de sistemas o cualquier otro tipo de daño cibernético.
Desgaste del equipo de seguridad: Un alto número de falsos positivos puede consumir recursos significativos, ya que el personal de seguridad tiene que revisar y verificar cada alerta.	**Confianza mal fundamentada:** Creer que un sistema es seguro cuando en realidad hay amenazas activas puede llevar a la complacencia y a una falta de preparación para posibles incidentes.
Insensibilización: Si las alertas de seguridad se perciben comúnmente como falsas alarmas, el personal puede comenzar a ignorarlas, lo que podría llevar a la omisión de alertas verdaderamente críticas.	

Llegados a este punto, conviene recordar dos de los desafíos más significativos que supone la utilización de la IA y el ML en ciberseguridad. En primer lugar, el relativo a la **calidad de datos**, sabiendo, como hemos dicho, que la precisión de los modelos de ML está directamente vinculada a la calidad de los datos con los que son entrenados, siendo así que datos no representativos o desequilibrados pueden llevar a tasas más altas de falsos positivos y negativos. En segundo lugar, lo relativo a los **modelos complejos**, lo que propicia que algunas técnicas avanzadas de ML, especialmente en aprendizaje profundo, puedan actuar como «cajas negras», haciendo difícil entender por qué ciertas decisiones se toman y, por lo tanto, complicando la tarea de ajustar el modelo para reducir estos errores.

Frente a estas realidades se hace preciso, por consiguiente, desarrollar **estrategias de mitigación del riesgo de la IA**, en concreto:

- **Entrenamiento constante:** Los modelos de ML deben ser reentrenados y ajustados regularmente con datos actualizados para mejorar su precisión.
- **Incorporación de feedback:** Al integrar retroalimentación humana, los sistemas pueden aprender de los errores y ajustar sus criterios de detección.
- **Combinación de técnicas:** Utilizar un enfoque híbrido que combine diferentes técnicas de detección puede ayudar a reducir tanto los falsos positivos como los falsos negativos.

A modo de **conclusión** debemos decir que, como hemos visto, los falsos positivos y negativos presentan desafíos significativos en la ciberseguridad, especialmente cuando se utilizan sistemas basados en IA. Aunque es difícil eliminarlos por completo, un entendimiento adecuado y una gestión efectiva de estos errores pueden minimizar su impacto y asegurar una defensa cibernética más robusta.

Los falsos positivos y negativos presentan desafíos significativos en la ciberseguridad, especialmente cuando se utilizan sistemas basados en IA

5.4 Privacidad y ética en la aplicación de la IA

La **privacidad y la ética** en la aplicación de la Inteligencia Artificial es un tema de creciente importancia, también en el contexto de la ciberseguridad, donde los datos y la información personal pueden estar en juego. Las soluciones de seguridad impulsadas por IA tienen el potencial de ser extremadamente efectivas, pero también presentan preocupaciones sobre cómo se recopilan, almacenan y utilizan los datos.

De entre la antedicha **problemática** podemos citar los siguientes aspectos:

EN RELACIÓN CON…	LOS PROBLEMAS PUEDEN SURGIR DE…
…la recopilación de datos:	**Sobredimensionamiento:** para entrenar y operar, los sistemas de IA requieren grandes conjuntos de datos. En el proceso, existe la posibilidad de que se recopilen más datos de los necesarios, lo que puede invadir la privacidad de los usuarios. **Consentimiento:** a menudo, los datos se recopilan sin el conocimiento o el consentimiento del usuario, lo que plantea preocupaciones éticas y legales.
… almacenamiento y uso de datos:	**Seguridad de los datos:** al almacenar grandes conjuntos de datos, las organizaciones se convierten en objetivos atractivos para los ciberdelincuentes. Un fallo en la seguridad podría exponer información personal y/o confidencial. **Perfilado:** con suficientes datos, la IA puede usarse para perfilar a los individuos basándose en su comportamiento en línea, lo que puede llevar a decisiones sesgadas o a discriminación.
… transparencia y toma de decisiones:	**Decisiones de «caja negra»:** muchos modelos de IA, especialmente aquellos basados en aprendizaje profundo, no proporcionan una visibilidad clara de cómo toman las decisiones. Esto puede derivar en falta de confianza y dificultades para verificar la justicia o la idoneidad de estas decisiones. **Sesgo y justicia:** si los datos utilizados para entrenar modelos de IA están sesgados, las decisiones que toma el modelo también lo estarán. Esto puede reforzar estereotipos o llevar a discriminación.

… vigilancia y supervisión:	**Abuso potencial:** las soluciones de ciberseguridad basadas en IA que monitorizan redes y sistemas para detectar amenazas también pueden ser usadas para vigilar el comportamiento de los usuarios con propósitos maliciosos o invasivos.
… rendición de cuentas y responsabilidad:	**Falta de responsabilidad:** Determinar la responsabilidad en caso de fallos o errores de un sistema basado en IA puede ser complicado, especialmente si no está claro cómo el sistema tomó una decisión particular.
… regulaciones y directrices éticas:	**Necesidad de marcos regulatorios:** Para garantizar que se aborden las preocupaciones éticas, es esencial contar con directrices y regulaciones claras que guíen el desarrollo y aplicación de soluciones de IA en ciberseguridad.

La ética y la regulación en torno a la IA y la ciberseguridad están evolucionando rápidamente a medida que las tecnologías avanzan y los posibles problemas y consecuencias se hacen más evidentes.

Por lo que toca a la **Unión Europea,** son dos los marcos regulatorios más significativos:

- **Reglamento General de Protección de Datos (RGPD) de la Unión Europea**[37]:

  Aunque no está diseñado específicamente para la IA, el RGPD ha establecido importantes estándares en materia de privacidad de datos y derechos de los individuos, como el derecho al olvido y la transparencia en el procesamiento de datos. Estos principios también se aplican cuando se utiliza IA en ciberseguridad.

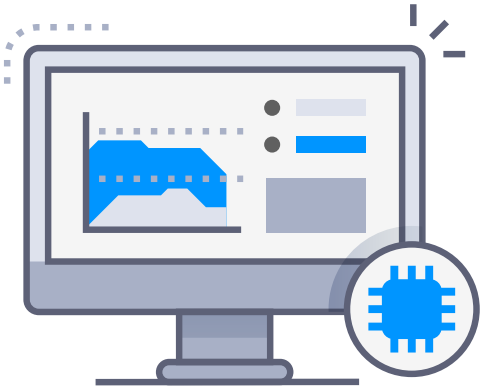
  El RGPD es una regulación fundamental para la privacidad y la protección de datos de todos los individuos dentro de la Unión Europea (UE). Entró en vigor el 25 de mayo de 2018. Estos son sus aspectos esenciales:

  - **Ámbito de aplicación territorial:** El RGPD se aplica no solo a organizaciones ubicadas dentro de la UE, sino también a organizaciones ubicadas fuera de la UE si ofrecen bienes o servicios a individuos en la UE o monitorizan el comportamiento de estos.
  - **Consentimiento:** Las organizaciones ya no pueden utilizar términos y condiciones largos y difíciles de entender. La solicitud de consentimiento debe darse en una forma fácilmente accesible y comprensible. Además, debe ser tan fácil retirar el consentimiento como darlo.

37 REGLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos).

- **Derechos del interesado:**
  - **Derecho de acceso:** Los individuos tienen el derecho de saber si se están procesando datos personales suyos y, de ser así, acceder a esos datos.
  - **Derecho a la rectificación:** Las personas tienen el derecho a corregir datos inexactos.
  - **Derecho al olvido:** También conocido como derecho de supresión, permite a las personas solicitar la eliminación de sus datos.
  - **Derecho a la portabilidad:** Las personas pueden obtener y reutilizar sus datos personales en diferentes servicios.
  - **Derecho a la limitación del tratamiento:** Las personas pueden solicitar que no se procesen sus datos, salvo para propósitos concretos, definidos, adecuados y legales.
  - **Derecho de oposición:** Las personas tienen el derecho de oponerse al procesamiento de sus datos en ciertas circunstancias.
- **Notificación de violación de datos:** En caso de una violación de datos, las organizaciones deben notificar a las autoridades de protección de datos pertinentes dentro de las 72 horas, a menos que la violación no represente un riesgo para los derechos y libertades de las personas. Si la violación supone un alto riesgo para los derechos y libertades de los individuos, también éstos deben ser notificados.
- **Responsabilidad del Responsable del Tratamiento (RT) y del Encargado del Tratamiento (ET):** Establece la responsabilidad de los RT y ET para garantizar el cumplimiento del RGPD, incluida la necesidad de mantener registros detallados de las actividades de procesamiento de datos.
- **Protección de datos desde el diseño y por defecto:** Las organizaciones deben considerar la protección de datos en el diseño de nuevos sistemas, procesos o productos y también garantizar que, por defecto, solo se procesen los datos necesarios para cada uso específico.
- **Delegados de Protección de Datos (DPD):** Las organizaciones deben designar un DPD si pertenecen a determinados grupos de entidades o realizan ciertos tipos de procesamiento de datos, como el procesamiento a gran escala de datos sensibles.

Los individuos tienen el derecho de saber si se están procesando datos personales suyos y, de ser así, acceder a esos datos



- **Transferencias Internacionales:** Se establecen condiciones más estrictas para la transferencia de datos personales fuera de la UE.
- **Sanciones:** Las organizaciones pueden ser multadas con hasta el 4% de su facturación anual global o 20 millones de euros (la cifra que sea mayor) por incumplimientos graves. Hay un sistema escalonado de multas para infracciones menos graves.

