PANORAMA DE LA INTELIGENCIA ARTIFICIAL EN EL DOMINIO DE LA CIBERSEGURIDAD

OVERVIEW OF ARTIFICIAL INTELLIGENCE IN THE DOMAIN OF CYBERSECURITY

Autor:

Santiago Portela. portela.santiago@gmail.com ORCID https://orcid.org/0000-0003-0610-6618

Resumen:

La Inteligencia Artificial está entrando de forma acelerada en el dominio de la ciberseguridad, de la mano de los grandes actores tecnológicos y startups especializadas que realizan grandes inversiones impulsadas por planes estratégicos entre los que destacan China y EEUU con uno de sus focos de confrontación centrado en la generación de talento. El mercado ya ha incorporado generalizadamente en sus productos de protección técnicas de *Machine Learning*. Las agencias de seguridad y estándares analizan y clasifican la aplicación de la IA a la seguridad, destacando en Europa los informes de Europol y ENISA. La IA como fuente de amenaza también evoluciona creándose *malware* elusivo, ataques a la propia cadena de suministro de soluciones IA, y analítica del comportamiento humano con un potencial elevado de impacto en las técnicas de subversión política.

Abstract:

Artificial Intelligence is rapidly entering the domain of cybersecurity, leaded by big technology players and specialized startups that make large investments driven by strategic plans, among which China and the US stand out, one of their sources of confrontation centered on talent generation. Markets have widely incorporated machine learning techniques into their security products. Security and standards agencies analyze and classify the application of AI to security, highlighting in Europe the reports of Europol and ENISA. AI as a threat source is also evolving, creating elusive *malware*, attacks on the AI supply chain itself, and human behavior analytics with a high potential impact on political subversion techniques.

Palabras clave:

Ciberseguridad; Inteligencia artificial; Aprendizaje automático

Keywords:

Cybersecurity; Artificial Intelligence; Deep Learning

> — *¿Usted cree que queda mucho por descubrir en el campo de la electricidad?*
> — *Es un campo de campos. No podría decirle, pero guarda el secreto que reorganizará la vida del mundo.*
> — *Ha descubierto usted mucho al respecto.*
> — *Sí, y sin embargo muy poco en comparación con las posibilidades que ofrece.*

Esta entrevista de Orison Swett a Thomas Edison en los albores del siglo XX ilustra la aparición de un avance tecnológico transversal con un poder disruptivo histórico. La electricidad transformó el mundo de formas inimaginables para sus coetáneos. A la ingeniería eléctrica siguió la electrónica, de telecomunicaciones y la radioquímica. Ese es el vértigo que produce el potencial de la Inteligencia Artificial en los inicios del siglo XXI. Así lo transmitía el informe presentado en abril de 2021 por la Comisión Nacional de Seguridad de la Inteligencia Artificial de Estados Unidos (NSCAI). Comenzaba con esta misma cita.

Las tecnologías en torno a la Inteligencia Artificial están ya presentes en muchas facetas de la realidad social y económica. También lo están en el dominio de la seguridad de la información, y en este estudio trataremos de mostrar el panorama del estado de la cuestión a principios del año 2022.

ESCENARIO DE INSEGURIDAD

En los últimos años la ciberseguridad demanda de forma creciente la introducción de recursos tecnológicos y organizativos en un escenario de riesgo creciente y exigencia normativa. El Centro Criptológico Nacional, en su informe de Ciberamenazas y Tendencias de diciembre de 2021[1] , considera que 2020 fue un año de inflexión,

[1] https://www.ccn-cert.cni.es/informes/informes-ccn-cert-publicos/6338-ccn-cert-ia-13-21-ciberamenazas-y-tendencias-edicion-2021-1/file.html

duplicando con 82.530 ataques los del año precedente. INCIBE a su vez [2] detectó en el mismo periodo 133.155 incidentes de los cuales 25.499 contra la red académica y de investigación RedIris. Las cifras son duras, pero más lo son las valoraciones cualitativas del CCN alertando de amenazas constantes para el sector sanitario y de un aumento de la superficie de ataque de 'la casa conectada' a partir de los cambios sociotécnicos derivados de la pandemia de COVID-19. El escenario de Ciberguerra puesto de relieve durante la agresión rusa a Ucrania empeora la amenaza con una mayor intención de ataque respaldado por estados.

Entre los responsables de ciberseguridad circula un mensaje pesimista: "Ser atacado es como andar en moto: están los que se han caído, y los que se caerán". Todos los informes sectoriales alertan de falsa percepción de seguridad, de cibercrimen creciente, de insuficiencia en la prevención y capacidad de reacción. Sin embargo, los fabricantes que impulsan la introducción de IA en la ciberseguridad hablan de un nuevo paradigma que reduce eficazmente la vulnerabilidad en el *endpoint*, es decir que reduce la superficie de exposición. ¿Esperanza, o mercadotecnia? Nos hemos habituado a cerrar los oídos ante el abuso comercial del término 'Inteligencia Artificial', pero el propio CCN reconoce que estos avances podrían acelerar considerablemente la identificación de nuevas amenazas y sus respuestas para frenar los ataques antes de que puedan propagarse.

Dentro de la diversidad de formas de ataque destaca la virulencia de los ataques de *ransomware*, agravados con un mecanismo de doble extorsión cuando se exfiltra información confidencial y se amenaza con su difusión, o con un propósito meramente destructivo que busca aniquilar la capacidad del atacado.

Dado que el principal vector de ataque para este caso se origina en las cuentas y dispositivos de usuario final, tomaremos este punto de partida para el análisis.

[2] https://www.incibe.es/sala-prensa/notas-prensa/incibe-gestiono-mas-130000-incidentes-ciberseguridad-durante-el-ano-2020

PROTECCIÓN DEL *ENDPOINT*

Según EY[3], el 70% de los incidentes se genera en el *endpoint*, es decir el dispositivo final conectado a la red, y dentro de estos los más comprometidos son el ordenador personal y el smartphone. La fórmula ampliamente desplegada en el pasado de agentes antivirus que consultaban diariamente las librerías de patrones de *malware* ha sido superada por la diversidad de amenazas, las técnicas de mutación y ocultamiento, y la virulencia de los ataques basados en vulnerabilidades desconocidas "Zero Day". Los modelos de detección y respuesta basados en el *endpoint* (EDR, XDR) integran los dispositivos en un proceso de observación y análisis permanente, habitualmente integrado con un sistema SIEM que correlaciona los eventos y analiza potenciales amenazas y es capaz de reaccionar bloqueando la ejecución y propagación del *malware*. Ni que decir tiene que esta modalidad consume recursos de comunicaciones y cómputo, una operación costosa 7x24, y una actualización permanente para mantener la efectividad. Este modelo de protección debe responder a tres desafíos de calado:

- Manejar un gran volumen de información en tiempo real
- Detectar ataques no catalogados previamente
- Evitar los falsos positivos

Manejar un gran volumen de información en tiempo real

Un sistema XDR efectivo precisa de una arquitectura *Big Data* para ingestar y procesar la información, pues el objeto analizado ya no son meramente las piezas de código que se cargan en la memoria del dispositivo para su ejecución, sino las peticiones de servicio llegadas desde la red, los procesos de identificación, los orígenes de los mensajes, los ficheros intercambiados, la monitorización de servidores y redes, etc. y además la secuencia temporal en que se producen los eventos. La complejidad y exigencia de estas arquitecturas lleva naturalmente a confiar su operación a empresas especializadas en Detección y Respuesta Gestionada (MDR). Gartner[4] estima que para 2025 el 50% de las organizaciones adoptarán esta solución, que esencialmente

[3] https://www.ey.com/en_se/giss/why-a-culture-change-program-is-key-to-effective-cybersecurity
[4] https://www.gartner.com/en/documents/4007295-market-guide-for-managed-detection-and-response-services

consiste en complementar la capacidad interna en ciberseguridad conectándose a Centros de Operaciones de Seguridad (SOC) que aprovisionan una plataforma capaz de orquestar la detección y respuestas (SOAR), operación 7x24, y capacidad especializada para una mejora continua de la defensa. En nuestro país el Centro Criptológico Nacional ha impulsado en diciembre de 2021 el despliegue de una Red Nacional de SOC[5] con el fin de mejorar las capacidades de protección y defensa del ciberespacio español.

Detectar ataques no catalogados previamente

Detectar ataques inéditos, sobre vulnerabilidades del sistema operativo o del software aún no catalogadas (Zero Day), conlleva un desafío tecnológico de primer orden. Naturalmente quienes elaboran *malware* conocen y prueban su eficacia para superar todos los sistemas de protección del mercado. Para prevenir y detectar estos ataques se utilizan sistemas de correlación que ya incorporan técnicas de Inteligencia Artificial, principalmente *Machine Learning* (ML). Pero si el código malicioso supera las líneas de defensa perimetral en la red, cortafuegos, filtros como WAF y políticas Zero Trust, la última línea de defensa reside en el agente de protección instalado en el sistema operativo del dispositivo. En un momento cualquiera hay en un PC más de 50 servicios corriendo, cada uno con su espectro de vulnerabilidades[6] [7] . Cuando el agente evalúa un código debe actuar, no como un portero en una caseta que comprueba una lista, sino como un guardameta o un perro guardián. Se requiere velocidad, acierto e intuición. ¿Amigo o enemigo? Hoy por hoy, una cámara de vigilancia identifica autónomamente si está viendo una cara humana o no; un robot identifica una pieza defectuosa en una cadena de montaje. ¿Es posible hacer lo mismo e identificar código dañino antes de cargarlo en la memoria? ¿Es posible hacerlo sin necesidad de consultarlo a un sistema remoto? Hay fabricantes que proclaman que sí, entrenando redes neuronales (NN) mediante tecnologías *Deep Learning* (DL) [8]. Se afirma que

[5] https://www.ccn.cni.es/index.php/es/soc/red-nacional-de-soc
[6] https://ayudaleyprotecciondatos.es/2021/10/14/zerologon/
[7] https://www.incibe.es/protege-tu-empresa/avisos-seguridad/vulnerabilidad-critica-print-spooler-microsoft-windows
[8] Real-Time Malware Process Detection and Automated Process Killing, Matilda Rhode et Al., Wiley 2021

estos nuevos agentes identifican al código dañino nuevo, sin usar firmas ni patrones, incluso al que explota vulnerabilidades desconocidas.

