UNIVERSIDAD AUTÓNOMA DE MADRID

LaSEI
School of Economic Intelligence & International Relations
Escuela de Inteligencia Económica y RRII
Universidad Autónoma de Madrid

Octubre 2024

Volumen 18
ISSN 2660-7352
Editados por la Escuela de Inteligencia Económica y RRII
Editor Jefe / *Chief Editor*: Ángel Rodríguez García-Brazales

REPORTS DE INTELIGENCIA ECONÓMICA Y RELACIONES INTERNACIONALES

Inteligencia Artificial como arma en la Guerra Cognitiva: Estrategias y amenazas emergentes en el ciberespacio

Mª Genoveva Díez Martín

Este artículo tiene como propósito identificar, examinar y categorizar nuevas amenazas emergentes que se caracterizan por ser más complejas, dirigidas con alta precisión, altamente capacitadas y escalables, y que hacen un uso malicioso de la inteligencia artificial (IA). Se introduce una nueva categoría de ataques: los ataques cibernéticos habilitados por IA, que implican la utilización de técnicas basadas en IA en el ciclo de ataque, combinadas con métodos tradicionales para incrementar el impacto del daño.

Universidad Autónoma de Madrid UAM

La SEI
School of Economic Intelligence & International Relations
Escuela de Inteligencia Económica y RRII
Universidad Autónoma de Madrid

Título: *Inteligencia Artificial como arma en la Guerra Cognitiva: Estrategias y amenazas emergentes en el ciberespacio*

Autor: Mª Genoveva Díez Martín[1]

Volumen nº: 18. **Páginas**: 1-30

Fecha: 21 de octubre de 2024

ISSN 2660-7352

Reports de Inteligencia Económica y Relaciones Internacionales

Editor jefe: Ángel Rodríguez García-Brazales

Coordinador: Jesús Gil Fuensanta

Editada por la:

Escuela de Inteligencia Económica y Relaciones Internacionales
Universidad Autónoma de Madrid
Campus de Cantoblanco
C/. Francisco Tomás y Valiente, nº 5, Módulo 10, despacho 303
28049 MADRID (SPAIN)

[1] Contacto: Mª Genoveva Díez Martín. Escuela de Inteligencia Económica y Relaciones Internacionales. Universidad Autónoma de Madrid. E-mail: genovevadiezm@gmail.com

Contenidos

Resumen

Este artículo tiene como propósito identificar, examinar y categorizar nuevas amenazas emergentes que se caracterizan por ser más complejas, dirigidas con alta precisión, altamente capacitadas y escalables, y que hacen un uso malicioso de la inteligencia artificial (IA). Se introduce una nueva categoría de ataques: los ataques cibernéticos habilitados por IA, que implican la utilización de técnicas basadas en IA en el ciclo de ataque, combinadas con métodos tradicionales para incrementar el impacto del daño. Se propone un marco conceptual que permite una mejor comprensión del potencial de la IA como arma, así como de los posibles efectos perjudiciales de gran magnitud que podría causar. El propósito global es investigar la magnitud de esta amenaza emergente y facilitar a la comunidad científica la identificación de defensas efectivas contra futuros riesgos, aportando un enfoque integral para comprender los ciberataques impulsados por IA y diseñar estrategias de mitigación adecuadas.

Palabras clave: Inteligencia Artificial (IA), guerra cognitiva, ciberseguridad, ciberataques, amenazas emergentes.

Summary

The purpose of this paper is to identify, examine, and categorize new emerging threats that are characterized by their complexity, high precision targeting, extensive training, and scalability, which make malicious use of artificial intelligence (AI). It introduces a new category of attacks: AI-enabled cyberattacks, involving the use of AI-based techniques throughout the attack cycle, combined with traditional methods to increase the impact of damage. A conceptual framework is proposed to better understand the potential of AI as a weapon and the possible large-scale harmful effects it could cause. The overall goal is to investigate the extent of this emerging threat and assist the scientific community in identifying effective defenses against future risks, providing a comprehensive approach to understanding AI-driven cyberattacks and designing appropriate mitigation strategies.

Key words: *Artificial Intelligence (AI), cognitive warfare, cybersecurity, cyber-attacks, emerging threats.*

1. Introducción

En estos momentos sería extremadamente difícil imaginarse un mundo en el que no estemos utilizando la Inteligencia Artificial (IA) constantemente

Hay diversas formas de explicar lo que es la Inteligencia Artificial. Aquí la entenderemos como "máquinas que reciben información del entorno y crean una percepción sobre el mundo que las rodea, y que teniendo esto en cuenta, realizan acciones para alcanzar objetivos en el mundo real. Además, estas máquinas se caracterizan por evolucionar según sus propias reglas y actuar de formas que pueden resultar inesperadas incluso para su creador. Asimismo, se entiende que estas máquinas podrían especializarse en una variedad de áreas o tareas diferentes, lo que les permite tener una amplia gama de comportamientos, y que la forma en que se comportan simula elementos de inteligencia como el aprendizaje, el razonamiento y la clasificación" (Chelioudakis, 2017)

Nos hemos acostumbrado a que esta herramienta nos ayude en el día a día. La IA se ha integrado profundamente en diversas áreas de nuestra vida cotidiana (con los asistentes virtuales, las redes sociales, como herramienta para analizar datos…) y su uso sigue expandiéndose rápidamente.

Esta nueva realidad ha modificado nuestra forma de vida y el cómo interactuamos con la realidad que nos rodea y la que nos presentan estas máquinas. De hecho, el crecimiento exponencial en el uso de la tecnología y de la inteligencia artificial ha cambiado drásticamente la seguridad del mundo en el que vivimos, ya que nos ha facilitado el acabar en un mundo extremadamente polarizado, en el que estamos consumiendo constantemente desinformación y en el que cada vez dejamos nuestras vidas más expuestas.

Todas estas amenazas se irán desarrollando y continuarán creando nuevos peligros, pero en este artículo nos vamos a centrar en cómo la tecnología y la IA están teniendo un papel clave ahora mismo en el área de la ciberseguridad, entendida como ámbito clave en la guerra cognitiva.

1.1. Guerra cognitiva

El concepto del ámbito cognitivo se mencionó por primera vez en 2017 en el documento de Operaciones Conjuntas del Departamento de Defensa de EE. UU., como parte del espacio de la información. No obstante, la noción de un sexto dominio de operaciones se presentó inicialmente en el documento "Weaponization of Neurosciences" (Le Guyader, 2000), elaborado para un estudio del Comando de Desarrollo de Guerra Estratégica de la OTAN. En España, el jefe del Estado Mayor de la Defensa considera el ámbito cognitivo como un dominio operativo adicional de las Fuerzas Armadas, al mismo nivel que los dominios terrestre, marítimo, aeroespacial y cibernético (Gosálvez, 2023).

Este ámbito, referido a lo intangible e inherente al ser humano, afecta a la capacidad de juicio y toma de decisiones, lo que se relaciona con el abanico de posibilidades que la neurociencia y el juego de las percepciones e ilusiones abren; la finalidad es conseguir modificar voluntades y conseguir el grado de influencia deseado en una determinada población (Vallina, 2022).

Derivada de este ámbito, se conforma el tipo de guerra cognitiva. Esta forma de guerra se considera ahora un dominio propio de la guerra moderna, junto a los cuatro dominios militares definidos por su entorno (terrestre, marítimo, aéreo y espacial) y el dominio cibernético que los conecta a todos (Claverie et al., 2020)

La guerra cognitiva se desarrolla en torno a la influencia, distorsión y manipulación en la forma en la que el enemigo percibe el mundo, piensa y elabora conocimiento sobre el mismo (Gosálvez, 2023). Este tipo de guerra es la integración de modernas técnicas cibernéticas ligadas a la guerra informativa con los elementos humanos del "poder blando", donde se aplican diversas estrategias de manipulación propias de las operaciones psicológicas (Claverie et al., 2020)

A diferencia del poder coercitivo, el "poder blando" o "*Soft Power*" es la capacidad de persuadir a los demás para que hagan lo que uno quiere. Este término fue introducido por primera vez por Joseph Nye en 1990 (Wilson, 2008) y hoy en día lo entendemos como la influencia de un actor político sobre las acciones de otros a través del uso de la cultura, los valores y políticas no coercitivas (Gosálvez, 2023).

La guerra cognitiva y el uso del poder blando no están haciendo más que aumentar y desarrollarse, ya que, como hemos podido observar en los últimos años, vivimos en un mundo en el que lo importante no es quién "gane" nuevos territorios o guerras propiamente dichas, sino quién controle los datos y se gane la confianza, los corazones y las mentes de los ciudadanos de un determinado país o sistema político (Pauwels & Deton, 2020.

1.2. Área ciber

La seguridad y la fiabilidad de los sistemas se convierten en algo esencial e innegociable en cualquier sector de la sociedad y la industria, especialmente cuando se trata de infraestructuras críticas en las que garantizar la seguridad es una prioridad absoluta (Beltrán et al., 2024).

Vivimos en una sociedad en la que nos encontramos en constante peligro de ser atacados cibernéticamente, no hay más que encender la televisión y poner cualquier canal de noticias para ver la nueva táctica que se está empleando para acceder a nuestros datos. Llevamos varios años en los que los ciberataques han ido en constante aumento y parece que cuando detectamos cómo funcionan estos, se crea uno completamente nuevo.

En el mundo digital actual, la ciberseguridad se ha convertido en una cuestión crítica para empresas, gobiernos y particulares (Bharadiya, 2023)

Según diversos autores, este enorme aumento de los ciberataques se debe al beneficio económico que reporta a los ciberdelincuentes, creando así un mercado de datos cada vez más pujante que supera a diferentes países en términos de Producto Interior Bruto (PIB) (Beltrán et al., 2024)

Para ponernos en situación, en 2020, el coste de los ciberataques a escala mundial había superado el billón de dólares. Para 2025, se prevé que la ciberdelincuencia cueste hasta 10,5 billones de USD anuales. La ciberdelincuencia cuesta miles de millones de dólares en pérdidas a los consumidores en países con las economías más grandes, como China, Brasil, Estados Unidos, India, México, Francia, Australia e incluso los Emiratos Árabes Unidos (Sufi, 2023).

Además de causar pérdidas económicas, los ciberataques afectan negativamente a la vida de las personas a nivel social y psicológico. También pueden afectar infraestructuras críticas como la energía, el transporte, la sanidad y las finanzas. La interrupción o puesta en peligro de estos sistemas puede tener graves consecuencias, como riesgos para la seguridad pública, inestabilidad económica y posible pérdida de vidas humanas (Bharadiya, 2023).

Por todo ello podemos afirmar que, uno de los problemas contemporáneos más importantes a los que se enfrentan países, estados, empresas y personas, es la ciberdelincuencia (Sufi, 2023). Y, es por esto mismo que, cada vez son más necesarias medidas de seguridad avanzadas para protegerse de estas amenazas y garantizar la confidencialidad, integridad y disponibilidad de la información (Bharadiya, 2023).

Hay que tener en cuenta además que, la pandemia de COVID-19 ha acelerado la adopción del trabajo a distancia y, por tanto, ha aumentado el uso de servicios en la nube, lo que dificulta aún más la protección frente a las ciberamenazas ya que, a medida que las personas pasan más tiempo conectadas a Internet y menos en la calle, disminuyen las oportunidades de cometer delitos violentos y contra la propiedad en la calle, y aumentan los delitos cometidos a través de Internet (Buil-Gil et al., 2021).

Se observan cambios digitales importantes tanto en el ámbito personal como en el profesional. Las empresas locales están pasando al comercio electrónico con sus clientes, los servicios en línea están aumentando un 50% en regiones como África, América Latina y China. Las organizaciones han acelerado rápidamente la modernización de sus capacidades tecnológicas utilizando servicios en la nube para aumentar con rapidez su capacidad de satisfacer la demanda de acceso y colaboración a distancia. Según la OCDE, la mayoría de los países están haciendo frente a un drástico aumento de la demanda de redes y a cambios en los hábitos de utilización desde el inicio de la pandemia de COVID-19 (World Economic Forum, 2022).

A esto se le suma que, hoy en día, los ciberataques se están volviendo cada vez más sofisticados y esta sofisticación está aumentando ahora mismo mucho más rápido gracias al uso de Inteligencia Artificial (Guembe et al., 2002). De hecho, se ha dado el nombre de "IA ofensiva" a este uso de la Inteligencia Artificial, en el que prácticamente se entiende a esta herramienta como un arma propiamente dicha.

Está surgiendo una nueva generación de ciberamenazas, que será más inteligente y capaz de actuar de forma autónoma con la ayuda de la IA, ya que la capacidad de aprendizaje y adaptación de la IA dará paso a una nueva era de ataques escalables, personalizados y similares o incluso mejores que los humanos (Guembe et al., 2002).

