FACULTAD DE ECONÓMICAS Y EMPRESARIALES
ICADE

EL IMPACTO DE LA INTELIGENCIA ARTIFICIAL EN LA DIFUSIÓN DE *FAKE NEWS* EN LAS REDES SOCIALES

Autora: Eugenia Fenellós Valero de Palma
Tutor: Antonio Tena Blázquez

MADRID | Junio de 2025

TEMA PRINCIPAL DEL TFG

EL IMPACTO DE LA INTELIGENCIA ARTIFICIAL EN LA DIFUSIÓN DE *FAKE NEWS* EN LAS REDES SOCIALES

RESUMEN

Este Trabajo de Fin de Grado (TFG) analiza el papel de la inteligencia artificial (IA) en la generación y difusión de noticias falsas (*fake news)* en redes sociales. A través de una revisión teórica y un estudio de campo aplicado a estudiantes universitarios, se examinan los mecanismos de la IA que contribuyen a crear entornos informativos cerrados. Se identifican varios tipos de desinformación y se analizan sus motivaciones desde perspectivas políticas, económicas, sociales y psicológicas. Por último, los resultados del estudio empírico revelan que la mayoría de los encuestados no son capaces de distinguir entre noticias reales y falsas, lo cual pone de manifiesto la urgencia de incluir formación en alfabetización mediática e implementar medidas de control algorítmico, verificación automatizada y regulación efectiva. El trabajo concluye con una reflexión crítica sobre los desafíos éticos y democráticos del uso masivo de IA en entornos informativos.

PALABRAS CLAVE DEL TFG

Inteligencia artificial, Internet de las cosas, *fake news*, *Big Data*, *deep learning y machine learning.*

ÍNDICE

1. INTRODUCCIÓN AL TFG

La expansión de la IA ha revolucionado la forma en la que producimos, compartimos y consumimos información. En los últimos años, la presencia de tecnologías basadas en IA en redes sociales ha crecido exponencialmente, modificando profundamente las dinámicas del entorno digital. Este TFG se propone analizar el impacto de la IA en la difusión de *fake news* en plataformas sociales, una cuestión clave tanto para la libertad de información como para la salud de las democracias contemporáneas.

El objetivo principal del trabajo es explorar cómo determinadas aplicaciones de la IA, como los algoritmos de recomendación, los *bots*, los *deepfakes* y los *influencers* virtuales, **contribuyen a la propagación de desinformación en redes sociales** como TikTok, Instagram o X (anteriormente Twitter). Se busca comprender cómo estos mecanismos generan entornos informativos cerrados, conocidos como burbujas de filtro, que refuerzan sesgos cognitivos y debilitan la capacidad crítica de los usuarios.

La **metodología** utilizada ha sido **mixta**. En primer lugar, se ha realizado un análisis teórico-conceptual a través de revisión bibliográfica académica y documental, donde se estudian las tipologías de desinformación, las motivaciones de su difusión, los efectos de los algoritmos y las estrategias regulatorias. En segundo lugar, se ha desarrollado un estudio de campo mediante encuesta estructurada aplicada a 200 estudiantes universitarios españoles, para observar empíricamente su capacidad de distinguir entre noticias reales y falsas.

Este trabajo se estructura en siete capítulos. Tras la introducción y contextualización teórica de la Cuarta Revolución Industrial y el auge de la IA, se exploran las dinámicas de desinformación, el papel de la IA en su propagación, las consecuencias sociales y éticas derivadas de ello, y las estrategias que podrían mitigar este fenómeno. Finalmente, se presenta un estudio de campo que aporta datos concretos sobre la alfabetización mediática en jóvenes universitarios y unas conclusiones reflexivas con propuestas prácticas y normativas.

2. ANTECEDENTES Y CONTEXTO: LA CUARTA REVOLUCIÓN INDUSTRIAL

2.1. EL SURGIMIENTO DE LA CUARTA REVOLUCIÓN INDUSTRIAL

El concepto de Cuarta Revolución Industrial, también conocida como **Industria 4.0**, hace referencia al reciente período histórico en el que se han fusionado las tecnologías digitales con las físicas y las biológicas, transformando radicalmente los procesos de producción, distribución y consumo. No sólo ha afectado de lleno a los modelos de negocio, sino también a las estructuras sociales (Schwab, K. (2016). *The Fourth Industrial Revolution*. Crown Business), y a la forma en la que nos comunicamos. Se caracteriza y se distingue de todas las revoluciones previas[1] por su exponencial crecimiento y su veloz expansión debido a la interoperabilidad[2] de los sistemas ciberfísicos y los distintos tipos de inteligencias artificiales. De forma que, en esta revolución, las máquinas y los sistemas informáticos pueden tomar decisiones autónomas gracias al análisis de grandes volúmenes de datos y a la interacción de plataformas digitales y físicas. De este modo, **el análisis masivo de datos (*Big Data*), el Internet de las cosas (IoT), la impresión 3D, la robótica avanzada, los *bots* y la inteligencia artificial (IA) se integran en cadenas de valor interconectadas, redefiniendo los sistemas de trabajo y la comunicación masiva** (Kagermann, H., Wahlster, W., & Helbig, J. (2013). *Recommendations for implementing the strategic initiative INDUSTRIE 4.0*. Final report of the Industrie 4.0 Working Group.).

El término de **Industria 4.0** fue presentado por primera vez en 2011 en Alemania la **Feria de Hannover**[3], uno de los primeros eventos industriales en los que se comenzó a hablar de la digitalización de los procesos de producción. El concepto fue concebido por una serie de científicos alemanes que presentaron un **informe**[4] que definió las bases del nuevo paradigma de la tecnología y los medios y formas de producción a nivel mundial. Lo que se pretendía era transformar los sistemas productivos tradicionales

[1] La primera revolución industrial (siglo XVIII) se basó en la mecanización gracias a la energía hidráulica y de vapor. La segunda revolución industrial (finales del siglo XIX y principios del siglo XX) se caracterizó por la producción en masa a través de la electricidad, principalmente. La tercera revolución industrial (siglo XX) introdujo la tecnología electrónica y la automatización que ésta supuso en los procesos productivos.

[2] La interoperabilidad es la capacidad que tienen diferentes sistemas para intercambiar datos y utilizar la información de forma recíproca. En el contexto de la Cuarta Revolución Industrial, la interoperabilidad permite la sincronización de diversos ecosistemas ciberfísicos. Existen varios tipos y niveles, que van desde lo puramente técnico – como la interoperabilidad fundacional que permite el intercambio de datos en tiempo real en sistemas de IoT – hasta lo organizativo, como la interoperabilidad que alinea distintos modelos de negocio (Moya, J. (2021). *Interoperabilidad en sistemas de información sanitaria*. Revista Iberoamericana de Tecnología, 18(2), 45–60.).

[3]La Feria de Hannover tuvo lugar por primera vez en 1947 como una iniciativa para reactivar la economía alemana tras la Segunda Guerra Mundial. Desde entonces, ha evolucionado hasta convertirse en un evento global clave para conocer los estándares industriales emergentes y la digitalización de los procesos productivos, puesto que reúne a miles de empresas, investigadores, gobiernos y expertos de sectores clave como la automatización industrial, la robótica, la energía, la logística, la IA y la digitalización (Klein, T. (2019). *Historia y evolución de la Hannover Messe*. Revista de Industria Global, 25(1), 12–17.).

[4] Informe titulado en alemán: *Industrie 4.0: Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution*

creando fábricas inteligentes, donde los procesos se optimizasen y coordinasen en tiempo real mediante sistemas ciberfísicos de monitorización y automatización y una **interconexión digital de las máquinas, los productos y las personas**. A partir de esta referencia, numerosos gobiernos y corporaciones empezaron a adoptar iniciativas de la Industria 4.0.

En el año 2015 el **Foro Económico Mundial** comenzó a incluir la Cuarta Revolución Industrial como eje de debate en sus reuniones, concepto que fue popularizado por el economista e ingeniero alemán **Klaus Schwab**[5] en su libro ***The Fourth Industrial Revolution*** (2016), que definió esta nueva era como una fusión de lo físico, digital y biológico. El alemán explicó que la IA, el IoT y la impresión 3D convergen para transformar industrias, mercados y sociedades a un ritmo sin precedentes, lo cual se ha dado efectivamente durante estos últimos años.

A partir de 2018 se aceleró la adopción de tecnologías clave como la **IA**, la **robótica**, el ***blockchain*** y el **5G** en sectores imprescindibles en nuestro día a día como la salud, la energía o los transportes. En 2020 la Industria 4.0 se convirtió en una realidad innegable: se produjo una digitalización masiva con el teletrabajo como consecuencia del del **COVID-19**, así como la automatización de aquellas tareas que no podían realizar los trabajadores al no acudir a los lugares de trabajo.

Aunque la Cuarta Revolución Industrial sea reciente, no surgió de manera espontánea, ya que sus bases se cimentaron en los avances del **siglo XX**. Uno de los pilares fundamentales fue el desarrollo de la **computación electrónica**. A partir de la invención del transistor en 1947 y del circuito integrado durante la década de 1950, la informática experimentó un crecimiento acelerado (Ceruzzi, P. E. (2003). *A history of modern computing* (2ª ed.). MIT Press.). La creación del **microprocesador** en 1971 permitió la miniaturización de los sistemas de procesamiento y la **Ley de Moore**, formulada en 1965 por el ingeniero estadounidense Gordon Moore, sirvió como hoja de ruta para los ingenieros y gobernantes en la planificación del desarrollo tecnológico de la segunda mitad del siglo (Moore's law: Past, present and future. *IEEE Spectrum, 34*(6), 52–59.). Asimismo, el surgimiento de **Internet** permitió el intercambio instantáneo y masivo de información, sentando las bases para la hiperconectividad global actual.

Tal y como señala Schwab, la Cuarta Revolución Industrial sigue en marcha: no tiene una fecha de fin como tal, sino que es considerada por muchos como un proceso continuo que evoluciona con la adopción masiva de la IA, la computación cuántica y la biotecnología.

[5] Klaus Schwab (1938) es un economista e ingeniero alemán, fundador y presidente del *World Economic Forum*, organización que reúne anualmente a líderes políticos, empresariales y académicos para debatir retos globales. Su libro *The Fourth Industrial Revolution* es el más famoso en cuanto a la digitalización de los procesos, pero ha publicado múltiples trabajos sobre la sostenibilidad y el papel de la sociedad civil en la economía y la política actuales.

2.2. LOS AVANCES DE LA CUARTA REVOLUCIÓN INDUSTRIAL

2.2.1. El *Big Data*

El *Big Data* es el conjunto de tecnologías, procesos y metodologías utilizadas para **capturar, almacenar, procesar y analizar grandes volúmenes de datos** que no podrían ser gestionados de manera eficiente y viable con herramientas tradicionales. Se caracteriza por las "3 Vs": i) **volumen** (gran cantidad de datos), ii) **velocidad** (generación y procesamiento en tiempo real) y iii) **variedad** (diferentes tipos de datos: estructurados, no estructurados y semi-estructurados). A estas 3 Vs, se le han añadido con el tiempo otras dos: **veracidad:** fiabilidad de los datos, importante para tomar decisiones precisas y **valor:** capacidad de extraer conocimiento útil a partir del análisis de los datos (Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.).