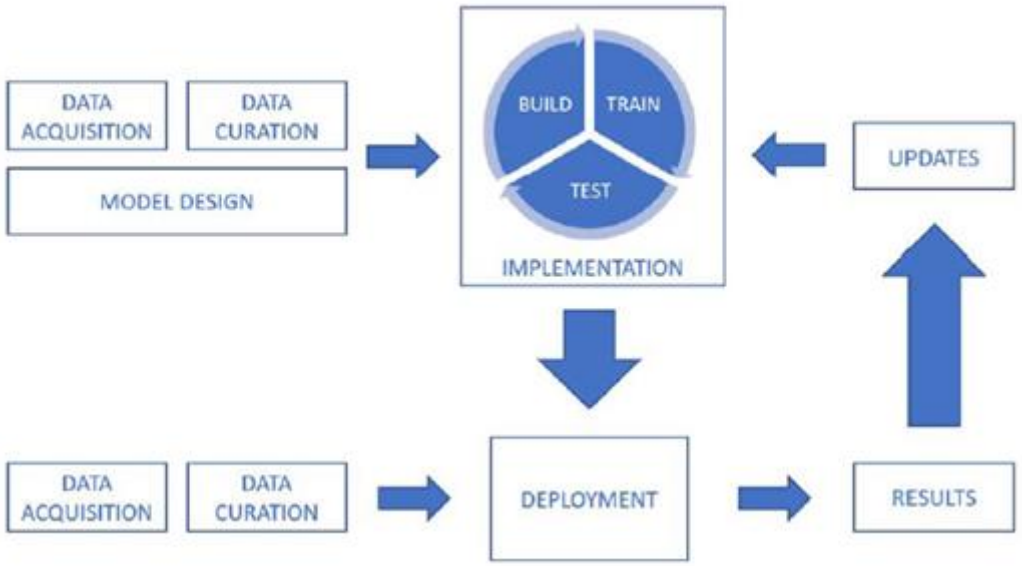

Figura 1
Proceso habitual de Machine Learning
Fuente: ETSI, Securing Artificial Intelligence (SAI); AI Threat Ontology 9

Evitar los falsos positivos

No nos importa matar mosquitas de *malware* a cañonazos de Inteligencia Artificial si es efectivo, siempre y cuanto se haga con eficacia y eficiencia. La eficacia implica acertar. Los fabricantes entrenan los modelos predictivos con ingentes cantidades de datos; la creación de redes neuronales mediante *Deep Learning* y otras técnicas, creando conjuntos de datos para el entrenamiento, para el *testing*, para el contraste y estableciendo mecanismos para la mejora continua del modelo. La tasa de eficacia tiene que superar la de los antivirus de firma y la de los sistemas EDR. Pero un Doberman que muerda a cualquier visitante no es eficiente, el auténtico desafío consiste en reducir la tasa de falsos positivos. Los falsos positivos generan costes técnicos y operativos, como por ejemplo tener que buscar un correo importante que fue marcado erróneamente como spam. Pero si los falsos positivos se reducen significativamente en los automatismos del *Endpoint*, entonces la función de protección cobra transparencia y se reduce el coste total (TCO) de la Seguridad. Si el

[9] ETSI, Securing Artificial Intelligence (SAI); AI Threat Ontology
https://www.etsi.org/deliver/etsi_gr/SAI/001_099/001/01.01.01_60/gr_SAI001v010101p.pdf

malware es atajado y no genera eventos, el SIEM no los ingesta, no consume almacenamiento ni computación. Si hay eficiencia en la detección se reduce la necesidad de *sandboxes*, de observación, alerta y asistencia humana. Esta es la piedra de toque que hace competitiva la aplicación de Inteligencia Artificial a la protección en el *Endpoin*t.

INTELIGENCIA ARTIFICIAL, *MACHINE LEARNING* Y *DEEP LEARNING*

El término 'Inteligencia Artificial' nacido en 1956 es tan inspirador que siempre ha perturbado la percepción de los dominios de la computación avanzada, la cibernética y las técnicas de análisis predictivo. Muchas de las técnicas analíticas que hoy aparecen bajo las siglas de IA se crearon hace décadas: análisis multivariante, K-means, Rain Forest, Redes Neuronales y muchas otras. Se trata de un *continuum* y debatir si un método es 'Inteligencia Artificial' o no es irrelevante.

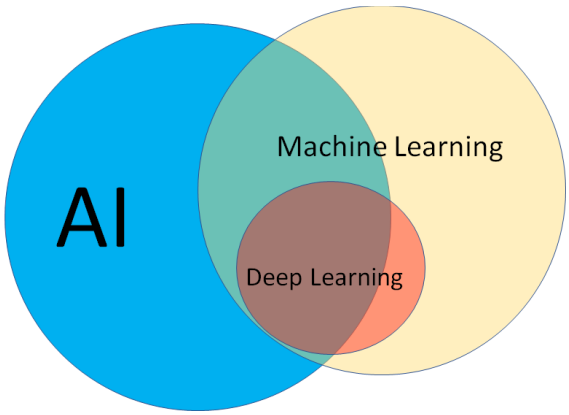

Figura 2
Relación entre AI, Machine Learning y Deep Learning
Fuente: elaboración propia

Queremos poner de relieve algunas características de esta disciplina:

- Requiere conocimientos sólidos de estadística, matemáticas y método científico.
- En general se manejan grandes volúmenes de datos con un alto coste computacional y mucho tiempo dedicado a la recolección, limpieza y clasificación por parte de los científicos de datos.

— Se requiere siempre una comprensión detallada del problema a modelar. La idea de que acumular datos lleva a una solución automáticamente es falaz.

Actualmente el dominio de la Inteligencia Artificial está dando muchos frutos en la rama de *Machine Learning* (ML). Muy sucintamente, ML elabora un modelo predictivo a partir de métodos analíticos y de clasificación que aprende a medida que se expone a más casos. Una de las formas de crear el modelo es mediante Redes Neuronales, que simulan varias capas de 'neuronas' que se conectan entre sí con pesos cuyo valor es entrenado.

Una forma avanzada de *Machine Learning* es *Deep Learning*, en el que la propia estructura del algoritmo de aprendizaje es modificada y evoluciona con la experiencia.

En general uno de los problemas de estos modelos es la inexplicabilidad. Una vez entrenados, si bien pueden resultar muy eficientes en la predicción, no se evidencia cómo lo hacen.

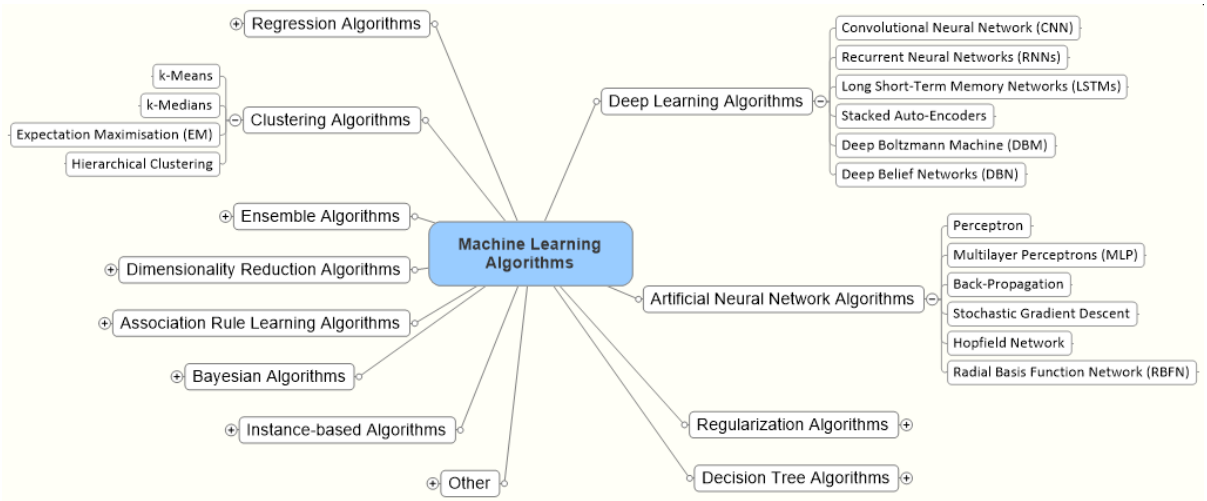

Figura 3
Clasificación de métodos de Machine Learning
Fuente: Jason Brown Lee. *"A Tour of Machine Learning Algorithms"*[10]

[10] Jason Brown Lee, "A Tour of Machine Learning Algorithms", https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Si el lector tiene curiosidad y ganas, puede acercarse por el blog de Philipp Drieger y jugar con el *Machine Learning Toolkit* de Splunk[11] . Alternativamente, la plataforma Kaggle[12] no tiene foco en la ciberseguridad, pero facilita herramientas y datasets abiertos para una aproximación a *Machine Learning*.