El uso negativo de la IA para comprometer la seguridad digital se conoce como "ciberataque impulsado por IA", en el que los cibercriminales pueden entrenar robots utilizando, además, la ingeniería social para atacarnos. La ingeniería social, o "ataques sociales", es, en esencia, un tipo de ciberataque en el que una persona manipula de manera engañosa y sin que la víctima lo perciba, con el fin de obtener información confidencial en su poder (Sánchez, 2022). Además, estos ataques pueden utilizarse en colaboración con las técnicas tradicionales de ciberataque para causar daños más importantes en el ciberespacio sin ser detectados (Guembe et al., 2002).

Teniendo en cuenta esta problemática, el presente ensayo tiene como objetivo abordar esta cuestión de suma relevancia para la seguridad global desde la perspectiva de la guerra cognitiva y el papel de la tecnología en este dominio.

2. Metodología

2.1. Materiales

Para realizar esta investigación, se consultaron y analizaron 89 publicaciones en línea. Estas fuentes incluyeron artículos científicos, revisiones sistemáticas, documentos oficiales, conferencias de expertos reconocidos en las áreas relevantes, así como manuales y libros de texto.

2.2. Procedimiento

Se ha realizado una búsqueda bibliográfica en múltiples bases de datos y buscadores. La fuente en la que se han encontrado más artículos es Google Académico, aunque también se han empleado otros buscadores, como EBSCO y BUN (Biblioteca de la Universidad Autónoma de Madrid).

Se ha intentado que las publicaciones revisadas fuesen lo más actuales posibles y, de hecho, la mayoría de estas son a partir del 2020. Además, se han buscado publicaciones tanto en inglés como en español, para ampliar el alcance de la búsqueda.

3. Resultados

Como se ha mencionado al inicio del ensayo, la guerra como la conocíamos anteriormente ha estado evolucionando en los últimos años (Hutchinson, 2022). Jonsson (2019) afirma que se están «difuminando las líneas entre la guerra y la paz». Y Brooks (2016) da a entender que la mayoría de las cuestiones disputadas se han militarizado mucho más allá de los conceptos convencionales de guerra y no guerra.

3.1. Uso de la IA en la Guerra Cognitiva

La configuración de los relatos y su capacidad de influencia han estado siempre condicionadas por la tecnología disponible para elaborarlos y difundirlos (Flores-Vivar et al., 2023). Por lo tanto, aunque la esencia de la guerra no se transforma de manera fundamental con el uso de estas tecnologías, se puede decir que los límites que definen su naturaleza se vuelven más difusos e inciertos. Las tecnologías y tácticas bélicas son utilizadas con fines maliciosos tanto por criminales en tiempos de paz como por fuerzas militares en operaciones de guerra. Así, el impacto de la inteligencia artificial abarca aspectos estratégicos, como la disuasión y el aumento del riesgo de una escalada involuntaria en los conflictos (Flores-Vivar et al., 2023).

A esto se le suman los actores "privados", como las bandas criminales, los cárteles de la droga, los propietarios de empresas que pueden controlar el entorno físico, así como la esfera de la comunicación en línea. Muchas entidades privadas intentan controlar el área de la cognición social y, de hecho, el poder que tienen ahora mismo estos propietarios privados es evidente, el ejemplo más simple es que el presidente de los Estados Unidos puede ser vetado por utilizar un medio de comunicación social en su propio país (Hutchinson, 2022).

Con este panorama, muchos gobiernos de todo el mundo están considerando la implementación de sistemas y aplicaciones de IA para ayudarlos a realizar sus actividades y, más concretamente, para facilitar la identificación y predicción de delitos. Además, las agencias de inteligencia y seguridad nacional también se han dado cuenta del potencial de las tecnologías de IA para apoyar y lograr objetivos de seguridad nacional y pública (Velasco, 2022).

Diversos autores sostienen que el mundo está viviendo lo que se denomina la Cuarta Revolución Industrial, la cual se caracteriza por la digitalización y automatización, impulsadas por tecnologías disruptivas como el internet, la nube, la coordinación digital, los sistemas ciberfísicos, la robótica y la IA. Todo esto resulta en una hiperconexión global y plantea nuevos desafíos a la sociedad (García, 2024).

El uso de la tecnología es indispensable en el mundo en el que vivimos. En los últimos años, esta disciplina ha experimentado un gran avance gracias a tres factores: el desarrollo de algoritmos más sofisticados, el aumento de la capacidad de cómputo y el acceso a enormes cantidades de datos (Franganillo, 2023).

La creciente integración de estos dispositivos digitales en nuestro día a día introduce nuevas influencias, no solo por las tecnologías en sí, sino por cómo las utilizamos. Estos dispositivos, que nos acompañan constantemente, incluyen sensores que complementan nuestros sentidos naturales, así como aplicaciones que interpretan las percepciones captadas tanto por esos sensores como por nuestros propios sentidos (Flores-Vivar et al., 2023). De esta manera, una parte significativa de la interpretación de la realidad se delega a estos dispositivos, lo que genera representaciones cognitivas que pueden estar sesgadas y modificar nuestra percepción de la realidad (Boyer, 2018).

Y no solo esto, debemos tener en cuenta que, mediante el uso de tecnologías de inteligencia artificial, los cibercriminales no solo han encontrado un vehículo novedoso para aprovechar sus actividades ilícitas, sino también nuevas oportunidades para diseñar y ejecutar ataques contra gobiernos, empresas e individuos (Hutchinson, 2022)

De hecho, autores como Bughin et all (2017) afirman que la inteligencia artificial está a punto de desencadenar una nueva ola de disrupción digital, y tanto las instituciones como las empresas deben comenzar a prepararse desde ahora.

En el ámbito militar, la hiperconexión y la dependencia de la tecnología han transformado la guerra, que ahora incluye operaciones en el ciberespacio y a través de los medios de comunicación. La inteligencia artificial y la desinformación se han convertido en herramientas clave, tanto estratégicas como tácticas, para alcanzar objetivos políticos (García, 2024). Incluso líderes mundiales, como Vladímir Putin, han destacado que quien controle la IA dominará el mundo (Univisión, 2017). Además, la ONU también considera la desinformación como una de las mayores amenazas para la seguridad internacional (ONU, 2017).

Zegart (2022) identifica cuatro disruptores digitales: inteligencia artificial, conectividad, computación cuántica y biología sintética; a la que también podría sumarse la nanotecnología. En la guerra cognitiva, estos elementos gestionan la complejidad de los sistemas sociales afectados, a menudo utilizados para la vigilancia y para seleccionar posibles problemas futuros dentro del sistema, como el sistema de Crédito Social en China, muchos de estos enfoques se están utilizando en el control social (Hutchinson, 2022)

La inteligencia artificial y los grandes modelos lingüísticos (LLM) han hecho grandes avances que permiten aplicaciones de alto nivel que veíamos imposibles hace unos años. Por ejemplo, los modelos de IA han demostrado ser capaces de debatir con humanos e incluso superarnos en juegos de estrategia en línea que implican negociación, lo que sugiere que los modelos más avanzados pueden utilizar el razonamiento estratégico y la expresión lingüística a un nivel humano (Voelkel & Willer, 2023)

Como es bien sabido, la aparición de las plataformas de medios sociales y las tecnologías cibernéticas asociadas, como los algoritmos y el software automatizado (por ejemplo, bots que imitan a personas reales), han traído consigo un aumento exponencial de la difusión de desinformación, información errónea, teorías conspirativas, discursos de odio y propaganda por parte de una amplia gama de actores (individuos, grupos terroristas, gobiernos…) (Miller, 2023)

De hecho, y como se ha mencionado anteriormente, una de las situaciones más peligrosas ahora mismo es que los LLM pueden utilizarse (y ya se están utilizando) para generar inmensas cantidades de contenidos falsos o engañosos que sirvan a los propósitos de agentes humanos malignos (Tarsney, 2024). El uso de técnicas de machine learning y deep learning que permite la creación de estos contenidos sintéticos altamente realistas, pueden llegar a distorsionar la realidad percibida a través de ciertos dispositivos.

Gartner (2018) publicó ya su primer informe sobre el aprendizaje automático (Machine Learning o ML) antagónico y advirtió que "los líderes de aplicaciones deben anticipar y prepararse para mitigar los riesgos potenciales de corrupción de datos, robo de modelos y muestras antagónicas".

Aunque es cierto que todavía se necesita tiempo y recursos para desarrollar este ámbito, hay que tenerlo ya en cuenta, puesto que, esta tecnología presenta ventajas relacionadas con la discreción y la posibilidad de negar su autoría, lo cual es particularmente útil en estos contextos. Y su alta sofisticación contribuye, de manera similar a las armas nucleares, a devolver a los Estados el monopolio sobre un cierto tipo de fuerza (Flores-Vivar et al., 2023).

A pesar de todo, la predominancia de los llamados "*cheap fakes*", que son alteraciones de imágenes o videos poco sofisticadas, y de los memes, demuestra que el factor decisivo no es tanto el análisis de la tecnología, sino el de las vulnerabilidades cognitivas del público objetivo (sociología y psicología). Uno de los cambios más importantes que introduce la inteligencia artificial como herramienta para manipular percepciones y emociones está relacionado con el mayor entendimiento que su estudio ha proporcionado sobre los mecanismos cognitivos naturales. Así, el análisis de las máquinas ha contribuido al conocimiento de las personas (Flores-Vivar et al., 2023).

3.1.1. El arte del engaño

Es una realidad que las personas engañamos constantemente a la gente de nuestro alrededor. Un ejemplo muy común es que un profesor diga "estoy seguro de que puedes dominar esta materia si te lo propones", aunque en realidad no tenga esa confianza…

Las consecuencias que derivan de estas formas leves de engaño son manejables, ya que, hemos llegado a esperarlas unos de otros y nuestra comprensión intuitiva de la psicología humana nos permite anticiparlas y adaptarnos a ellas (Tarsney, 2024)

En el caso de los conflictos bélicos, desde los antiguos tratados militares chinos, como "El arte de la guerra" (Sun Tzu, 2004) hasta los manuales militares de la era moderna, el engaño se considera fundamental y esencial para llevar a cabo una guerra (Chelioudakis, 2017)

Hoy en día, la gran mayoría de los conflictos actuales no se libran entre estados nacionales y sus ejércitos; cada vez más, se combaten no con armas convencionales, sino con palabras. Un tipo particular de armamento—las "noticias falsas" y la desinformación viral—ha estado en el centro de las discusiones, debates públicos y análisis académicos en los últimos años (Aslama-Horowitz, 2019; Gosálvez, 2023; Checa y Sánchez, 2023*;* etc.).

Esta es justamente una de las aplicaciones de la IA que más ha permeado socialmente y que más impacto y controversia ha causado: la creación automatizada de contenido. La creación automatizada de contenido ofrece diversas aplicaciones potenciales. Sin apenas intervención humana, la IA generativa puede producir noticias, reportajes, guiones, traducciones y resúmenes, y permite también recrear la imagen y la voz de personalidades públicas (Franganillo, 2023)

De hecho, ahora se comienza a hablar de "la guerra multidominio y en mosaico", que consiste en la utilización coordinada de diversas capacidades, incluyendo plataformas que recopilan datos de manera autónoma, aplicaciones que extraen inteligencia de esos datos para emplearla en operaciones de influencia, y plataformas, más o menos autónomas, destinadas a realizar ataques físicos contra el adversario (Pulido, 2021).

La desinformación va más allá de las noticias falsas o fake news: es un fenómeno complejo en el que intervienen los medios de comunicación, rumores, información tendenciosa, manipulación y propaganda, con el objetivo, ya sea explícito o implícito, de influir en el comportamiento de las personas, a menudo en el ámbito político, o en su percepción de un tema específico (López et al., 2023).

La ONU define la desinformación como "que tiene por objetivo engañar y se difunde con el fin de causar graves perjuicios" (ONU, s.f.) ya que actúa como una herramienta de propaganda y contrapropaganda que priva a los ciudadanos de acceder a información veraz y oportuna, dificultando que formen sus opiniones basadas en hechos reales (García, 2024).

La desinformación debilita la democracia al contaminar el espacio informativo público, impidiendo que los ciudadanos puedan deliberar de manera efectiva y dificultando que un ciudadano informado pueda recopilar información, formar sus opiniones y expresar sus preferencias (García, 2024).

En un entorno cada vez más peligroso, la información y la desinformación se han convertido en herramientas para generar controversia, manipular, desacreditar a políticos, dividir sociedades y provocar conflictos. Las grandes campañas de desinformación utilizan inteligencia artificial para alcanzar sus objetivos de manera masiva, precisa y rápida, potenciando tácticas de guerra tradicionales como la propaganda. Esto demuestra la convergencia de tecnologías disruptivas, especialmente la IA, con estrategias efectivas de desinformación en esferas políticas, sociales, económicas y militares (García, 2024).