El uso del *Big Data* se ha expandido exponencialmente y a día de hoy es empleado por una gran cantidad de empresas e industrias. En el campo del marketing, por ejemplo, el *Big Data* puede resultar muy útil para que las empresas puedan conocer con mayor exactitud las necesidades y preferencias de los consumidores. Mediante el procesamiento de información proveniente de múltiples fuentes, como el historial de las compras, la actividad en las redes sociales, la ubicación geográfica e, incluso, el clima, es posible construir perfiles detallados de cada cliente, permitiendo la segmentación de la audiencia en grupos específicos y el diseño de estrategias de marketing personalizadas para cada uno. De esta forma, las empresas pueden recomendar productos acordes al comportamiento del cliente en tiempo real – tal y como hace Amazon cuando sugiere añadir productos similares a la lista de la compra – o generar contenidos ajustados a los gustos personales, como hace Spotify con las preferencias musicales de cada usuario.

2.2.2. El IoT

El término **Internet de las Cosas** fue acuñado en 1999 por el pionero tecnológico Kevin Ashton en el Instituto Tecnológico de Massachussets, refiriéndose al uso de la **identificación por radiofrecuencia** (RFID)[6] y a los **sensores** para integrar objetos físicos en redes digitales. De modo que el IoT extiende la conectividad de la red más allá de los ordenadores y los *smartphones:* abarca todo tipo de **objetos físicos** (electrodomésticos,

[6] La identificación por radiofrecuencia (RFID) es una tecnología de identificación automática que emplea ondas de radio para comunicar datos entre una etiqueta y un lector sin contacto visual directo. Cada etiqueta RFID contiene un microchip unido a una antena; cuando el lector emite un campo electromagnético, la etiqueta responde modulando la señal, lo que permite transmitir un identificador único. Aunque los primeros prototipos se remontan a experimentos militares, han ido evolucionando hacia aplicaciones comerciales como peajes automáticos y control de inventarios. Hoy en día el RFID es clave en la logística y trazabilidad de las cadenas de suministros, la gestión de activos, la autenticación de productos, etc (Finkenzeller, K. (2010). *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication*. Wiley.)

vehículos, infraestructuras, etc.) dotándolos de **capacidad de medición, comunicación y actuación remotas**.

En 1991 el informático estadounidense Mark Weiser introdujo por primera vez el concepto de **computación ubicua**, al describir dispositivos conectados que cooperan sin la intervención humana directa. Gracias a los avances en sensores, redes inalámbricas e IA, la computación ubicua se aplica hoy a todo tipo de cosas. Existen hogares inteligentes en las que los dispositivos tecnológicos están controlados mediante sistemas que gestionan de forma autómata y remota funciones como la iluminación, la climatización, la seguridad de la casa, los electrodomésticos o los dispositivos de entretenimiento. Estos sistemas se basan en sensores, redes inalámbricas e IA para responder al comportamiento del usuario y al contexto ambiental, como puede ser el ajuste automático de la temperatura según la hora del día o la detección de movimientos inusuales para activar alarmas (Chan, M., Estève, D., Escriba, C., & Campo, E. (2008). *A review of smart homes—Present state and future challenges*. Computer Methods and Programs in Biomedicine.). En el contexto urbano, cada vez es más común encontrar ciudades con espacios inteligentes. En Barcelona, por ejemplo, se emplean sensores para la gestión del tráfico y para el alumbrado adaptativo en las calles y los parques. Además, ya se están creando ciudades desde cero donde las redes ubicuas controlan el tráfico, la energía y la seguridad desde un centro de operaciones unificado. Es el caso de la ciudad surcoreana de Songdo.

Figura 1[7]

La evolución de estas herramientas es tal que cada vez hay más posibilidades de que exista una conectividad total y absoluta entre las personas y los objetos inteligentes. Se puede producir una transformación radical de la manera en la que vivimos, trabajamos y empleamos nuestro tiempo libre, de forma que incluso cambiemos la manera en la que nos relacionamos con otras personas.

[7] Fuente: Google imágenes

2.2.3. La impresión 3D

La impresión 3D, también conocida como **fabricación aditiva**, es una tecnología que permite crear objetos tridimensionales a partir de modelos digitales, añadiendo material capa por capa. Se caracteriza por la creciente aplicación en industrias clave como la medicina, la automoción, la moda o la construcción.

La impresión 3D surgió a principios de la década de 1980, con el invento de la **estereolitografía (SLA)**, una técnica que utiliza luz ultravioleta para solidificar resina líquida capa por capa ((Hull, C. W. (1986). *Apparatus for production of three-dimensional objects by stereolithography* (U.S. Patent No. 4,575,330). U.S. Patent and Trademark Office.). Desde 2010, la impresión 3D se ha ido integrando en sectores como la **medicina personalizada**, con prótesis o implantes; **la tecnología aeroespacial**, con piezas para motores; **la moda y el arte**, con diseños de ropa y texturas; y la **construcción**, con las primeras casas impresas en hormigón capa por capa.

Figura 2[8]

Cabe destacar especialmente la utilidad de la impresión 3D durante el COVID-19, que permitió la rápida fabricación de material sanitario como viseras, mascarillas y piezas de respiradores. La precisión, velocidad y escalabilidad de la impresión 3D han mejorado considerablemente, y esta tecnología se ha consolidado como parte esencial de la Industria 4.0.

[8] Casa impresa por Power2Build en Luanda (Angola). El coste total del material de los muros de hormigón fue inferior a 1.000$.
Fuente: página web de una empresa de construcción (*Lugon. (2023, enero). Casas impresas 3D.)*

2.2.4. La automatización de los procesos: el surgimiento de la robótica y los *bots*

La automatización de los procesos ha transformado radicalmente la producción, los servicios y el uso que le damos a las máquinas. Desde los autómatas mecánicos antiguos hasta los *bots* inteligentes actuales, su evolución refleja el avance en ingeniería, informática e IA. Aunque la automatización de ciertas actividades tiene comienzo hace muchos siglos[9], si hablamos del **primer robot funcional moderno**, hay que referirse a ***Elektro*** (1939), un robot humanoide presentado que podía caminar, hablar mediante un sistema de discos fonográficos, y realizar gestos simples (Gurney, J. (2013). *Robots: From Science Fiction to Technological Fact*. MIT Press.).

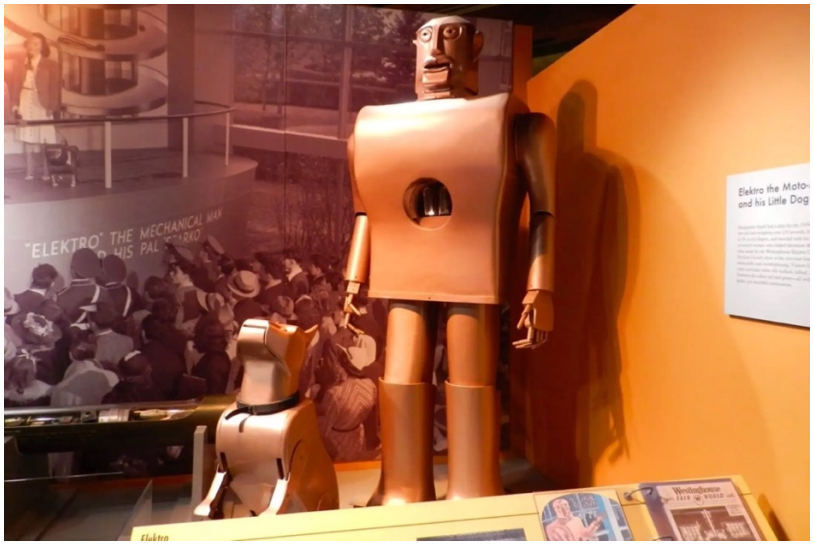

Figura 3[10]

En 1961 se creó **el primer robot industrial comercial llamado *Unimate*** en una planta de *General Motors*. Los robots que se crearon durante esta época eran electromecánicos: ejecutaban movimientos repetitivos y se utilizaban principalmente en tareas como la soldadura, la pintura o la manipulación de piezas en fábricas automotrices. Paralelamente a estos avances en la robótica física, los ***bots* de software** comenzaron a desarrollarse.

[9] Inventores como Herón de Alejandría (siglo I d.C.) crearon dispositivos automáticos usando principios de presión de agua y vapor. Más tarde, en los siglos XII y XIII, el ingeniero turco del medievo Al-Jazari recogió en su manuscrito *Libro del conocimiento de ingeniosos dispositivos mecánicos* (1206) una serie de dispositivos mecánicos con funciones autómatas como relojes de agua con figuras móviles o las especie de máquinas o aparatos autómatas que creó y que simulaban tareas reales humanas, marcando el inicio de la robótica funcional: servían bebidas, tocaban ciertos instrumentos musicales de percusión como el tambor o los platillos y lavaban las manos de los comensales (Hill, D. R. (1974). *The Book of Knowledge of Ingenious Mechanical Devices*. Dordrecht: D. Reidel Publishing Company.) (Hill, D. R. (1974). *Studies in Medieval Islamic Technology: From Philo to Al-Jazari*. Variorum Reprints.). Durante ante el Renacimiento, Leonardo da Vinci diseñó su famoso caballero mecánico, capaz de mover los brazos, el cuello y la mandíbula mediante sistemas de poleas y engranajes.

[10] Foto tomada en la Feria Mundial de Nueva York en 1939, donde Elektro fue presentado por primera vez al público Fuente: medio digital de comunicación especializado en temas de tecnología, ciencia, cultura digital y entretenimiento (Hipertextual. (2022, diciembre). *Elektro, el primer robot humanoide que sorprendió al mundo en los años 30*.)

Los *bots* son programas informáticos diseñados para realizar y automatizar tareas repetitivas y predefinidas en Internet. Pueden operar de forma autónoma y su comportamiento depende del propósito con el que hayan sido programado (Shawar, B. A., & Atwell, E. (2007). Chatbots: Are they really useful? *LDV Forum*, 22(1), 29–49.). A finales del siglo XX, los *bots* automatizaban tareas simples como recopilar información o responder a comandos básicos, pero la automatización de los procesos ha avanzado hacia una **automatización inteligente apoyada por la IA** en la que los *bots* son capaces de ejecutar tareas más complejas y tomar decisiones, puesto que pueden imitar tareas humanas como enviar correos, mover archivos o rellenar formularios. Hay varios tipos de *bots,* pero lo más comunes son los *webcrawlers*, los *chatbots*, los *bots* de redes sociales, los de compras y los "maliciosos" .

Los ***webcrawlers***[11] exploran páginas web para indexarlas en buscadores como Google, Bing o Yahoo. Algunos de sus usos principales son la recopilación de información para que los motores de búsqueda puedan mostrar resultados relevantes o la verificación de que los enlaces a las páginas webs sean correctos. Los ***chatbots*** simulan conversaciones con humanos, por ejemplo, los chats que hay en las páginas web o en las *apps*[12] que conversan con el cliente de forma instantánea. Un ejemplo pionero fue ***ELIZA***, un *chatbot* creado en 1966 que simulaba conversaciones con un terapeuta utilizando reglas gramaticales simples (Weizenbaum, J. (1966). ELIZA — A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45.). Entre los *bots* que están diseñados para realizar una serie de tareas concretas destacan principalmente los ***bots* de redes sociales**, que automatizan las publicaciones, siguen a cuentas y generan interacciones, y los ***bots* de compras**, que se usan para adquirir productos automáticamente, como pueden ser las entradas de conciertos cuando estén de oferta. Por último, los ***bots* maliciosos** realizan actividades que llegan a ser perjudiciales para los consumidores y usuarios, como enviar correos "basura" (***spambots***), extraer información sin permiso (***bots* de *scraping***) o participan en ataques cibernéticos[13] (Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96–104.).

2.2.5. El surgimiento de la IA

La IA ha experimentado una evolución notable desde sus inicios, pasando de simples programas que realizaban tareas específicas y bien definidas, a sistemas avanzados capaces de aprender, adaptarse y tomar decisiones complejas. Esta evolución se ha visto impulsada por avances en algoritmos, la disponibilidad de grandes cantidades de datos y mejoras significativas en la capacidad de cómputo.