Una historia inspiradora acerca de la Inteligencia Artificial es el recorrido de Google, Deepmind y AlphaGo. Google adquirió la británica Deepmind en 2014, con el llamativo compromiso de mantener una comisión ética para prevenir malos usos de la tecnología IA. Deepmind desarrolló *Deep Learning* avanzado en el terreno del juego del GO, venciendo al campeón mundial en 2016. Posteriormente Alpha Zero se basó puramente en algoritmos de autoaprendizaje jugando contra sí misma, prescindiendo de librerías de jugadas y desarrollando la capacidad para aprender y dominar cualquier juego. Todo esto se narra en un premiado documental de Netflix disponible en abierto. La expresión del campeón Lee Sedol en la rueda de prensa, pidiendo disculpas por no haber vencido a la máquina, apela a nuestra más humana empatía.[13]

TENDENCIAS EN EL USO DE IA EN CIBERSEGURIDAD

En un estudio realizado en 2019 entre 850 responsables en ciberseguridad de 10 países [14] , Capgemini Research Institute identificó una fuerte tendencia a incorporar tecnologías AI en la ciberseguridad. Al menos un 69% de las empresas tenía planes de hacerlo a lo largo de 2020, en cinco casos de uso: Detección de Intrusión, Clasificación del Riesgo en la Red, Detección del Fraude, Análisis del comportamiento de usuarios y dispositivos, y detección de *malware*. Cada uno de estos escenarios tiene una dificultad y un impacto distinto en la seguridad, pero ante todo la decisión de incorporar la tecnología IA implica sus propios riesgos y oportunidades y debe analizarse con cuidado[15] :

[11] https://www.splunk.com/en_us/blog/author/pdrieger.html
[12] https://www.kaggle.com/
[13] AlphaGo - The Movie | Full award-winning documentary https://www.youtube.com/watch?v=WXuK6gekU1Y
[14] https://www.capgemini.com/wp-content/uploads/2019/07/AI-in-Cybersecurity_Report_20190711_V06.pdf
[15] https://www.gartner.com/smarterwithgartner/5-questions-to-cut-through-the-ai-security-hype

- ¿Cómo gestionamos el talento interno? ¿Qué cualificación en Inteligencia Artificial necesita el equipo humano? Es conocida la escasez de perfiles cualificados en estas disciplinas en el mercado laboral. ¿Nos lo podemos permitir y mantener?

- ¿El proveedor introduce un diferencial respecto a tecnologías ya consolidadas? ¿Se usa realmente el potencial de IA? Si se despliega un agente inteligente, ¿Es autónomo y aporta una protección eficiente sin estar conectado a un servicio remoto?

- Si el equipo humano no comprende el modelo, su adopción se entorpece mucho[16]. ¿Qué grado de visibilidad y explicabilidad se tiene sobre la solución?

- La adopción temprana de tecnología, pues es el caso, requiere un enfoque basado en el análisis del riesgo y en la ecuación coste/beneficio, para lo que se requiere una gestión madura que integre cumplimiento, gobernanza y Seguridad.

La realidad es que todos los proveedores de relevancia en el mercado de la seguridad están invirtiendo en incorporar la Inteligencia Artificial. Hagamos un recorrido no exhaustivo para ilustrar la variedad y madurez de las soluciones en el momento presente.

CHECK POINT

Nacida en 1993 y especializada en cortafuegos, la evolución de la israelí Check Point hacia la protección integral enfatiza el rol de Check Point Lab coordinando la actualización permanente de aprendizaje de sus motores de ML. Su servicio centralizado Campaing Hunting explora todos los puntos de la red y analiza anomalías; Huntress es una *sandbox* en la que cualquier software es escrutiñador previamente a su uso o despliegue. Todo esto crea una plataforma de protección desde el cloud.

[16] https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020

CROWDSTRIKE

Holding de California muy capitalizado, enfoca análisis de comportamiento del usuario y dispositivo (UEBA) para ser capaz de identificar virus, *malware*, robo de credenciales y amenaza interna. El fundamento de este tipo de protección es crear usando técnicas ML un modelo de actividad normal, '*baseline*', e identificar en tiempo real las desviaciones respecto al modelo y actuar preventivamente. La arquitectura, como en SPLUNK, implica una integración en cloud.

CYLANCE - BLACKBERRY

Blackberry, reenfocada completamente hacia el mercado de la seguridad, adquirió Cylance en 2019, con lo que completa sus arquitecturas de protección MDR con un agente de Endpoint entrenado con ML desde 2012. Estamos hablando de tecnología avanzada, exigente y costosa. El esfuerzo para crear algoritmos IA eficaces usando Machine Learning y Deep Learning no está al alcance de presupuestos pequeños ni de plazos ajustados. Las estrategias de entrenamiento ML y NN exigen una potencia computacional enorme que frecuentemente requieren el uso de hardware capaz de paralelismo masivo, típicamente granjas de tarjetas procesadoras NVIDIA con sus lenguajes y métodos de programación específicos (CUDA). Las compañías que ahora están comercializando herramientas basadas en IA para la protección del *Endpoint* han invertido durante años. Por ejemplo, Cylance va por la 7ª generación de su motor ML. Para ello ha analizado más de 8 Pbte de software y un equipo de cerca de 50 matemáticos y científicos de datos con una inversión en R&D de 148M$.

DARKTRACE

También propone una plataforma que modela una baseline (*pattern of life*). Enterprise Inmune System está más enfocada a la prevención de intrusión en la red WAN, LAN y WIFI. Sus mecanismos de Machine Learning afinan el modelo continuamente sin intervención humana, presumiblemente adaptándose a la idiosincrasia del cliente y mejorando la capacidad de defensa indefinidamente.

DEEP INSTINCT

Fundada en 2015 con el foco puesto específicamente en crear una plataforma Deep Learning para la prevención en el *endpoint*, ha cosechado numerosas menciones y

premios y un respaldo financiero de casi 180M$. Su objetivo de reducir por debajo de los 20 ms el tiempo de reacción a una amenaza en el *endpoint* y bajar del 0,1% los falsos positivos que progresan hacia el SIEM ponen de relieve el potencial alcance de la tecnología DL. DI ha puesto el foco en un impresionante esfuerzo de 5 años para entrenar su red neuronal y protegerla con patentes [17]. Al producir un agente desplegable en dispositivos de todo tipo, es capaz de integrarse en plataformas EDR y ecosistemas SIEM de forma complementaria.

FIREEYE

Fundada en 2004, FireEye es una compañía veterana en ciberseguridad que ha aportado importantes contribuciones en la identificación de los cibergrupos de atacantes. FireEye usa técnicas de ML para el análisis y categorización de información procedente de la redes y ciberataques documentados para categorizar e identificar los grupos de cibercriminales[18]. FireEye, fusionada con McAfee, es ahora TRELLIX y su plataforma XDR se llama Helix. Helix usa *sandboxes* para analizar el *malware* y aplica ML extensivamente para automatizar la respuesta sin olvidar la participación humana. La rama de FireEye especializada en respuesta a incidentes, Mandan, célebre por su investigación de los ciberataques organizados en 2014 desde el gobierno chino y por la investigación en verano de 2021 del ciberataque al oleoducto Colonial en EEUU, acaba de ser adquirida por 5.400 millones de dólares por … Google. Continuará.

FORTINET

Con una larga trayectoria desde 2000, Fortinet es conocida por su firewall hardware, Forigare. Menos conocido es su largo esfuerzo de investigación a través de FortiGuard Labs. Desde 2012 ha desarrollado una gama de productos de seguridad basados en IA. Partiendo de Machine Learning alimentado por millones nodos, incorporando Deep Learning supervisado, luego autónomo, y creando Deep Neural Networks que se han paqueteado en su producto estrella FortiAI para ofrecer un sistema dotado de

[17] https://patents.justia.com/patent/10193902
[18] https://www.fireeye.com/blog/products-and-services/2020/06/machine-learning-support-for-cyber-threat-attribution-at-fireeye.html

autoaprendizaje[19]. Una característica diferencial es que FortiAI no se conecta a servicios de internet para evolucionar y adaptarse.

IBM

La reconocida experiencia de IBM en IA con su plataforma Watson se enriqueció en noviembre de 2021 con la adquisición de la compañía holandesa ReaQta, especializada en protección del *Endpoint* mediante IA. Así IBM amplia la capacidad en torno a Qradar como SIEM, EDR, XDR y MDR. Su producto más avanzado, ReaQta-Hive, instala el agente NanoOS que se oculta en una máquina virtual en el *endpoint* para ser liviano a indetectable al *malware.* Este agente interactúa con el EDR (ahora el ecosistema Qradar) para brindar mediante técnicas ML una respuesta rápida y predictiva.