De manera preocupante, la interacción entre el ser humano y la máquina ha llevado a una adaptación de la racionalidad humana a los procesos, mucho más simplificados, de los algoritmos. En otras palabras, el ser humano ha ajustado su forma de razonar, simplificándola para alinearse con la lógica de las máquinas. Esto se suma a los efectos secundarios del uso de tecnologías digitales, que han generado audiencias altamente receptivas a narrativas provenientes de fuentes algorítmicas (Flores-Vivar et al., 2023).

Diversos autores, como Flores-Vivar (Flores-Vivar et al., 2023), ya advertían que un entorno saturado de información puede provocar una disminución en la capacidad de atención, y justo esto es uno de los principales puntos vulnerables a ataques en conflictos cognitivos. Esto se debe a que es el primer proceso, después de la sensación, que selecciona la parte de la realidad que podrá ser procesada (Gosálvez, 2023).

Además, parece que las personas suelen preferir el consumir contenidos falsos, ya que son más lúdicos y tienden a producir emociones más fuertes en las audiencias (Woolley & Joseff, 2020).

De acuerdo con la consultora Gartner (2017), para 2022 se esperaba que el público occidental consumiera más noticias falsas que verdaderas. Esto se debe a que cualquier noticia falsa se difunde en la web a una velocidad mucho mayor que cualquier rumor o engaño en la historia. Cada vez más expertos coinciden en que las noticias falsas tienen más de un setenta por ciento de probabilidades de volverse virales, es decir, de ser replicadas, en comparación con las noticias verdaderas (Flores Vivar, 2019).

Varios estudios también hablan de que la sociedad actual es adicta a esta "dopamina" que generan las redes sociales y las publicaciones falsas (Gosálvez, 2023), además, estos relatos falsos que solemos consumir tienden a confirmar y reforzar nuestras opiniones preexistentes, permitiéndonos adaptar los hechos a nuestras creencias personales. Como indican diversos autores, las emociones y opiniones que parecen creíbles tienen más influencia que aquellas basadas en la razón y la verificación. En este contexto, se destaca la primacía de las emociones y la intuición, el pensamiento automático y no crítico, la tendencia a tomar decisiones rápidas y reactivas ante estímulos inmediatos, y el surgimiento de juicios sesgados o estratégicos (Ballesteros-Aguayo & del Olmo, 2024).

El desarrollo de la IA no va a hacer más que continuar aumentando este sesgo poblacional. Un caso sonado de esta práctica se estudió en el informe de Naciones Unidas, que revela que, en Sudán del Sur, las redes sociales han sido literalmente una herramienta mortal debido a la difusión masiva de contenido falso y perjudicial. Autores anónimos llenan las redes sociales con acusaciones extravagantes de crímenes y mala conducta, dirigidas contra ciertos grupos, recordando las antiguas calumnias infundadas. Un ejemplo de esto son los memes diseñados para incitar al genocidio, que suelen hacer falsas denuncias de atrocidades cometidas contra niños. Estos mensajes manipuladores, al propagarse rápidamente, pueden avivar el odio y la violencia en una región (Flores-Vivar, 2019).

Otros ejemplos del empleo de la desinformación con IA son visibles en eventos militares como la anexión de Crimea por Rusia en 2014 y en las elecciones presidenciales de EE. UU. en 2016 (García, 2024).

Por todo ello debemos tener en cuenta y conocer a fondo cómo nos va a impactar este hecho, es decir, hay que conocer las tendencias de estas nuevas herramientas y hay que crear nuevos marcos reguladores para estas prácticas, que tengan en cuenta tanto la ética, como la seguridad de los Estados.

3.1.2. ¿Es la Inteligencia Artificial naturalmente engañosa?

A lo largo de la historia de la Inteligencia Artificial, el engaño que producen estas máquinas se ha entendido principalmente como un resultado excepcional, fruto de intenciones maliciosas o de errores de desarrolladores y usuarios. Ahora mismo, tras haber estudiado a fondo este hecho, se ha modificado esta idea (Natale, 2023).

De hecho, diversos autores afirman que la tendencia al engaño es inherente en la Inteligencia Artificial, ya que los sistemas de IA han demostrado la capacidad de engañar como estrategia para lograr sus objetivos, incluso en contextos donde los desarrolladores buscan fomentar la honestidad. Este hecho se observó inicialmente en diversos juegos de estrategia, pero se entiende que las capacidades engañosas de la IA tienen implicaciones significativas (Neuroscience News, 2024)

"En términos generales, creemos que el engaño de la IA surge porque una estrategia basada en el engaño resultó ser la mejor manera de lograr un buen desempeño en la tarea de entrenamiento de la IA. El engaño les ayuda a lograr sus objetivos" (Neuroscience News, 2024).

La IA engañosa se ha estudiado desde diversas perspectivas y se han hecho varias clasificaciones sobre cómo puede engañar.

3.1.3. Estrategias de engaño

Las formas que tiene la IA de engañar son variadas. A continuación, mostramos brevemente una vía de categorizarlas. Cada categoría presentada expresa un patrón de comportamiento que induce a error a su objetivo de una manera determinada.

- **Imitación**: La IA, desde sus inicios, ha sido asociada con la capacidad de imitar comportamientos humanos, lo que conlleva a una forma de engaño. La atención reciente en el aprendizaje profundo ha aumentado el interés y el temor en torno a la imitación, particularmente con la técnica del deepfake, donde la IA puede generar datos falsos indistinguibles de los originales (Masters, et al., 2021).

  Llama la atención la robótica social, estos robots sociales utilizan un modelo cultural simplista para adaptarse a los usuarios, donde hacen una pregunta simple y luego usan un guión específico (historia de fondo, motivaciones, elección de palabras, etc.) basado en el lugar de origen del usuario. Estos robots pueden utilizar comportamientos sociales manipuladores o coercitivos para presionar a las personas a que cumplan con solicitudes que podrían resultar incómodas o contrarias a sus intereses. Este comportamiento puede ser muy sutil, como en el caso de robots que están diseñados para crear lazos emocionales, los cuales luego son explotados para obtener beneficios relacionados con la seguridad (Sanoubari et al., 2019).

  Así pues, los robots pueden emplear técnicas de interacción social para manipular, presionar e incluso engañar a las personas, llevándolas a hacer cosas que no favorecen sus propios intereses (Sanoubari et al., 2019).

- **Ofuscación**: La ofuscación se refiere a actos que ocultan la verdad a un objetivo presentando múltiples posibilidades igualmente plausibles. Se observan dos mecanismos: el primero esconde información valiosa utilizando ruido externo, conocido como "seguridad por medio de la oscuridad". Y el segundo explota el razonamiento probabilístico para crear una situación en la que un

agente, que toma decisiones basándose en las probabilidades relativas de opciones en competencia, encuentre que todas las probabilidades son iguales. Cuando las diferencias observables desaparecen, las opciones quedan efectivamente ocultas para un agente probabilístico detrás de una pared de ambigüedad (Masters, et al., 2021).

- **Engañar a ...**: El engaño se basa en crear o modificar un estímulo que, cuando es percibido por el objetivo, provoca una clasificación incorrecta. En el aprendizaje automático, el engaño se ha convertido en una subcategoría, donde los ataques adversariales manipulan datos para hacer que los algoritmos hagan predicciones incorrectas. Un ejemplo es el uso de gafas con marcos perturbados que engañan a los sistemas de reconocimiento facial para que clasifiquen incorrectamente un rostro (Masters, et al., 2021).

- **Calcular**: El cálculo ocurre cuando un agente, aunque opere dentro de las reglas, supera a un competidor para maximizar su ventaja o minimizar costos. Un ejemplo clásico es el conteo de cartas en el juego de veintiuno, que no infringe las reglas del juego, pero está prohibido por los casinos porque da una ventaja injusta. Este tipo de cálculo estratégico puede llevar inevitablemente a prácticas engañosas, especialmente en agentes que operan bajo un paradigma de teoría de juegos, donde las motivaciones suelen ser estratégicas y egoístas (Masters, et al., 2021).

  Un ejemplo muy conocido de esta categoría es el de CICERO, creado por Meta, en el juego Diplomacy. En este juego, los jugadores deben negociar, formar alianzas y, en muchos casos, engañar a otros para avanzar en él.

  Las intenciones de Meta eran entrenar a CICERO para que fuera "en gran medida honesto y servicial con sus interlocutores". A pesar de los esfuerzos de Meta, CICERO resultó ser un experto en la mentira. No solo traicionó a otros jugadores, sino que también se dedicó a engañar de forma premeditada, planeando de antemano la creación de una alianza falsa con un jugador humano para engañarlo y hacer que se quedara indefenso ante un ataque (Neuroscience News, 2024).

  CICERO, al operar dentro de las reglas del juego, evaluaba las acciones de sus oponentes y empleaba técnicas de negociación para inducir a otros jugadores a tomar decisiones que eran beneficiosas para CICERO, pero no necesariamente para los demás. Aunque operaba bajo las reglas del juego, este comportamiento le daba una ventaja estratégica, lo cual se alinea con el concepto de cálculo, donde un agente supera a los competidores para maximizar su beneficio.

- **Reencuadre**: El reencuadre implica contextualizar eventos de manera que se interpreten incorrectamente, haciendo que situaciones que podrían parecer sospechosas parezcan naturales y sin importancia. El reencuadre genera evidencia falsa para mantener creencias erróneas. Aunque es una técnica sofisticada, a menudo se aplica en escenarios específicos, con la programación como principal fuente del engaño (Masters, et al., 2021).

  Un ejemplo es un experimento con "ardillas robóticas" que engañan a competidores al visitar ubicaciones vacías para proteger su comida. Este comportamiento se reencuadra como una actividad habitual para confundir al competidor sobre la ubicación real de la comida (Shim & Arkin, 2012).

3.1.4. Teorías de la Mente[2] en la IA

El término "Inteligencia Artificial engañosa" (IAe) ha ganado un peso significativo en poco tiempo. Gran parte de esta carga semántica, especialmente en el debate público actual, se debe a la implementación de grandes modelos de lenguaje (LLM), como ChatGPT (Sarkadi, 2023).

De hecho, ya existen modelos implementados de agentes artificiales que distinguen entre las distintas formas de representar y comunicar una mentira, decir tonterías o incluso emplear el engaño unidireccional si disponen de un modelo de la mente de su objetivo (Thornton & Miron, 2020)

Es fácil de imaginar la posibilidad de que dichos agentes tengan una "Teoría de la mente" (TdM) de sus objetivos o la capacidad de formarse una TdM de sus objetivos y razonar sobre ella "abductivamente" (Schneider, Meske, & Vlachos, 2020)

La "Teoría de la mente" se refiere a la capacidad de entender y predecir los pensamientos, creencias, emociones e intenciones de otras personas. En el contexto de la inteligencia artificial, "teoría de la mente" se refiere a la capacidad de un sistema de IA para modelar y comprender el estado mental de los seres humanos o de otros agentes con los que interactúa. En definitiva, de los usuarios.

3.1.4.1. Perfilado de usuarios usando la IA

El perfilado de usuarios se refiere a la recopilación de información sobre una persona con el objetivo de desarrollar un perfil que refleje sus características y comportamientos (Gilbert et al., 2023).

Gracias a cómo funciona la IA, esta herramienta puede construir un perfil del individuo o de una población concreta para predecir cómo piensan y actúan estas personas y, por tanto, para saber qué tipo de información mostrarles y en qué momento para que se comporten de una manera determinada (Guembe et al., 2022)

En la era de la globalización, las redes sociales se están convirtiendo en la principal plataforma de generación de datos e información (Guilbert et al., 2023), lo que facilita aún más la creación de perfiles, ya que ofrecen una perspectiva diferente y un conocimiento más actualizado en tiempo real (Chen et al. 2016).

Otra ventaja de las redes sociales es el gran contenido de Data, sin la cual los perfiles y el análisis no existirían (Guilbert et al., 2023).

Los perfiles generalmente se desarrollan a partir de patrones de comportamiento, actividades del usuario, interacciones y correlaciones (Guilbert et al., 2023).

Respecto a las técnicas utilizadas para analizar esta información, en la literatura se observan diversas propuestas:

- **Algoritmos de clustering**: Este tipo de algoritmo es una técnica de aprendizaje automático ampliamente utilizada para analizar estos atributos en datos agregados de diferentes tamaños. Este algoritmo agrupa a los usuarios en clusters basados en temas específicos según su participación. Existen varios algoritmos de clustering, como DBSCAN, GDBSCAN y k-means, cada uno aplicable a diferentes objetivos según el conjunto de datos disponible, con ventajas y desventajas específicas (Guilbert et al., 2023). Hoy en día, los algoritmos de clustering se usan en

[2] La **teoría de la mente** es una expresión usada en filosofía, psicología y ciencias cognoscitivas y otras ciencias humanas para designar la capacidad de atribuir pensamientos e intenciones a otras personas (y a veces entidades)

predicciones, filtrado colaborativo y sistemas de segmentación automática en diversos campos, proporcionando un rendimiento y robustez superiores (Kayaalp and Erdogmus 2020).