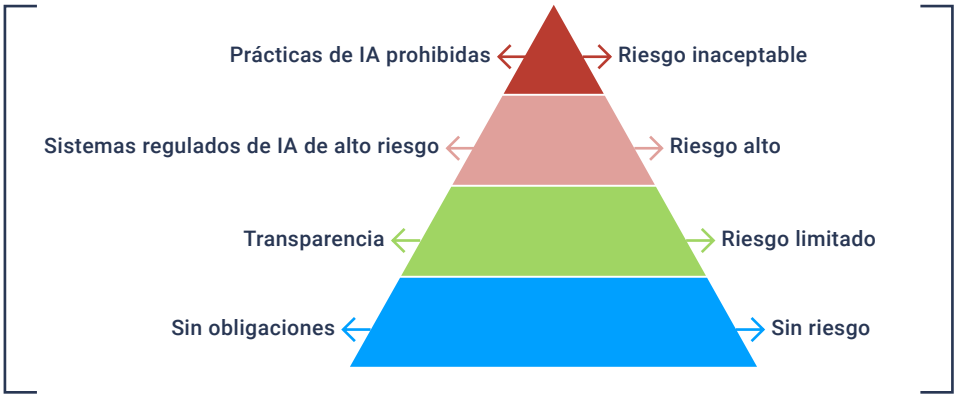
Propuesta de regulación de la IA de la Comisión Europea (2021)[38]:

La Comisión Europea presentó en abril de 2021 una propuesta de marco regulador de la UE sobre inteligencia artificial (IA)[39]. El proyecto de ley sobre IA es el primer intento de promulgar una normativa horizontal al respecto. El marco jurídico propuesto se centra en la utilización específica de los sistemas de IA y los riesgos asociados.

En el texto, la Comisión propone establecer una definición tecnológicamente neutra de los sistemas de IA en la legislación de la UE y establecer una clasificación para los sistemas de IA con diferentes requisitos y obligaciones adaptados a un «enfoque basado en el riesgo».

Así, se prohibirían algunos sistemas de IA que presenten riesgos «inaceptables»; se autorizaría una amplia gama de sistemas de IA de «alto riesgo», pero sujetos a una serie de requisitos y obligaciones para acceder al mercado de la UE. Los sistemas de IA que sólo presenten un «riesgo limitado» estarán sujetos a obligaciones de transparencia muy leves.

PIRÁMIDE DE RIESGOS



38 Propuesta de REGLAMENTO DEL PARLAMENTO EUROPEO Y DEL CONSEJO POR EL QUE SE ESTABLECEN NORMAS ARMONIZADAS EN MATERIA DE INTELIGENCIA ARTIFICIAL (LEY DE INTELIGENCIA ARTIFICIAL) Y SE MODIFICAN DETERMINADOS ACTOS LEGISLATIVOS DE LA UNIÓN. (Bruselas, 21.4.2021).

39 Fuente: European Parliament. Artificial intelligence act. Briefing. EU Legislation in Progress. (2023).

El Consejo aprobó la posición general de los Estados miembros de la UE en diciembre de 2021. El Parlamento votó su posición en junio de 2023.

Al tiempo de redactar estas líneas, los legisladores de la UE inician ahora las negociaciones para ultimar la nueva legislación, con modificaciones sustanciales de la propuesta de la Comisión, que incluyen la revisión de la definición de sistemas de IA, la ampliación de la lista de sistemas de IA prohibidos y la imposición de obligaciones a la IA de propósito general y a los modelos de IA generativa como ChatGPT.

La Propuesta de Reglamento sobre la Inteligencia Artificial (IA) marca un paso significativo hacia la regulación de las aplicaciones de la IA en la Unión Europea. Estos son sus elementos esenciales:

- **Objetivo:** La propuesta tiene como objetivo garantizar que la IA se utilice de una manera que sea segura y respete los derechos fundamentales de los ciudadanos de la UE.
- **Clasificación de riesgos:** Las aplicaciones de IA se clasifican según el nivel de riesgo que presentan:
  - **Riesgo inaceptable:** Algunas prácticas estarían completamente prohibidas debido a un claro potencial para perjudicar los derechos de las personas. Esto incluye, por ejemplo, sistemas de IA que distorsionen el comportamiento humano.
  - **Riesgo alto:** Aplicaciones en áreas críticas, como sistemas biométricos de identificación y sistemas de infraestructura crítica. Estos sistemas estarán sujetos a estrictas regulaciones y requerirán una evaluación antes de su implementación, llegando a estar, en algunos casos, totalmente prohibidos.
  - **Riesgo limitado:** Las aplicaciones deben seguir requisitos específicos de transparencia. Por ejemplo, los chatbots deberían declararse como tales para que los usuarios sepan que están interactuando con una máquina.
- **Transparencia:** La propuesta enfatiza la transparencia en el uso de sistemas de IA, especialmente en áreas como los deepfakes o las interacciones con chatbots.
- **Creación de un Comité Europeo de IA:** Se propone la creación de un comité para ayudar a implementar y actualizar el Reglamento.

Los legisladores de la UE inician ahora las negociaciones para ultimar la nueva legislación, que incluyen la revisión de la definición de sistemas de IA, la ampliación de la lista de sistemas de IA prohibidos y la imposición de obligaciones a la IA de propósito general y a los modelos de IA generativa como ChatGPT

- **Sanciones:** La propuesta establece sanciones considerables para las empresas que no cumplan con la normativa, incluidas multas de hasta el 6% de su facturación global anual.
- **Aplicabilidad:** El Reglamento se aplicará no solo a los proveedores de sistemas de IA establecidos en la UE, sino también a los proveedores que ofrezcan sus sistemas en el mercado de la UE.
- **Innovación y apoyo:** Aunque la propuesta tiene un enfoque regulatorio, también destaca la importancia de fomentar la innovación en el campo de la IA y apoyar el desarrollo de capacidades de IA en la UE.

Marcos Éticos

Podemos mencionar los siguientes:

1. **Principios de Asilomar sobre Inteligencia Artificial**[40]**:**
   Estos principios, establecidos en la conferencia de IA de 2017, cubren áreas como la investigación para hacer que la IA sea segura, la ética en su aplicación y la necesidad de que beneficie a todos.
2. **Principios de IA de OpenAI**[41]**:**
   Esta organización de investigación en IA ha establecido principios que buscan asegurar que la IA beneficie a toda la humanidad, priorizando la seguridad y la cooperación a largo plazo.
3. **Principios de IA de Google**[42]**:**
   Aunque provienen de una empresa específica, estos principios han sido influyentes. Abordan la seguridad, la justicia, la transparencia y la responsabilidad, entre otros aspectos.
4. **Ética de la IA del IEEE**[43]**:**
   El IEEE, una de las organizaciones profesionales más grandes del mundo para el avance de la tecnología, ha establecido estándares éticos para la IA y la robótica que se centran en incorporar valores humanos en el diseño y la operación.

Así pues, como **conclusión** debemos señalar que mientras que la regulación y los marcos éticos son esenciales para guiar la aplicación de la IA en la ciberseguridad, es crucial que estas directrices se mantengan actualizadas y sean flexibles para adaptarse a la rápida evolución de la tecnología.

40 https://futureoflife.org/open-letter/ai-principles/

41 https://openai.com/charter

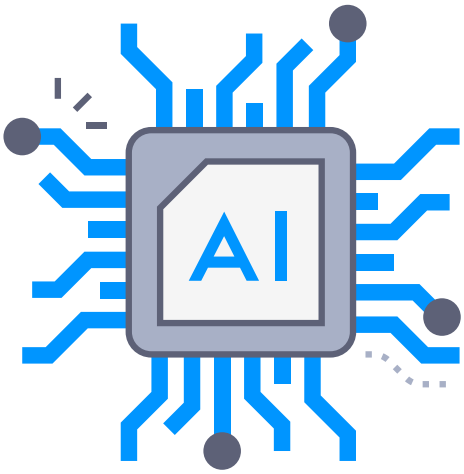
42 https://ai.google/responsibility/principles/

43 https://standards.ieee.org/news/get-program-ai-ethics/

6. Futuro de la IA en la Ciberseguridad

A medida que nos aventuramos en las próximas décadas, la integración de la IA en la ciberseguridad se tornará aún más profunda y compleja, prometiendo transformaciones significativas en cómo detectamos, respondemos y prevenimos ciberamenazas.

En este epígrafe, exploraremos las proyecciones y tendencias que definirán el futuro de la IA en la ciberseguridad. Partiendo de los sistemas de defensa autónomos y aprendizaje continuo y hasta los desafíos éticos y la necesidad de marcos regulativos robustos, abordaremos las expectativas y preocupaciones que envuelven a este horizonte tecnológico. Además, destacaremos cómo las innovaciones actuales pueden esbozar los contornos de futuras soluciones y cómo la comunidad global puede prepararse y adaptarse a estos cambios inminentes.



6.1 Tendencias emergentes

Se presentan seguidamente, de forma sumaria, algunas de las cuestiones sobre las que actualmente se está trabajando y que podrán marcar el futuro de la IA aplicada a ciberseguridad.

Autodefensa cibernética autónoma

La autodefensa cibernética autónoma se refiere al uso de tecnologías avanzadas, particularmente la inteligencia artificial y el aprendizaje automático, para permitir que los sistemas informáticos y las redes detecten, respondan y mitiguen automáticamente las amenazas en tiempo real, sin intervención humana.

Sus características son:

1. **Detección proactiva:** Tradicionalmente, muchos sistemas de seguridad operan de un modo reactivo, respondiendo a las amenazas después de que ocurren. La autodefensa autónoma, en cambio, se centra en detectar patrones y anomalías en tiempo real, permitiendo una respuesta casi inmediata.
2. **Respuesta y contención inmediata:** Una vez detectada una amenaza, los sistemas autónomos pueden tomar medidas para contenerla, lo que puede incluir el aislamiento de un dispositivo comprometido, bloquear una dirección IP sospechosa o limitar el acceso a ciertas partes de la red.
3. **Adaptabilidad:** Dada la naturaleza siempre cambiante de las ciberamenazas, la autodefensa cibernética autónoma está diseñada para aprender y adaptarse constantemente. Esto significa que, con cada amenaza detectada, el sistema se vuelve más inteligente y eficaz en su respuesta.
4. **Reducción de la carga humana:** Con una respuesta automatizada, se reduce la necesidad de intervención humana constante, permitiendo que los equipos de seguridad se centren en amenazas más complejas o en la estrategia de ciberseguridad a largo plazo.