[11] También conocidos como arañas web, *bots web* o *spiders*

[12] Una *app* (abreviación de *application* o aplicación) es un programa de software diseñado para ejecutar una función o conjunto de funciones específicas en dispositivos digitales, como teléfonos inteligentes, tabletas, computadoras o incluso relojes inteligentes.

[13] Hay *bots* que operan de forma coordinada, formando una *botnet* o *bot network*.

En sus inicios, la IA se centraba en la lógica simbólica[14] y en sistemas diseñados para **emular la capacidad de decisión humana en áreas específicas mediante reglas predefinidas**. El término de inteligencia artificial fue acuñado en la década de 1950, concretamente en la **Conferencia de Dartmouth**[15], organizada por una serie de académicos estadounidenses que expusieron formalmente el ámbito de la IA consistente en diseñar sistemas computacionales que reproducen los procesos cognitivos humanos, como la resolución de problemas o la toma de decisiones de forma razonada (McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence*. Dartmouth College.). En esta misma década, el matemático británico Alan Turing creó el denominado **Test de Turing**[16], que evaluaba la capacidad de una máquina para exhibir un comportamiento que sea indistinguible del de un humano. Paralelamente se desarrolló el primer programa de IA capaz de demostrar **teoremas lógicos mediante reglas "si-entonces**" (Newell, A., & Simon, H. A. (1956). *The logic theory machine – A complex information processing system*. IRE-Transactions on Information Theory, 2(3), 61–79.). Es decir, que si a la IA le aportamos instrucciones muy detalladas, ésta se crea una especie de mapa mental mediante el que sabe cómo tiene que ir actuando en función de estas reglas o instrucciones Entre 1960 y 1980 se produjo el desarrollo de **sistemas expertos**, programas diseñados para emular la toma de decisiones de especialistas humanos. Ejemplos emblemáticos son los sistemas expertos DENDRAL y MYCIN[17], que demostraron un rendimiento cercano al de expertos humanos en tareas concretas. Si algo salía mal, era fácil saber cuál había sido la instrucción equivocada, ahora bien, al ser una IA inicial tan básica, **sólo funciona en ecosistemas bien definidos y se atasca si aparecen situaciones nuevas**.

En la década de 1990 se trató de superar la gran dependencia de la IA del conocimiento humano explícito y la IA se reorientó hacia el **aprendizaje automático (*machine learning*).** Se trata de la rama de la IA centrada en el desarrollo de algoritmos y modelos que sean capaces de aprender patrones a partir de datos y de mejorar su rendimiento en

[14] La lógica simbólica es una rama de la lógica formal que utiliza símbolos y notación matemática para representar proposiciones, argumentos y relaciones lógicas de forma precisa y estructurada, eliminando ambigüedades del lenguaje natural. La lógica simbólica presenta argumentos en fórmulas compuestas por variables, conectivos lógicos (como ∧, ∨, →, ¬) y cuantificadores (∀, ∃) que permiten operar sobre ellos como si fueran expresiones matemáticas (Smith, P. (2020). *An Introduction to Formal Logic* (2nd ed.). Cambridge University Press.).

[15] La Conferencia de Dartmouth - oficialmente el *Dartmouth Summer Research Project on Artificial Intelligence* - se celebró el 18 de junio de 1956 y se prolongó aproximadamente ocho semanas en Hanover (New Hampshire).

[16] El test de Turing es considerado uno de los padres de la computación moderna. Consiste en que una persona (el interrogador) mantiene una conversación con dos interlocutores – un humano y un programa – sin verlos físicamente. Si tras varias rondas el interrogador no logra identificar cuál es la máquina, ésta ha pasado la prueba. Con sus investigaciones, Turing demostró que las máquinas pueden, en principio, simular cualquier proceso de razonamiento humano, sentando las bases filosóficas y lógicas de la IA.

[17] El sistema experto DENDRAL fue creado por científicos de la universidad de Stanford como el primer sistema experto aplicado a la resolución de problemas reales en química orgánica mediante reglas "si–entonces". En pruebas de laboratorio, DENDRAL llegó a proponer menos de veinte candidatos plausibles frente a los cientos que un químico experto debía filtrar manualmente, alcanzando así un rendimiento comparable al de investigadores humanos.
MYCIN fue un sistema experto para el diagnóstico de infecciones bacterianas y la recomendación de tratamientos antibióticos ajustados al peso del paciente. MYCIN incorporaba más de 450 reglas de la siguiente forma: "si infección = X y recuento glóbulos blancos > Y, entonces recomendar antibiótico = Z". En un experimento ciego, las recomendaciones de MYCIN fueron evaluadas por especialistas en enfermedades infecciosas y calificadas como apropiadas en el 65 % de los casos.

tareas específicas sin ser programados explícitamente para cada una de ellas. A diferencia de los sistemas expertos, que operan con reglas definidas manualmente, el aprendizaje automático construye su conocimiento de manera inductiva, extrayendo relaciones y estructuras desde grandes volúmenes de datos (Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press).

En la década de los 2000, con el auge de Internet, se fueron recopilando datos digitales a mayor escala y se desarrollaron **algoritmos de clasificación de texto e imágenes, algoritmos de recomendación** y los primeros **algoritmos de aprendizaje profundo** (***deep learning*****),** aunque aún con ciertas limitaciones. El aprendizaje profundo es una rama del aprendizaje automático que utiliza redes neuronales profundas para aprender representaciones complejas directamente desde los datos. Gracias a sus múltiples capas, puede identificar patrones en imágenes, texto o audio sin la necesidad de programar manualmente características.

Entre 2010 y 2020 se produjo el **boom la IA moderna con el auge del *deep learning***, que consolidó gracias al uso de GPU[18] para entrenamiento masivo y al *Big Data*. También han aparecido modelos de lenguaje avanzados como BERT (2018) y GPT-2 (2019), capaces de resumir y traducir textos, así como de responder a preguntas de forma coherente. Desde 2020 se ha vivido un salto cualitativo y exponencial con la **IA generativa** y los **modelos fundacionales**, caracterizados por su entrenamiento en grandes volúmenes de texto e imágenes. Entre ellos destacan los dos grandes modelos de lenguaje desarrollados por *OpenAI*[19]: el GPT-3 y el GPT-4, que han dado lugar a *DALL·E* o *ChatGPT* y modelos multimodales que combinan texto, imágenes, audio y video, como *Gemini* o *Flamingo*[20], de *Google DeepMind.*

Como mencionado supra, la idea de crear una máquina pensante se remonta al siglo pasado con Alan Turing y su fórmula para medir la capacidad de una máquina de imitar el raciocinio humano. Durante la segunda mitad del siglo XX y a principios del siglo XXI

[18] Una GPU (*Graphics Processing Unit* o unidad de procesamiento gráfico) es un tipo de chip que originalmente se diseñó para mostrar imágenes y gráficos en ordenadores y videojuegos. Actualmente se está usando en IA porque puede hacer millones de cálculos al mismo tiempo más rápidamente que un procesador normal (Owens, J. D., Houston, M., Luebke, D., Green, S., Stone, J. E., & Phillips, J. C. (2008). GPU computing. *Proceedings of the IEEE*, 96(5), 879–899.).

[19] *OpenAI* es una organización de investigación y desarrollo sobre IA fundada en 2015. Su enfoque combina avances científicos con aplicaciones prácticas y es conocida por desarrollar modelos de lenguaje (Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016).

[20] *DALL·E* es un modelo que genera imágenes a partir de descripciones en lenguaje natural. Por ejemplo, si alguien escribe "un perro subido a un árbol", DALL·E puede crear una imagen original con ese contenido. Se basa en entrenamientos con *Big Data*.
ChatGPT es un modelo de lenguaje diseñado para conversar con personas de forma coherente (entiende el contexto al trabajar con *Big Data*) y ayudar con tareas.
Gemini combina capacidades de texto e imágenes y sus funciones son similares a las de *ChatGPT*.
Flamingo es otro modelo de *Google DeepMind,* especializado en combinar texto e imágenes. Está diseñado para tareas en las que se necesita razonar sobre imágenes con ayuda del lenguaje, como describir lo que hay en una foto o responder preguntas sobre una escena visual (OpenAI. (2020). *GPT-3* [Modelo de lenguaje].).

la IA ha ido evolucionado pasando de la lógica formal y los sistemas simbólicos a los sistemas expertos, el avance en las capacidades computacionales y el *Big Data*. Pero sin duda el auge de la IA se ha dado en los últimos años con el desarrollo del aprendizaje automático – que ha permitido que los sistemas aprendan del *Big Data* sin la necesidad de ser programados explícitamente para una tarea específica – y el aprendizaje profundo, que ha logrado que un enorme nivel de detalle en dicho aprendizaje.

Con el tiempo, la IA ha ido evolucionando en varias direcciones y se ha diversificado en variantes como la IA analítica, la interactiva, la difusora y la generativa. Su vertiginosa expansión ha abierto debates éticos sobre la transparencia, los sesgos, el trabajo y el control sobre la IA.

3. LA IA: VARIANTES, EVOLUCIÓN Y AUGE

La IA ha irrumpido en nuestra sociedad como una de las revoluciones tecnológicas más transformadoras del siglo XXI. Desde sus inicios teóricos hasta sus aplicaciones actuales, ha evolucionado en cuanto a su complejidad, sofisticación y accesibilidad. Hoy en día, no hablamos de la IA como un único campo, sino que hay múltiples variantes que reflejan el amplio abanico de posibilidades tecnológicas que, junto con algoritmos de aprendizaje automático, están modificando cómo interactuamos con el mundo digital.

3.1. LAS VARIANTES CONTEMPORÁNEAS DE LA IA

A día de hoy podemos clasificar la IA en distintas variantes que responden a diferentes objetivos tecnológicos y sociales. Estas formas de IA no solo se diferencian en sus métodos, sino en sus campos de aplicación.

La IA analítica se centra en el procesamiento y en el análisis de datos. Su propósito es descubrir patrones, realizar predicciones y generar modelos predictivos útiles para la toma de decisiones. El ejemplo más básico es el sistema de recomendación de plataformas como Netflix, que analiza el comportamiento de los usuarios para sugerir contenido personalizado. Pero también es muy útil en otros campos, como en el de la salud. En este ámbito, la IA analítica permite detectar enfermedades, puesto que los algoritmos de *deep learning* analizan - en el caso de la detección de cáncer de mama, por ejemplo – miles de mamografías para aprender a identificar patrones asociados con tumores malignos.

La **IA interactiva**, por su parte, busca mejorar la interacción entre las personas y las máquinas. Utiliza tecnologías como el procesamiento del lenguaje natural (PLN)[21], el reconocimiento de voz y la síntesis de texto para **facilitar la comunicación**. Aplicaciones como Siri de Apple, Alexa de Amazon y los asistentes conversacionales en páginas web son ejemplos tangibles.

La **IA generativa** es la más reciente y quizás la más mediática. Se basa en modelos capaces de **crear contenido nuevo como textos, imágenes, música o códigos**. Herramientas como GPT-4, DALL·E y *Midjourney*[22] son ejemplos paradigmáticos. En arte, se han generado composiciones musicales y obras visuales que llegan a ser indistinguibles de las humanas y en programación, *GitHub Copilot* asiste a los desarrolladores escribiendo códigos automáticamente (Castillo Martínez, K., Aguilar Rodríguez, J. A., & Madrigal Rentería, A. S. (2024). *Desafíos éticos de la inteligencia*

[21] El procesamiento del lenguaje natural (o *Natural Language Processing*, NLP por sus siglas en inglés) es el campo de la IA que estudia cómo hacer que las computadoras entiendan, interpreten, generen y respondan al lenguaje humano de forma útil. Su objetivo es permitir que las máquinas trabajen con el lenguaje que usamos las personas cuando hablamos, escribimos o leemos para realizar tareas como traducir textos, responder preguntas, resumir información, conversar e incluso detectar sentimientos (Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Draft. Stanford University.).