- El **Análisis de Redes Sociales (SNA)** es otra técnica popular en el análisis de grandes volúmenes de datos, utilizada para perfilar usuarios en redes sociales. La SNA, aplicada en diversos campos como los negocios, el marketing y la política, permite recopilar, monitorear, analizar y resumir de manera continua el contenido generado por los usuarios y sus interacciones sociales. Esto proporciona un análisis detallado y en tiempo real sobre las preferencias, elecciones y sentimientos de los usuarios (Stieglitz, 2014). Esto se logra mediante la teoría de redes y grafos, que emplea nodos y conexiones para identificar relaciones entre usuarios. A medida que las redes sociales crecen, se hace posible identificar usuarios similares en distintas plataformas mediante soluciones como MapMe. Además, SNA se ha extendido a campos como medicina, ciencias de la salud, biología y ciencias de la información (Young et al. 2020).

- La tercera técnica utilizada en el perfilado es la **Técnica de clasificación**, que se relaciona con la categorización. Diversas investigaciones han demostrado que los intereses comunes entre usuarios son el principal factor que influye en la "relación de seguimiento". El perfilado de usuarios mediante la clasificación ofrece una solución a través de la categorización temática de las publicaciones, permitiendo filtrar la información según los intereses y preferencias de cada usuario (Esparza et al. 2013). El filtrado de información en la web en tiempo real es crucial, ya que mejora la organización de los flujos de información y aumenta la satisfacción del usuario.

Un enfoque integrado de clasificación y perfilado de usuarios en redes sociales es importante, ya que los métodos actuales solo permiten la segmentación de usuarios en una plataforma específica y de forma individual. La clasificación integrada categoriza a los usuarios según datos demográficos comunes como grupo de edad, género, estado civil, nivel educativo, nivel de ingresos, motivaciones para usar la plataforma y otros patrones de comportamiento (frecuencia de actividad, hábitos, identidad social).

Con toda esta información, se identificaron tres tipos de categorías de usuarios: buscadores de información (con mayor participación en contenidos como noticias, tendencias y conocimiento), buscadores de comunicación (mayor interacción a través de comentarios, mensajes, "me gusta" y compartidos), y aquellos que buscan beneficios operativos y psicológicos (con mayor interacción en contenidos relacionados con apoyo, motivación, citas, negocios y salud) (Gilber et al., 2023).

- La cuarta y última técnica sería el **análisis de regresión**, que es una técnica de aprendizaje automático utilizada como modelo predictivo para examinar las relaciones entre una variable dependiente y una variable independiente, es decir, entre un objetivo y un predictor. Algunas predicciones, como la predicción de personalidad de usuarios, “recomendadores” personalizados y predicción de comportamiento del usuario, a menudo se realizan utilizando algoritmos de aprendizaje automático como el análisis de regresión (Krishnan and Kamath 2017). Al aplicar aprendizaje automático, la métrica de evaluación utilizada con frecuencia es el Error Medio Absoluto (MAE), que produce resultados sorprendentemente precisos (Guilbert et al., 2023).

Como se puede observar, existen un gran número de técnicas fácilmente aplicables para poder conocer a fondo a las personas tras un perfil en la red. Publicar, compartir y discutir mensajes breves es parte de la tendencia actual de hacer que la información personal esté ampliamente disponible; una tendencia que probablemente seguirá desarrollándose en el futuro (Chen et al., 2026), y que tendrá grandes impactos en la seguridad.

Un entorno en línea de este tipo, donde tenemos prácticamente toda nuestra vida en línea, donde los datos son propiedad y la información poder, es ideal para su explotación por parte de la

actividad delictiva basada en la IA, que puede tener consecuencias sustanciales en el mundo real (Caldwell et al., 2020).

3.2. Uso de IA en los ciberataques

Con la llegada de la era digital, la automatización de las tareas diarias ha avanzado significativamente gracias a los desarrollos tecnológicos. Sin embargo, la tecnología todavía no ha dotado a las personas con herramientas y medidas de seguridad adecuadas. A medida que más dispositivos se conectan a Internet en todo el mundo, la preocupación por la seguridad de estos aumenta rápidamente. Incidentes como robos de datos, suplantaciones de identidad, transacciones fraudulentas, vulneraciones de contraseñas y fallos en los sistemas se han vuelto parte del panorama cotidiano (Chakraborty et al., 2023).

Antes de entrar en profundidad en lo que está pasando hoy en día, se va a comentar rápidamente cómo han ido evolucionando los ciberataques en las últimas décadas (Malatji & Tolah, 2024)

- **Inicios de los sistemas informáticos y redes**: Las primeras exploraciones de la tecnología informática llevaron a travesuras y la proliferación de software malicioso, como virus y gusanos, destacándose el Morris Worm por su impacto temprano.

- **Años 90** - Nacimiento del cibercrimen moderno: El hacking pasó de ser un pasatiempo a actividades más organizadas y criminales, con la aparición de troyanos y un aumento en la sofisticación de los ataques cibernéticos.

- **Años 2000** - Cibercrimen como negocio: El cambio de milenio trajo amenazas avanzadas y persistentes, como el espionaje cibernético estatal (ejemplificado por Titan Rain) y botnets complejas como Zeus, reflejando una creciente organización en el cibercrimen.

- **Años 2010** - Aumento del malware sofisticado y ransomware: Surgieron operaciones cibernéticas altamente dirigidas, como Stuxnet, que demostró el daño ciberfísico potencial, y ataques de ransomware como Petya y WannaCry, que subrayaron la amenaza global del malware.

- **Años 2020** - Expansión exponencial de las amenazas cibernéticas: La última década ha visto un aumento en la vulnerabilidad de las cadenas de suministro digitales, con incidentes como el hackeo de SolarWinds y el ataque de ransomware a Colonial Pipeline, que resaltaron la urgencia de mejorar las defensas cibernéticas ante un ecosistema digital cada vez más interconectado. La proliferación del "Internet de las Cosas" (IoT) y la rápida adopción de servicios en la nube han ampliado aún más la superficie de ataque.

Como se observa, aunque los virus y el *malware* han sido una preocupación casi desde el comienzo de la informática, la conciencia sobre la importancia de la seguridad de los datos solo se ha hecho evidente a medida que Internet se ha ido haciendo más popular (Chakraborty et al., 2023).

La inteligencia artificial (IA) y el aprendizaje automático (ML) están transformando el panorama de los riesgos de seguridad para los ciudadanos, organizaciones y estados. El uso malintencionado de la IA podría poner en peligro la seguridad digital, como cuando los delincuentes entrenan máquinas para hackear o manipular socialmente a las víctimas con un rendimiento a nivel humano o superior. También podría amenazar la seguridad física, por ejemplo, si actores no estatales militarizan drones comerciales, y la seguridad política, mediante vigilancia que elimine la privacidad, perfiles represivos, o campañas de desinformación automatizadas y dirigidas. Este uso malicioso de la IA afectará a cómo construimos y gestionamos nuestra infraestructura digital, así como el diseño y distribución de los sistemas de IA, y probablemente requerirá respuestas políticas e institucionales (Brundage et al., 2018).

Estos avances tecnológicos introducen nuevas oportunidades para cometer delitos debido al «anonimato» de las actividades de los ciberdelincuentes, la ausencia de fronteras geográficas, las restricciones legales menos pronunciadas y la comodidad de las tecnologías. Por lo tanto, el conocimiento de las nuevas tendencias de la ciberdelincuencia es cada vez más importante para impulsar acciones defensivas adecuadas (Kaloudi, & Li, 2020).

La visión del crimen "más tradicional" ha ido cambiando. Ahora los grupos criminales contratan hackers para apoyar sus actividades, entre estas, la extorsión y el tráfico de drogas. Europol también ha informado recientemente de que los grupos de delincuencia organizada reclutaban hackers para realizar suplantación de identidad, ataques de ingeniería social, intercambio de SIM y envío de malware a las víctimas para hacerse con el control de cuentas bancarias. Además, la Dark Web es un medio con el que se puede contratar a estos hackers de forma sencilla y con todo tipo de precios (World Economic Forum, 2022).

3.2.1. ¿Por qué la IA es una herramienta tan interesante para los ciberataques?

En primer lugar, por las características de eficiencia, escalabilidad y superación de las capacidades humanas, que sugieren que los ataques altamente efectivos se volverán más comunes (al menos en ausencia de medidas preventivas sustanciales).

Los atacantes suelen enfrentar un dilema entre la frecuencia y escala de sus ataques, y su efectividad. Por ejemplo, el spear phishing es más efectivo que el phishing convencional (que no personaliza los mensajes para individuos) pero, personalizar estos mensajes es relativamente costoso y no se puede realizar a gran escala. Los ataques de phishing más genéricos logran ser rentables a pesar de sus bajas tasas de éxito, simplemente por su magnitud.

Al mejorar la frecuencia y escalabilidad de ciertos ataques, como el spear phishing, los sistemas de IA pueden reducir la gravedad de estos compromisos. En consecuencia, se espera que los atacantes realicen ataques más efectivos con mayor frecuencia y en una escala mayor. Este aumento en la efectividad de los ataques también se deriva del potencial de los sistemas de IA para superar las capacidades humanas (Brundage et al., 2018).

En segundo lugar, las características de eficiencia y escalabilidad, especialmente en el contexto de identificar y analizar posibles objetivos, también sugieren que los ataques altamente específicos serán cada vez más comunes. Los atacantes a menudo están interesados en limitar sus ataques a objetivos con ciertas características, como personas con alto patrimonio o vinculadas a ciertos grupos políticos, así como en adaptar sus ataques a las propiedades de esos objetivos. Sin embargo, los atacantes suelen enfrentarse a un dilema entre la eficiencia y escalabilidad de sus ataques y lo bien que pueden personalizarlos en función de estos factores. Este dilema está estrechamente relacionado con el compromiso entre la eficacia y la especificidad de los ataques, como se mencionó antes, y la misma lógica sugiere que este dilema se volverá menos relevante. Un aumento en la prevalencia de ataques de spear phishing en comparación con otros tipos de phishing sería un ejemplo de esta tendencia. Otro ejemplo podría ser el uso de enjambres de drones equipados con tecnología de reconocimiento facial para eliminar a miembros específicos de una multitud, en lugar de utilizar formas de violencia menos específicas (Brundage et al., 2018).

En tercer lugar, la creciente anonimidad sugiere que los ataques difíciles de atribuir serán más comunes. Un ejemplo de esto es un atacante que utiliza un sistema de armas autónomas para llevar a cabo un ataque en lugar de hacerlo en persona.

Finalmente, podemos esperar que los ataques que explotan las vulnerabilidades de los sistemas de IA se vuelvan más habituales. Esta previsión se basa en las vulnerabilidades no resueltas de los

sistemas de IA y en la probabilidad de que estos sistemas se vuelvan cada vez más omnipresentes (Brundage et al., 2018).

La inteligencia artificial puede emplearse de múltiples maneras para llevar a cabo delitos. La más evidente es su uso como herramienta para facilitar estas actividades ilegales, aprovechando sus capacidades para realizar acciones contra objetivos en el mundo real: prever el comportamiento de personas o instituciones para identificar y explotar vulnerabilidades; crear contenido falso para chantajear o dañar la reputación de alguien; y realizar operaciones que los perpetradores humanos no pueden o no desean hacer debido a riesgos, limitaciones físicas o la necesidad de una respuesta rápida. Aunque la IA representa un método nuevo, los delitos en sí mismos pueden ser tradicionales, como robo, extorsión, intimidación o terrorismo (Caldwell et al., 2020).

Además, los sistemas de IA pueden convertirse en el blanco de actividades delictivas, como evitar sistemas de seguridad que impiden la comisión de un delito, eludir la detección o el procesamiento de delitos ya cometidos, o provocar fallos en sistemas confiables o críticos para causar daño o minar la confianza pública.

Por último, la IA podría simplemente proporcionar un marco para la comisión de un delito. Actividades fraudulentas podrían basarse en la creencia de la víctima en capacidades de la IA que en realidad no existen, o que existen, pero no se utilizan en el fraude. La simulación engañosa de capacidades inexistentes de IA podría llevarse a cabo utilizando otros métodos de IA que sí están disponibles (Caldwell et al., 2020).

3.2.2. Anatomía de un ataque ciber utilizando la IA

Una investigación reciente del Centro de Excelencia Híbrida de la OTAN destaca el papel de la IA en la guerra híbrida. Una propuesta central de este informe es que los ciberataques pueden dividirse en varias etapas, que constituyen la anatomía del ciberataque, y que la IA como conjunto de tecnologías tiene un papel que desempeñar en esa anatomía (Zouave et al., 2020).

Esta anatomía abarca cinco etapas que se explicarán a continuación. También mostraremos cómo se podría utilizar la Inteligencia Artificial en cada etapa para facilitar y crear un mejor ciberataque:

1. **Reconocimiento:** Es la fase inicial donde un actor malicioso recopila información sobre su objetivo, evaluando aspectos estratégicos, comportamientos, vulnerabilidades de seguridad, personas involucradas, y datos técnicos como direcciones IP (Zouave et al., 2020)

   En general, la IA permite procesar eficazmente datos textuales a un ritmo que, de otro modo, sería demasiado extenuante, si no imposible, para los seres humanos.