5. **Desafíos:** A pesar de sus ventajas, la autodefensa cibernética autónoma no está exenta de desafíos, que incluyen la posibilidad de respuestas excesivas o erróneas, la complejidad en la implementación y el mantenimiento, y la dependencia de la IA, que puede ser susceptible a ataques específicos, como ataques adversarios.
6. **Aplicaciones en la vida real:** Actualmente, hay soluciones en el mercado que ofrecen capacidades de respuesta autónoma, especialmente en el ámbito de la detección y respuesta de endpoints (EDR). Estas soluciones pueden identificar comportamientos maliciosos en los dispositivos de la red y tomar medidas inmediatas para neutralizar la amenaza

EJEMPLO

Darktrace (https://www.darktrace.com/)

Esta empresa posee su «Enterprise Immune System», una solución que utiliza algoritmos de aprendizaje automático para detectar, responder y mitigar amenazas cibernéticas en tiempo real. Su sistema aprende y entiende el «patrón de vida» normal de cada usuario y dispositivo en la red, lo que le permite detectar desviaciones significativas que indican posibles amenazas. Además, su producto «Darktrace Antigena» actúa como un «anticuerpo digital», tomando decisiones autónomas sobre cómo responder a amenazas específicas sin intervención humana.
https://es.darktrace.com/resources/autonomous-response-darktrace-antigena

Aprendizaje federado

El aprendizaje federado es un enfoque de entrenamiento de modelos de aprendizaje automático en el que múltiples dispositivos o servidores retienen sus datos localmente y comparten únicamente las actualizaciones del modelo con un servidor central, en lugar de compartir los datos en sí mismos.

Sus características esenciales son:

- Cada dispositivo entrena un modelo localmente, utilizando sus propios datos.
- Una vez que un dispositivo ha procesado su lote local de datos y actualizado el modelo, envía solo las actualizaciones o el resumen del modelo al servidor central.
- El servidor central agrega las actualizaciones de todos los dispositivos para formar un modelo global actualizado.
- Este modelo global se envía de vuelta a todos los dispositivos para la siguiente ronda de entrenamiento.
- Este proceso se repite hasta que el modelo converge o satisface ciertos criterios de parada.

Como decimos, una de las ventajas de este modelo es que, dado que los datos en bruto nunca salen del dispositivo local, hay menos riesgos de exposición, lo que es especialmente útil para datos sensibles o personales. Además, se reduce el ancho de banda necesario, ya que solo se comparten las actualizaciones del modelo, que suelen ser mucho más pequeñas en tamaño que el conjunto de datos completo, lo que también es muy adecuado para escenarios donde los datos están distribuidos, tales como dispositivos móviles o ubicaciones geográficas dispersas.

Respecto de sus **aplicaciones en ciberseguridad**, podemos destacar las dos siguientes:

- **Detección de amenazas en redes distribuidas:** Al permitir que cada nodo o dispositivo de la red aprenda localmente sobre amenazas y comparta sus actualizaciones con un servidor central, se puede construir un modelo global de detección sin comprometer la privacidad de los datos en cada nodo.
- **Actualizaciones de modelos en tiempo real:** Los dispositivos en una red pueden adaptarse rápidamente a nuevas amenazas, aprendiendo localmente, y actualizando después un modelo global.

Sistemas de IA explicables (XAI)

A medida que la IA desempeña roles más críticos en la ciberseguridad, es esencial que todas las partes implicadas (especialmente, los equipos de ciberseguridad) puedan comprender y confiar en las decisiones tomadas por estos sistemas. El acrónimo XAI se refiere a los métodos y técnicas en la investigación de IA que hacen que los resultados de los algoritmos sean comprensibles para los humanos.

Con la popularidad de modelos de aprendizaje profundo, como las redes neuronales, la IA ha alcanzado significativos niveles de precisión en muchas tareas. Sin embargo, como hemos venido repitiendo, estos modelos a menudo actúan como «cajas negras», donde incluso los expertos tienen dificultades para entender por qué se tomó una decisión específica. Esta falta de transparencia puede ser problemática, especialmente en campos como la medicina, el derecho y la banca, donde las decisiones incorrectas pueden tener graves consecuencias y donde se requiere, incluso como imperativo legal, una justificación.

Los modelos XAI pueden desarrollarse en base a diferentes enfoques:

1. **Interpretabilidad Local vs. Global:** La interpretabilidad puede enfocarse en entender decisiones individuales (local) o en entender cómo funciona generalmente el modelo (global).
2. **Modelos Intrínsecamente Interpretables:** Estos modelos, como los árboles de decisión o la regresión lineal, son naturalmente explicables. Sin embargo, pueden no ser tan precisos como los modelos complejos.

EJEMPLOS

TensorFlow Federated (TFF):
Es una plataforma de código abierto desarrollada por Google que permite a los desarrolladores usar APIs para implementar aprendizaje federado. Está construido sobre TensorFlow y proporciona herramientas para simular la formación federada en datos distribuidos.
(https://www.tensorflow.org/federated?hl=es-419)

PySyft:
Es una extensión flexible de PyTorch para el aprendizaje federado y otras técnicas de privacidad en el aprendizaje automático. Se centra en la descentralización y ofrece herramientas para la computación segura multiparte, entre otros.
(https://github.com/OpenMined/PySyft)

Federated AI Technology Enabler (FATE):
Es una plataforma de código abierto que proporciona un marco seguro y conveniente para el entrenamiento colaborativo y el aprendizaje federado. Fue una iniciativa de WeBank y ha logrado contribuciones de muchas otras empresas y organizaciones.
(https://fate.fedai.org/)

Mientras que estas herramientas ofrecen marcos y bibliotecas para el aprendizaje federado, es importante destacar que muchas empresas tecnológicas grandes, como **Google** y **Apple**, ya están implementando aprendizaje federado en algunos de sus productos para mejorar la privacidad del usuario. Un ejemplo clásico es la predicción de texto en teclados de smartphones, donde el modelo se entrena localmente en el dispositivo del usuario, en base a sus entradas, sin enviar los datos reales a servidores centrales.

3. **Métodos Post-hoc:** Estos métodos se aplican después de que el modelo ha sido entrenado. Pueden ser visualizaciones, como mapas de calor, o técnicas que descomponen las decisiones del modelo, como LIME (Local Interpretable Model-agnostic Explanations) o SHAP (SHapley Additive exPlanations).

4. **Técnicas de Descomposición de Atributos:** Estas técnicas intentan explicar la contribución de cada característica a una decisión específica, dando una idea de qué características son las más influyentes.

Sea como fuere, los beneficios del enfoque XAI son claros: permiten sustentar la **confianza** (cuando los usuarios, especialmente aquellos que no son expertos en IA, entienden cómo funciona un sistema, es más probable que confíen en él); ayudan a mantener la **responsabilidad** (la XAI puede ayudar a asegurar que los sistemas de IA actúen de manera responsable y justa, reduciendo sesgos y errores); facilita la **mejora y la depuración** del modelo (al comprender cómo un modelo toma decisiones, es más fácil identificar y corregir errores o inexactitudes); y facilita el **cumplimiento regulatorio** (puesto que en algunas jurisdicciones, como es el caso de la UE, se está requiriendo que los sistemas de toma de decisiones automatizados sean transparentes y puedan justificar sus decisiones).

Naturalmente, la aplicación de los modelos XAI no está exenta de desafíos, tales como el **compromiso precisión-interpretabilidad** (sabiendo que, a menudo, hay un equilibrio entre la precisión del modelo y su interpretabilidad, los modelos más simples pueden ser más fáciles de entender pero menos precisos); la **subjetividad** (puesto que la «explicabilidad» puede ser subjetiva; es decir, lo que es claro y comprensible para un experto técnico puede no serlo para otro o para una persona no especializada); o la **generalización** (puesto que las explicaciones generadas para una instancia específica pueden no generalizarse bien para otras instancias).

EJEMPLOS

LIME (Local Interpretable Model-agnostic Explanations) :
es una técnica que puede explicar las predicciones de cualquier clasificador o regresor de una manera que es comprensible para el humano. Funciona creando un modelo interpretable que es localmente fiel a las predicciones del modelo original.
(https://github.com/marcotcr/lime)

SHAP (SHapley Additive exPlanations):
Se basa en la teoría de juegos para explicar la salida de cualquier modelo de máquina. Es una medida unificada de la importancia de las características.
(https://github.com/slundberg/shap)

DeepLIFT (Deep Learning Important FeaTures):
Presenta un enfoque para descomponer las salidas de las redes neuronales y calcular la importancia de cada entrada para la salida. Es especialmente útil para redes neuronales profundas.
(https://github.com/kundajelab/deeplift)

AI Explainability 360:
Es un conjunto de herramientas que incluye algoritmos, bibliotecas y tutoriales para ayudar a los desarrolladores a entender, explicar y visualizar las decisiones tomadas por modelos de IA.
(https://aix360.mybluemix.net/)

InterpretML:
Es una biblioteca open-source de Microsoft para interpretar modelos de máquina. Proporciona una variedad de técnicas y herramientas para la interpretación de modelos.
(https://interpret.ml/)

Adopción de Blockchain para seguridad

Aunque blockchain es más conocido por su uso en criptomonedas, tiene aplicaciones en ciberseguridad, especialmente en la gestión de identidades y en la garantía de la integridad de los datos.

Se resumen seguidamente las formas en las que blockchain está participando en la ciberseguridad:

1. **Integridad de datos y autenticación:** Blockchain proporciona un registro inmutable y transparente de datos. Una vez que un bloque se añade a la cadena, no puede ser alterado sin alterar todos los bloques siguientes, lo que se considera extraordinariamente difícil debido a la naturaleza descentralizada de la red blockchain. Esto garantiza la integridad de los datos y previene la manipulación.
2. **Descentralización:** La ciberseguridad tradicional depende de servidores centralizados, que son puntos de ataque vulnerables. Blockchain es intrínsecamente descentralizado, lo que significa que no hay un único punto de compromiso potencial, lo que, dependiendo de la seguridad adoptada para cada nodo, puede ser una ventaja o una fuente de riesgo potencial.
3. **Identidad segura:** Los sistemas basados en blockchain pueden proporcionar soluciones de identidad digital en las que la identidad de los usuarios se verifica y almacena en la blockchain.
4. **Comunicaciones seguras:** Las soluciones blockchain pueden garantizar comunicaciones seguras y autenticadas entre dispositivos en Internet de las cosas (IoT). Estos dispositivos suelen ser vulnerables a ataques, pero con una gestión de identidad basada en blockchain, podrían validarse y comunicarse de forma más segura.
5. **Auditoría y trazabilidad:** Blockchain proporciona una traza clara y verificable de todas las transacciones. Esto es muy valioso para las operaciones de auditoría y facilita la transparencia y la rendición de cuentas.
6. **Resistencia a la censura y disponibilidad:** Debido a su naturaleza descentralizada, las redes blockchain son resistentes a la censura y las interrupciones. Es difícil cerrar o censurar una red blockchain sin el consenso de la mayoría de sus participantes.
7. **Smart Contracts para la automatización segura:** Los contratos inteligentes son programas autónomos que se ejecutan en blockchain cuando se cumplen ciertas condiciones preestablecidas, y pueden ser utilizados para automatizar y validar acuerdos y operaciones sin intermediarios, reduciendo así la posibilidad de fraude o intervención maliciosa.