[22] *Midjourney* es el modelo de IA que genera imágenes a partir de texto, como DALL·E, pero de la empresa *Midjourney*, *Inc*. Destaca por su estilo artístico y se accede principalmente a través de la plataforma digital de comunicación *Discord*.

artificial generativa en las nuevas formas organizacionales. Revista Digital de Tecnologías Informáticas y Sistemas, 8(1).

Por último, **la IA difusora se asocia a la capacidad que tiene una IA de expandir información, conocimientos o contenidos a través de diferentes medios**. Si se interpreta en línea con los modelos de difusión (en el sentido generativo)[23], puede referirse a IA que genera y propaga dichos contenidos generados, especialmente en redes sociales, medios digitales o incluso en la personalización masiva. También puede entenderse como IA usada para comunicar, amplificar o expandir información preexistente.

3.2. LA EVOLUCIÓN Y EL AUGE DE LA IA

La IA ha evolucionado desde programas basados en reglas hasta sistemas avanzados capaces de aprender y tomar decisiones complejas. Inicialmente centrada en la lógica simbólica y los sistemas expertos, dio un gran salto con el desarrollo del aprendizaje automático y el profundo, impulsados por el *Big Data*. En la última década, destacan los modelos generativos como GPT-4 y *Gemini*, lo que ha diversificado la IA y planteado nuevos retos éticos sobre su uso y control.

3.2.1. El aprendizaje automático y el aprendizaje profundo

Como se ha mencionado supra, el *Big Data* supuso el inicio del auge de la IA, puesto que sin el análisis de estos grandes volúmenes de datos, el aprendizaje automático que actualmente tiene la IA sería inconcebible. El avance en las capacidades computacionales que ha permitido el aprendizaje automático y el aprendizaje profundo de la IA ha supuesto una auténtica revolución digital, puesto que estos algoritmos de aprendizaje son **los principales motores del auge de la IA**.

El aprendizaje automático permite que los sistemas aprendan de los datos sin la necesidad de ser programados explícitamente, lo cual ha revolucionado sectores como el de la salud - con ejemplos como los mencionados en el punto anterior - el comercio y la industria. Amazon, por ejemplo, emplea estos algoritmos para analizar el comportamiento de compra, historial de navegación, valoraciones y patrones de consumo de millones de usuarios. Con esta información, genera **recomendaciones personalizadas de productos** en tiempo real, beneficiando al consumidor, porque mejora su experiencia

[23] Los modelos de difusión generativos son una serie de sistemas que aprenden a crear datos nuevos a través de un proceso que imita cómo se difumina y luego se recupera la información. Primero, toman datos reales (por ejemplo, una imagen) y les agregan ruido poco a poco hasta que se vuelven irreconocibles. Luego, aprenden a hacer el proceso inverso: empezar desde ruido puro y quitarlo poco a poco hasta generar una imagen clara y nueva. (Ho, Jain & Abbeel, 2020).También se usa en combinación con texto, lo que permite generar imágenes a partir de descripciones escritas con un alto grado de coherencia semántica (Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents*.).

como cliente, pero también a la empresa, porque le puede permitir maximizar los ingresos por ventas cruzadas y adicionales. La multinacional industrial y tecnológica General Electric aplica estos algoritmos en sensores de maquinaria para **analizar en tiempo real** datos como las vibraciones, la temperatura o la presión, pudiendo así **predecir fallos mecánicos antes de que ocurran**. Según Lee, J., Bagheri, B., & Kao, H. A. (2014). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3, 18–23., empresas como ésta han transformado la industria y optimizado los procesos empleando algoritmos de aprendizaje automático.

El aprendizaje profundo o *deep learning* permite un aprendizaje jerárquico y efica que va más allá: se enseña a la IA a aprender por sí misma aportándole *Big Data* como si a un robot se le tratase de enseñar a distinguir a un perro mostrándole muchas fotos de perros. Así, la IA puede reconocer voces, leer textos o ver lo que hay en una imagen sin la necesidad de una supervisión humana. Estos algoritmos han dado lugar a avances como como los sistemas de reconocimiento facial o los diagnósticos médicos asistidos por computadora. Por ejemplo, el sistema *AlphaFold*, desarrollado por *DeepMind*, predice con alta precisión la estructura de proteínas, lo cual representa un avance extraordinario en biología estructural (De Francisco, A. L. M. (2021). *Innovación tecnológica: La inteligencia artificial (IA)*. Nefrología al Día.).

3.2.2. Desafíos éticos y sociales

A diferencia de lo que muchos pueden pensar, como hemos visto la IA es tremendamente variada y va más allá de plataformas Chat-GPT. A día de hoy, prácticamente todos los *smartphones* cuentan con asistentes virtuales como Siri o Google *Assistant,* y plataformas de redes sociales como TikTok o de *streaming*[24] como Netflix que emplean recomendaciones personalizadas son usadas constantemente por millones de usuarios. Asimismo, cada vez más empresas van integrando la IA en sus operacione: *apps* de educación como *Duolingo* emplean algoritmos de aprendizaje automático para adaptar las lecciones al usuario (el sistema traza un perfil individual de aprendizaje, ajustando la dificultad, el vocabulario y el ritmo de las lecciones según el rendimiento del usuario); y bancos como el BBVA y el Santander usan modelos de IA para detectar fraudes en tiempo real o para personalizar sus servicios. Por tanto, cabe preguntarse los desafíos éticos con los que nos encontramos al respecto.

La privacidad de los datos de los usuarios que es recopilada (*Big Data*) y almacenada para su posterior análisis (*machine y deep learning*), **los sesgos algorítmicos** a los que los usuarios están expuestos o **la automatización de ciertas tareas laborales** e, incluso,

[24] El *Streaming* es el término que se refiere a la transmisión continua de archivos de audio o video que se puede disfrutar en todo tipo de dispositivos electrónicos con acceso a internet, no siendo necesario descargarse el archivo completo primero. El *streaming* puede ser en tiempo real, como en transmisiones en vivo de eventos, o bajo demanda, como en servicios que ofrecen películas y series que puedes empezar a ver en cualquier momento.

la sustitución de puestos de trabajo, son algunos de los retos a los que la sociedad actual debe hacer frente como consecuencia del auge de la IA. En la Unión Europea (UE) se lleva tiempo trabajando en el desarrollando de unos marcos regulatorios sólidos al respecto que puedan evitar o paliar los efectos negativos de estos avances tecnológicos. El ***AI Act*** (Reglamento de Inteligencia Artificial de la Unión Europea)[25] es la **primera legislación integral del mundo que regula el desarrollo, comercialización y uso de la IA** en función de los riesgos que presenta para los derechos fundamentales y la seguridad de las personas.

3.3. CONCLUSIONES

La IA en sus múltiples variantes, representa una herramienta transformadora con implicaciones profundas en casi todas las dimensiones de la vida humana. Desde sus fundamentos lógicos hasta las aplicaciones generativas actuales, la IA ha experimentado una evolución exponencial, debido a la conjunción de avances técnicos, acceso a *Big Data* y capacidad computacional. Los cambios sustanciales que ha provocado en sectores básicos como la sanidad y la educación y en los procesos de producción conlleva a plantearnos los retos que nos genera como sociedad. Es necesario un marco ético que englobe los beneficios y perjuicios que suponen estos cambios y tecnologías para los usuarios, así como una regulación legal y educativa a nivel global que garantice un uso justo y transparente de la IA.

4. LA DESINFORMACIÓN

[25] European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final

4.1. TIPOS DE DESINFORMACIÓN

En la era digital actual, la circulación de información a través de redes sociales, medios digitales y plataformas en línea ha aumentado significativamente. Sin embargo, este fenómeno ha traído consigo un incremento en la propagación de contenidos falsos o manipulados, lo que ha motivado a investigadores y organismos especializados a clasificar los distintos tipos de desinformación en tres grandes categorías: desinformación, malinformación e información desacreditada (Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Council of Europe report.)

La **desinformación** se refiere a aquella información completamente falsa o alterada que se crea y se difunde con la **intención deliberada de engañar**, manipular o causar un daño. Su propósito, por tanto, es estratégico, y generalmente busca influir en la opinión pública, sembrar desconfianza en ciertas teorías o personajes públicos o, por el contrario, beneficiar a determinados actores políticos o impulsar ciertos pensamientos. Por ejemplo, durante las elecciones presidenciales en Estados Unidos en 2016, se difundieron masivamente en redes sociales noticias falsas que atribuían crímenes a ciertos candidatos sin evidencia alguna (Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236.), por lo que se puede afirmar que dicha desinformación fue diseñada con el objetivo de alterar la percepción pública con respecto a los mismos.

La **malinformación**, por el contrario, corresponde a datos verídicos, pero que son utilizados de forma distorsionada o son sacados de contexto con intenciones maliciosas, como dañar la reputación de una persona, institución o grupo. Un ejemplo de malinformación es cuando se filtran correos electrónicos auténticos, pero se extraen fragmentos fuera de su contexto original para desacreditar a alguien. Tal fue el caso del escándalo del *Climategate,* vinculado a la uiversidad de East Anglia (UEA)**,** en el cual se filtraron públicamente mensajes tergiversados entre científicos que sugerían que podían haber manipulado datos en cuanto al cambio climático. Pese a que posteriormente se demostró que no había existido tal fraude (Painter, J. (2013). *Climate change in the media: Reporting risk and uncertainty*. I.B. Tauris.), la opinión de mucha gente con respecto a dichos científicos e, incluso, con respecto al cambio climático, se vio afectada.

Por último, la **información desacreditada** alude a **información errónea que se comparte sin que haya intención de engañar**. En estos casos, las personas que la difunden creen genuinamente que la información es cierta sin serlo. Este tipo es común en contextos de crisis sanitaria o desastres naturales, cuando circulan remedios falsos o recomendaciones sin respaldo científico. Durante la pandemia del COVID-19, por ejemplo, se viralizó el uso del dióxido de cloro como cura, promovido por usuarios que no necesariamente buscaban dañar, pero que reproducían contenidos desmentidos por la

comunidad médica (Brennen, J. S., Simon, F., Howard, P. N., & Nielsen, R. K. (2020). Types, sources, and claims of COVID-19 misinformation. *Reuters Institute*, University of Oxford.).

En suma, aunque estos tres tipos de información falsa pueden parecer similares, sus diferencias radican en la intencionalidad y la veracidad del contenido difundido. La desinformación se basa en la falsedad y la manipulación con fines maliciosos; la malinformación se sustenta en verdades sacadas de contexto; mientras que la información desacreditada es incorrecta pero no busca intencionalmente causar daño. Comprender estas diferencias es fundamental para promover la alfabetización mediática y el pensamiento crítico frente al creciente flujo informativo en entornos digitales.

Tipo	**¿Es falsa?**	**¿Es intencional?**	**¿Es dañina?**
Desinformación	Sí	Sí	Sí
Malinformación	No	Sí	Sí
Información desacreditada	Sí	No	Puede ser

Tabla 1

4.2. MOTIVACIONES PARA LA DIFUSIÓN DE INFORMACIÓN FALSA

Las motivaciones para la difusión de información falsa son múltiples y pueden variar dependiendo del contexto social, político y tecnológico en el que se producen. En términos generales, estas motivaciones pueden agruparse en cuatro grandes categorías: políticas, económicas, sociales y psicológicas.