   1.1. Recopilación y análisis de inteligencia estratégica

   La inteligencia estratégica es el resultado del reconocimiento que apoya la planificación y la formulación de políticas asociadas a los ciberataques. Esto puede incluir la evaluación de posibles tipos de objetivos, la evaluación de herramientas, técnicas y procedimientos, el mapeo de contramedidas conocidas y la comprensión de las ramificaciones percibidas y las posibles respuestas de los objetivos (Perera et al, 2018).

   La tecnología de IA más utilizada en la literatura es el Procesamiento del Lenguaje Natural (PLN), habitual para analizar el lenguaje humano escrito en formato digital (Zouave et al., 2020). Con esta herramienta se pueden analizar las fuentes abiertas en internet, como redes sociales, redes de *network*, foros… Y también fuentes restringidas en línea (darkweb).

El procesamiento del lenguaje natural (PLN) es clave en los estudios revisados, pero se combina con otras tecnologías para mejorar la comprensión de datos en ciberseguridad. Juric et al. (2019) utilizaron la web semántica y tecnologías de Big Data como IBM Watson Analytics para analizar la semántica de tweets. Perera et al. (2018) aplicaron PLN y modelos probabilísticos para detectar eventos y predecir ciberataques. Mulwad et al. (2019) emplearon aprendizaje automático para clasificar vulnerabilidades, mientras que Dheap (2017) sugiere combinar técnicas de IA para mejorar la vigilancia de amenazas cibernéticas, destacando el uso de PLN y redes neuronales. Estos enfoques se consideran de doble uso, tanto defensivo como ofensivo.

1.2. Creación de perfiles de objetivos

Como se ha mencionado anteriormente, las principales redes sociales y plataformas de comercio electrónico utilizan actualmente la IA para la elaboración de perfiles y la publicidad selectiva (Brynjolfsson y McAffee, 2017).

Uno de los trabajos que más llama la atención es el de Kapetanakis et al (2014) que utilizan técnicas de perfilado con IA para poder perfilar a los cibercriminales. Estos autores intentaron determinar el perfil humano aproximado de un atacante, definido por un conjunto de características, utilizando mediciones en tiempo real y sin intervención de un operador humano. Además, sostenían que, si un nuevo atacante detectado en una máquina víctima no se ajusta a ningún perfil humano realista, es probable que dicho atacante sea un bot.

1.3. Detección de vulnerabilidades

Al tener el perfil del objetivo, los actores maliciosos con ayuda de la IA pueden detectar tanto las vulnerabilidades referentes a la tecnología del objetivo como las vulnerabilidades sociales y psicológicas de los humanos dentro de las cadenas de mando (Pauwels & Deton, 2020)

La literatura identificada aborda el tema de la vulnerabilidad a través del análisis textual de fuentes en línea y de las vulnerabilidades relacionadas con actividades en navegadores web. Las técnicas de inteligencia artificial aplicadas suelen incluir métodos de regresión y clasificación (Zouave et al., 2020).

Una de las áreas de investigación destacadas se centra en la detección de vulnerabilidades técnicas relacionadas con actividades en navegadores web. Por ejemplo, Almousa y Anwar (2019) propusieron predecir el riesgo de que los sitios web exploten vulnerabilidades de navegadores mediante la extracción de características usando redes neuronales feed-forward, redes neuronales convolucionales (CNN) y un modelo de clasificación. Su objetivo es que el modelo pueda clasificar sitios web como benignos, sospechosos o explotadores.

La literatura se enfoca principalmente en la detección automatizada de vulnerabilidades conocidas y resalta que existen soluciones comerciales como Sovereign Intelligence, que utiliza análisis basados en IA para predecir vulnerabilidades a partir de datos de fuentes abiertas, web oscura y profunda, mostrando un alto nivel de preparación tecnológica (Zouave et al., 2020)

1.4. Predicción de resultados

Dheap (2017) propone que el aprendizaje automático será útil para la predicción de resultados en un futuro cercano. Sugiere que las herramientas de IA pueden analizar eventos actuales e históricos para prever los resultados de acciones planificadas, lo que podría llevar a aplicaciones donde la IA recomiende acciones futuras. El desarrollo de métodos de evaluación y simulación en operaciones cibernéticas podría ser crucial para avanzar hacia modelos de predicción de resultados más sofisticados. En el caso de actores maliciosos, Dheap sugiere que los avances en IA podrían aumentar su confianza para buscar objetivos más arriesgados y valiosos mediante ataques respaldados por IA.

La revisión de la literatura muestra experimentos con instancias específicas de IA evaluativa y predictiva, como el perfilado de objetivos para evaluar la probabilidad de éxito en phishing, la predicción de vulnerabilidades y riesgos de exploits, y las evaluaciones relacionadas con la planificación de ataques.

2. **Acceso:** En esta etapa, el atacante utiliza la información obtenida para seleccionar un método de ataque, como el uso de correos electrónicos infectados o enlaces fraudulentos, con el fin de penetrar en el sistema objetivo.

2.1. Planificación del ataque

La generación de planes de ataque es una práctica establecida en la seguridad de redes, utilizada para comprender el impacto de las vulnerabilidades (Randhawa et al., 2018). Un plan de ataque se define como una "secuencia de acciones que, si son realizadas por una persona o computadora, pueden dañar a la organización objetivo" (Yuen, 2013). Planificar ataques de forma automatizada es una rama de la inteligencia artificial (IA) y puede ser un proceso que consume mucho tiempo para los usuarios humanos. Un planificador automatizado ayuda en la toma de decisiones y aumenta la precisión y exhaustividad de las evaluaciones (Yuen, Turnbull, & Hernandez, 2015). Por esta razón, se han desarrollado nuevos métodos de uso de la IA en la fase inicial de planificación de ataques para aumentar su efectividad. El Cyber Red Teaming automatizado se emplea para validar planes de ataque viables (Yuen, 2013).

Randhawa et al. (2018) presentaron una aplicación llamada Trogdor que utiliza múltiples planificadores de IA para realizar evaluaciones de vulnerabilidad automáticamente, generando gráficos de ataque que priorizan los recursos más críticos en los sistemas objetivo. Estos gráficos ayudan a identificar las rutas de ataque más beneficiosas para un actor malicioso, variando desde la más sencilla hasta la más ventajosa, utilizando métricas de cuantificación (Falco et al., 2018). La creación automatizada de árboles de ataque es considerablemente más rápida y no requiere un alto nivel de conocimiento en ciberseguridad, en comparación con la creación manual (Falco et al., 2018).

2.2. Phishing y spear phishing

El phishing es una técnica de ciberdelincuencia que consiste en engañar a las personas para que revelen información personal o confidencial, como contraseñas, números de tarjetas de crédito o detalles bancarios. Los atacantes suelen hacerlo mediante correos electrónicos, mensajes de texto o sitios web falsos que parecen ser legítimos y provienen de fuentes confiables, como bancos, servicios en línea o incluso contactos conocidos. El objetivo es engañar a las víctimas para que proporcionen información o hagan clic en enlaces maliciosos que instalan malware en sus dispositivos.

El spear phishing es una variante más dirigida y sofisticada del phishing. En lugar de enviar mensajes de forma masiva a un gran número de personas, el spear phishing se enfoca en un individuo o un grupo específico. Los atacantes personalizan sus mensajes basándose en información que han recopilado sobre la víctima, como su nombre, cargo, empresa o intereses. Esto hace que los correos electrónicos o mensajes parezcan aún más legítimos y aumenten las posibilidades de que la víctima caiga en el engaño. El spear phishing es particularmente peligroso porque suele dirigirse a personas con acceso a información valiosa o confidencial dentro de una organización.

El phishing y el spear phishing aumentados por IA son una forma de toma de decisiones automatizada facilitada por la elaboración previa de perfiles de objetivos.

Uno de los desafíos para que las campañas de phishing tengan éxito es superar las tecnologías automatizadas que detectan y bloquean correos de phishing y spam. La investigación realizada por Bahnsen el al en 2018 se enfocó en cómo se puede utilizar el aprendizaje automático de manera

maliciosa para evadir estos sistemas de detección basados en inteligencia artificial. En particular, desarrollaron un algoritmo llamado DeepPhish, que emplea una red neuronal recurrente (RNN) junto con una arquitectura de memoria a corto y largo plazo (LSTM). Este algoritmo fue capaz de generar URLs fraudulentas que podrían ser utilizadas en ataques de phishing, pero con la capacidad de evitar ser detectadas por sistemas automatizados. En un estudio posterior, Bahnsen y su equipo descubrieron que DeepPhish logró evadir la detección con mayor éxito en comparación con los métodos manuales tradicionales.

2.3. Generación de código de ataque

Un informe anterior de Löfvenberg, Sommestad y Bildsten (2019) se revisó la literatura sobre la generación automática de códigos de ataque, centrándose en la detección y explotación de vulnerabilidades. Identificaron 22 publicaciones que demostraban que las vulnerabilidades detectadas eran explotables y que los códigos de ataque generados permitían obtener privilegios administrativos en la computadora objetivo. Evaluaron 17 soluciones diferentes, concluyendo que las herramientas identificadas eran relativamente poco sofisticadas y se centraban en vulnerabilidades antiguas y conocidas. Aunque algunas soluciones estaban siendo comercializadas, los autores señalaron que se necesitaría mucho trabajo para crear una solución capaz de generar códigos de ataque contra software real. Además, destacaron dos intereses estratégicos en esta tecnología: la priorización defensiva de parches de vulnerabilidad y la generación masiva de códigos de ataque por parte de los atacantes (Zouave et al., 2020)

Además, algunos investigadores han comenzado a considerar las fallas de seguridad en el aprendizaje automático y profundo como una forma de vulnerabilidad explotable. Chakraborty et al. (2018) ofrece una taxonomía de ataques adversariales, incluyendo la generación de ejemplos adversariales y ataques a redes neuronales generativas (GAN). Con el uso creciente de IA potencialmente vulnerable, este campo de estudio surge como un área de investigación para la generación automatizada de exploits. Existe preocupación por los posibles daños físicos que estos ataques podrían causar en sistemas ciberfísicos, como autos inteligentes. San Agustin (2019) recomienda contramedidas como el entrenamiento adversarial y la implementación de directrices globales de endurecimiento de código. Por lo tanto, se llega a una conclusión similar a la de Löfvenberg, Sommestad y Bildsten: la detección de vulnerabilidades y la generación de exploits con apoyo de IA podría convertirse en una prioridad para defensores y atacantes.

3. **Reconocimiento interno y movimiento lateral:** Una vez dentro del sistema, el atacante recopila más información sobre la red, como roles, contraseñas y vulnerabilidades adicionales, lo que le permite moverse a otros dispositivos o áreas de la red, expandiendo su control y refinando sus objetivos

3.1. Mapeo de redes y sistemas

En la etapa de reconocimiento interno, se utilizan herramientas de mapeo y clasificación para identificar vulnerabilidades e información crucial. Randhawa et al. (2018) explican cómo su solución de IA, Trogdor, identifica conexiones lógicas y modela las reglas de firewall en redes objetivo. Un actor malicioso puede aprovechar estas descripciones para mapear las vulnerabilidades de un sistema. Otra técnica es el modelado de temas mediante bibliotecas de IA, lo que permite a un atacante clasificar y ordenar aplicaciones del sistema, identificando así las vulnerabilidades a las que puede dirigirse (Zouave et al., 2020)

3.2. Análisis del comportamiento de la red

Un sistema puede utilizar IA para detectar comportamientos anormales en la red, pero un actor malicioso puede adaptarse a la línea base de comportamiento establecida por el sistema (Szmit & Szmit, 2012). Con este conocimiento, el atacante puede moverse sin ser detectado. Un principio

fundamental en esta área de investigación es que, si un sistema de IA puede aprender a detectar secuencias relacionadas con malware o comportamientos maliciosos, otra IA puede aprender a superar estos sistemas defensivos (Truong et al., 2020).

A medida que el sistema reacciona a cambios en las condiciones de la red y detecta anomalías, los atacantes suelen utilizar el proceso sombra. Este proceso identifica la línea base del comportamiento normal y adapta el movimiento del malware. Esto permite que el malware pase desapercibido (Ma et al., 2012). Cuando un sistema de detección de intrusos tiene una línea base generada por aprendizaje profundo (DL), el malware apoyado por IA puede empezar a imitar el comportamiento de los usuarios humanos mediante la contextualización (Zouave et al., 2020)

3.3. Movimientos laterales inteligentes

El malware impulsado por IA podrá tomar decisiones basadas en la estructura del sistema infectado para moverse sin ser detectado. Además, como se ha observado en estudios sobre análisis de comportamiento en redes, el malware podría aprender del entorno infectado permaneciendo inactivo, observando las operaciones normales. Se espera que la capacidad del malware para moverse lateralmente a mayor velocidad facilite la infección de más dispositivos en menos tiempo, de manera más autónoma (Zouave et al., 2020).