Aunque blockchain es más conocido por su uso en criptomonedas, tiene aplicaciones en ciberseguridad, especialmente en la gestión de identidades y en la garantía de la integridad de los datos

A pesar de los beneficios de la tecnología blockchain para la ciberseguridad, también existen desafíos. Por ejemplo, aunque la blockchain es inmutable y las transacciones no pueden ser alteradas una vez validadas, si un atacante logra hacerse con el control de la mayoría de la red (un ataque del 51%), podría potencialmente validar transacciones fraudulentas. Además, como cualquier tecnología emergente, la implementación práctica de blockchain en la ciberseguridad aún está en desarrollo, y las organizaciones deben ser cautelosas y diligentes al adoptarla.

EJEMPLOS

Guardtime:
Utiliza la tecnología blockchain para asegurar la integridad y autenticidad de los datos.
(https://www.guardtime.com/)

Civic:
Es una solución de identidad segura basada en blockchain. Ofrece a las empresas y a los individuos herramientas para controlar y proteger las identidades. A través de su plataforma descentralizada, Civic permite la autenticación sin la necesidad de contraseñas tradicionales.
(https://www.civic.com/)

REMME:
Es una solución que tiene como objetivo eliminar los ataques de phishing, las contraseñas y los certificados. Utiliza blockchain para autenticar usuarios y dispositivos en lugar de una contraseña.
(https://remme.io/)

Chain of Things (CoT):
Investiga y desarrolla aplicaciones que combinan blockchain con Internet de las cosas (IoT). Esto tiene aplicaciones en áreas como la energía, el transporte y la logística, donde la integridad y la seguridad de los datos recopilados por los dispositivos IoT son críticas.
(https://www.chainofthings.com/)

Modelos de IA basados en el comportamiento del usuario

En lugar de depender únicamente de contraseñas o datos biométricos, la IA podría analizar patrones de comportamiento continuo (como la forma en que alguien escribe o mueve el mouse) para autenticar usuarios y detectar comportamientos anómalos.

Los modelos de IA basados en el comportamiento del usuario son una tendencia emergente en ciberseguridad y representan una de las maneras más avanzadas de detectar anomalías y actividades sospechosas en un sistema. Como decimos, estos modelos se entrenan para aprender patrones normales de comportamiento del usuario, al objeto de identificar cualquier desviación de esos patrones como actividad potencialmente sospechosa.

Algunas de las **aplicaciones** más comunes de este tipo de modelo serían:

1. **Detección de fraude:** En el sector financiero, por ejemplo, si un usuario realiza transacciones de grandes sumas de dinero repentinamente o en patrones que no coinciden con su comportamiento histórico, la IA puede generar una alerta.
2. **Control de acceso y autenticación:** Si el comportamiento del inicio de sesión de un usuario cambia repentinamente (por ejemplo, inicia sesión en horarios inusuales o desde ubicaciones geográficas desconocidas), podría ser una señal de que alguien más está utilizando sus credenciales.
3. **Protección contra amenazas internas:** Los empleados descontentos o malintencionados pueden representar amenazas para las organizaciones. Si comienzan a acceder a archivos o sistemas que normalmente no utilizan, puede detectarse por sistemas basados en el comportamiento.

La utilización de este tipo de modelos presenta indudables **ventajas**, entre ellas: la **personalización** (puesto que se basan en el comportamiento individual del usuario, son altamente personalizados y adaptativos); la **detección proactiva** (pueden detectar amenazas en tiempo real, permitiendo una respuesta más rápida) y la **reducción de falsos positivos** (al estar más afinados con el comportamiento real del usuario, tienden a generar menos alertas por actividad legítima que se desvía ligeramente de la norma).

No obstante, el uso de estos modelos comporta también asumir ciertos **desafíos**, tales como: **requerir tiempo para aprender** (puesto que, como cualquier sistema de aprendizaje automático, estos modelos necesitan tiempo para aprender los patrones de comportamiento de los usuarios); **cambios en el comportamiento del usuario** (de forma que si un usuario cambia de rol o responsabilidades, su comportamiento puede cambiar, lo que puede llevar a falsos positivos hasta que el sistema se adapte) o la **privacidad** (puesto que estos sistemas recopilan y analizan mucha información sobre el comportamiento del usuario, lo que plantea preocupaciones sobre la privacidad y cómo se manejan y almacenan esos datos).

EJEMPLOS

Darktrace:
Ya reseñado con anterioridad, utiliza lo que llama una «Enterprise Immune System» para aprender del 'patrón de vida' normal de cada usuario y dispositivo en una red y luego identificar comportamientos anómalos. (https://www.darktrace.com)

Cylance:
Como hemos señalado, utiliza IA para ofrecer prevención de amenazas de endpoint basada en el comportamiento. Su plataforma se centra en detener malware, ransomware y otras amenazas basándose en la identificación de comportamientos sospechosos en lugar de firmas conocidas de virus. (https://www.cylance.com)

Exabeam:
Es una plataforma de gestión de información y eventos de seguridad (SIEM) que utiliza el aprendizaje automático para analizar el comportamiento del usuario y detectar amenazas. (https://www.exabeam.com)

UserInsight de Rapid7:
Se centra específicamente en la detección de comportamientos anómalos de los usuarios y los atacantes dentro de una red. Puede identificar cuando las credenciales de un usuario han sido comprometidas y están siendo utilizadas por un atacante. (https://www.rapid7.com)

Estas herramientas combinan técnicas tradicionales de seguridad con aprendizaje automático avanzado para detectar amenazas en tiempo real basadas en el comportamiento del usuario.

IA cuántica

A medida que la computación cuántica se vaya convirtiendo en una realidad más cotidiana, es probable que asistamos a desarrollos específicos en IA cuántica. Estos sistemas podrían ser capaces de procesar información a velocidades exponencialmente más rápidas y manejar problemas de seguridad que son actualmente intratables para los sistemas convencionales. La «IA cuántica» es un campo emergente que combina técnicas y conceptos de la inteligencia artificial con la mecánica y la computación cuánticas. Aunque todavía está en sus primeras etapas, este ámbito promete revolucionar las capacidades y el rendimiento de los sistemas de IA.

Como es sabido, la computación cuántica se basa en la mecánica cuántica, una teoría de la física que describe cómo funcionan las partículas subatómicas. A diferencia de la computación clásica, que utiliza bits (0s y 1s) para representar y procesar información, la computación cuántica utiliza «qubits». Los qubits tienen la asombrosa capacidad de representar múltiples estados a la vez (superposición) y de estar «entrelazados», lo que significa que el estado de un qubit puede depender del estado de otro, independientemente de la distancia que los separe.

La utilización de modelos de IA cuántica tendría las siguientes **ventajas:**

- **Velocidad y Escalabilidad:** Los algoritmos cuánticos pueden realizar ciertas operaciones mucho más rápidamente que sus contrapartes clásicas. En teoría, la IA cuántica podría abordar problemas que actualmente son inabordables para las computadoras clásicas debido a su complejidad.
- **Optimización:** Problemas como la optimización combinatoria, que son cruciales en campos como la logística, la economía y muchas aplicaciones de IA, podrían beneficiarse enormemente de la capacidad de las computadoras cuánticas para explorar múltiples soluciones simultáneamente.
- **Aprendizaje profundo y entrenamiento de modelos:** Las computadoras cuánticas tienen el potencial de acelerar significativamente el entrenamiento de modelos de IA complejos, lo que podría revolucionar áreas como el aprendizaje profundo.

No obstante, como siempre, la utilización de estos modelos plantea ciertos **desafíos**, a saber:

- **Hardware:** Aunque se han logrado avances significativos en la construcción de computadoras cuánticas, todavía enfrentamos desafíos en términos de estabilidad, coherencia y escalabilidad.
- **Algoritmos cuánticos:** La adaptación de los algoritmos clásicos de IA a la computación cuántica sigue siendo un área activa de investigación. No todos los problemas se beneficiarán de una solución cuántica.
- **Interacción Cuántico-Clásica:** Integrar sistemas de computación cuántica con infraestructuras y algoritmos clásicos es un desafío considerable.

EJEMPLOS

IBM Q Experience:
IBM ha sido un líder en el campo de la computación cuántica, y a través de IBM Q, ofrece acceso a computadoras cuánticas reales para que los investigadores y desarrolladores experimenten y desarrollen algoritmos cuánticos. Aunque no es exclusivamente para IA, es una plataforma que podría usarse para experimentar con algoritmos de IA cuántica. (https://quantum-computing.ibm.com/)

D-Wave Systems:
D-Wave es conocido por sus sistemas cuánticos avanzados, diferentes de los modelos de computadoras gate-based. Han trabajado en optimización y machine learning utilizando sus sistemas cuánticos. (https://www.dwavesys.com/).

Google AI Quantum:
Google ha estado investigando activamente en el campo de la computación cuántica y ha hecho avances significativos. Aunque se centran en muchos aspectos de la computación cuántica, la inteligencia artificial es una de las aplicaciones que están explorando. (https://quantumai.google/).