En primer lugar, las **motivaciones políticas** son quizá las más evidentes, especialmente en contextos electorales o de conflicto, donde actores con intereses partidistas o personales difunden información falsa con el fin de manipular la opinión pública, desacreditar a oponentes o reforzar narrativas ideológicas. Estudios como los de *Wardle* y *Derakhshan* (Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe report.), promovidos por el Consejo de Europa, subrayan cómo la desinformación es utilizada estratégicamente por gobiernos, grupos de presión o movimientos extremistas para obtener ventajas políticas.

Las **motivaciones económicas** también juegan un papel importante. En muchos casos, la información falsa se produce y disemina con el fin de obtener beneficios financieros a través de la monetización del tráfico web. Plataformas digitales permiten que el contenido viral, aunque sea falso, genere ingresos mediante publicidad. Esto ha sido analizado en profundidad en Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, 6(2), 137–153., donde se evidencia cómo la desinformación se utiliza como estrategia comercial en el ecosistema digital.

En tercer lugar, las **motivaciones sociales** se relacionan con la necesidad de pertenencia, validación o influencia dentro de determinados grupos. Las personas pueden compartir información falsa como una forma de reafirmar su identidad social, fortalecer lazos grupales o ganar visibilidad en redes sociales. Los entornos digitales facilitan la creación de comunidades ideológicamente homogéneas donde la desinformación se propaga como forma de cohesión social.

Finalmente, las **motivaciones psicológicas**, como el sesgo de confirmación, el pensamiento motivado o la necesidad de sentido ante la incertidumbre, también son fundamentales. De acuerdo Pennycook, G., & Rand, D. G. (2018). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings. *Management Science*, 66(11), 4944–4957, las personas tienden a creer y compartir información que refuerza sus creencias preexistentes, incluso cuando es falsa, debido a procesos cognitivos automáticos que favorecen la coherencia interna sobre la veracidad objetiva.
Por tanto, la difusión de información falsa responde a una compleja interacción de factores políticos, económicos, sociales y psicológicos, todos los cuales se ven potenciados por la estructura y dinámica de los medios digitales contemporáneos y de las redes sociales.

4.3. DIFUSIÓN DE *FAKE NEWS* EN REDES SOCIALES

4.3.1. Las *fake news*

Entendemos por ***fake news*** como **aquellos contenidos informativos que son intencional y verificablemente falsos**. Se difunden con la intención de desinformar y manipular, normalmente para obtener beneficios económicos o influir en la opinión pública y, por tanto, en acciones y decisiones sociales y políticas.

Las características de las *fake news* son: i) la difusión de una **información falsa, distorsionada o tergiversada**; ii) la **capacidad de verificar la falsedad**; iii) la **intencionalidad del engaño**, ya que no se trata de errores periodísticos, sino de mentiras o manipulaciones conscientes; iv) la **viralidad**, pues se expanden rápida y

exponencialmente en las redes sociales y las distintas plataformas digitales; v) la **apariencia de veracidad**: normalmente el creador usa formatos y estilos similares a medios y/o informaciones fiables para parecer creíble.

4.3.2. Las redes sociales

Las redes sociales son plataformas digitales que **permiten la creación, el intercambio y la difusión de los contenidos generados por los usuarios en tiempo real**. Estas plataformas recogen todo tipo de comunicación audiovisual: textos, imágenes, videos y enlaces (los usuarios pueden compartir todo tipo de información, como su ubicación en tiempo real) Estas plataformas, como Facebook, X, Instagram o TikTok, han transformado profundamente el ecosistema mediático al permitir que cualquier persona con acceso a internet pueda transmitirle información a miles, e incluso, millones de personas: sin la necesidad de los intermediarios o las trabas tradicionales, solamente con tener acceso a un smartphone y a internet ya es posible convertirse en emisor a escala mundial.

Como se ha visto supra, una de las características principales de las *fake news* es la **viralidad**, haciendo que el papel que tienen las redes sociales en su difusión sea fundamental. Éstas actúan como canales para la **rápida circulación** de la desinformación, **sin los filtros editoriales o la verificación de hechos** que caracterizan a los medios de comunicación convencionales. Las redes sociales permiten una difusión viral de noticias falsas debido a su estructura algorítmica, que prioriza contenidos llamativos o emocionalmente impactantes por encima de su veracidad, favoreciendo el alcance de desinformación sobre la información verificada. Además, las redes sociales potencian fenómenos como las cámaras de eco o las burbujas de filtro[26], en las que los usuarios tienden a exponerse únicamente a contenidos que confirman sus creencias preexistentes, dificultando la corrección de noticias falsas una vez que han sido aceptadas como verdaderas. Esto ha sido evidenciado por estudios como los del Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3), 554–559., quienes demostraron que **los usuarios tienden a compartir y creer información dentro de comunidades ideológicamente homogéneas,** incrementando la polarización y la credibilidad de la desinformación.

[26] Una cámara de eco (*echo chamber*) es un entorno comunicativo donde las personas solo escuchan, comparten o reciben ideas que refuerzan sus propias creencias, mientras se excluyen o desacreditan las opiniones contrarias. Esto genera un efecto de retroalimentación ideológica, en el que los mensajes se repiten y amplifican dentro del mismo grupo, reduciendo la exposición al disenso. En estas cámaras, la percepción de consenso puede ser ilusoria y contribuir a la polarización (Jamieson, K. H., & Cappella, J. N. (2008). *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press.). La burbuja de filtro (*filter bubble*), término propuesto por Eli Pariser (2011), se refiere al resultado de algoritmos que personalizan la información que ve un usuario en función de su comportamiento previo, como clics, búsquedas o interacciones. Ambos conceptos están relacionados, pero mientras la cámara de eco es más social (personas que interactúan en grupos cerrados), la burbuja de filtro es más tecnológica (algoritmos que seleccionan lo que se muestra).

De acuerdo con Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151, **las noticias falsas se difunden significativamente más rápido que las verdaderas**, especialmente cuando apelan a emociones como la sorpresa o la indignación.

4.4. IMPACTO EN LA SOCIEDAD Y CONCLUSIONES

Como hemos visto, las dinámicas de desinformación en el entorno digital contemporáneo representan uno de los mayores desafíos comunicativos y sociales de la actualidad. La propagación de distintos tipos de información falsa - desinformación, malinformación e información desacreditada - no ocurre de forma aleatoria, sino que responde a diversas motivaciones estructurales y psicológicas: intereses políticos que buscan moldear la opinión pública y **socavar la confianza en las instituciones y los gobiernos**; **motivaciones económicas que convierten a la viralidad en un modelo de negocio; necesidades sociales de validación y pertenencia a un grupo; y sesgos psicológicos como el de confirmación, que hacen que las personas tiendan a aceptar como verdadera aquella información que refuerza sus creencias.** En este ecosistema, las redes sociales juegan un papel crucial al ser el principal vehículo para la circulación de *fake news*. Plataformas como TikTok e Instagram han transformado la lógica de la comunicación pública al permitir que cualquier usuario, sin intermediarios, pueda convertirse en emisor de contenidos capaces de llegar a miles o millones de personas en cuestión de minutos. Además, promueven la viralización de contenidos falsos y contribuye a la formación de burbujas de filtro en las que los usuarios son expuestos únicamente a contenidos que confirman sus ideas previas, lo que refuerza la polarización y reduce la posibilidad de contrastar la información.

En definitiva, la difusión de *fake news* en las redes sociales a través de sistmas y algortimos de IA no es solo un problema de contenido, sino de estructura comunicativa, alimentada por intereses estratégicos, por una arquitectura digital que prioriza el impacto emocional sobre la verdad, y por una falta de herramientas críticas en los usuarios para enfrentar esta complejidad. Reconocer la existencia de estos fenómenos, comprender su lógica y fortalecer la alfabetización mediática son pasos fundamentales para construir una ciudadanía digital informada, crítica y resiliente frente a los riesgos de la manipulación informativa.

5. LA INFLUENCIA DE LA INTELIGENCIA ARTIFICIAL EN LA DIFUSIÓN DE *FAKE NEWS*

5.1. LOS ALGORITMOS DE RECOMENDACIÓN Y LAS BURBUJAS DE FILTRO (*COOKIES*)

5.1.1. ¿Qué son?

Los algoritmos de recomendación y las burbujas de filtro -conocidas comúnmente como *cookies*- son componentes esenciales en la estructura de un gran número de plataformas *online* actuales. Sirven para la difusión de información y, por tanto, también de desinformación.

Los **algoritmos de recomendación** son fórmulas o programas informáticos diseñados para predecir la preferencia de los usuarios en base a su comportamiento previo. Se trata de algoritmos que analizan datos tales como el historial de compras o las interacciones a contenidos previos, para sugerir productos, servicios o contenidos similares. Su objetivo es, esencialmente, identificar los patrones de comportamiento de cada usuario para poder recomendarle elementos o información que otros usuarios con criterios similares han acogido positivamente. Mediante un riguroso análisis de grandes cantidades de información, estos modelos de aprendizaje en línea son capaces de prever las preferencias de cada usuario de forma particular y subjetiva.

Las **burbujas de filtro o *cookies*** han acompañado a la web prácticamente desde sus comienzos, siendo *Netscape Navigator* el primer navegador en implementarlas con la finalidad de facilitar al usuario una experiencia de compra más continua en aplicaciones de *e-commerce*[27]. De este modo, mientras los consumidores exploraban una tienda online, los productos elegidos permanecían en su carrito de compras, incluso al cambiar de página. A día de hoy, siguen existiendo este tipo de *cookies* temporales, cuya función principal es mantener activa la sesión del usuario en un sitio web mientras éste lo navega. Pero a diferencia de éstas, que se eliminan al cerrar el navegador, existen *cookies* que permanecen almacenadas en el navegador hasta que el usuario decide eliminarlas o hasta que expiran. Se emplean normalmente para guardar información de registros e historiales de compras, siendo un efecto secundario de los algoritmos de recomendación. Debido a que estos algoritmos limitan la diversidad de contenido que un usuario ve en línea puesto que se basan únicamente en sus preferencias pasadas, el contenido que se muestra al usuario, especialmente en redes sociales y plataformas de *streaming*, se personaliza extremadamente para coincidir con sus intereses anteriores. Comercios de venta online como, por ejemplo, Amazon, usan estas *cookies* para sugerir productos en función de las

[27] *E-commerce* es el concepto que hace referencia a aplicaciones o sitios web que permiten la compra y venta de productos y servicios a través de internet.

compras previas y las búsquedas realizadas. Plataformas como Netflix o redes sociales como Instagram, YouTube o Tiktok personalizan lo que el usuario ve basándose en interacciones pasadas.

5.1.2. Impacto en la sociedad y ejemplos de casos destacados

Aunque la publicidad personalizada resulte tan útil para la parte vendedora y los anuncios específicamente dirigidos a ciertos usuarios basados en sus comportamientos en línea previos sean eficaces, no hay que dejar de lado la burbuja de información que ello supone, **limitando la exposición del usuario a puntos de vista divergentes y, por ende, aumentando la polarización de la sociedad.**

En definitiva, los algoritmos de recomendación y las burbujas de filtro han transformado la manera en la que las personas interactuamos con el contenido digital, ofreciendo experiencias altamente personalizadas y eficientes desde el punto de vista del consumo. Sin embargo, este avance tecnológico trae consigo desafíos sociales significativos, especialmente con respecto a la información sesgada a la que el usuario queda limitado. El problema de dicho sesgo no es sólo la limitación a la diversidad, que también, sino la problemática que supone la expansión de cierto contenido incorrecto. Una sociedad polarizada no es sólo una sociedad poco informada, sino desinformada o malinformada. Así, es necesario el debate que poco a poco ha ido floreciendo en el último lustro sobre la necesidad de una mayor transparencia en el diseño de los algoritmos para mitigar la formación de burbujas de filtro, introduciendo elementos aleatorios y diferentes en las recomendaciones.