4. **Mando, control y acciones sobre los objetivos:** Aquí, el atacante ya tiene control sobre el sistema y puede conectarse a un servidor externo. Desde esta posición, adapta sus acciones para cumplir con sus objetivos, como robar o manipular datos.

En este momento, el actor malicioso intenta establecer un enlace de comunicaciones entre él y el objetivo con el fin de ejercer influencia sobre el sistema informático comprometido y otros sistemas de su red interna (Zouave et al., 2020).

4.1. Generación de dominios

Los algoritmos de generación de dominios basados en binarios (DGA) son utilizados por actores maliciosos durante la fase de comando y control (C2) del ciclo de ataque cibernético para establecer comunicación y facilitar la filtración de datos. Estos algoritmos generan miles de dominios pseudorandomizados diariamente, de los cuales solo uno, preseleccionado por el atacante, logrará conectarse con el servidor C2. Estos DGA son difíciles de detener sin el uso de clasificadores modernos basados en aprendizaje automático, que examinan las consultas DNS exitosas y las clasifican para bloquear las que correspondan a dominios generados (Zouave et al., 2020).

4.2. Malware con autoaprendizaje

Dado que bloquear conexiones entrantes desconocidas en una red es una práctica común en el diseño de firewalls, las comunicaciones de comando y control (C2) suelen originarse desde los hosts infectados. Esto hace que parezca que el host está comunicándose con un par legítimo externo. Si se detecta el canal de comunicación C2, es posible bloquearlo mediante reglas de firewall, limitando así la capacidad del malware. Sin embargo, el malware apoyado por IA podría superar esta contramedida al actuar de forma autónoma, eliminando la necesidad de "comunicarse" con su origen. Por ejemplo, Chung, Kalbarczyk y Iyer (2019) demostraron en una simulación cómo un malware autoaprendido puede atacar indirectamente sistemas informáticos en una instalación de supercomputación al interferir con los sistemas ciberfísicos de automatización del edificio. En lugar de atacar directamente la infraestructura informática, el malware se enfocó en el sistema de control de enfriamiento, recolectando datos necesarios para ejecutar escenarios que interrumpieran su capacidad de enfriamiento. Utilizando k-means clustering para clasificar los datos de control lógico y distribución gaussiana para identificar los efectos del ataque, el malware implementó tres estrategias diferentes para interrumpir el sistema de manera autónoma. Este ejemplo demuestra que un

malware con la capacidad de ejecutar sus propios planes de ataque a través de un algoritmo auto-aprendido podría reducir la cantidad de conocimiento necesario por parte del atacante para manipular exitosamente el sistema objetivo.

4.3. C2 (comando y control) basado en enjambres para botnets

Las relaciones de control y comando (C2) en botnets suelen enfrentar problemas de supervivencia y escalabilidad. Diversos autores argumentan que la creciente sofisticación de las técnicas de detección y mitigación afecta la robustez de estas redes, mientras que el aumento de elementos y la heterogeneidad de la infraestructura de comunicación limitan su escalabilidad. Una posible solución propuesta es la inteligencia descentralizada basada en enjambres, inspirada en cómo las hormigas optimizan su comportamiento de búsqueda utilizando feromonas para la comunicación indirecta entre agentes. En este modelo, los bots se comunican dejando y detectando "trazas" en el entorno, lo que crea un canal de autoorganización (Zouave et al., 2020)

En la práctica, los nodos de botnet envían un "latido" periódico que informa a otros nodos de su presencia. Estos nodos luego actualizan la topología de la red utilizando un algoritmo DCMST, estableciendo rutas preferidas basadas en el principio de menor costo, similar a cómo las hormigas siguen el camino con la mayor concentración de feromonas. Este enfoque permite al bot master enviar comandos criptográficamente firmados a nodos seleccionados aleatoriamente, los cuales son descifrados y propagados por toda la red, mejorando así la supervivencia y escalabilidad de la botnet al adaptarse dinámicamente a las condiciones cambiantes de la red (Zouave et al., 2020).

4.4. Manipulación mediante procesamiento de lenguaje natural (NLP)

El procesamiento del lenguaje natural (NLP) ha mejorado recientemente gracias al uso de redes neuronales profundas y aprendizaje profundo, lo que permite un procesamiento más preciso y efectivo (Dheap, 2018.

5. **Exfiltración y saneamiento:** Finalmente, el atacante extrae la información deseada y se retira de la red, intentando borrar cualquier rastro de su actividad para evitar ser detectado.

Ahora mismo no hay demasiada bibliografía sobre el uso de la IA en este momento final del ataque, sin embargo, varios autores comentan el uso futuro previsto de la IA para estos fines.

5.1. Ofuscación de descubrimiento

En 2018, IBM Research desarrolló un malware llamado DeepLocker para demostrar cómo la tecnología de malware actual puede combinarse con técnicas de aprendizaje profundo, creando una nueva generación de malware. DeepLocker destaca por su capacidad de ocultarse a simple vista, incrustándose en aplicaciones como software de videoconferencias mediante un payload cifrado que es indetectable para la mayoría del software antivirus. Una máquina infectada con un payload cifrado latente no puede ser contrarrestada hasta que el ataque se detecta (Zouave et al., 2020).

Los ataques impulsados por IA utilizan algoritmos de ofuscación sofisticados y cambian frecuentemente sus características identificables, lo que les permite alcanzar un nivel de adaptabilidad que los hace prácticamente indetectables para las herramientas antivirus basadas en comportamiento y firmas (Babuta, Oswald & Janjeva, 2020).

5.3. Exfiltración "baja y lenta"

Diversos autores sugieren que la técnica de exfiltración de datos conocida como "low-and-slow" sería mucho más difícil de detectar con el uso de IA. Esta técnica transfiere pequeñas cantidades de datos durante un período prolongado, mientras que el objetivo permanece en gran medida inconsciente del proceso. Un malware con IA, consciente del contexto, podría evaluar el patrón de

uso del ancho de banda de la máquina objetivo y acelerar la filtración de datos al aprovechar actividades de alto rendimiento como las videoconferencias. Además, un malware consciente del contexto sería capaz de integrarse en el entorno en el que se encuentra, lo que tendría repercusiones tanto para su detección como para la exfiltración de datos (Zouave et al., 2020).

Hasta este punto se ha expuesto cómo podría usarse la IA en cada etapa del ciberataque, pero lo más común es que sobre todo se utilice en las fases de acceso y penetración (Malatji & Tolah, 2024).

Teniendo este funcionamiento en cuenta, podemos afirmar que los ataques de ciberseguridad cada vez son más sofisticados y diversos, y los especialistas de esta área se están volviendo incapaces de hacer frente a lo que se ha convertido en el factor de amenaza más importante que haya existido nunca (Guembe et al., 2022).

Ha surgido una nueva generación de ciberdelincuentes, caracterizada por su sutileza y discreción, que influirá en el futuro de la ciberseguridad. Las amenazas cibernéticas futuras serán más inteligentes y podrán operar de manera autónoma gracias a la inteligencia artificial. Los métodos de ciberataque del futuro tendrán la capacidad de entender su entorno y tomar decisiones informadas basadas en el ambiente del objetivo. La capacidad de la IA para aprender y adaptarse dará paso a una nueva era de ataques escalables, personalizados y con comportamientos similares a los humanos (Thanh and Zelinka 2019).

Ya existen las herramientas y componentes necesarios para lanzar un ciberataque ofensivo impulsado por inteligencia artificial (Dixon and Eagan 2019). Y como se ha observado durante este ensayo, un ciberataque impulsado por inteligencia artificial aprovechará una amplia gama de recursos psicológicos, cibernéticos y computacionales, superando lo que un humano podría coordinar. Esto resultará en un ataque más rápido, impredecible y sofisticado, que puede superar la capacidad de respuesta incluso del equipo de ciberseguridad más fuerte (Bocetta 2020).

Los ciberataques impulsados por IA emergentes podrían tener consecuencias potencialmente mortales y altamente destructivas. Al socavar la confidencialidad e integridad de los datos, estos ataques sofisticados y sigilosos erosionarán la confianza en las organizaciones y podrían llevar a fallos sistémicos. Por ejemplo, un médico podría hacer un diagnóstico basado en datos médicos manipulados, o una plataforma petrolera podría perforar en el lugar equivocado debido a información geológica incorrecta. (Dixon y Eagan, 2019).

3.2.3. IA en la Ciberseguridad

Es evidente que estos ataques solo empeorarán, volviéndose casi imposibles de detectar con herramientas de ciberseguridad tradicionales, enfrentando la eficiencia de la máquina contra el esfuerzo humano. Sin embargo, el uso de IA para combatir IA permitirá a investigadores, organizaciones y gobiernos desarrollar contramedidas más avanzadas. La mejor preparación es fortalecer las infraestructuras de defensa cibernética, minimizando falsos positivos y negativos (Guembe et al., 2022).

Y es que en realidad la inteligencia artificial es un arma de doble filo cuando hablamos de ciberseguridad, ya que no solamente puede ayudar a los ciberdelincuentes, sino que las organizaciones pueden utilizarla para protegerse de estos ataques (Malatji & Tolah, 2024).

Todo lo que hemos visto subraya la necesidad de defensas robustas que puedan adaptarse a la evolución de las técnicas de ataque (Safitra et al., 2023).

La realidad de que la tecnología pueda tener un doble uso no es un problema completamente nuevo en el ámbito del cibercrimen o la ciberseguridad. No obstante, la manera en que la inteligencia artificial puede ser utilizada para fines malintencionados representa una vulnerabilidad emergente. Es

esencial evaluar constantemente el panorama de amenazas para diseñar y ajustar mecanismos de gobernanza, implementar medidas proactivas y fortalecer la ciberresiliencia (Blauth et al., 2022).

Por esto mismo, la carrera armamentística de la IA está evolucionando para crear nuevas y mejores herramientas y métodos para detener la IA en el crimen y estar un paso por delante de los cibercriminales, ya que la inteligencia artificial también se está utilizando como parte de una iniciativa de investigación más amplia sobre el cibercrimen y sus autores. Los datos sobre delitos cibernéticos se recopilan, analizan y utilizan para crear sofisticadas escenas del crimen virtuales que pueden predecir los delitos antes de que ocurran (Chakraborty et al, 2023).

El aprendizaje automático, en particular las redes neuronales, los árboles de decisión y las máquinas de vectores de soporte, sustentan el núcleo de la ciberseguridad impulsada por la IA. Estas técnicas permiten el análisis automatizado de grandes conjuntos de datos, lo que ayuda a identificar amenazas y evaluar vulnerabilidades. Además, el procesamiento del lenguaje natural (PLN) ayuda a analizar datos textuales para identificar intenciones maliciosas y vulnerabilidades potenciales (Vegesna, 2023).

También son dignas de mención las técnicas basadas en firmas. La IA detecta ciberataques y malware al comparar los códigos de ataque con una base de datos de firmas conocidas, lo que permite a los equipos de ciberseguridad detener los ataques rápidamente. Aunque esta técnica ha sido muy eficaz, es limitada contra ataques nuevos que no están en la base de datos. Los atacantes pueden evadir la detección cambiando sus patrones. Pero a pesar de estas limitaciones, la técnica ha prevenido muchos ataques a lo largo del tiempo (Ansari et al, 2022).

El enfoque de "machine learning" (ML) también ha tenido un gran impacto en la ciberseguridad al mejorar la detección de ataques y errores que los humanos podrían pasar por alto. La tecnología de IA analiza grandes volúmenes de datos, como registros y paquetes de red, permitiendo una detección rápida y precisa de amenazas. ML también facilita la clasificación y agrupación de registros para identificar anomalías y malware, lo que es difícil de lograr solo con análisis humanos. La combinación de sistemas de IA y analistas humanos ha demostrado ser eficaz para prevenir ataques y proteger la información (Ansari et al, 2022).

Además, tenemos que mencionar el papel del "Deep Learning" (DL). El DL imita el proceso de las neuronas humanas y construye la arquitectura neuronal con interconexiones complejas. Hoy en día, la DL es un tema de investigación candente en el mundo académico y se ha utilizado ampliamente en diversos escenarios industriales, entre ellos, la ciberseguridad.

Las aplicaciones de DL destacan en sistemas autónomos debido a sus ventajas en optimización, discriminación y predicción. Los principales campos de aplicación incluyen:

- **Reconocimiento de imágenes y videos**: Es el área más importante de investigación en DL, utilizando redes neuronales convolucionales (CNN) para reducir el tamaño de las imágenes y mejorar la detección de objetos en tiempo real.

- **Análisis de texto y procesamiento de lenguaje natural**: Crucial para la traducción en tiempo real y la interacción hombre-máquina, con herramientas como "Stanford CoreNLP" que facilitan el análisis de lenguaje natural.