Rigetti Computing:
Es una empresa emergente que se centra en la construcción de computadoras cuánticas y también proporciona una plataforma en la nube para que los desarrolladores y científicos experimenten con la computación cuántica. Aunque la plataforma no está dedicada exclusivamente a la IA cuántica, sí ofrece el potencial para explorar aplicaciones en ese dominio. (https://www.rigetti.com/)

Es importante señalar que muchas de estas herramientas y sistemas están diseñadas para ser plataformas de computación cuántica en general, no específicamente para IA cuántica. Sin embargo, dado que la computación cuántica tiene aplicaciones potenciales en IA, estas plataformas pueden desempeñar un papel crucial en el desarrollo futuro de la IA cuántica.

Colaboración entre humanos y máquinas

A pesar de los avances en IA, los humanos seguirán siendo esenciales en ciberseguridad. La tendencia emergente será la de sistemas donde humanos y máquinas trabajen juntos, complementando sus respectivas fortalezas.

Efectivamente, la colaboración entre humanos y máquinas es un enfoque en el que se aprovechan las capacidades únicas tanto de los seres humanos como de las máquinas, especialmente en el contexto de la inteligencia artificial, para mejorar la toma de decisiones, la eficiencia y los resultados generales. La idea principal detrás de esta colaboración es que, mientras que las máquinas son excelentes en el cálculo, el análisis y el procesamiento de grandes volúmenes de datos, los seres humanos poseen la intuición, el entendimiento contextual, la empatía y la creatividad.

Esta pretendida simbiosis hombre-máquina presenta algunos **aspectos clave** que conviene tener en cuenta:

1. **Complementariedad:** La IA y los seres humanos tienen puntos complementarios. Por ejemplo, la IA puede manejar grandes cantidades de datos, hacer operaciones repetitivas y cálculos complejos a velocidades que los humanos no pueden alcanzar. Sin embargo, los humanos aportan creatividad, juicio y experiencia basada en la intuición.
2. **Interacción natural:** El desarrollo de interfaces intuitivas y naturales, como la conversación por voz o la interacción gestual, permite una colaboración más fluida entre máquinas y humanos. Estos sistemas se basan en el procesamiento del lenguaje natural, el reconocimiento de voz, y el reconocimiento de gestos para entender y responder adecuadamente a las interacciones humanas.
3. **Aprendizaje bidireccional:** Mientras que la máquina aprende del comportamiento humano para mejorar sus predicciones y acciones, el ser humano también aprende a confiar y comprender cómo funciona la máquina, estableciendo un ciclo de retroalimentación y mejora constante.
4. **Transparencia y explicabilidad:** Para que los humanos confíen en las decisiones tomadas o sugeridas por las máquinas, es esencial que las máquinas puedan ofrecer explicaciones comprensibles de sus decisiones.

La tendencia emergente será la de sistemas donde humanos y máquinas trabajen juntos, complementando sus respectivas fortalezas

5. **Intervención humana:** En muchos sistemas de colaboración, se incorpora un mecanismo que permite la intervención humana en ciertos escenarios. Por ejemplo, en un sistema de conducción autónoma, puede haber situaciones en las que el sistema pida al conductor humano que tome el control.

EJEMPLOS

IBM Watson:
Uno de los usos más notables de Watson es en el campo de la medicina. Watson Health ayuda a los profesionales de la salud a tomar decisiones informadas sobre el tratamiento de los pacientes al analizar grandes cantidades de datos médicos y literatura científica. Aunque Watson proporciona recomendaciones, siempre es el médico quién toma la decisión final.
(https://www.merative.com/)

Google's DeepMind AlphaGo:
Aunque es más conocido por vencer a campeones humanos en el juego de Go, el verdadero logro aquí es cómo los humanos y la máquina aprendieron mutuamente, propiciando que los jugadores de Go estudien las jugadas de AlphaGo para mejorar sus propias estrategias.
(https://www.deepmind.com/research/highlighted-research/alphago)

KUKA LBR iiwa:
Es un robot colaborativo diseñado para trabajar junto a humanos en un entorno industrial. Estos «cobots» son sensibles al tacto y pueden detenerse o ralentizarse si detectan un objeto o persona en su camino, lo que permite a humanos y robots trabajar lado a lado en la misma tarea.
(https://www.kuka.com/en-us/products/robotics-systems/industrial-robots/lbr-iiwa)

OpenAI Codex:
Es una plataforma basada en IA que ayuda a los programadores a escribir código. Puede generar fragmentos de código en función de las descripciones proporcionadas por el usuario, actuando como un asistente de programación.
(https://openai.com/blog/openai-codex)

Adobe Sensei:
Integrado en las herramientas de Adobe, Sensei utiliza la IA y el aprendizaje automático para asistir en tareas creativas, desde la edición de fotos y videos hasta el diseño y la ilustración. Aunque automatiza algunas funciones, el creativo tiene el control final y utiliza la herramienta para mejorar y acelerar el proceso de diseño.
(https://www.adobe.com/sensei.html)

Estos ejemplos representan una variedad de aplicaciones en diferentes industrias, mostrando cómo las herramientas basadas en IA pueden trabajar en colaboración con los humanos para mejorar la eficiencia, la precisión y la creatividad.

IA en el borde (Edge AI)

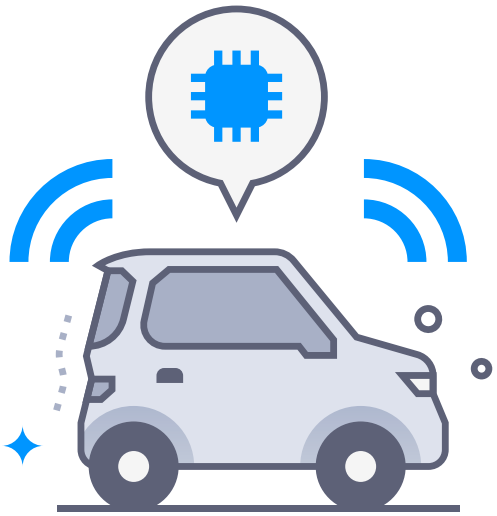
En lugar de depender de centros de datos centralizados, la IA podría procesarse en el dispositivo (como teléfonos móviles, IoT, etc.). Esto tiene implicaciones significativas para la ciberseguridad, ya que permite respuestas más rápidas y reduce los riesgos asociados con la transmisión de datos.

La utilización de este tipo de modelos presenta las siguientes **ventajas:**

1. **Latencia reducida:** Al procesar datos directamente en un dispositivo, se elimina la necesidad de enviar esos datos a un servidor centralizado para el procesamiento, lo que a su vez reduce la latencia. Esto es especialmente crucial en aplicaciones en tiempo real como vehículos autónomos.
2. **Privacidad y seguridad:** Mantener el procesamiento de datos en el dispositivo puede minimizar los riesgos de seguridad asociados con la transmisión de datos y garantizar que los datos sensibles no abandonen el dispositivo.
3. **Operación sin conexión:** Los dispositivos con capacidades de Edge AI pueden operar y tomar decisiones sin necesidad de una conexión activa a la nube o al servidor central.
4. **Eficiencia de ancho de banda:** Al reducir la necesidad de transmitir grandes cantidades de datos a la nube, se ahorra ancho de banda.
5. **Consumo de energía:** Aunque los dispositivos de Edge AI pueden requerir más potencia que los dispositivos sin capacidades de IA, a menudo consumen menos energía que la necesaria para comunicarse constantemente con un servidor central.

Las aplicaciones de este modelo serían de utilidad en varios contextos: **vehículos autónomos** (puesto que los vehículos necesitan procesar rápidamente enormes cantidades de datos de sus sensores para navegar con seguridad, la latencia en la toma de decisiones puede tener consecuencias graves); en **electrodomésticos inteligentes** (refrigeradores, aspiradoras, hornos y otros dispositivos que utilizan IA para la optimización y la toma de decisiones); en **dispositivos de salud portátiles** (monitores de ritmo cardíaco, dispositivos de glucosa y otros dispositivos médicos que necesitan procesar datos en tiempo real); en **cámaras de seguridad** (que detectan actividades anómalas o reconocen caras y toman decisiones basadas en tales datos) o en **drones** (utilizados para la navegación, detección de objetos y decisiones en tiempo real).

En lugar de depender de centros de datos centralizados, la IA podría procesarse en el dispositivo (como teléfonos móviles, IoT, etc.)



En el otro platillo de la balanza, los **desafíos** que se ven involucrados en este tipo de modelos son:

- **Limitaciones de hardware:** Aunque los dispositivos Edge están evolucionando rápidamente, todavía hay limitaciones en términos de capacidad de procesamiento, memoria y almacenamiento, en comparación con los centros de datos centralizados.
- **Gestión y actualización:** Mantener y actualizar modelos de IA en numerosos dispositivos dispersos puede ser un desafío.
- **Desarrollo de modelos:** A menudo, es necesario optimizar y adaptar los modelos para que sean lo suficientemente pequeños y eficientes para ejecutarse en dispositivos Edge sin sacrificar excesivamente la precisión o la capacidad.

Todas estas tendencias emergentes en la IA para la ciberseguridad representan sólo una pequeña parte de lo que sin duda vendrá.

EJEMPLOS

TensorFlow Lite:
Es una solución de Google diseñada para llevar los modelos de machine learning a dispositivos móviles y Edge. Proporciona herramientas para convertir y optimizar modelos TensorFlow estándar para ser eficientes en estos dispositivos.
(https://www.tensorflow.org/lite)

ONNX Runtime:
Comprende una biblioteca de inferencia de machine learning para modelos ONNX (Open Neural Network Exchange). ONNX Runtime es ligero y puede usarse en diferentes plataformas, incluyendo dispositivos Edge.
(https://onnxruntime.ai/)

Intel OpenVINO (Open Visual Inferencing & Neural Network Optimization) Toolkit:
Una herramienta de Intel para acelerar y optimizar modelos de IA, específicamente diseñada para ejecutarse eficientemente en hardware de Intel, incluyendo chips diseñados para dispositivos Edge.
(https://software.intel.com/content/www/us/en/develop/tools/openvino-toolkit.html)

NVIDIA Jetson Platform:
Es una serie de sistemas embebidos diseñados por NVIDIA que integran una GPU y están optimizados para tareas de inferencia de IA Edge.
(https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/)

Azure IoT Edge:
Es una solución de Microsoft que permite desplegar cargas de trabajo en la nube directamente en dispositivos IoT (Internet of Things) y en otros dispositivos Edge, incluidos modelos de machine learning.
(https://azure.microsoft.com/en-us/services/iot-edge/)

Estas herramientas y plataformas representan solo una fracción de la creciente industria de la IA Edge.