La académica y profesora de la Universidad de California, Safiya Noble, refleja a la perfección en Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press, cómo los algoritmos de búsqueda en internet, particularmente los de Google, pueden perpetuar y reforzar la discriminación social y racial. Noble señala numerosos ejemplos muy ilustrativos, como los resultados divergentes e, incluso, opuestos, que aparecen en Google al buscar “chicas negras” en contraposición con “chicas blancas”, no en cuanto a la raza, sino diferencias discriminatorias hacia las primeras. La autora explica que, en el primer caso, los términos principales de búsqueda incluyen expresiones sexuales o sugerencias de sitios pornográficos, lo que no ocurre de forma tan explícita en las búsquedas de chicas blancas.

A diferencia de lo que en muchas ocasiones se presupone, **los algoritmos de búsqueda no son neutrales ni objetivos**, y tal y como señala Noble, hay casos donde las búsquedas de términos asociados a ciertas minorías étnicas arrojan **resultados negativos o estereotipados**. Los algoritmos de búsqueda influyen e, incluso, llegan a moldear la percepción pública sobre diferentes grupos sociales ya que las representaciones sesgadas en los resultados de búsqueda pueden reforzar ciertos estereotipos negativos y prejuicios en la sociedad.

Aunque teóricamente internet y las redes sociales permiten el acceso a todo punto de vista que está plasmado en la red, en la práctica los algoritmos que personalizan contenidos tienden a crear las burbujas de filtro que mencionábamos supra. Hemos asumido que toda la información que está en línea es completamente disponible y está "al alcance de nuestra mano", pero ¿**realmente es todo este contenido tan accesible como nos han hecho creer**? La realidad es que los algoritmos de personalización limitan a los usuarios a un estrecho rango de opiniones, impidiéndoles acceder a perspectivas contrarias que puedan desafiar sus creencias preexistentes. Esto mismo lo señalan los académicos Engin Bozdag y Jeroen van den Hoven (Bozdag, E., & van den Hoven, J. (2015). "Breaking the filter bubble: Democracy and design." (4), 249-265.) al analizar la problemática desde una perspectiva liberal en los sistemas democráticos. En una democracia los ciudadanos han de ser capaces de considerar y confrontar ideas opuestas para tomar decisiones racionales y libres. Sin embargo, los sistemas algorítmicos de recomendación tienden a crear entornos informativos homogéneos, donde las personas no encuentran contenidos que confronten con sus visiones del mundo, debilitando así uno de los principios fundamentales del liberalismo democrático.

Estos ecosistemas informativos cerrados **contribuyen a la polarización y a la rápida viralización de desinformación**. Las **últimas elecciones en Brasil** estuvieron marcadas por una oleada de desinformación en redes y *apps.* Las *fake news* difundidas incluían ataques personales, distorsiones constantes de las propuestas políticas reales y teorías conspirativas como las relacionadas con el comunismo o el fraude electoral. Un estudio identificó que las burbujas de filtro generadas por el algoritmo de Facebook potenciaron contenidos afines ideológicamente, dificultando la exposición a otras perspectivas (*Acosta, D.G. & Masjuán, M.E.G. (2022). Fake news en tiempos de posverdad. Academia.edu).* Recientemente se ha realizado un estudio sobre **el comportamiento de estudiantes de comunicación frente al consumo de noticias en TikTok**, destacando que los usuarios quedaban atrapados en **burbujas informativas**, donde TikTok no sólo destaca por la segmentación algorítmica en su sistema de recomendaciones, sino que además prioriza **contenido en función de las interacciones que genera** (como likes, comentarios y retención**) independientemente de su veracidad.**

Por tanto, estas modalidades de la IA están contribuyendo a que el espacio público sea cada vez menos diverso y, por ende, esté cada vez más sesgado, siendo esto perjudicial tanto para la autonomía del individuo, como también para el rigor de la deliberación pública y, por tanto, para la salud de las democracias.

5.2. LOS *BOTS* Y LAS CUENTAS AUTOMATIZADAS

5.2.1. Impacto en la sociedad y ejemplos de casos destacados

Como se ha expuesto en puntos anteriores, los ***bots*** son programas informáticos diseñados para ejecutar **tareas automáticas, repetitivas y predefinidas sin intervención humana directa**. En contextos académicos y técnicos, se reconocen distintos tipos de bots según su función, como los *chatbots*, los web *crawlers* o los que forman parte de *botnets* en ataques cibernéticos.

En el ámbito de la comunicación digital, los *bots* son relevantes por su capacidad para **influir en las opiniones y conversaciones públicas, por automatizar la difusión de contenido y amplificar mensajes políticos o comerciales.** Su uso en redes sociales plantea desafíos importantes, ya que pueden imitar la actividad humana prácticamente a la perfección manipulando opiniones, propagando desinformación y alterar artificialmente tendencias políticas, sociales y económicas (Woolley, S. C., & Howard, P. N. (2016). *Automation, algorithms, and politics: Political communication, computational propaganda, and autonomous agents*. Oxford Internet Institute.). Desde una perspectiva académica, diversos estudios (Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96–104.) han documentado cómo los *bots* afectan negativamente la calidad del discurso público puesto que difunden contenido polarizador y fragmentan comunidades a través de las redes sociales. Su capacidad para publicar de manera masiva, coordinada y sin descanso les permite superar en visibilidad a los usuarios reales, lo que les otorga un **poder desproporcionado en la viralización de *fake news***.

Uno de los casos más significativos es el de las **elecciones presidenciales de Estados Unidos en 2016**, donde investigaciones posteriores, como las de los norteamericanos Bessi y Ferrara[28], evidenciaron que aproximadamente un 19% del contenido relacionado con la campaña electoral en Twitter (ahora X) fue generado por *bots*, muchos de los cuales promovían ataques coordinados sin importar la veracidad del contenido. Otro caso notable fue el **referéndum del Brexit en 2016**, donde también se identificó el uso intensivo de *bots* para difundir mensajes nacionalistas, de antiinmigración y teorías conspirativas (Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2017). Junk News and Bots during the U.S. Election: What Were Michigan Voters Sharing? *Oxford Internet Institute*.).

De modo que, el uso de *bots* para difundir *fake news* se ha vuelto un arma cada vez más común para distorsionar la opinión pública por medio de las redes sociales, lo que supone una trágica amenaza para el debate democrático,

[28] Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, 21(11).

5.3. LOS INFLUENCERS VIRTUALES Y LOS *DEEPFAKES*

5.3.1. Los *influencers* virtuales: qué son y su impacto en la sociedad

Un *influencer* es una persona que, a través de plataformas digitales como las redes sociales, tiene una audiencia significativa y ejerce una capacidad de persuasión sobre sus seguidores, especialmente en lo que respecta a opiniones, comportamientos de consumo y tendencias culturales. Su influencia radica en la credibilidad, autenticidad y, principalmente, en la conexión que su comunidad siente con él/ella. De esta forma, los *influencers* se han convertido en intermediarios valiosísimos para el marketing y la publicidad de las marcas, pero también para los movimientos sociales y las campañas políticas[29].

Los *influencers* virtuales son personajes digitales generados mediante IA. Son diseñados por equipos de creativos y programadores para **interactuar con el público en redes sociales como si fueran personas reales**. Son utilizados principalmente con fines comerciales y de marketing, pero su capacidad para generar contenido emocionalmente persuasivo también los convierte en agentes potenciales de difusión de desinformación.

Estos *influencers* operan mediante sistemas de IA como los algoritmos de aprendizaje automático que les permiten aprender de las reacciones del público, adaptar su lenguaje y estilo, y generar publicaciones, comentarios o videos que simulan autenticidad y empatía. Un caso representativo es **Lil Miquela**, una *influencer* virtual creada en 2016 por la empresa tecnológica Brud que acumula millones de seguidores en Instagram y colabora con marcas de moda de alto perfil.

Figura 4[30]

[29] En las últimas elecciones de Estados Unidos Kamala Harris contó con el apoyo de celebridades de alto perfil como Taylor Swift o Jennifer Lopez, y Trump fue respaldado por numerosos *influencers* en sus plataformas de redes sociales.

[30] Se trata de la *influencer* virtual Lil Miquela junto con Bella Hadid, una de las modelos más conocidas y prestigiosas internacionalmente.
Fuente: Google imágenes.

Aunque en principio se utilizan para campañas publicitarias, su presencia puede ser instrumentalizada para otros fines, incluyendo la difusión de contenidos ideológicos, políticos o falsos. El impacto social de estos influencers virtuales gestionados por IA es complejo y ambivalente. Por un lado, plantean desafíos éticos en torno a la autenticidad, la manipulación emocional y la transparencia: muchos usuarios no son plenamente conscientes de que están interactuando con entidades artificiales. Según estudios como los Mavridis, N., & Kameas, A. (2020). Artificial agents as social influencers: the effect of virtual characters on human behavior. *AI & Society*, 35(3), 639–651., esta ambigüedad puede debilitar la confianza pública en los contenidos digitales y abrir la puerta a nuevas formas de manipulación simbólica, especialmente cuando se utilizan para difundir *fake news* de manera deliberada o como parte de campañas automatizadas. Como advierte el informe de la European Parliamentary Research Service (EPRS). (2020). *Tackling disinformation online: The EU's code of practice on disinformation*. European Parliament, estas tecnologías, al combinar automatización, simulación de emociones y estrategias de *microtargeting*, representan una amenaza potencial puesto que pueden manipular la opinión pública de forma efectiva y con costes bajos.

En definitiva, los *influencers* virtuales son herramientas poderosas de persuasión digital que, si no se regulan adecuadamente, pueden convertirse en vectores significativos de desinformación.

5.3.2. Los *deepfakes*: qué son y su impacto en la sociedad

Los *deepfakes* son, básicamente, vídeos, audios o imágenes generadas mediante tecnologías de aprendizaje profundo que imitan la apariencia y/o voz de una persona real. Mediante sistemas de IA[31] aprenden cómo se ve y suena una persona en diferentes situaciones y luego replican esos patrones y lo hacen de manera muy convincente. Es decir, a través de la recopilación de grandes cantidades de datos visuales y auditivos de la persona objetivo, los sistemas de IA pueden generar **representaciones falsas pero realistas que son difíciles de distinguir de la realidad** para quienes no estén fijándose en si se trata realmene de una IA.

Se usan en multitud de campos: la educación, el arte, el entretenimiento, etc. Por ejemplo, se emplean para recrear actores fallecidos y doblar películas en otros idiomas conservando el movimiento labial. Sin que esto último tenga un impacto negativo masivo puesto que no afecta a la desinformación, es indicativo de la sociedad con la que nos estamos encontrando donde las personas son cada vez más sustituidas por IA.

El principal impacto negativo a nivel social son sus usos en la **desinformación política**, ya que pueden simular discursos o acciones de figuras públicas, erosionando la confianza en los medios y en la veracidad de los registros audiovisuales.

[31] Utilizan redes neuronales como los *autoencoders* y las redes generativas antagónicas (Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, 63(2), 135–146.)

Más allá de la política, los *deepfakes* han sido usados para la **pornografía no consentida**, especialmente contra mujeres representando una **forma de violencia digital y una amenaza a la privacidad y la dignidad**. En los últimos años se han difundido numerosas imágenes de pornografía falsa de rostros muy famosos en nuestro país como Rosalía o Ester Expósito. Otro bastante común en nuestro día a día y del que es cada vez más difícil de huir es los *deepfakes* empleados para **fraudes de identidad o estafas** mediante suplantación de voz o rostro en entornos digitales o bancarios.