MICROSOFT

La relevancia de Microsoft en el panorama de la ciberseguridad es inmensa. Sus productos están en prácticamente toda la superficie de exposición. Las mayores cibercrisis a nivel planetario se han originado en torno a vulnerabilidades de los productos de este gigante tecnológico (Exchange, Wannacry, etc.). Por ello su inversión y despliegue de soluciones para la ciberseguridad se ha disparado en los últimos años. Microsoft ya ofrece una arquitectura de protección XDR/SIEM con los servicios Sentinel, 365 Defender y Defender for Cloud[20]. Desde el punto de vista de la aplicación de tecnologías IA la seguridad, Microsoft ofrece dos facetas bien diferenciadas. Por un lado, la protección del *Endpoint*, en esencia el sistema operativo Windows, apoyándose en Defender ATP como agente; Por otro, el propio stack de servicios de *Machine Learning* en Azure, que es en sí mismo objetivo de ataques Adversarial AI. Defender ATP utiliza *Deep Learning* para evolucionar su capacidad de identificar amenazas en el árbol de procesos y en el flujo temporal de la ejecución[21]. En octubre de 2021 Microsoft ha publicado un excelente informe, “Microsoft Digital Defense Report”[22], donde pormenoriza el panorama de ataques observados, identifica

[19] https://www.fortinet.com/products/fortiai
[20] https://www.microsoft.com/en-us/security/business/threat-protection
[21] https://www.microsoft.com/security/blog/2020/07/23/seeing-the-big-picture-deep-learning-based-fusion-of-behavior-signals-for-threat-detection/
[22] https://www.microsoft.com/en-us/security/business/microsoft-digital-defense-report?rtc=1

los ataques procedentes de naciones estado y las tendencias cibercriminales. Sus cifras son astronómicas: en un año ha bloqueado 32.000.000.000 ataques al correo electrónico…

PALO ALTO

La californiana Palo Alto Networks es una de las compañías de ciberseguridad más capitalizadas del mundo y tiene una fuerte presencia en nuestro país. En 2017 adquirió LightCyber, startup especializada en análisis del comportamiento, y en 2019 anunció CORTEX como plataforma XDR basada en Machine Learning supervisado y no supervisado. CORTEX absorbe el flujo de información obtenido a través de SIEM, y traza una baseline de comportamiento contra la que contrasta todos los eventos en el perímetro ofreciendo lo que llama 'seguridad continua'. Palo Alto tiene una actividad destacada en cibervigilancia a través de Unit 42, rama especializada en el análisis de amenazas. En 2014 fundó la Cyber Threat Alliance (CTA) [23] , liderando la colaboración de los principales proveedores de ciberseguridad

SOPHOS

Otra histórica compañía fundada en UK en 1985 enfocada a la producción de antivirus, fue finalmente adquirida en marzo de 2020 por el fondo estadounidense Thoma Bravo. Su acercamiento a la IA se intensificó en 2016 con la adquisición de la irlandesa Barricada y en 2017 la norteamericana Invincea. Con esto acumuló una robusta experiencia en tecnologías de *sandbox virtual, Machine Learning,* y *Deep Learning* para la protección del *Endpoint*, que ofrece a través de sus productos Intercepta X y XG Firewall. La red neuronal distribuida en Intercept X reacciona a las anomalías sin necesidad de consultar una base central de firmas y protege el *Endpoint* generando respuestas de alerta y aislamiento. El mismo principio se aplica en el firewall hardware XG.

SPLUNK

Solución SIEM bien posicionada y líder en analítica y correlación de eventos, adquirió en 2015 la empresa Caspida, especializada en ML, y desde 2016 mantiene un kit de

[23] https://www.cyberthreatalliance.org/

ML y DL que permite a sus clientes y partners desarrollar modelos integrados. La librería Splunk MLKit[24] ofrece un modelo cloud contenerizado, soporte a lenguajes y librerías populares como Spark, Tensorflow y PyTorch, y capacidad para entrenar los modelos haciendo uso de computación masivamente paralela usando GPU y CUDA.

SYMANTEC

Symantec se fundó en 1982 enfocada en proyectos de Inteligencia Artificial. Eran otros tiempos en que era difícil llevar los productos resultantes al mercado. Adquirida en noviembre de 2019 por Broadcom, su nombre desaparece en la historia. Pero no el resultado de su investigación, en forma de numerosas patentes y productos de protección que ahora se integran en la plataforma de Broadcom. Una de sus aportaciones originales es la aplicación de Machine Learning para la clasificación y detección de ataques dirigidos (Targeted Attack Analytics), una aproximación original y específica[25] que en 2017 tuvo su momento de gloria en la detección de ataques a infraestructuras energéticas en Europa y US[26]. Puede verse la contribución de Symantec en la clasificación de grupos de cibercrimen ofrecida por Mitre. Si se quiere tener una perspectiva actualizada de los grupos de ataque y su arsenal, la sección de Groups en Mitre no defrauda[27].

VECTRA AI

Vectra asegura que la prevención es insuficiente, y ha obtenido una potente financiación para crear un ecosistema de análisis y vigilancia integral desde el cloud aplicando metodologías de Machine Learning para acelerar los procesos de detección y respuesta. Su objetivo es evitar que las amenazas y ataques progresen hasta producir una brecha, investigando tanto el *endpoint* como la red y sobre todo los servicios cloud. Vectra intenta reducir a cuestión de horas el tiempo promedio de remediación. Para ilustrar sus dominios de protección y detección usando la base de conocimientos MITRE patrocinada por la NSA.

[24] https://github.com/splunk/deep-learning-toolkit
[25] https://symantec-enterprise-blogs.security.com/blogs/feature-stories/targeted-attacks-game-has-changed
[26] https://fortune.com/2017/09/06/hack-energy-grid-symantec/
[27] https://attack.mitre.org/groups/G0074/

VIRUSTOTAL

La legendaria compañía malagueña fundada en 2002 y adquirida por Google en 2012 anunció en 2020 su acuerdo con la israelí Cynet, incorporando a su robusta plataforma de análisis de *malware* la capacidad de Machine Learning que Cynet viene optimizando desde 2015. La inmensa acumulación de datos y análisis de *malware* de VirusTotal enriquecen los motores ML Cynet para identificar amenazas y reducir los falsos positivos. El hecho de que VirusTotal ofrece su servicio desde la infraestructura de Google mitiga el impacto de los continuos ataques DDOS y de intrusión que recibe.

CAPITALIZACIÓN Y PATENTES

Compañía	Origen	Fundada	Productos	Capitalización	US Patents
Checkpoint	Israel	1993	Harmony Endpoint	$ 18.170.000.000	208
CrowdStrike	California	2011	CrowdStrike Falcon	$ 52.110.000.000	24
Cylance (comprada en 2018 por Blackberry)	California	2012	Blackberry Protect	$ 1.800.000.000	15
Darktrace	UK	2013	Enterprise Immune System	$ 3.413.000.000	1
DeepInstinct	New York	2015	Deep Instinct Prevention Platform	$ 260.000.000	4
FireEye (fusionada con McAfee en 2022 creando TRELLIX)	California	2004	Helix	$ 4.130.000.000	130
Fortinet	California	2000	FortiAI	$ 52.710.000.000	186
Palo Alto Networks	California	2005	Cortex XDR	$ 61.300.000.000	137
ReaQTA (comprada en 2021 por IBM)	Holanda	2014	Reaqta-Hive	$ 3.000.000	
Sophos (comprada en 2019 por Thoma Bravo)	UK	1985	Intercept X	$ 3.400.000.000	78
Splunk	California	2003	Splunk Insights , Splunk Phantom	$ 21.590.000.000	57
Symantec (Comprada en 2019 por Broadcom)	California	1982	Symantec Endpoint Security	$ 14.000.000.000	322
Vectra AI	California	2012	Plataforma Cognito	$ 1.200.000.000	4

Figura 4
Capitalización y número de patentes US relacionadas con ciberseguridad y ML
Fuente*: Elaboración propia*

Sobre esta muestra caben varias observaciones. La tendencia a la concentración mediante fusiones y adquisiciones es acorde a lo que ocurre en otros sectores industriales y notablemente en torno a las tecnologías de la información. El capital inversor, o el destino final de las compañías, es abrumadoramente mayoritario a favor de EEUU, aunque algunas hayan iniciado su andadura en Europa (por ejemplo, Sophos en UK y Reaqta en Holanda). Lamentablemente esto es coherente con la falta de visión y fuelle en la UE, cuyo Plan AI expuesto en 2021[28] ni siquiera menciona la aplicación de la IA a la ciberseguridad. El elocuente ensayo de Luis Moreno y Andrés Pedreño [29] profundiza en los desafíos que afronta Europa para no caer en la irrelevancia ante la emergencia de la tecnología IA.

[28] https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review
[29] https://www.prevenireldeclive.com/

Por otro lado, los productos de ciber protección basados en tecnologías de inteligencia artificial han tenido un desarrollo de muchos años, y grandes compañías (como Google e IBM) han optado por adquirir firmas especializadas para acelerar su posicionamiento comercial. No son raros recorridos de seis u ocho años para presentar productos maduros al mercado (véase FortiAI). Estamos ante inversiones altas y largos esfuerzos, que además se protegen con un registro minucioso de patentes y respaldo financiero robusto. No es un mercado para pequeños emprendedores.

Además, estamos ante tendencias aún no adoptadas masivamente, caras, con curva de adopción pronunciada e impacto global por medir. Los esfuerzos realizados responden desgraciadamente a un escenario asimétrico: un fabricante que nos ofrezca un sistema de protección debe hacerlo para TODAS las amenazas: todos los virus, todo el *malware*, todos los *exploits*, todas las veces. Hay ahora unas 400.000 variantes activas en la red. Mientras que un atacante se especializa y sólo necesita penetrar una vez.

Todas estas firmas colaboran con DARPA, CISA, NSA y prácticamente con todas las agencias gubernamentales occidentales, en respuesta a la necesidad, impuesta por la dura realidad, de una colaboración público-privado en las tareas de ciber protección. Además, la ingente cantidad de datos con que se alimentan los proyectos analíticos con que se entrenan sus algoritmos proceden del espacio común, y también aumenta disponibilidad de datasets y plataformas de acceso público para estandarizar los métodos de defensa y ataque.