6.2 Investigaciones en curso

El ámbito de la ciberseguridad es un terreno fértil para la investigación, especialmente con la integración de tecnologías emergentes como la Inteligencia Artificial. En este epígrafe se exploran las áreas más recientes de investigación.

Aprendizaje profundo contra ataques de red avanzados:	Los ataques de red están evolucionando para ser más sofisticados y difíciles de detectar. Las investigaciones actuales están enfocadas en cómo utilizar técnicas de aprendizaje profundo, como las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN), para identificar patrones ocultos en el tráfico de red que puedan indicar un ataque.
Técnicas adversarias para robustecer los sistemas de IA:	La idea es utilizar técnicas adversarias (ataques diseñados específicamente para engañar a los modelos de IA) en un entorno controlado para mejorar la robustez de los modelos de IA en ciberseguridad. Esto implica entrenar modelos con datos perturbados para hacerlos más resistentes a ataques adversarios en escenarios del mundo real.
Seguridad de la cadena de suministro de IA:	A medida que la IA se convierte en una parte esencial de muchos sistemas, garantizar la integridad de toda la cadena de suministro de IA (desde el entrenamiento hasta la implementación) se ha vuelto crucial. Las investigaciones se centran en cómo se pueden infiltrar y comprometer estos sistemas y cómo defenderlos contra tales vulnerabilidades.
Detección de deepfakes y, en general, contenido sintético:	Con la creciente capacidad de las herramientas de IA para crear contenido sintético realista (como las deepfakes de audio o video, por ejemplo), la investigación en la detección automática de dicho contenido se ha intensificado. Esto tiene aplicaciones directas en la ciberseguridad, particularmente en áreas como la autenticación y la protección contra la desinformación.
Automatización de la respuesta a incidentes:	En lugar de simplemente detectar amenazas, se están desarrollando sistemas de IA que pueden responder automáticamente a incidentes de seguridad, tomando decisiones en tiempo real sobre cómo mitigar o neutralizar una amenaza.

Comprendiendo la «caja negra» de la IA:	Una gran área de investigación es la IA explicable (XAI). En el contexto de la ciberseguridad, es esencial comprender por qué un sistema de IA toma una decisión particular, especialmente si está relacionada con la detección de amenazas o la respuesta a incidentes.
IA para protección contra amenazas internas:	Las amenazas internas, ya sean maliciosas o inadvertidas, siguen siendo un desafío importante en la ciberseguridad. La IA puede desempeñar un papel en la monitorización del comportamiento del usuario para identificar actividades sospechosas o desviaciones de la norma.
Metodologías de entrenamiento seguras:	Investigar formas de entrenar modelos de IA sin exponer datos sensibles, como el aprendizaje federado o el aprendizaje con privacidad diferencial.

Estas áreas de investigación muestran la continua evolución y adaptación de la IA en el ámbito de la ciberseguridad. A medida que las amenazas cambian y se vuelven más sofisticadas, es esencial que la investigación en este campo mantenga el ritmo para proporcionar soluciones efectivas y proactivas.

EJEMPLOS

Aprendizaje profundo contra ataques de red avanzados:
«DeepDefense», una plataforma que usa técnicas de aprendizaje profundo para detectar ataques en tiempo real. (https://ieeexplore.ieee.org/document/7946998)

Técnicas adversarias para robustecer los sistemas de IA:
Proyecto «CleverHans» de Google. (GitHub.com/tensorflow/cleverhans)

Seguridad de la cadena de suministro de IA:
MITRE Corporation ha estado investigando en áreas relacionadas. (mitre.org).

Detección de deepfakes y contenido sintético:
«Deepware Scanner» de la compañía Cyabra.
(cyabra.com)

Automatización de la respuesta a incidentes:
«Cortex XSOAR» de Palo Alto Networks. (paloaltonetworks.com/cortex/xsoar)

Comprendiendo la «caja negra» de la IA:
LIME (Local Interpretable Model-Agnostic Explanations)». (github.com/marcotcr/lime)

IA para protección contra amenazas internas:
«Insider Threat Solution» de Varonis. (varonis.com/solutions/insider-threat-detection)

Metodologías de entrenamiento seguras:
«TensorFlow Federated» para el aprendizaje federado. (tensorflow.org/federated)

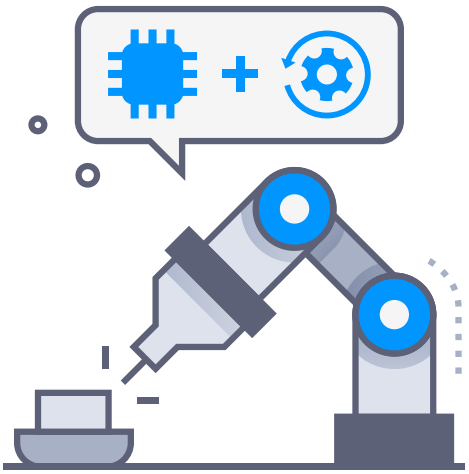
6.3 Impacto potencial en la industria y la sociedad

El avance de la IA en la ciberseguridad tiene el potencial de redefinir muchos aspectos de la industria y ejercer una influencia notable en la sociedad. A medida que los sistemas autónomos se vuelven más sofisticados, las **repercusiones** se hacen sentir en una variedad de niveles, a saber:

1. **Optimización de la seguridad empresarial:**
   - Las empresas pueden esperar una mayor protección contra amenazas con sistemas que pueden aprender y adaptarse en tiempo real.
   - Se anticipa una reducción en los tiempos de respuesta a incidentes y una capacidad mejorada para prevenir brechas antes de que sucedan.
2. **Transformación del trabajo de los profesionales de la ciberseguridad:**
   - La IA puede manejar tareas rutinarias, permitiendo que los profesionales de la ciberseguridad se centren en tareas más estratégicas o complejas.
   - Esto podría resultar en una reestructuración de roles y responsabilidades, y también en la necesidad de nuevas habilidades y formación.
3. **Sociedad digitalmente más segura:**
   - A medida que las soluciones basadas en IA se integren en más plataformas y servicios, el ciudadano promedio puede beneficiarse de una seguridad digital más robusta en su vida diaria.
   - Las transacciones en línea, el almacenamiento de datos personales y otras actividades digitales podrían experimentar menos riesgos asociados.

El avance de la IA en la ciberseguridad tiene el potencial de redefinir muchos aspectos de la industria y ejercer una influencia notable en la sociedad

4. **Cambios en la naturaleza de las amenazas:**
   - Los atacantes también evolucionarán y adaptarán sus métodos en respuesta a sistemas de defensa más avanzados.
   - Podríamos ver un aumento en los ataques altamente sofisticados y dirigidos que utilizan la IA para encontrar y explotar vulnerabilidades.
5. **Desafíos éticos y privacidad:**
   - A medida que la IA se convierte en una herramienta común en ciberseguridad, surgirán preocupaciones sobre el uso y abuso de datos personales.
   - Las empresas y organizaciones deberán ser transparentes sobre cómo usan la IA y garantizar que se respeten los derechos de privacidad.
6. **Economía y mercado laboral:**
   - La adopción masiva de soluciones de IA podría influir en la demanda de profesionales de ciberseguridad, posiblemente aumentando la demanda de expertos especializados en IA y disminuyendo la necesidad de roles más tradicionales.
   - Las startups y empresas que desarrollan soluciones de IA para ciberseguridad podrían experimentar un crecimiento significativo, influenciando la economía y creando nuevas oportunidades de mercado.
7. **Normativas y regulaciones:**
   - Los gobiernos y organismos reguladores de todo el mundo —como está sucediendo en la Unión Europea en la actualidad— podrían introducir nuevas regulaciones para garantizar que la IA se utilice de manera responsable, también en la ciberseguridad.
   - Estas regulaciones podrían influir en cómo las empresas desarrollan, implementan y usan soluciones de IA.



El impacto de la IA en la ciberseguridad es vasto y abarca muchas facetas de la industria y la sociedad. Es esencial que las partes interesadas, desde profesionales hasta reguladores, estén informadas y preparadas para abordar estos cambios de manera proactiva y ética.

7. Recomendaciones y buenas prácticas

A medida que la Inteligencia Artificial (IA) cobra cada vez más relevancia en el mundo de la ciberseguridad, tanto para la defensa como para el ataque, se vuelve imperativo que las organizaciones adopten estrategias informadas para su implementación. Sin embargo, la IA no es una solución mágica que pueda ser aplicada sin contemplaciones; su uso efectivo requiere una comprensión matizada y un enfoque estratégico.

En este epígrafe, exploraremos recomendaciones y buenas prácticas que las organizaciones deben considerar al integrar soluciones de IA en sus sistemas de ciberseguridad. Desde la selección y el entrenamiento de modelos hasta su implementación y monitorización en tiempo real, es esencial que las organizaciones estén equipadas con las mejores prácticas para garantizar que la IA sea una ventaja y no un punto débil.

Abordaremos cómo garantizar la robustez y fiabilidad de los sistemas de IA, cómo manejar y proteger los datos que alimentan estos sistemas, y cómo asegurar que la ética y la transparencia sean centrales en cualquier implementación de IA. Asimismo, se destacará la importancia de la formación continua y la adaptabilidad, dada la naturaleza cambiante de las ciberamenazas.

A medida que la Inteligencia Artificial (IA) cobra cada vez más relevancia en el mundo de la ciberseguridad, tanto para la defensa como para el ataque, se vuelve imperativo que las organizaciones adopten estrategias informadas para su implementación

7.1 Integración de equipos de ciberseguridad y equipos de IA

La convergencia efectiva de la inteligencia artificial y la ciberseguridad requiere no solo la combinación de tecnologías, sino también la colaboración entre los expertos en ambas áreas. La integración efectiva de estos equipos puede potenciar las capacidades de ciberdefensa de una organización y garantizar que las soluciones de IA sean robustas, confiables y adecuadas para enfrentar ciberamenazas reales.