Por tanto, nos encontramos ante una situación en la que los contenidos reales pueden ser puestos en duda y los que no lo son realmente no se distinguen, lo que debilita la confianza en la evidencia audiovisual y complica la verificación de la información veraz.

5.4. CONCLUSIONES

La polarización de la opinión pública ha aumentado por el uso de algoritmos de recomendación, los *bots* y los *deepfakes*, entre otros. Los algoritmos refuerzan creencias previas al mostrar solo contenido afín, creando burbujas de filtro que limitan la exposición a ideas diferentes, los *bots* amplifican discursos extremistas y desinformación, simulando consenso y aumentando la confrontación entre grupos y los *deepfakes*, al falsificar evidencia audiovisual, erosionan la confianza pública y pueden manipular elecciones o desacreditar personas. En conjunto, estos elementos debilitan el debate democrático y aumentan la fragmentación social.

La capacidad de acceder a información no sesgada y diversa ayuda a los ciudadanos a desarrollar juicios más informados y autónomos, esenciales para una participación democrática efectiva. Por tanto, se podría decir que las sociedades democráticas solamente tienen sentido si sus ciudadanos están expuestos a opiniones y opciones diversas e, incluso, contrarias, para que puedan tomar decisiones basadas en un análisis un tanto profundo y racional. Los algoritmos que filtran el contenido interfieren en la autonomía de los usuarios y, por ende, de los votantes de democracia, y en su capacidad de juzgar sus propios intereses. De modo que, **¿podríamos decir que nos encontramos ante una limitación e, incluso, coacción de la capacidad de razonamiento de los ciudadanos y, por tanto, ante una disminución de la libertad de pensamiento?**

Debemos tener en cuenta que los algoritmos no son neutros ni objetivos. Así, deerñiamos empezar a cuestionar el diseño de los sistemas tecnológicos, planteando que los algoritmos deberían ser diseñados considerando valores democráticos como la equidad, la diversidad y la transparencia. Cada vez resulta más necesaria una visión ética de los usos de la IA que priorice el bienestar de los usuarios y fomente la exposición a diversidad informativa. En este último sentido cabría preguntarse si hay que ir más allá de la libertad individual de cada persona y si se le debería otorgar a los usuarios la posibilidad de

acceder a variedad de contenidos o si su exposición a esta diversidad debería ser obligatoria, incluso si esto fuese en contra de sus preferencias individuales.

6. INFLUENCIA DE LA IA EN LA DESINFORMACIÓN

6.1. DIFICULTADES EN LA DETECCIÓN Y VERIFICACIÓN DEL CONTENIDO

La IA juega un papel dual en el ámbito de la desinformación: por un lado, puede facilitar la creación y propagación de contenido falso o manipulado; por otro lado, es una herramienta valiosa para detectar y verificar la autenticidad del contenido. Sin embargo, las dificultades en la detección y verificación de contenido generado por IA son significativas y presentan desafíos únicos.

En primer lugar, cabe destacar el **constante perfeccionamiento de la tecnología.** A medida que las técnicas de generación de contenido como los *deepfakes* y otros medios manipulados mejoran, se vuelven más difíciles de distinguir de los contenidos auténticos. La capacidad de la IA para imitar sutilezas humanas en vídeos, imágenes y textos puede engañar incluso a los observadores más astutos. Los desarrolladores de IA se encuentran en una carrera interminable contra los creadores de herramientas de detección, lo que puede resultar en una escalada tecnológica continua donde cada mejora en las técnicas de generación de contenido es seguida por una innovación en la detección, y viceversa.

El segundo principal problema es las limitaciones de las herramientas de detección actuales. Se renueven a mucha menor escala. Las herramientas basadas en IA que se utilizan para detectar contenido falso pueden ser susceptibles a errores, especialmente cuando se enfrentan a técnicas nuevas o mejoradas de desinformación. Los atacantes pueden diseñar contenido falso específicamente para eludir los modelos de detección existentes, utilizando técnicas que explotan las debilidades de estos sistemas.

El tercer gran reto es la **escalabilidad** y el **volumen** masivo de contenido generado continuamente en línea. Escalar las herramientas de detección para examinar y verificar todo el contenido relevante en tiempo real presenta enormes desafíos logísticos y de recursos. Desarrollar, entrenar y mantener sistemas de IA para la detección de desinformación requiere inversiones significativas, lo que puede ser un obstáculo para organizaciones más pequeñas o países con menos recursos.

6.2. CAMBIOS EN LA REGULACIÓN Y POLÍTICAS DE PLATAFORMAS

Los cambios en la regulación y en las políticas de plataformas digitales han sido una respuesta creciente ante las consecuencias negativas de la difusión de *fake news*, a través de la IA. A nivel global, gobiernos y empresas tecnológicas han implementado diversas

medidas para mitigar la propagación de desinformación, proteger los procesos democráticos y salvaguardar a los usuarios.

La Unión Europea ha sido pionera con iniciativas como el ***Digital Services Act***[32], que obliga a las plataformas a identificar y eliminar contenidos ilícitos, mejorar la transparencia de sus algoritmos y colaborar con verificadores de datos independientes. Por parte de las plataformas, compañías como Meta, YouTube, TikTok y X han adoptado políticas de contenido más estrictas. Estas incluyen **etiquetas de advertencia sobre contenido falso, reducción de la visibilidad de publicaciones desinformativas, eliminación de cuentas automatizadas maliciosas, y la colaboración con verificadores externos**. Además, se han desarrollado sistemas de detección automatizada de *deepfakes* y generación sintética de contenido, aunque con resultados aún limitados.

Académicamente, se señala que estas medidas, aunque necesarias, no son suficientes sin mayor **transparencia algorítmica**, **participación ciudadana y control democrático sobre las infraestructuras digitales** (Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.). El desafío principal es equilibrar la lucha contra la desinformación con la protección de la libertad de expresión y evitar que las plataformas actúen como árbitros opacos de la verdad.

En resumen, los cambios regulatorios y las políticas internas de las plataformas buscan reducir el impacto negativo de la IA en la difusión de *fake news*, pero requieren una implementación más robusta, cooperativa y ética para ser verdaderamente eficaces.

6.3. PARTE PRÁCTICA: DESAFÍOS ÉTICOS Y DE PRIVACIDAD Y ESTUDIO AL RESPECTO

6.3.1. Introducción teórica

La desinformación en redes sociales plantea desafíos éticos y de privacidad de gran magnitud, especialmente en contextos donde el acceso a la información está mediado casi exclusivamente por plataformas digitales.

Desde una **perspectiva ética**, uno de los principales problemas es la **manipulación cognitiva y emocional** de los usuarios a través de contenidos falsos, diseñados para generar **reacciones rápidas, reforzar sesgos y polarizar la opinión pública**. Esto debilita la autonomía informativa de los ciudadanos, al dificultar su capacidad para tomar decisiones basadas en datos verificados. Además, los algoritmos que determinan qué información se muestra a cada usuario operan con opacidad y sin control democrático, lo

[32] European Commission. (2022). *Digital Services Act*

cual supone un conflicto ético con principios básicos como la transparencia, la equidad y la rendición de cuentas.

En cuanto a la **privacidad**, el modelo de negocio de muchas plataformas se basa en la **extracción y comercialización de los datos personales de los usuarios**, lo que permite segmentarlos con gran precisión. Esto no sólo limita o acaba con la privacidad del consumidor (se suele emplear en estrategias de marketing que benefician a las empresas), sino que además puede enviar a los usuarios contenidos manipulativos que, según este análisis de *Big Data*, sean más propensos a aceptar.

6.3.2. Estudio

Para complementar el análisis, se ha diseñado y distribuido una **encuesta**[33] a través de Google *Forms* con el objetivo de recopilar datos sobre la **percepción que tienen los jóvenes universitarios frente a la desinformación digital.** La intención es evaluar el nivel de alfabetización mediática y de pensamiento crítico de una muestra de 200 estudiantes universitarios frente a la desinformación, determinando su habilidad para diferenciar entre noticias verificadas y *fake news* presentadas con formatos reales de redes sociales.

- **Tipo de estudio:** cuantitativo, exploratorio, con diseño experimental de tipo encuesta estructurada.
- **Población objetivo:** estudiantes universitarios de carreras variadas, pertenecientes a instituciones públicas y privadas en España.
- **Tamaño de la muestra:** 200 estudiantes seleccionados aleatoriamente
- **Procedimiento**: se les han presentado 10 titulares de noticias con apariencia real (basados en formatos gráficos de TikTok e Instagram). 5 de estos titulares correspondían a noticias verídicas y verificadas. 5 eran ejemplos de *fake news* intencionalmente falsificadas, diseñadas con alta credibilidad visual. Los participantes debían indicar si cada noticia era verdadera o falsa
- **Variables observadas:**
  - Nivel de acierto (respuestas correctas vs. incorrectas por noticia)
  - Nivel socioeconómico (alta, media, baja)
  - Carrera universitaria
- **Herramientas de análisis:**
  - Google Forms
  - Registro y tabulación de respuestas en Excel (archivo generado)
  - Clasificación binaria por noticia (correcto = 1; incorrecto = 0)
- **Limitaciones del estudio:**

[33] Anexo 1: Resultados de la encuesta (Link al archivo excel)

- No se han incluido métricas fisiológicas o psicológicas (emocionalidad, nivel de conocimiento político o científico).

El estudio demuestra que una gran mayoría de los jóvenes universitarios, independientemente de su clase social o su campo de estudio académico, **no posee herramientas sólidas para identificar *fake news* en entornos digitales**. La apariencia visual, la viralidad y la familiaridad con el estilo gráfico tienen más peso en su juicio que la veracidad de la información. Esto revela una vulnerabilidad crítica frente a la desinformación, y resalta la necesidad urgente de **incorporar formación en alfabetización mediática en todos los niveles universitarios**, con énfasis en la verificación de fuentes y el análisis crítico del discurso.

6.4. CONCLUSIONES

El análisis de las dificultades en la detección y verificación de contenido, los desafíos éticos y de privacidad, y los cambios regulatorios frente a la desinformación generada por la IA permite concluir que nos enfrentamos como sociedad ante una transformación profunda en sus dinámicas informativas y democráticas. En primer lugar, la creciente sofisticación de los *deepfakes* y la producción automática de textos compromete la capacidad de los ciudadanos y los sistemas tecnológicos para distinguir entre información veraz y falsificada, generando desconfianza generalizada y una percepción de vulnerabilidad informativa. **La carrera entre creadores de desinformación y desarrolladores de sistemas de detección representa un ciclo interminable donde parece que la veracidad siempre se queda atrás.** Aunque los marcos regulatorios como el *Digital Services Act* y las nuevas políticas de plataformas son pasos positivos, siguen siendo insuficientes si no se garantizan la transparencia, la rendición de cuentas y la participación ciudadana en su diseño y aplicación. Además, los desafíos éticos derivados del *Big Data*, la falta de transparencia de los algoritmos y el impacto en la privacidad y la autonomía individual son preocupaciones crecientes. Estos problemas no solo amenazan derechos fundamentales, sino que también afectan la equidad social y la confianza en las instituciones tecnológicas y gubernamentales. Los datos del estudio llevado a cabo confirman una tendencia preocupante: casi la mitad de los estudiantes validan como verdaderas varias *fake news*, y más de tres cuartas partes desconfían de información real. En conjunto, **el impacto en la sociedad se manifiesta en una mayor polarización, un debilitamiento del debate público informado y la necesidad urgente de reforzar la gobernanza ética de la IA**.