Una de las referencias de mayor éxito en los últimos años la aportan los servicios MITRE ATT&CK[30] y MITRE DEF3ND[31] mantenidos por MITRE desde 2015, que taxonomizan los dominios de la ciberseguridad y que las compañías están utilizando para identificar sus capacidades. Otra aportación de gran utilidad es la librería YARA[32]

[30] https://attack.mitre.org/
[31] https://d3fend.mitre.org/
[32] https://support.virustotal.com/hc/en-us/articles/115002178945-YARA

creada por VIRUSTOTAL para la clasificación y definición de *malware* mediante reglas y adoptada como estándar de facto.

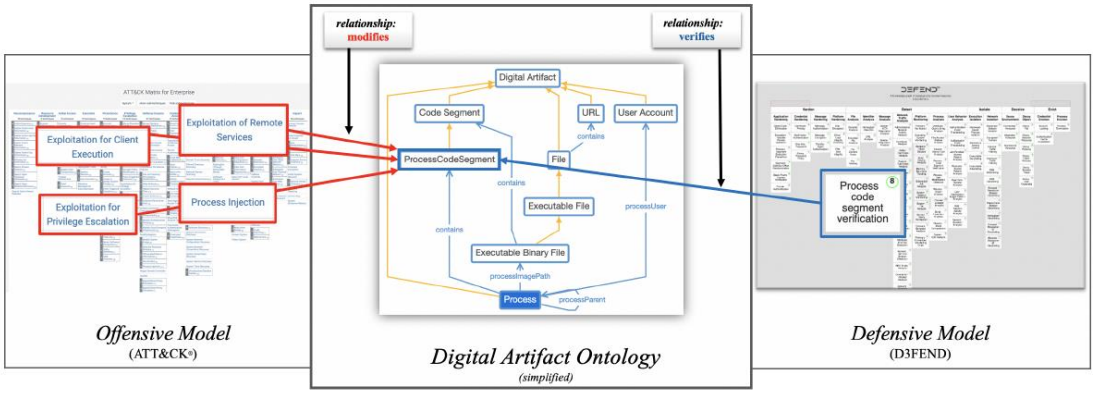

Figura 5
Estructura de modelos de MITRE D3fend
Fuente: MITRE D3fend 2022 [33]

SECTOR PÚBLICO

Como hemos mencionado, el progreso de la colaboración público-privado se manifiesta en nuestro país con iniciativas como la recientemente anunciada Red Nacional de SOC por parte del Centro Criptológico Nacional. Utilizando las herramientas de interoperabilidad LUCIA y REYES creadas por el CCN-CERT, se constituirá una plataforma nacional para la notificación y seguimiento de ciber incidentes. Esto sienta las bases para una capacidad de observación y análisis (además del propósito inmediato de aumentar la capacidad de respuesta), que alimentará los proyectos relacionados con la analítica avanzada que indudablemente aparecerán. Además, el CCN-CERT cuenta entre sus soluciones para el sector público con la herramienta CARMEN, para la detección y protección de amenazas persistentes, y con el software de protección del *endpoint* CLAUDIA/microCLAUDIA desarrollado por la empresa S2 Group. Estos componentes son susceptibles de evolucionar e integrarse con las propuestas de los fabricantes que mencionábamos en el apartado anterior, siempre que el CCN lo encuentre apropiado. Para ello las soluciones a integrar deben pasar un exigente proceso de certificación y cualificación en el Catálogo de Productos STIC del CCN. La buena noticia es que varias de las

[33] https://d3fend.mitre.org/resources/D3FEND.pdf

soluciones mencionadas ya figuran en dicho catálogo. La reciente colaboración entre CRUE y CCN emitiendo una guía del Esquema Nacional de Seguridad para Universidades[34] puede dar paso a colaboraciones aún más ambiciosas en el ámbito de la investigación y de la formación.

OTRAS LATITUDES

Veamos unas pinceladas que ilustran el estado de la aplicación de IA al dominio de la ciberseguridad en otras regiones.

Japón puja por mantener su liderazgo en robótica y ser una potencia en Inteligencia Artificial. Tras los ciberataques a Mitsubishi en 2020[35], el Ministerio de Defensa de Japón impulsó un proyecto de 237M$ para desarrollar sistemas de IA para detectar, analizar y responder a ciberataques.

Australia, al contrario, tiene planes estratégicos diferentes para la ciberseguridad y para el desarrollo en Inteligencia Artificial, aunque ambos se refieren mutuamente.

La extraordinaria apuesta de **Israel** por la innovación, que ha hecho del país una reconocida 'Startup Nation' [36], ha dado pie a la creación de un auténtico 'Silicon Valley' enfocado a la ciberseguridad en la ciudad de Beerseba, con más de 400 empresas especializadas. De allí proceden Check Point y CyberArk (CyberArk tiene un producto AI enfocado a proteger la identidad digital).

¿Y qué hace **China**? Tomar la delantera con contundencia. Sus planes trienales para IA desde 2017 [37] ya identifican la necesidad de aplicar IA a la ciberseguridad. Consecuentemente lanzó el programa "World-Class Cybersecurity Schools (WCCS)" por el que al menos 11 de las mejores universidades chinas ofrecen formación conjunta de IA+Ciberseguridad, preparando un contingente de expertos que harán honor a la visión de su presidente Xi: "la competencia en el ciberespacio es, a fin de

[34] https://www.ccn-cert.cni.es/seguridad-al-dia/novedades-ccn-cert/11567-nueva-guia-ccn-stic-881-de-adecuacion-al-ens-para-universidades.html
[35] https://www.asahi.com/ajw/articles/13948123
[36] https://www2.deloitte.com/il/en/pages/innovation/article/the_israeli_technological_eco-system.html
[37] https://opengovasia.com/chinese-government-sets-out-specific-targets-in-three-year-action-plan-on-artificial-intelligence/

cuentas, competencia por el talento". Lo mismo concluye el extenso informe de la National Security Commission on Artificial Intelligence de EEUU[38] al establecer las áreas de acción urgente: Liderazgo, Talento, Hardware, Inversión.

Así pues, la siguiente frontera definida por las dos grandes potencias se encuentra en la capacitación simultánea en Ciberseguridad e Inteligencia Artificial. EEUU constata que China ya le lleva la delantera también en esto, y lanza a su vez un refuerzo a su programa CAE-Cyber para definir certificaciones y currículos académicos equiparables[39].

En España se ha presentado una Estrategia Nacional de Inteligencia Artificial que identifica como eje estratégico la formación en IA, aunque no se dota económicamente. Sin ir mucho más allá de los Pirineos, en la estrategia de Inteligencia Artificial de Francia (2020[40] se dotan 700M€ para la formación en IA (2.000 graduados, 1.500 master, 200 tesis al año, hasta 2025). Ninguno de los dos países pone el foco en la IA aplicada a la ciberseguridad. Sin embargo, dado el éxito y respaldo que tiene actualmente la oferta académica en nuestro país para los estudios de BigData/IA y el buen prospecto profesional que ofrece, es previsible que tanto desde el sector público como desde el privado se lance oferta formativa que integre disciplinas de IA y Ciberseguridad. En esto Europa tiene una ventaja en el sistema de credenciales académicas armonizado desde 2010 en toda la UE por el proceso de Bolonia y sus mecanismos de acreditación.

DESAFIOS DE SEGURIDAD INTRODUCIDOS POR LA INTELIGENCIA ARTIFICIAL

El potencial de doble uso de las tecnologías de Inteligencia Artificial es desconocido y debemos presuponer que al menos igual de peligroso que el de otros avances tecnológicos. Los enormes esfuerzos presupuestarios y organizativos de largo recorrido que conlleva crear los sistemas descritos en los apartados anteriores no están al alcance de los grupos APT, ni siquiera de las divisiones de ciberguerra

[38] https://bookstore.gpo.gov/products/national-security-commission-artificial-intelligence-final-report
[39] China's CyberAI Talent Pipeline - CSET Policy Brief https://cset.georgetown.edu/publication/chinas-cyberai-talent-pipeline/
[40] https://www.economie.gouv.fr/la-strategie-nationale-pour-lintelligence-artificielle#

patrocinadas por estados. La experiencia de la última década ha identificado una especie de 'Ley de Moore' del esfuerzo para entrenar modelos con *Deep Learning*, por la que cada cuatro meses se dobla el número de *petaflops* a analizar diariamente cuando se quiere entrenar un modelo relevante[41].

Figura 6
Incremento exponencial del coste computacional de crear modelos avanzados
Fuente: *OPENAI, "AI and Compute",* 2018

El esfuerzo del cibercrimen es de corto recorrido y alta rentabilidad. Podemos usar como símil la industria nuclear: no es concebible la producción de armas nucleares por parte del criminal, aunque sí el robo, sabotaje o subversión de armas o materiales radiactivos. Por lo tanto, cabe más esperar desarrollos menores, robo de herramientas peligrosas, sabotaje a la producción de modelos IA, y aprovechamiento de las debilidades específicas de las tecnologías IA. Por el mismo razonamiento aparece más peligrosa la implicación de los estados en el uso ofensivo de la IA. En los últimos años el ascenso del *ransomware* ha capitalizado nuestras preocupaciones, pero la realidad es más ominosa. El *malware* NotPetya intentaba en 2017 borrar literalmente la Administración ucraniana; en enero de 2022 Toyota sufrió un ataque que la obligó a paralizar 14 plantas. La AMENAZA, compuesta por capacidad y voluntad, ha aumentado y con ella el RIESGO, al margen de las tecnologías usadas.