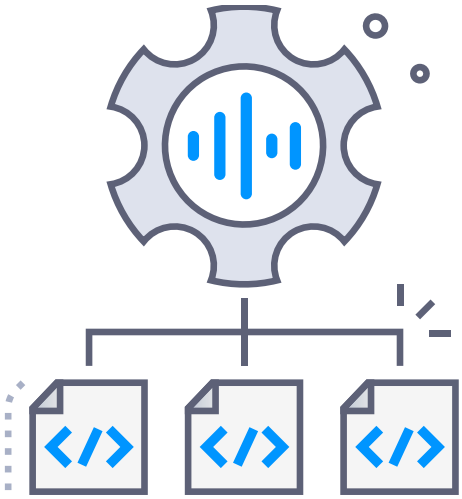
Elementos esenciales para el éxito:

1. **Comunicación y colaboración:** Es fundamental que haya un flujo de comunicación constante y eficiente entre los equipos de ciberseguridad y los de IA. Los desafíos, objetivos y soluciones de ambos campos deben ser compartidos y comprendidos mutuamente.
2. **Formación cruzada:** Impartir formación sobre los fundamentos de ciberseguridad al equipo de IA y, recíprocamente, sobre los principios básicos de la IA al equipo de ciberseguridad puede construir un entendimiento mutuo y mejorar la colaboración.
3. **Desarrollo conjunto de soluciones:** En lugar de trabajar en silos, los equipos de IA y ciberseguridad deben colaborar en el diseño, desarrollo y despliegue de soluciones. Esto garantiza que las soluciones de IA sean relevantes y estén alineadas con los objetivos de ciberseguridad.
4. **Revisiones regulares y feedback:** Las soluciones de IA en ciberseguridad deben ser sometidas a revisiones regulares por ambos equipos. Estas revisiones pueden identificar deficiencias, áreas de mejora y las adaptaciones necesarias para enfrentar amenazas emergentes.
5. **Pruebas conjuntas:** Al igual que en el desarrollo, la prueba de las soluciones también debe ser una actividad conjunta. Esto puede ayudar a identificar y rectificar fallos antes de que sean explotados por adversarios.

La convergencia efectiva de la inteligencia artificial y la ciberseguridad requiere no solo la combinación de tecnologías, sino también la colaboración entre los expertos en ambas áreas

6. **Integración de herramientas y plataformas:** Utilizar herramientas y plataformas que permitan la integración y colaboración entre ambos equipos puede ser clave, contemplando la utilización de plataformas de desarrollo colaborativo, sistemas de gestión de proyectos y herramientas de comunicación.
7. **Respeto y valoración mutua:** Para una colaboración efectiva, es esencial que ambos equipos reconozcan y valoren las competencias y contribuciones del otro. La inteligencia artificial y la ciberseguridad son disciplinas complejas y especializadas, y es esencial que se respeten mutuamente para garantizar una colaboración efectiva.
8. **Planificación de escenarios de crisis:** En el caso de un incidente de seguridad, es fundamental tener planes establecidos sobre cómo los equipos trabajarán juntos. Esto incluye la determinación de roles, responsabilidades y flujos de comunicación.
9. **Actualizaciones y formación continua:** Dado que tanto el campo de la IA como el de la ciberseguridad están en constante evolución, es esencial que ambos equipos se mantengan actualizados sobre las últimas tendencias, técnicas y amenazas en sus respectivas áreas.

En definitiva, la integración efectiva de los equipos de ciberseguridad y de IA no es simplemente una opción, sino una necesidad en el mundo actual. Las ciberamenazas están evolucionando rápidamente, y la combinación de experiencia en ciberseguridad con capacidades avanzadas de IA puede ofrecer una defensa robusta contra adversarios cada vez más sofisticados. Sin embargo, para que esta colaboración sea fructífera, es esencial que se adopten prácticas adecuadas que fomenten la comunicación, la comprensión y la colaboración entre estos equipos especializados.



7.2 Formación continua

En un mundo donde la tecnología y las amenazas cibernéticas evolucionan a un ritmo vertiginoso, la capacitación y la formación continuas son esenciales para mantenerse al día y garantizar una defensa efectiva contra los adversarios. Este imperativo no solo se aplica a los profesionales de la ciberseguridad, sino también a aquellos que trabajan en la intersección de la IA y la ciberseguridad.

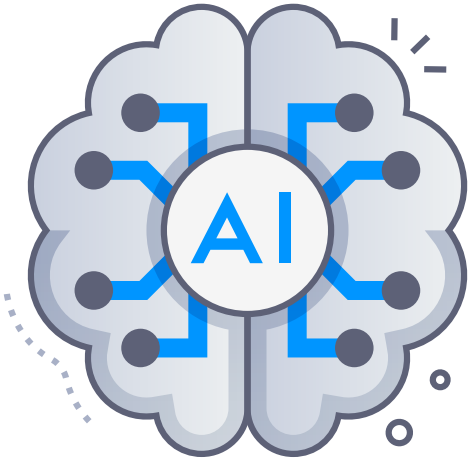
Para lograr todo ello, es conveniente tener en cuenta:

1. **Programas de formación actualizados:** Las instituciones educativas y los organismos de formación deben revisar y actualizar regularmente sus currículos para reflejar los desarrollos más recientes en ciberseguridad y IA. Esto garantizará que los nuevos profesionales posean el conocimiento más actualizado.
2. **Workshops y seminarios:** Organizar o asistir a talleres y seminarios sobre temas emergentes puede proporcionar una visión profunda de técnicas específicas, amenazas emergentes o nuevos enfoques en ciberseguridad y IA.
3. **Certificaciones profesionales:** Las certificaciones como CISSP, CISM, y otras relacionadas con IA y/o ciberseguridad, pueden ayudar a los profesionales a validar y mejorar sus habilidades. Estas certificaciones a menudo requieren educación continua, garantizando que los profesionales se mantienen actualizados.
4. **Capacitación en el trabajo:** Las organizaciones deben fomentar una cultura de aprendizaje continuo, ofreciendo oportunidades para que los empleados se formen en nuevas herramientas, técnicas y mejores prácticas.
5. **Participación en comunidades y foros:** Ser miembro activo en comunidades en línea y foros especializados puede proporcionar una plataforma para aprender de los colegas, compartir conocimientos y estar al tanto de los últimos desarrollos y desafíos.

En un mundo donde la tecnología y las amenazas cibernéticas evolucionan a un ritmo vertiginoso, la capacitación y la formación continuas son esenciales para mantenerse al día y garantizar una defensa efectiva contra los adversarios

6. **Simulacros y ejercicios prácticos:** La realización de simulacros y ejercicios prácticos, como «capture the flag» o wargames, puede ayudar a los profesionales a aplicar sus conocimientos en escenarios del mundo real, mejorar sus habilidades y aprender de sus errores.
7. **Recursos en línea y MOOCs:** Con la proliferación de plataformas de educación en línea, hay una abundancia de cursos (a menudo gratuitos) que cubren una amplia variedad de temas en IA y ciberseguridad. Estos pueden ser una excelente manera de aprender a su propio ritmo.
8. **Asistencia a conferencias:** Las conferencias como Black-Hat, DEF-CON, Jornadas STIC CCN-CERT y otras específicas de IA, ofrecen una oportunidad para aprender sobre investigaciones recientes, descubrimientos y desarrollos en el campo.

En resumen, la formación y educación continua en ciberseguridad y IA no es un lujo, sino una necesidad. Para organizar una defensa eficaz contra las amenazas actuales y futuras, los profesionales deben estar constantemente actualizando y ampliando sus conocimientos y habilidades. Las organizaciones y profesionales que invierten en educación continua estarán mejor posicionados para enfrentar y mitigar los riesgos en el cambiante panorama de la ciberseguridad.



Se muestra seguidamente una lista de referencias formativas que abarcan tanto la Inteligencia Artificial (IA) como la ciberseguridad. Se trata de un pequeño número de referencias que incluyen cursos, certificaciones y recursos.

1. **Cursos y Especializaciones:**
   - **Coursera:**
     - Especialización en Ciberseguridad
     - Introducción a la Inteligencia Artificial
     - Deep Learning Specialization
   - **edX:**
     - Fundamentos de Ciberseguridad
     - Principios de Inteligencia Artificial
   - **Udacity:**
     - Nanodegree en Ciberseguridad
     - Nanodegree en Inteligencia Artificial

2. **Certificaciones:**
   - **Certified Information Systems Security Professional (CISSP):**
     Una certificación reconocida globalmente que demuestra la capacidad y el conocimiento en ciberseguridad. Más información en el sitio oficial de ISC^2.
   - **Certified Information Security Manager (CISM):**
     Ofrecida por ISACA, esta certificación es esencial para la gestión de la seguridad de la información. Más detalles en el sitio oficial de ISACA.
   - **TensorFlow Developer Certificate:**
     Una certificación centrada en la IA y el aprendizaje profundo. Encuentra más información en el sitio oficial de TensorFlow.

3. **Recursos Adicionales:**
   - **MIT OpenCourseWare:**
     - Introducción a Deep Learning
     - Computación y Seguridad de Redes
   - **Cybrary:**
     Una plataforma que ofrece cursos gratuitos en ciberseguridad y otros campos relacionados. Visita el sitio web de Cybrary.
   - **ArXiv:**
     Un recurso invaluable para investigadores, ArXiv es un repositorio de artículos preprint en varios campos, incluidos la IA y la ciberseguridad. Revisa ArXiv para estar al día con la investigación actual.
   - **Plataforma ANGELES**, del CCN-CERT
     https://angeles.ccn-cert.cni.es/es/

7.3 Diseño de sistemas robustos y resilientes

La robustez y resiliencia en el diseño de sistemas, especialmente aquellos integrados con Inteligencia Artificial, son esenciales para garantizar que sigan funcionando de manera óptima bajo condiciones adversas y puedan recuperarse rápidamente de fallos o ataques.

Entendemos por **robustez** la capacidad de un sistema para resistir perturbaciones y continuar funcionando correctamente sin degradación. Un sistema robusto puede manejar condiciones imprevistas y variaciones en el entorno operativo.

Por su parte, entendemos por **resiliencia** la capacidad de un sistema para recuperarse rápidamente de fallos, adaptándose a nuevas condiciones y restableciendo su funcionamiento normal.