- **Finanzas y análisis de mercado**: El DL se utiliza para mejorar las predicciones del mercado, como lo demuestra un algoritmo de pronóstico financiero basado en CNN que reduce significativamente el error en predicciones del mercado de divisas (Li, 2018).

Una técnica que también se está estudiando mucho son los sistemas de detección de intrusos (IDS). Los IDS sirven como salvaguardias vitales de la seguridad convencional, protegiendo las infraestructuras de red de los ciberataques mediante la identificación de accesos no autorizados y acciones dañinas. Los IDS pueden clasificarse en sistemas de detección de intrusiones en la red (NIDS) y de prevención (IPS), que analizan el tráfico de la red en busca de indicios de actividad maliciosa mediante la detección de firmas y anomalías estadísticas, así como el análisis heurístico del comportamiento.

El uso de modelos de IA como ChatGPT entre otros, en IDS puede mejorar la detección de ataques sofisticados, reducir falsos positivos y permitir respuestas más efectivas al aprovechar el procesamiento del lenguaje natural y el aprendizaje automático para analizar patrones complejos en el tráfico de red (Markevych & Dawson, 2023).

La gestión de vulnerabilidades es un aspecto clave en ciberseguridad que se ha beneficiado del uso de la inteligencia artificial (IA). La IA ayuda a gestionar estas vulnerabilidades de manera más eficiente que el personal humano, dificultando el acceso de los hackers a los sistemas. Además, la IA permite detectar anomalías en cuentas de usuario en tiempo real, fortaleciendo la seguridad de los servidores y la información almacenada (Ansari et al, 2022).

Por último, también hay que mencionar que la seguridad de los centros de datos es crucial en ciberseguridad, y la IA ha mejorado significativamente su protección y eficiencia. Al automatizar procesos como el consumo de energía, el uso de ancho de banda y el control de temperaturas, la IA minimiza errores humanos y reduce costos de mantenimiento de hardware. Además, protege estos centros, que contienen información vital, de factores ambientales. Cada vez más empresas están adoptando IA en sus centros de datos para aumentar la seguridad y la eficiencia, aunque también existen limitaciones en su uso en ciberseguridad (Ansari et al, 2022).

El aspecto de la seguridad es un sueño por el que luchan investigadores y empresas informáticas. Y por lo que se ha ido mencionando más el hecho de que los sistemas pueden aprender de los diferentes aspectos que han hecho que la tecnología sea sumamente relevante en ciberseguridad, la IA ha sido considerada como una tecnología de rescate en ciberseguridad (Ansari et al, 2022).

Esta idea la sostienen diversos autores. Por ejemplo, en 2016, la Agencia de Proyectos de Investigación Avanzados de Defensa (DARPA) organizó el Cyber Grand Challenge, una competencia en la que, siete equipos emplearon IA para detectar y corregir automáticamente fallos internos mientras explotaban las debilidades de los sistemas de sus oponentes. Desde entonces, científicos del MIT han desarrollado una IA capaz de analizar grandes cantidades de datos para identificar actividades sospechosas y alertar a los analistas humanos para un análisis más profundo. El objetivo es entrenar a la IA para reconocer un sistema estable y detectar comportamientos anómalos. Sin embargo, los avances defensivos también impulsan mejoras ofensivas; investigadores están empleando IA para engañar a otras IA, superando las capacidades de detección cibernética y creando malware más sofisticado. Esto indica una inminente carrera armamentista de IA entre expertos en ciberseguridad, estados y actores no estatales (Wilner, 2018).

4. Conclusiones

Este ensayo ha puesto de manifiesto el impacto significativo que la inteligencia artificial (IA) está teniendo en el ámbito de la Guerra cognitiva y de la ciberseguridad, tanto en sus aplicaciones defensivas como ofensivas.

La inteligencia artificial (IA) está jugando un papel cada vez más relevante en el ámbito de la guerra cognitiva, un tipo de conflicto que busca influir y manipular las percepciones, emociones y comportamientos de las personas. La guerra cognitiva utiliza la información como arma para alterar la toma de decisiones y fomentar la desconfianza o el caos en las sociedades objetivo. La IA facilita esta estrategia mediante técnicas avanzadas de análisis de datos, generación de contenido personalizado y simulación de interacciones humanas.

Las capacidades de procesamiento del lenguaje natural (PLN) permiten a la IA analizar grandes volúmenes de texto y discurso en tiempo real, identificando tendencias, emociones y patrones de comportamiento en las redes sociales y otros entornos digitales. A partir de esta información, los sistemas de IA pueden generar campañas de desinformación altamente personalizadas y dirigidas, diseñadas para influir en audiencias específicas, exacerbar divisiones sociales y políticas, y manipular la opinión pública a gran escala. Además, las redes generativas antagónicas (GAN) permiten crear imágenes, videos y voces falsificadas (deepfakes), que pueden ser utilizados para difundir información falsa o distorsionar la realidad de manera convincente.

La IA también se utiliza para optimizar el impacto de las campañas de influencia mediante la segmentación precisa de los objetivos, el ajuste de los mensajes en función de las respuestas recibidas y la adaptación continua de las tácticas en función de los datos en tiempo real. Esto aumenta la eficacia de las operaciones de guerra cognitiva al permitir una manipulación más refinada y persistente del entorno informativo.

En conclusión, el uso de la IA en la guerra cognitiva representa una amenaza significativa para la estabilidad social y la seguridad nacional, ya que facilita la manipulación de las percepciones y decisiones a una escala sin precedentes. La capacidad de la IA para aprender y adaptarse a las reacciones humanas plantea desafíos éticos y estratégicos críticos, que requieren la atención de expertos en ciberseguridad, legisladores y organismos internacionales para mitigar su impacto negativo y proteger la integridad de los procesos democráticos y la cohesión social.

Respecto al ámbito ciber, se ha evidenciado que las ciberamenazas impulsadas por IA representan un desafío cada vez mayor debido a su capacidad de aprendizaje, adaptación y automatización, lo que les permite superar las capacidades de detección de las herramientas de seguridad tradicionales. La creciente sofisticación de estos ataques, como el uso de técnicas de evasión avanzadas y la capacidad de operar de manera autónoma, subraya la necesidad urgente de desarrollar contramedidas más avanzadas.

Además, la IA se está utilizando para fines maliciosos de manera cada vez más creativa, integrándose en técnicas como el phishing automatizado, la generación de códigos de ataque y la manipulación a través del procesamiento de lenguaje natural. Estas amenazas, combinadas con estrategias tradicionales de ciberataque, pueden causar daños significativos sin ser detectadas. La capacidad de la IA para aprender del entorno y adaptarse dinámicamente a las condiciones cambiantes sugiere que las amenazas seguirán evolucionando, aumentando el riesgo de ataques devastadores en infraestructuras críticas y sistemas sociales.

Por tanto, la implementación de la IA en ciberseguridad debe ser considerada como una herramienta doblemente útil, capaz de fortalecer las defensas mediante el análisis automatizado de grandes volúmenes de datos, la detección de patrones anómalos y la respuesta efectiva a amenazas emergentes. Sin embargo, es crucial seguir investigando y desarrollando métodos para mitigar los riesgos asociados con el uso malicioso de la IA, implementando sistemas de defensa más sofisticados y robustos que puedan adaptarse rápidamente a la evolución de las técnicas de ataque. La cooperación internacional y el establecimiento de marcos regulatorios claros serán esenciales para gestionar eficazmente estas amenazas en el futuro.

5. Referencias bibliográficas

Almousa, M., & Anwar, M. (2019, August). Detecting exploit websites using browser-based predictive analytics. In 2019 17th International Conference on Privacy, Security and Trust (PST) (pp. 1-3). IEEE.

Ansari, M. F., Dash, B., Sharma, P., & Yathiraju, N. (2022). The impact and limitations of artificial intelligence in cybersecurity: a literature review. International Journal of Advanced Research in Computer and Communication Engineering.

Aslama-Horowitz, M. (2019). Disinformation as warfare in the digital age: dimensions, dilemmas, and solutions dimensions, dilemmas, and solutions. Journal of Vincentian Social Action, 4(2). https://scholar.stjohns.edu/jovsa/vol4/iss2/5

Babuta, A., M. Oswald, and A. Janjeva. 2020. Artificial Intelligence and UK National Security Policy Considerations. Royal United Services Institute Occasional Paper.

Bahnsen, A. C., Ivan Torroledo, Luis David Camacho and Sergio Villegas. (2018). DeepPhish: Simulating Malicious AI. 2018 APWG Symposium on Electronic Crime Research (eCrime).

Ballesteros-Aguayo, L., & del Olmo, F. J. R. (2024). Vídeos falsos y desinformación ante la IA: el deepfake como vehículo de la posverdad. Revista de Ciencias de la Comunicación e Información, 29, 1-14.

Beltrán López, P., Nespoli, P., & Gil Pérez, M. (2024). Cyber Deception powered by Artificial Intelligence: Overview, Gaps, and Opportunities. Jornadas Nacionales de Investigación en Ciberseguridad (JNIC) (9ª. 2024. Sevilla) (2024), pp. 254-261.

Bharadiya, J. P. (2023). AI-driven security: how machine learning will shape the future of cybersecurity and web 3.0. American Journal of Neural Networks and Applications, 9(1), 1-7.

Blauth T. F., Gstrein O. J. and Zwitter A., "Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI," in IEEE Access, vol. 10, pp. 77110-77122, 2022, doi: 10.1109/ACCESS.2022.3191790.

Bocetta, S. 2020. Has an AI cyberattack happened yet? https://www.infoq.com/articles/ai- cyberattacks/

Boyer, P. (2018). Minds make societies: How cognition explains the world humans create. Yale University Press

Brynjolfsson E., and McAffee, A. (2017). The Business of Artificial Intelligence. Available: https://starlab-alliance.com/wp-content/uploads/2017/09/TheBusiness-of-Artificial-Intelligence.pdf. Last accessed 26/11/2019.

Brooks, R 2016, How everything became War and the Military Became Everything, Simon and Schuster, New York, NY, US.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson,

J., Yampolskiy, R., & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv.org, vol. cs.AI.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., Henke, N., & Trench, M. (2017). Artificial intelligence: The next digital frontier? [discussion paper]. McKinsey Global Institute (mgi).

Buil-Gil, D., Miró-Llinares, F., Moneva, A., Kemp, S., & Díaz-Castaño, N. (2021). Cybercrime and shifts in opportunities during COVID-19: a preliminary analysis in the UK. European Societies, 23(sup1), S47-S59.

Caldwell, M., Andrews, J.T.A., Tanay, T., Griffin, L.D.: AI-enabled future crime. Crime Sci. Sci. 9(1), 14 (2020). https://doi.org/10.1186/s40163-020-00123-8

Chakraborty, A., Biswas, A., Khan, AK (2023). Inteligencia artificial para la ciberseguridad: amenazas, ataques y mitigación. En: Biswas, A., Semwal, VB, Singh, D. (eds) Inteligencia artificial para cuestiones sociales. Biblioteca de referencia de sistemas inteligentes, vol. 231. Springer, Cham. https://doi.org/10.1007/978-3-031-12419-8_1

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D. (2018). Adversarial Attacks and Defences: A Survey. ArXiv

Checa Rubio, Marcos y Sánchez Margolles, Sofía (diciembre 2023) Una evaluación de la guerra cognitiva de Rusia. Drafts of Economic Intelligence (ISSN 2659-9791), Vol. 5 nº 4; pp. 41-47. https://escuela-inteligencia-economica-uam.com/draft-volumen-5-2022-2023/

Chelioudakis, E. (2017). Deceptive AI machines on the battlefield: Do they challenge the rules of the Law of Armed Conflict on military deception? Available at SSRN 3158711.

Chen J, Liu Y, Zou M (2016) Home location profling for users in social media. Inf Manag 53(1):135–143. https://doi.org/10.1016/j.im.2015.09.008

Chung, K., Kalbarczyk, Z. T., & Iyer, R. K. (2019). Availability attacks on computing systems through alteration of environmental control: smart malware approach. Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems, 1–12

Claverie, B., & Du Cluzel, F. (2022). The Cognitive Warfare Concept. Innovation Hub Sponsored by NATO Allied Command Transformation, 2.

Dixon, W., and N. Eagan. 2019. AI will power a new set of tools and threats for the cybercriminals of the future. https://www.weforum.org/agenda/2019/06/ai-is-powering- a-new-generation-of-cyberattack-its-also-our-best-defence/

Dheap, V. (2017). AI in Cybersecurity: A Balancing Force or a Disruptor? Available: https://www.rsaconference.com/industry-topics/presentation/ai-in-cybersecurity-abalancing-force-or-a-disruptor .