[41] https://openai.com/blog/ai-and-compute/

Así como MITRE mantiene el mapa general de la ciberseguridad, tres instituciones europeas se han ocupado de identificar y sistematizar el panorama de las amenazas de seguridad relacionadas con la Inteligencia Artificial. EUROPOL[42] publicó en 2020 el informe "Malicious Uses and Abuses of Artificial Intelligence"; ENISA[43] difundió en el mismo año "AI CYBERSECURITY CHALLENGES" en el que hace una taxonomía detallada de las amenazas de relacionadas con la Inteligencia Artificial; por último, el grupo de trabajo "ETSI Industry Specification Group on Securing Artificial Intelligence (ETSI ISG SAI)" de ETSI ha iniciado el camino para estandarizar tanto la clasificación de las amenazas (acaban de publicar una Ontología [44]) como estudios sistemáticos en los dominios identificados por ENISA. ETSI vendrá respaldada en esta labor gracias al impulso alemán para la estandarización de IA a través de las organizaciones DIN y DKE[45]. Consciente del riesgo emergente al introducir la tecnología, la Comisión Europea ha publicado una lista de autoevaluación, elaborada por un comité de expertos independientes, en aspectos de confianza para la implantación de soluciones IA, "ALTAI"[46] . Es llamativa la similitud y coincidencia en el tiempo con la recomendación de la NSCAI de EEUU de encomendar a NIST la estandarización de métricas para garantizar la confianza en los sistemas IA.

Complementario a este enfoque tenemos los análisis de la OTAN, que abordan el impacto de la IA en Seguridad desde el punto de vista militar y de la subversión política.

Sintetizando estas aportaciones, el uso observado de IA como ciberamenaza se puede clasificar así:

- Optimización del *malware*
- Ataque a la cadena de suministro (al propio ciclo de uso de la tecnología IA)
- Subversión Política

[42] https://www.europol.europa.eu/publications-events/publications/malicious-uses-and-abuses-of-artificial-intelligence
[43] https://www.enisa.europa.eu/news/publications/artificial-intelligence-cybersecurity-challenges
[44] https://www.etsi.org/technologies/securing-artificial-intelligence#mytoc3
[45] https://knowledge4policy.ec.europa.eu/ai-watch/germany-ai-strategy-report_en
[46] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Optimización del *malware*

La Agencia Sueca de Defensa emitió un informe en marzo de 2020 clasificando la anatomía de fases de un ciberataque a partir de 96 ciberataques AI documentados, y creó una escala de madurez para las tecnologías utilizadas. En general, AI está bastante madura para las etapas de reconocimiento y en estado de prototipo para el resto.

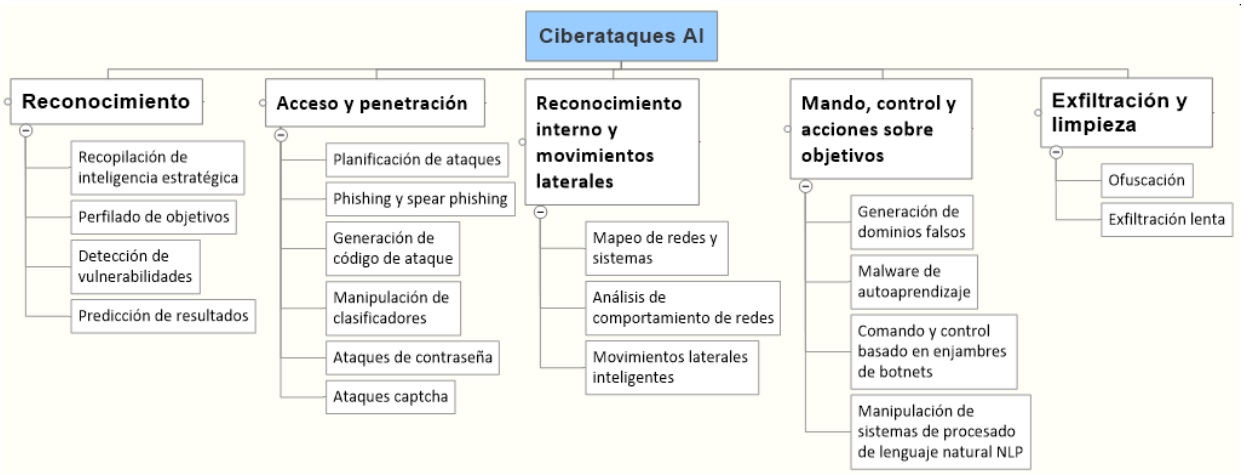

Figura 7
Desglose de etapas de un ciberataque
Fuente[47] : Swedish Defence Research Agency

Hagamos un recorrido por las distintas formas de ataque documentadas que utilizan Inteligencia Artificial:

1. Las técnicas de *Machine Learning* se utilizan para superar las barreras *antispam* y *anti phishing*, mediante el análisis de grandes volúmenes de correos históricos para generar mensajes gramática y sintácticamente correctos e indiferenciables de los humanos.
2. Mediante ML se han generado también algoritmos utilizados para romper contraseñas (Generative Adversarial Networks - GAN), que mejoran mucho los resultados de herramientas previamente existentes.
3. Igualmente, se entrenan algoritmos para responder a los CAPCHA que protegen el acceso abusivo a servicios online.

[47] Erik Zouave, et al. (Mar. 4, 2020). Swedish Defence Research Agency (FOI). “Artificially Intelligent Cyberattacks.” https://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf.

4. También se usa ML para estudiar el comportamiento de los sistemas antivirus y extraer su patrón de detección inherente.
5. Hay una debilidad implícita en los algoritmos IA que se puede explotar desde el *malware*: las redes neuronales adolecen de inexplicabilidad, es decir que la distribución de pesos en sus capas genera resultados, pero no puede trazarse como una secuencia lógica ni deducirse su propósito. Diversos analistas[48] han demostrado que se pueden ocultar en redes neuronales de forma sencilla patrones de ataque, condiciones de activación, y otras características que los sistemas de detección de *malware* podrían detectar. Una variación de esta técnica permite al *malware* cifrar u ocultar las comunicaciones entre elementos atacantes o incorporar esteganografía para camuflar los mensajes.
6. El análisis de los algoritmos IA con fines maliciosos (Adversarial AI) permite encontrar este tipo de debilidades. Por ejemplo, unos investigadores de Carnige Melon nos muestran cómo eludir el sistema de identificación facial en cámaras de videovigilancia, suministrando patrones de rostros concebidos para engañar al algoritmo[49].

Figura 8
Se engaña a la IA al superponer unas gafas con un patrón gráfico que responde a las características que analiza la red neuronal de la cámara, codificando un rostro erróneo
Fuente: Mahmood Sharif &Al

7. El propio *malware* no tiene por qué incluir algoritmos IA, ya que puede hacer uso de librerías online de AWS, Azure, IBM, OpenIA, etc.

[48] Dhilung Kirat, Jiyong Jang, and Marc Ph. Stoecklin. (Aug. 9, 2018). Black Hat USA. “DeepLocker — Concealing Targeted Attacks with AI Locksmithing”
[49] Real and stealthy attacks on state-of-the-art face recognition. Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter

8. Una forma llamativa de ataque, que no conlleva necesariamente el uso de tecnología IA, consiste en tomar el control de altavoces conectados para emitir comandos de voz destinados a los propios asistentes como Alexa o Siri que sí se apoyan en servicios de IA. Además, en algunas situaciones se pueden utilizar comandos inaudibles.
9. En la Darknet y en los foros de hacker aparecen herramientas, originalmente enfocadas a demostraciones, pruebas de concepto y análisis para la defensa, reenfocadas al servicio de los análisis orientados a la intrusión: DeepHack, DeepExploit, Metasploit...
10. Una de las amenazas con mayor potencial es la creciente aplicación de Inteligencia Artificial a los ataques de ingeniería social. Escribir correos maliciosos verosímiles y adivinar contraseñas parecen logros menores cuando se constata las capacidades de que hoy se dispone. Por citar algunos ejemplos de los instrumentos que ya son una realidad:
    - Hay herramientas (EgleEye) capaces de localizar todas las cuentas de un individuo en redes sociales, y de aplicar reconocimiento facial para localizar su fotografía en posts y canales
    - Es viable analizar y clonar la voz de un sujeto
    - La librería GPT-3, del laboratorio OpenAI, permite generar texto indistintinguble del escrito por humanos
    - Hay numerosos servicios lúdicos y librerías experimentales Deep Fake (Zao, FakeApp, FaceApp etc), que permiten imitar o reemplazar de forma realista rostros humanos. En general los humanos somos bastante fáciles de engañar con biometrías falsas (imagen, audio, texto). Estas herramientas ya se han usado en ataques reputacionales
    - EUROPOL alerta del uso de ML para generar fotografías de pasaporte que mezclan fotos de dos individuos de forma que se engaña tanto a la IA como al humano.

Para un recorrido detallado sobre el alcance actual de los ciberataques basados en IA recomendamos el informe de EUROPOL.

ATAQUE A LA CADENA DE SUMINISTRO

Se puede decir que el ataque a Solarwinds[50] de 2020 marcó un antes y un después. El éxito en colocar *malware* en la propia distribución de software de un proveedor de confianza en servicios de ciber protección para gobiernos y agencias de todo el mundo atrajo inmediatamente la atención sobre el flanco, bastante descubierto, de la producción de productos tecnológicos. El *malware* ya no solo procede de las redes externas a la organización, sino que nos lo suministran nuestros proveedores escondido en sus productos porque meses antes han sido hackeados.