Para lograr sistemas robustos y resilientes es necesario considerar los **elementos siguientes:**

Principios de Diseño	**Minimización del Área de Ataque:**	Reducir la superficie de ataque limitando los puntos de entrada al sistema y eliminando componentes innecesarios.
	Redundancia:	Implementar sistemas y componentes duplicados que puedan asumir carga de trabajo en caso de fallo de un componente primario.
	Segmentación/segregación:	Dividir el sistema en segmentos más pequeños e independientes, para que un fallo o ataque en un segmento no comprometa todo el sistema.
	Monitorización/vigilancia continua:	Utilizar herramientas y soluciones de monitorización y vigilancia para detectar rápidamente cualquier irregularidad o anomalía.
	Actualizaciones Regulares:	Mantener el software y hardware actualizados para corregir vulnerabilidades conocidas.

Consideraciones para IA	**Datos de Entrenamiento Robustos:**	Garantizar que los modelos de IA se entrenan con conjuntos de datos variados y actualizados para manejar diversas situaciones.
	Validación Rigurosa:	Evaluar y validar modelos en varios escenarios y condiciones.
	Defensa contra Ataques Adversarios:	Implementar técnicas como la regularización y el incremento de datos para proteger los modelos de IA de ataques que persigan explotar sus debilidades.
	Transparencia y Explicabilidad:	Usar modelos de IA que puedan ser interpretados y auditados para entender su funcionamiento y toma de decisiones.
Pruebas y Simulaciones	**Monitorización Continua:**	Realizar pruebas regulares en entornos controlados para identificar y corregir vulnerabilidades.
	Simulaciones de Fallos:	Simular fallos en diferentes partes del sistema para evaluar la respuesta y el tiempo de recuperación.
	Ejercicios de Respuesta a Incidentes:	Practicar, conforme a la normativa que resulte de aplicación, la respuesta a posibles incidentes de seguridad para mejorar la eficiencia y eficacia en situaciones reales.
Cultura de Mejora Continua	**Fomentar una mentalidad en la que se esté constantemente buscando mejorar la robustez y resiliencia del sistema, aprendiendo de los incidentes y adaptándose a las nuevas amenazas y desafíos.**	

En resumen, el diseño de sistemas robustos y resilientes es esencial para garantizar que los sistemas, especialmente aquellos integrados con IA, puedan manejar y recuperarse rápidamente de condiciones adversas. Esta es una tarea continua que requiere una combinación de técnicas de diseño, pruebas rigurosas y una cultura de mejora constante.

8. Conclusión

8.1 Reflexiones finales sobre el estado actual y el futuro de la IA en la ciberseguridad

La evolución de la ciberseguridad y la inteligencia artificial ha demostrado ser un binomio fascinante, pero también desafiante. Ambas disciplinas, por separado, tienen trayectorias complejas, y su intersección ha provocado tanto revoluciones como dilemas. Al reflexionar sobre su estado actual y el panorama futuro, podemos extraer varias **consideraciones clave:**

1. **Interdependencia Creciente:**
   La ciberseguridad ya no puede considerarse una disciplina independiente de la IA. La vastedad y complejidad del ciberespacio, combinada con la cantidad abrumadora de datos generados, hace que las soluciones basadas en IA sean esenciales para una defensa efectiva.
2. **Desafíos Cambiantes:**
   A medida que la IA se vuelve más avanzada, también lo hacen las amenazas. Los actores maliciosos adoptan rápidamente nuevas tecnologías para mejorar sus tácticas. Es un juego constante de gato y ratón, donde la defensa y el ataque evolucionan en paralelo.

3. **Relevancia del Factor Humano:**
   A pesar de la automatización y las capacidades avanzadas que la IA aporta, el factor humano sigue siendo insustituible. Las decisiones éticas, la interpretación de datos y la comprensión del contexto siguen siendo responsabilidad humana. La colaboración entre humanos y máquinas será fundamental para el éxito de la ciberseguridad en el futuro.
4. **Desafíos Éticos y Regulatorios:**
   La adopción de IA en la ciberseguridad trae consigo desafíos éticos y regulatorios, como el caso del Reglamento Europeo en materia de IA al que nos hemos referido en este trabajo en varias ocasiones. La privacidad, el consentimiento y la transparencia son áreas que deben ser abordadas con precaución y responsabilidad, especialmente cuando se equilibran con la necesidad de seguridad.
5. **Potencial sin Explotar:**
   Si bien hemos visto avances impresionantes en la IA aplicada a la ciberseguridad, todavía hay un vasto potencial sin explotar. Tecnologías emergentes, como la IA cuántica y el aprendizaje federado, pueden remodelar aún más el panorama de la ciberseguridad en la próxima década.
6. **Preparación para el Futuro:**
   Las organizaciones y los profesionales de la ciberseguridad deben estar preparados para adaptarse rápidamente. La educación continua, la investigación y la colaboración interdisciplinar serán esenciales para mantenerse al día con las últimas tendencias y amenazas.
7. **Visión Holística:**
   La ciberseguridad, en su esencia, es una disciplina holística. Ya no se trata solo de tecnología, sino también de procesos, personas y políticas. La adopción de la IA debe ser vista como parte de un enfoque más amplio y estratégico para proteger el ciberespacio.

En conclusión, el entrelazamiento de la IA y la ciberseguridad está redefiniendo el futuro de la seguridad digital. Si bien presenta oportunidades sin precedentes para una defensa más eficaz y una detección más rápida, también introduce desafíos complejos que deben ser abordados con prudencia, innovación y colaboración. La trayectoria futura de esta intersección será sin duda apasionante y determinante para el futuro digital de la humanidad.

La colaboración entre humanos y máquinas será fundamental para el éxito de la ciberseguridad en el futuro

8.2 Acciones subsiguientes y recomendaciones para futuras investigaciones

El paisaje en evolución de la ciberseguridad y la inteligencia artificial no solo exige una reflexión sobre lo que hemos aprendido hasta ahora, sino también una visión clara de los pasos a seguir. A medida que avanzamos hacia un futuro más digitalizado e interconectado, es esencial que la comunidad global —desde investigadores y profesionales hasta legisladores y ciudadanos comunes— se una en la misión de asegurar nuestro ciberespacio.

Se formulan seguidamente algunas recomendaciones y acciones subsiguientes:

1. **Establecimiento de Centros de Investigación Colaborativos:**
   Es esencial establecer más centros y plataformas que permitan la colaboración interdisciplinar en ciberseguridad y IA. Estos centros pueden actuar como puntos focales para la investigación innovadora, reuniendo a expertos en IA, ciberseguridad, derecho y otros campos relacionados.
2. **Promover la Educación y Formación Especializada:**
   Hay una necesidad urgente de desarrollar programas educativos y de formación que se centren en la intersección de la IA y la ciberseguridad. Esta cuestión no solo pretende abordar la escasez de habilidades en este campo, sino que también garantiza que los futuros profesionales posean el conocimiento necesario.
3. **Normativas y Estándares Globales:**
   La comunidad internacional debería trabajar conjuntamente para desarrollar estándares y regulaciones en torno a la aplicación de la IA en ciberseguridad. Estas normativas no solo proporcionarán un marco de referencia, sino que también asegurarán que la tecnología se utilice de manera ética y responsable. Para ello se considera

A medida que avanzamos hacia un futuro más digitalizado e interconectado, es esencial que la comunidad global —desde investigadores y profesionales hasta legisladores y ciudadanos comunes— se una en la misión de asegurar nuestro ciberespacio

imprescindible formalizar marcos de certificación para tecnologías, productos y servicios IA confiables[44].

4. **Investigación en Amenazas Emergentes:**
   Con la rápida evolución de la IA, las amenazas también cambian y se adaptan. Es crucial financiar y priorizar la investigación sobre amenazas emergentes, especialmente aquellas que se originan a partir de los más recientes avances tecnológicos.

5. **Desarrollo de Herramientas de IA Explicables:**
   El mundo necesita más investigación sobre herramientas y métodos que hagan que la IA sea más transparente y comprensible. Las decisiones tomadas por algoritmos de IA en el ámbito de la ciberseguridad pueden tener repercusiones significativas, por lo que es esencial que se puedan explicar y entender.

6. **Promoción de la Privacidad y la Ética:**
   Las futuras investigaciones no solo deben centrarse en la efectividad y eficiencia de las soluciones de IA en ciberseguridad, sino también en su impacto ético y en la privacidad. La privacidad no debe ser un sacrificio para alcanzar la seguridad.

7. **Pruebas y Validaciones Rigurosas:**
   Antes de implementar soluciones basadas en IA en entornos reales, es crucial realizar pruebas y validaciones exhaustivas. Esto asegurará que las soluciones sean robustas y confiables ante amenazas del mundo real.

8. **Incentivos para la Innovación:**
   Los gobiernos y organizaciones privadas deben ofrecer incentivos para la innovación en ciberseguridad y IA. Esto puede tomar la forma de subvenciones, competencias o reconocimientos.

En resumen, estamos en un punto crucial en la intersección de la ciberseguridad y la IA. A medida que estas disciplinas continúan evolucionando y entrelazándose, es esencial que adoptemos un enfoque proactivo, colaborativo, regulatorio y ético para enfrentar los desafíos del futuro. La llamada a la acción es clara: debemos unirnos en la misión de asegurar nuestro futuro digital, protegiendo nuestros datos y nuestras infraestructuras; en última instancia: nuestras sociedades.

44 Sobre el particular, véase el excelente trabajo de Rand Corporation. *Labelling Initiatives, codes of conduct and othres self-regulatory mechanisms for artificial intelligence applications*. (2022). https://www.rand.org/pubs/research_reports/RRA1773-1.html

centro criptológico nacional

centro criptológico nacional

www.ccn.cni.es

www.ccn-cert.cni.es

https://oc.ccn.cni.es/


ccn-cert
centro criptológico nacional
CCN
centro criptológico nacional


PUBLICACIONES
Pd
DE DEFENSA


GOBIERNO
DE ESPAÑA
MINISTERIO
DE DEFENSA
SUBSECRETARÍA DE DEFENSA
SECRETARÍA GENERAL TÉCNICA
SUBDIRECCIÓN GENERAL
DE PUBLICACIONES
Y PATRIMONIO CULTURAL