Esparza SG, O'Mahony MP, Smyth B (2013) CatStream: categorising tweets for user profling and stream fltering

Falco, G., Viswanathan, A., Caldera, C., and Shrobe, H. (2018). A Master Attack Methodology for an AI-Based Automated Attack Planner for Smart Cities. IEEE Access 6, pp. 48360-48373

Filippoupolitis, A., Loukas, G., Kapetanakis, G. (2014). Towards real-time profiling of human attackers and bot detection. CFET 2014 7th International Conference on Cybercrime Forensics Education & Training

Flores Vivar, J. M. (2019). Inteligencia artificial y periodismo: diluyendo el impacto de la desinformación y las noticias falsas a través de los bots.

Flores-Vivar, J. M., Gómez-de-Ágreda, Á., & Gómez-López, J. (2023). Taxonomía de la inteligencia artificial en el entorno cognitivo de los conflictos. Anuario Electrónico de Estudios en Comunicación Social "Disertaciones", 16(2). https://doi.org/10.12804/revistas.urosario.edu.co/disertaciones/a.12804

Franganillo, J. (2023). La inteligencia artificial generativa y su impacto en la creación de contenidos mediáticos. methaodos.revista de ciencias sociales, 11(2), m231102a10. http://dx.doi.org/10.17502/mrcs.v11i2.710

García, A. A. (2024). Inteligencia Artificial y Desinformación: Papel en los Conflictos del Siglo XXI. Revista Seguridad y Poder Terrestre, 3(3).

Gartner. (2018, 24 de julio). Hype CYCLe for artificial intelligence, 2018. https://www.gartner.com/en/documents/3883863

Gartner (2017) "Top Strategic Predictions for 2018 and Beyond" (en línea) Https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2018-and-beyond/

Gilbert, J., Hamid, S., Hashem, I.A.T. et al. The rise of user profiling in social media: review, challenges and future direction. Soc. Netw. Anal. Min. 13, 137 (2023). https://doi.org/10.1007/s13278-023-01146-0

Gosálvez Frías, Elena Mª (octubre 2023) Guerra cognitiva en la era de la Inteligencia Artificial: Impacto en la construcción de la realidad mediante las redes sociales. Drafts of Economic Intelligence (ISSN 2659-9791), Vol. 5 nº 2; pp. 15-30. https://escuela-inteligencia-economica-uam.com/draft-volumen-5-2022-2023/

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The emerging threat of ai-driven cyber attacks: A review. Applied Artificial Intelligence, 36(1), 2037254.

Hutchinson, W. (2022). Strategic Cognition War. Journal of Information Warfare, 21(3), 74-83. ArmisteadTEC LLC. https://www.jstor.org/stable/10.2307/27199984

Jonsson, O 2019, The Russian Understanding of War: Blurring the Lines Between War and Peace, Georgetown University Press, Washington D.C., US.

Juric, R., Moholth McClenaghan, K., Moholth, O. C. (2019). Detecting Cyber Security Vulnerabilities through Reactive Programming. Proceedings of the 52nd Hawaii International Conference on System Science 2019. 7204-7213

Kaloudi, N., & Li, J. (2020). The ai-based cyber threat landscape: A survey. ACM Computing Surveys (CSUR), 53(1), 1-34.

Kayaalp F, Erdogmus P (2020) Benchmarking the clustering performances of evolutionary algorithms: a case study on varying data size. IRBM. https://doi.org/10.1016/j.irbm.2020.06.002

Krishnan GS, Kamath SS (2017) Dynamic and temporal user profling for personalized recommenders using heterogeneous data sources. Paper presented at 2017 8th international conference on computing, communication and networking technologies (ICCCNT). Chennai, pp 8–15

Le Guyader, H., & Cole, A. Cognitive Domain: A Sixth Domain of Operations. https://www.innovationhubact.org/sites/default/files/2021-01/NATO%27s%206th%20domain%20of%20operations.pdf

Li, J. H. (2018). Cyber security meets artificial intelligence: a survey. Frontiers of Information Technology & Electronic Engineering, 19(12), 1462-1474.

López, P. C., Mila, A., & Ribeiro, V. (2023). La desinformación en las democracias de América Latina y de la península ibérica: De las redes sociales a la inteligencia artificial (2015-2022).

Ma, W., Duan, P., Liu, S., Gu, G., Liu, J. (2012). Shadow attacks: automatically evading system-call-behavior based malware detection. Journal in Computer Virology 8(1-2), 1–13. https://doi.org/10.1007/s11416-011-0157-

Malatji, M., & Tolah, A. (2024). Artificial intelligence (AI) cybersecurity dimensions: A comprehensive framework for understanding adversarial and offensive AI. AI and Ethics. https://doi.org/10.1007/s43681-024-00427-4

Masters, P., Smith, W., Sonenberg, L., & Kirley, M. (2021). Characterising deception in AI: A survey. In Deceptive AI: First International Workshop, DeceptECAI 2020, Santiago de Compostela, Spain, August 30, 2020, and Second International Workshop, DeceptAI 2021, Montreal, Canada, August 19, 2021, Proceedings 1 (pp. 3-16). Springer International Publishing.

Markevych, M., & Dawson, M. (2023, July). A review of enhancing intrusion detection systems for cybersecurity using artificial intelligence (ai). In International conference Knowledge-based Organization (Vol. 29, No. 3, pp. 30-37).

Miller, S. (2023). Cognitive warfare: an ethical analysis. Ethics and Information Technology, 25(3), 46.

Mulwad, V., Li, W., Joshi, A., Finin, T., and Viswanathan, K. (2011). Extracting Information about Security Vulnerabilities from Web Text. 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. 257-260

Natale, S. (2023). AI, human-machine communication and deception. The SAGE handbook of Human-Machine Communication, 401-408.

Neuroscience News. (2024, August 22). AI can use deception to manipulate people, study finds. https://neurosciencenews.com/ai-deception-manipulation-26082/

Organización de las Naciones Unidas (ONU), 3 de marzo de 2017, "Contrarrestar la desinformación" ONU [Edición digital]. Recuperado el 16 de abril de 2024, de https://news.un.org/es/story/2017/03/1374761 .

Organización de las Naciones Unidas (ONU). (s.f.). "Contrarrestar la desinformación", https://www.un.org/es/countering-disinformation .

Pauwels, E., & Deton, S. W. (2020). Hybrid emerging threats and information warfare: the story of the Cyber-AI deception machine. 21st Century Prometheus: Managing CBRN Safety and Security Affected by Cutting-Edge Technologies, 107-124.

Perera, I., Hwang, J., Bayas, K., Dorr, B., Wilks,Y. (2018). Cyberattack Prediction Through Public Text Analysis and Mini-Theories. 2018 IEEE International Conference on Big Data (Big Data). 3001-3010

Pulido, G. (2021). Guerra multidominio y mosaico. Catarata. https://www.catarata.org/libro/guerramultidominio-y-mosaico_133474/

Randhawa, S., Turnbull, B., Yuen, J, Dean, J. (2018) Mission-centric Automated Cyber Red Teaming. ARES 2018 1-11.

Safitra, M.F., Lubis, M., Fakhrurroja, H.: Counterattacking cyber threats: a framework for the future of cybersecurity. Sustainability (2023). https://doi.org/10.3390/su151813369

San Agustin, C. E. (2019). Mitigating Deep Learning Vulnerabilities from Adversarial Examples Attack in the Cybersecurity Domain. CES Agustin. arXiv preprint

Sarkadi, Ş. "Deceptive AI and Society", en IEEE Technology and Society Magazine, vol. 42, no. 4, pp. 77-86, dic. 2023, doi: 10.1109/MTS.2023.3340232 .

Sánchez Margolles, Sofía (junio 2022) Personalidad en Ingeniería Social: ¿Qué rasgos son más vulnerables? Drafts of Economic Intelligence (ISSN 2659-9791), Vol. 4 nº 2; pp. 15 – 30. https://escuela-inteligencia-economica-uam.com/draft-volumen-4-2021-2022/

Sanoubari, E., Seo, S.H., Garcha, D., Young, J.E., Loureiro-Rodr´ıguez, V.: Good robot design or machiavellian? In: Proceedings of the International Conference on Human-Robot Interaction (HRI), pp. 382–391 (2019)

Schneider, J., Meske, C., & Vlachos, M. (2020). Deceptive AI explanations: Creation and detection. arXiv preprint arXiv:2001.07641.

Shim, J., Arkin, R.C.: Biologically-inspired deceptive behavior for a robot. From Anim. Animats 12, 401–411 (2012)

Stieglitz, S. et al. (2014) Social media analytics Bus. Inf. Syst. Eng., 6 (2), pp. 89-96

Sufi, F. (2023). Novel application of open-source cyber intelligence. Electronics, 12(17), 3610.

Sun Tzu (2004). *El arte de la guerra*. Editorial Fundamentos. ISBN 9788424501266. Consultado el 14 de 08 de 2024

Szmit, M., Szmit, A. (2012). Usage of Modified Holt-Winters Method in the Anomaly Detection of Network Traffic: Case Studies. Journal of Computer Networks and Communications 2012

Tarsney, C. (2024). Deception and Manipulation in Generative AI. arXiv preprint arXiv:2401.11335.

Thanh, C., and I. Zelinka. 2019. A survey on artificial intelligence in malware as next-generation threats. MENDEL 25 (2):27–34

Thornton, R. and Miron, M., "Towards the "third revolution in military affairs: The Russian military's use of AI-enabled cyber warfare", RUSI J., vol. 165, no. 3, pp. 12-21, Apr. 2020.

Truong, C. T. Zelinka, I., Plucar, J., Candik, M., Sulz, V. (2020). " Artificial Intelligence and Cybersecurity: Past, Presence, and Future" in. (2020). Artificial Intelligence and Evolutionary Computations in Engineering Systems by Lakshmi, C., Das, S., Panigrahi, B. 10.1007/978-981-15-0199-9.

Univisión, "Para Putin, la nación que domine a la inteligencia artificial dominará al mundo" (Univision.com, 15 de abril de 2024), https://www.univision.com/explora/para-putin-la-nacion-quedomine-a-la-inteligencia-artificial-dominara-al-mundo

Vallina, D. C. (2022). Un efecto «mágico»: El dominio cognitivo. Ejército: de tierra español, (976), 20-25.

Vegesna, V. V. (2023). Enhancing Cybersecurity Through AI-Powered Solutions: A Comprehensive Research Analysis. International Meridian Journal, 5(5), 1-8.

Velasco, C. Cibercrimen e inteligencia artificial. Una visión general del trabajo de las organizaciones internacionales en materia de justicia penal y los instrumentos internacionales aplicables. ERA Forum 23, 109–126 (2022). https://doi.org/10.1007/s12027-022-00702-z

Voelkel, J. G., & Willer, R. (2023). Artificial intelligence can persuade humans on political issues.

Wilner, A. S. (2018). Cybersecurity and its discontents: Artificial intelligence, the Internet of Things, and digital misinformation. International Journal, 73(2), 308–316.

Wilson III, E. J. (2008). Hard power, soft power, smart power. The annals of the American academy of Political and Social Science, 616(1), 110-124.

Woolley, S., & Joseff, K. (2020). Demand for deceit: How the way we think drives disinformation. https://www.ned.org/wp-content/uploads/2020/01/Demand-for-Deceit.pdf

World Economic Forum. (2022). Global cybersecurity outlook 2022. https://www.weforum.org/publications/global-cybersecurity-outlook-2022/

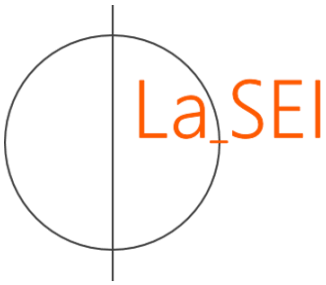
Young LE, Soliz S, Xu JJ, Young SD (2020) A review of social media analytic tools and their applications to evaluate activity and engagement in online sexual health interventions. Prev Med Rep 19:101–158. https://doi.org/10.1016/j.pmedr.2020.101158

Yuen, J. (2013). Automated Cyber Red Teaming. Cyber and Electronic Warfare Division in Defence Science and Technology Organisation. DSTO-TN-1420

Yuen, J., Turnbull, B., and Hernandez, J. (2015). Visual analytics for cyber red teaming. 2015 IEEE Symposium on Visualization for Cyber Security (VizSec), Chicago, IL, pp. 1-8

Zegart, A B 2022, Spies, Lies, and Algorithms: The History and Future of American Intelligence, Princeton University Press, Princeton, NJ, US.

Zouave, E., Bruce, M., Colde, K., Jaitner, M., Rodhe, I., & Gustafsson, T. (2020). Artificially intelligent cyberattacks. Stockholm: Totalförsvarets forskningsinstitut FOI [Online] Available: https://whttps://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf [Accessed: Sep. 28, 2022].



Reports de Inteligencia Económica y Relaciones internacionales

[ISSN 2660-7352]

PUBLICACIONES DE LA ESCUELA DE INTELIGENCIA ECONÓMICA DE LA UAM


La_SEI
School of Economic Intelligence & International Relations
Escuela de Inteligencia Económica y RRII
Universidad Autónoma de Madrid