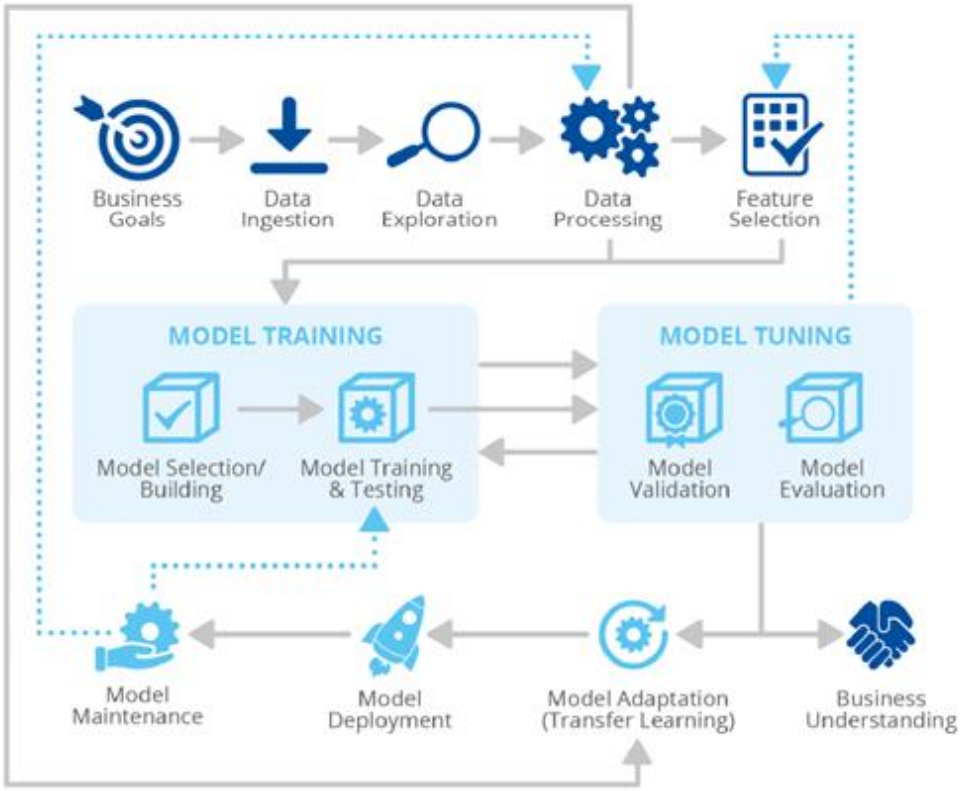

Figura 9
Modelo general de la cadena de suministro de sistemas de Inteligencia Artificial
Fuente : ENISA, "Artificial Intelligence Cybersecurity Challenges"

[50] https://www.zdnet.com/article/microsoft-fireeye-confirm-solarwinds-supply-chain-attack/

ENISA y ETSI han mapeado las amenazas y los ciclos de producción de la tecnología IA para facilitar la labor de proteger la cadena de suministro de lo que se ha llamado ADVERSARIAL AI. Todas las etapas de la cadena tienen sus vulnerabilidades:

- Ataque a la integridad o autenticidad de las fuentes de datos que alimentan los motores analíticos
- Robo o exfiltración de los datos de base
- Envenenamiento o corrupción de los datos de base que alimentan el análisis
- Robo del propio modelo y utilización para entregar Adversarial AI
- Manipulación con *malware* del modelo en su etapa de despliegue
- Diseño de medidas de elusión a partir del conocimiento del modelo

El informe de ENISA es extenso y desarrolla sistemáticamente toda la casuística.

Para ilustrar cuan inesperado o imaginativo puede ser un ataque que utiliza el conocimiento de un servicio basado en IA, Simon Wreck realizó esta performance en las calles de Berlín. Arrastrando un carrito con 99 móviles encendidos y conectados a Google Maps, provocó un falso atasco en el servicio de Google.

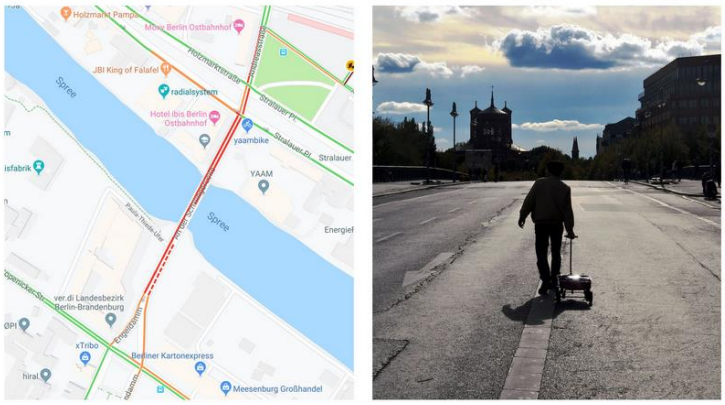

Figura 10
Engañando a Google Maps
Fuente: Simon Wreck, “Google Maps hacks” [51]

[51] http://www.simonweckert.com/googlemapshacks.html

SUBVERSIÓN POLÍTICA

Hasta aquí nos hemos centrado en la protección o ataque de sistemas tecnológicos, pero el eslabón más frágil en la cadena de la seguridad es el propio ser humano y el riesgo que implica su manipulación eficaz es inconmensurable.

El analista de la OTAN Edward Hunter [52] lo expresa así: "*Las principales vulnerabilidades en el espacio de información política se derivan de rasgos psicológicos humanos naturales que pueden exacerbarse y explotarse en entornos en línea*".

La rica huella digital que dejan los individuos en su actividad en las redes sociales ha permitido consolidar modelos del comportamiento y capacidades predictivas. Se ha evidenciado la estrecha relación entre las preferencias que se manifiestan en los entornos digitales y las características personales íntimas. Facebook ha llegado a medir que con 150 *likes* un individuo se hace tan predecible como para un miembro de su familia. Se sistematizan patrones como OCEAN (curiosidad, rigor, extraversión, amabilidad, neuroticismo) para describir de forma tabulada la personalidad del individuo. Estos modelos y la enorme capacidad de agregación de datos para un análisis predictivo en las plataformas con miles de millones de usuarios han facilitado la elaboración de mensajes publicitarios personalizados, pero de forma más inquietante, también la elaboración de mensajes políticos individualizados. En 2018 se acusó a Cambridge Analytica de usar ilícitamente datos procedentes de Facebook para influir en las elecciones presidenciales de EEUU en 2016 y 2018, y de México y de Argentina. La periodista Carole Cadwalladr denuncia sin reservas el papel jugado por Cambridge Analytica y Facebook[53]: "Si no fuera por Facebook, no hubiéramos tenido Brexit". Con 1.900 millones de usuarios y un comportamiento irrespetuoso con su privacidad, Facebook todavía no nos ha dado el último disgusto.

[52] Artificial Intelligence in Tomorrow's Political Information Space. Edward Hunter CHRISTIE. Defence Economist. NATO HQ (i)

[53] TED Talk: "El papel de Facebook en el Brexit y la amenaza a la democracia", Carole Cadwalladr, https://www.ted.com/talks/carole_cadwalladr_facebook_s_role_in_brexit_and_the_threat_to_democracy

Partiendo de estas capacidades, se dan diversas formas de subversión documentadas

- Mensajes políticos personalizados basados en analítica predictiva y psicometría, apoyándose en facilidades de la propia plataforma y falsificando identidades remitentes.
- Dinámicas de grupo tendentes a crear cajas de resonancia para la amplificación y polarización de la opinión política. Se usan técnicas como exposición selectiva, sesgo confirmativo, y creación de grupos o facciones ficticios sin seguidores reales o ampliando la visibilidad de grupos extremistas muy minoritarios.
- Polarización por sesgos algorítmicos de plataforma. Dado que un porcentaje creciente de ciudadanos se informa a través de plataformas online como Youtube, Facebook, Twitter, Instagram etc., en lugar de medios dotados de control editorial y sometidos a escrutinio y transparencia, los sesgos confirmativos y la selección de la información causada por los algoritmos de preferencia y fidelización acaban llevando al individuo a opiniones extremas y cerradas.
- Difusión de falsas noticias (*Fake News*). Las tácticas de difusión de rumores y panfletos han existido siempre, pero las redes sociales tienen ahora capacidad para la difusión y entrega personalizada y eficiente a individuos seleccionados, de forma masiva. Se facilita la desinformación y la manipulación.

Figura 11
Ejemplos de mensajes personalizados de contenido político
Fuente: Edward Hunter, “Artificial Intelligence in Tomorrow’s Political Information Space”

Para dar un atisbo del alcance que puede tener este recorrido, recordemos la impactante ponencia que impartió el historiador Juval Harari en el foro de Davos en enero de 2020: "*El conocimiento de la biología, multiplicado por potencia computacional, multiplicado por datos, generan la habilidad para hackear el cerebro humano […] Un sistema que nos comprende mejor de lo que nos comprendemos a nosotros mismos puede predecir nuestros sentimientos y decisiones, puede manipular nuestros sentimientos y decisiones y, en última instancia, puede tomar decisiones por nosotros.*"[54]

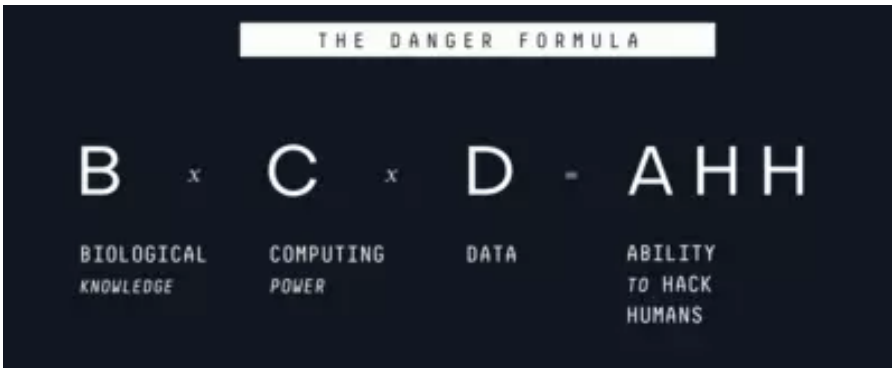

Figura 12
La fórmula del riesgo de manipulación mental
Fuente : World Economic Forum, "Yuval Harari's blistering warning to Davos", 2020

CONCLUSIONES ¿QUÉ SE PUEDE HACER DESDE LA UNIVERSIDAD?