7. ESTRATEGIAS PARA ABORDAR EL PROBLEMA

7.1. Estrategias vigentes

Para abordar el problema de las *fake news* en redes sociales mediante IA se han propuesto diversas estrategias tecnológicas, normativas y educativas que buscan mitigar la propagación de desinformación y fortalecer la integridad informativa en entornos digitales.

La mayoría de plataformas emplea IA para detectar cuentas falsas o redes de *bots* que difunden contenido manipulado. Esto incluye la identificación de comportamientos anómalos y patrones coordinados, aunque sigue siendo un reto en contextos multilingües o de baja visibilidad. Plataformas como X (antes Twitter), Meta, TikTok y YouTube aplican etiquetas para advertir sobre contenido dudoso o desmentido. En Estados Unidos Meta ha sustituido los verificadores profesionales por el sistema de ***Community Notes***, mientras que X lo mantiene desde 2021. Estas etiquetas reducen la difusión de desinformación, pero dependen de la colaboración activa de usuarios y pueden presentar sesgos o retrasos. También se utilizan algoritmos entrenados para detectar patrones de desinformación en texto, imagen y video, que aunque son útiles para filtrar contenidos, estas herramientas aún son limitadas frente a nuevas técnicas de manipulación, especialmente los *deepfakes* avanzados. Además, las plataformas de redes sociales tratan de **minimizar la visibilidad de los contenidos falsos, reduciendo su alcance algorítmico:** ocultan las publicaciones de los *feeds* o las quitan de las opciones de promoción e incluso llegan a eliminarlas según la normativa de cada *app*.

7.2. Propuestas y conclusiones

Más allá de la alfabetización digital clásica, sería necesario incluir formación sobre cómo operan los algoritmos de recomendación, la personalización informativa y los sesgos de la IA, ya que todavía existe mucho desconocimiento sobre su funcionamiento.

Además, cabe destacar la figura de los **verificadores profesionales**. Se trata de periodistas, investigadores o analistas que trabajan en organizaciones especializadas en la identificación, análisis y desmentido de información falsa o engañosa. Su labor se basa en metodologías rigurosas de *fact-checking* que incluyen la revisión de fuentes primarias, el contraste de datos con información oficial y científica, y la consulta de expertos independientes. Estas organizaciones operan bajo principios éticos y estándares internacionales de transparencia, imparcialidad y trazabilidad. Entre los más conocidos a nivel internacional están PolitiFact, FactCheck.org, AFP Fact Check o Snopes y Maldita.es en España. Muchas de estas entidades están agrupadas en la red *International Fact-Checking Network* que certifica su cumplimiento con criterios éticos y metodológicos, como la imparcialidad, la transparencia en la financiación y la corrección

de errores. Por tanto, cabría preguntarse porqué plataformas como TikTok no hacen uso de estos y Meta las ha sustituido por los verificadores personales que, como mencionado supra, dependen de la voluntad y los sesgos de los usuarios. En este sentido, se podría crear un sistema mixto, basado en modelos como los *Community Notes*, pero con una cierta profesionalización. Es decir, se podrían crear sistemas de "**verificadores ciudadanos certificados"** que reciban formación específica en ética digital, análisis crítico y desinformación. Así, su trabajo sería incentivado y supervisado públicamente para mantener su legitimidad y transparencia (Graves, L. (2018). *Understanding the Promise and Limits of Automated Fact-Checking*. Reuters Institute.).

También se puede crear un sistema de **huella digital informativa** basada en tecnología *blockchain* que registre el origen, ediciones y validaciones de una noticia o imagen. Esto permitiría verificar la autenticidad y trazabilidad de los contenidos compartidos, impidiendo su manipulación sin registro. Autores como Lemieux (Lemieux, V. L. (2016). Trusting Records: Is Blockchain Technology the Answer? *Records Management Journal*, 26(2), 110–139) sostienen que blockchain puede aumentar la confianza en la información mediante registros inalterables.

Las plataformas deberían estar legalmente obligadas a señalar de forma clara cualquier contenido generado o modificado con IA (*deepfakes*, textos automatizados, imágenes sintéticas), utilizando un lenguaje comprensible, visualmente prominente y sin ambigüedades, tal y como se obliga a los *influencers* a especificar cuando están generando contenido publicitario. Esto evitaría que los usuarios sean engañados por representaciones artificiales sin saberlo, que es la problemática principal con los *deepfakes.*

Por último, la creación de organismos independientes que auditen periódicamente los algoritmos de recomendación de las plataformas, incluyendo participación ciudadana y representación académica. Estas auditorías deberían ser públicas y obligatorias por ley, tal como sugiere la European Commission en su Digital Services Act (2022).

8. CONCLUSIONES FINALES

Este trabajo ha evidenciado que la IA tiene un rol ambivalente en el fenómeno de la desinformación digital: si bien es capaz de detectar contenidos manipulados, su uso actual - predominantemente dirigido a la personalización y la rentabilidad - favorece la difusión masiva de contenidos falsos, tendenciosos o polarizadores. Los algoritmos de recomendación operan bajo una lógica de maximización del tiempo de atención del usuario, mostrando contenidos que refuerzan sus creencias previas y evitando aquellos que podrían cuestionarlas. Esto genera cámaras de eco y burbujas de filtro, donde los usuarios se aíslan informativamente y se alejan de posiciones críticas o contrarias, lo cual debilita la deliberación democrática y refuerza la polarización social.

Además, el desarrollo de tecnologías como los *deepfakes* y los *bots* automatizados plantea riesgos inéditos. La generación de evidencia audiovisual falsa no solo mina la confianza en la información pública, sino que dificulta enormemente la verificación y puede emplearse para manipular procesos electorales, desprestigiar figuras públicas o promover discursos de odio. Del mismo modo, los *influencers* virtuales gestionados por IA pueden inducir creencias o comportamientos sin que los usuarios sean conscientes de que están interactuando con una entidad no humana, lo que vulnera principios éticos básicos como la transparencia y el consentimiento informado. Los resultados del estudio de campo realizado refuerzan estas preocupaciones: sólo el 0,5% de los estudiantes encuestados identificó correctamente todas las noticias presentadas, un 47,5% confundió al menos tres *fake news* como reales, y un alarmante 78% marcó como falsas dos noticias verdaderas. Esto refleja una crisis de pensamiento crítico, especialmente preocupante entre una generación que consume información casi exclusivamente en redes sociales. La apariencia visual y la viralidad de los contenidos pesan más que su veracidad, y existe una clara falta de competencias para evaluar fuentes y verificar información.

Desde un punto de vista ético y regulatorio, esto exige una actuación inmediata. Es necesario establecer políticas públicas que obliguen a las plataformas a transparentar sus algoritmos y ofrecer a los usuarios herramientas claras para distinguir entre contenido real y manipulado. De igual modo, debe impulsarse la alfabetización mediática desde edades tempranas, no solo para enseñar a "buscar en Google", sino para comprender cómo se construyen los discursos digitales y qué intereses hay detrás de cada recomendación. La educación crítica debe ser tan prioritaria como la alfabetización matemática o lingüística en la sociedad actual.

Finalmente, este TFG invita a una reflexión más profunda sobre el modelo de sociedad que estamos construyendo. ¿Queremos una ciudadanía que consuma contenido filtrado por sistemas opacos que priorizan el beneficio económico? ¿Es legítimo que algoritmos, cuyo funcionamiento no controlamos, condicionen nuestras opiniones políticas, nuestras emociones o nuestras decisiones de voto? ¿Es compatible esta dinámica con la noción clásica de libertad de pensamiento y deliberación racional? La respuesta no es sencilla.

La IA, como toda tecnología, no es intrínsecamente negativa ni positiva: su impacto depende y dependerá de cómo se regule, de qué valores se prioricen en su diseño, y de la educación que los ciudadanos reciban para interactuar con ella. Si no actuamos pronto, corremos el riesgo de que la verdad se convierta en un producto más, sometido a los mismos algoritmos que hoy deciden qué música escuchamos o qué anuncios vemos. Pero si somos capaces de integrar la ética, la transparencia y la educación crítica en su desarrollo, la IA puede convertirse en una herramienta poderosa para una sociedad más informada, plural y libre.

9. BIBLIOGRAFÍA

Acosta, D. G., & Masjuán, M. E. G. (2022). *Fake news en tiempos de posverdad. Análisis de informaciones falsas publicadas en Facebook durante procesos políticos en Brasil y México 2018*. Estudios sobre el mensaje periodístico.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236.

Brennen, J. S., Simon, F., Howard, P. N., & Nielsen, R. K. (2020). *Types, sources, and claims of COVID-19 misinformation*. Reuters Institute, University of Oxford.

Bozdag, E., & van den Hoven, J. (2015). Breaking the filter bubble: Democracy and design. *Ethics and Information Technology*, 17(4), 249–265.

Castillo Martínez, K., Aguilar Rodríguez, J. A., & Madrigal Rentería, A. S. (2024). Desafíos éticos de la inteligencia artificial generativa en las nuevas formas organizacionales. *Revista Digital de Tecnologías Informáticas y Sistemas*, 8(1).

Ceruzzi, P. E. (2003). *A history of modern computing* (2ª ed.). MIT Press.

Chan, M., Estève, D., Escriba, C., & Campo, E. (2008). A review of smart homes—Present state and future challenges. *Computer Methods and Programs in Biomedicine*, 91(1), 55–81.

De Francisco, A. L. M. (2021). Innovación tecnológica: La inteligencia artificial (IA). *Nefrología al Día*.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3), 554–559.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96–104.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Graves, L. (2018). *Understanding the Promise and Limits of Automated Fact-Checking*. Reuters Institute.

Gurney, J. (2013). *Robots: From Science Fiction to Technological Fact*. MIT Press.

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2017). *Junk News and Bots during the U.S. Election: What Were Michigan Voters Sharing?* Oxford Internet Institute.

Hull, C. W. (1986). Apparatus for production of three-dimensional objects by stereolithography (U.S. Patent No. 4,575,330). U.S. Patent and Trademark Office.
Kagermann, H., Wahlster, W., & Helbig, J. (2013). *Recommendations for implementing the strategic initiative INDUSTRIE 4.0*. Final report of the Industrie 4.0 Working Group.

Lee, J., Bagheri, B., & Kao, H. A. (2014). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3, 18–23.

Lemieux, V. L. (2016). Trusting Records: Is Blockchain Technology the Answer? *Records Management Journal*, 26(2), 110–139.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence*. Dartmouth College.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mavridis, N., & Kameas, A. (2020). Artificial agents as social influencers: the effect of virtual characters on human behavior. *AI & Society*, 35(3), 639–651.

Newell, A., & Simon, H. A. (1956). The logic theory machine – A complex information processing system. *IRE-Transactions on Information Theory*, 2(3), 61–79.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Painter, J. (2013). *Climate change in the media: Reporting risk and uncertainty*. I.B. Tauris.

Pennycook, G., & Rand, D. G. (2018). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings. *Management Science*, 66(11), 4944–4957.

Schwab, K. (2016). *The Fourth Industrial Revolution*. Crown Business.

Shawar, B. A., & Atwell, E. (2007). Chatbots: Are they really useful? *LDV Forum*, 22(1), 29–49.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining “Fake News”: A typology of scholarly definitions. *Digital Journalism*, 6(2), 137–153.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.

Weizenbaum, J. (1966). ELIZA — A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.

Woolley, S. C., & Howard, P. N. (2016). Automation, algorithms, and politics: Political communication, computational propaganda, and autonomous agents. *Oxford Internet Institute*.