Tanto la tendencia geopolítica como la industria y la dinámica de la ciberseguridad están introduciendo la Inteligencia Artificial con celeridad en el escenario.

La complejidad de esta tecnología, la carencia de especialistas, su rápida evolución, el coste de las soluciones, hacen inviable una aproximación interna o en solitario. Hay que agradecer el oportuno impulso del Centro Criptológico Nacional a la creación de la Red Nacional de SOC orientada a la gestión de ciber incidentes, que viene a completar un ecosistema de una veintena de aplicaciones para la protección, análisis, auditoría y formación. Las universidades de nuestro país que han sufrido ciberataques

[54] https://www.weforum.org/agenda/2020/01/yuval-hararis-warning-davos-speech-future-predications/

recientemente (UCLM, UOC, UCO, UAB, UCA)[55] no han contado con el apoyo suficiente de los organismos encargados de dar respuesta a nivel nacional, y esta iniciativa debería corregirlo.

Mientras tanto diversas universidades del país ya están adoptando la fórmula MDR estableciendo una relación con un proveedor de confianza especializado en Seguridad, que complementa con sus recursos, su SOC, sus herramientas, y su ecosistema de proveedores, al equipo interno especializado en los procesos universitarios (Sophos, Telefónica, Secure&IT, PaloAlto y otros). Son estos actores los que introducirán los productos y servicios de próxima generación en la protección.

Pero la auténtica aportación de la Universidad se hará desde su propia naturaleza: se necesitan titulaciones que aúnen las disciplinas IA y ciberseguridad; se necesita investigación en múltiples dominios y no sólo informáticos (por ejemplo, la analítica del comportamiento y la ética aplicada). Se necesita emprendedurismo en torno a la Ciberseguridad y en torno a la Inteligencia Artificial. Y desde su esencia educativa, tienen la responsabilidad de trasladar a la sociedad a través de su alumnado habilidades para evitar ser víctimas de manipulación en los medios digitales.

REFERENCIAS

- AGUIAR, Alberto R. *Ciberataque a la Universidad de Castilla la Mancha contado desde dentro*. Business Insider. 21 de Mayo de 2021. Disponible en https://www.businessinsider.es/ciberataque-universidad-castilla-mancha-contado-dentro-869181

- *AlphaGo*. INT. DEEPMIND. *Netflix. Youtube*, s.f. Disponible en https://youtu.be/WXuK6gekU1Y

- CAPGEMINI. *Reinventando la ciberseguridad con inteligencia artificial* . 2021. Disponible en https://www.capgemini.com/wp-content/uploads/2019/07/AI-in-Cybersecurity_Report_20190711_V06.pdf

[55] https://www.businessinsider.es/ciberataque-universidad-castilla-mancha-contado-dentro-869181

- CCN-CERT. *Ciberamenazas y Tendencias 2021*. 2021. Disponible en https://www.ccn-cert.cni.es/informes/informes-ccn-cert-publicos/6338-ccn-cert-ia-13-21-ciberamenazas-y-tendencias-edicion-2021-1/file.html
- CCN-CERT. *Red Nacional de SOC*. 2021. Disponible en https://www.ccn.cni.es/index.php/es/soc/red-nacional-de-soc
- DELOITTE ISRAEL. *The Israeli Technological Eco-system*. 2021. Disponible en https://www2.deloitte.com/il/en/pages/innovation/article/the_israeli_technological_eco-system.html
- DHILUNG KIRAT, Jiyong Jang ; STOECKLIN, Marc Ph.. DeepLocker - Concealing Targeted Attacks with AI Locksmithing. En *Black Hat Conference 2018*. 2018. Disponible en https://www.cisoplatform.com/profiles/blogs/deeplocker-concealing-targeted-attacks-with-ai-locksmithing-black
- ENISA. *Artificial Intelligence Cybersecurity Challenges*. 2020. Disponible en https://www.enisa.europa.eu/news/publications/artificial-intelligence-cybersecurity-challenges
- ETSI. *Securing Artificial Intelligence*. 2020. Disponible en https://www.etsi.org/technologies/securing-artificial-intelligence#mytoc3
- EUROPEAN COMMISSION. *Coordinated Plan on Artificial Intelligence 2021 Review*. 2021. Disponible en https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review
- EUROPOL. *Malicious Uses and Abuses of Artificial Intelligence*. 2021. Disponible en https://www.europol.europa.eu/publications-events/publications/malicious-uses-and-abuses-of-artificial-intelligence

- EY. *Why a culture change program is key to effective cybersecurity*. 2020. Disponible en https://www.ey.com/en_se/giss/why-a-culture-change-program-is-key-to-effective-cybersecurity

- GARTNER. *Gartner market guide for Managed Detection and Response Services*. 2021. Disponible en https://www.gartner.com/en/documents/4007295-market-guide-for-managed-detection-and-response-services

- GOBIERNO FEDERAL ALEMÁN. *Germany AI Strategy Report*. 2018. Disponible en https://knowledge4policy.ec.europa.eu/ai-watch/germany-ai-strategy-report_en

- GOUVERNEMENT DE LA FRANCE. *La stratégie nationale pour l'intelligence artificielle*. 2021. Disponible en https://www.economie.gouv.fr/la-strategie-nationale-pour-lintelligence-artificielle#

- HUNTER, Edward. Artificial Intelligence in Tomorrow's Political Information Space. En *Big Data and Artificial Intelligence for Military Decision Making*. Ed. NATO. 2018. Disponible en https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-IST-160/MP-IST-160-PT-3.pdf

- MAHMOOD SHARIF, y otros. Real and Stealthy Attacks on State-of-the-Art Face Recognition. En *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016) p. 1528–1540. Disponible en https://dl.acm.org/doi/10.1145/2976749.2978392

- RHODE, Matilda; BURNAP, Pete; WEDGBURY, Adam. *Real-Time Malware Process Detection and Automated Process Killing.* 2015. Disponible en https://www.hindawi.com/journals/scn/2021/8933681

- MCKINSEY. *The state of AI in 2020*. 2020. Disponible en https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020

- MICROSOFT. *Microsoft Digital Defense Report*. 10 de 2021. Disponible en https://www.microsoft.com/en-us/security/business/microsoft-digital-defense-report

- PEDREÑO, Andrés y Moreno, Luis. *Prevenir el declive en la era de la Inteligencia Artificial, Europa frente a EEUU y China*. 2021. Disponible en https://www.prevenireldeclive.com

- BHUNIA, Priyankar. *Chinese Government sets out specific targets in three-year action plan on artificial intelligence*. 2017. Disponible en https://opengovasia.com/chinese-government-sets-out-specific-targets-in-three-year-action-plan-on-artificial-intelligence

- US FEDERAL BOARDS & COMMISSIONS. *National Security Commission on Artificial Intelligence Final Report*. 2021. Disponible en https://bookstore.gpo.gov/products/national-security-commission-artificial-intelligence-final-report

- *Yuval Harari's blistering warning to Davos*. Enero de 2020. Disponible en https://www.weforum.org/agenda/2020/01/yuval-hararis-warning-davos-speech-future-predications

GLOSARIO

Adversarial AI	Inteligencia Artificial adversa
APT	Advanced Persistent Threat, Amenaza Persistente Avanzada
CASB	Cloud Access Security Brokers, Agente de seguridad en el acceso a la nube
CCN-CERT	Centro Criptológico Nacional, Centro de alerta y Respuesta Temprana
CRUE	Conferencia de Rectores de las Universidades Españolas
DIN	Instituto Alemán para la Normalización
DL	Deep Learning, técnica de aprendizaje profundo

EDR	Endpoint Detection and Response, detección y respuesta en el dispositivo
Endpoint	Extremo vulnerable tal como un ordenador personal o un movil
ENISA	European Union Agency for Cybersecurity
ENS	Esquema Nacional de Seguridad
ETSI	European Telecommunications Standards Institute, Instituto europeo de estándares de telecomunicaciones
INCIBE	Instituto Nacional de Ciberseguridad
MDR	Managed Detection and Response, Detección y respuesta gestionadas
ML	Machine Learning, Aprendizaje de máquina
NIST	National Institute of Standards and Technology, Instituto norteamericano de estándares de tecnología
NN	Neural Networks, Redes neuronales
Sandbox	Caja de arena, entorno seguro para probar *malware*
SIEM	Security Information and Event Management, Sistema de gestión de eventos de seguridad e información
SOAR	Security Orchestration, Automation and Response, Orquestación automatización y respuesta en Seguridad
SOC	Security Operations Center, Centro de operaciones de seguridad
UEBA	User and Entity Behavioral Analysis, análisis de comportamiento del usuario y del dispositivo
WAF	Web Application Firewall, cortafuegos de aplicaciones web
XDR	eXtended Detection and Response, Detección y respuesta extendida
Zero Day	Vulnerabilidad que todavía es desconocida
Zero Trust	Estrategia de presuponer que nada ni nadie es de fiar