30

INFORME ESPAÑA 2023

CÁTEDRA
JOSÉ MARÍA MARTÍN
PATINO DE LA CULTURA
DEL ENCUENTRO

COMILLAS
CIHS

Servicio de Biblioteca. Universidad Pontificia Comillas de Madrid

INFORME España 2023 / Cátedra José María Martín Patino de la Cultura del Encuentro ; [coordinación y edición Agustín Blanco, Sebastián Mora y José Antonio López-Ruiz]. --

Madrid : Universidad Pontificia Comillas, Cátedra J.M. Martín Patino, 2023.

508 p.

En la portada: 30.

Es continuación de la colección CECS publicada por la Fundación Encuentro ISSN 1137-6228.

D.L. M 32106-2023. -- ISBN 978-84-8468-605-7

1. Democracia. 2. Situación social. 3. Inteligencia artificial. 4. Educación sexual. 5. Formación profesional. 6. Cambios climáticos. 7. Estado social. 8. España. I. Blanco Martín, Agustín, editor literario. II. López-Ruiz, José Antonio (1968-), editor literario. III. Mora Rosado, Sebastián (1966-), editor literario.

Coordinación y edición: Agustín Blanco, Sebastián Mora
y José Antonio López-Ruiz

Edita: Universidad Pontificia Comillas
Cátedra J. M. Martín Patino

ISBN: 978-84-8468-605-7
Depósito Legal: M-32106-2023

Imprenta Kadmos
Salamanca

Gracias a la Fundación Ramón Areces, la Cátedra José María Martín Patino de la Cultura del Encuentro elabora este informe. En él ofrecemos una interpretación global y comprensiva de la realidad social española, de las tendencias y procesos más relevantes y significativos del cambio.

El informe quiere contribuir a la formación de la autoconciencia colectiva, ser un punto de referencia para el debate público que ayude a compartir los principios básicos de los intereses generales.

ÍNDICE

PARTE TERCERA: DESARROLLO E INTEGRACIÓN SOCIAL

Capítulo 3
EL ESTADO DE BIENESTAR Y SU FINANCIACIÓN EN LA TERCERA DÉCADA DEL SIGLO XXI

Jesús Ruiz-Huerta y Javier Loscos

Capítulo 4
LAS PERSONAS SIN HOGAR EN ESPAÑA: EL ALOJAMIENTO Y LA VIVIENDA COMO DERECHO SOCIAL

Pedro José Cabrera Cabrera

Capítulo 5
LA TRANSICIÓN NECESARIA EN LA GESTIÓN DE LA SALUD: DE LA GESTIÓN DE PERSONAS A LA GESTIÓN DE POBLACIONES

Ángel Asúnsolo del Barco

Parte Segunda
LA REVOLUCIÓN DE LA INTELIGENCIA ARTIFICIAL

Sara Lumbreras
Universidad Pontificia Comillas

Alex Rayón
Universidad de Deusto

1. El nuevo pincel

El año pasado, 2022, fue el año en el que la Inteligencia Artificial (IA) sorprendió al mundo mostrando, por primera vez, una creatividad que fuimos incapaces de distinguir de la humana. Desató un sinfín de titulares lo ocurrido en la Feria de Arte del Estado de Colorado (Roose, 2022), certamen en el que resultó ganador un cuadro titulado *Teatro de la Ópera Espacial*, donde se mostraba una escena evocadora en la que personajes ataviados con ropajes entre lo futurista y lo medieval se asoman a una inmensa ventana sobre un escenario de proporciones planetarias. La luz se difumina en el polvo de la enorme sala y parece posible intuir pinceladas precisas pero fluidas.

Los jueces la escogieron, sin dudar, como la ganadora en la categoría de arte digital, en la que artistas humanos emplean todo tipo de medios, tanto físicos (posteriormente digitalizados) como informáticos, para modificar las imágenes respondiendo a su intención creativa. Lo que los jueces no sabían es que el artista que firmaba la obra, Jason M. Allen, no había llegado siquiera a tomar un pincel o a tocar un ratón. *Teatro de la Ópera Espacial* había sido completamente realizado por una IA generativa, en particular, Midjourney. Este *software* permite la generación de imágenes a través de la descripción de un texto, que puede incluir no sólo la especificación objetiva del contenido sino también características de su estilo –por ejemplo, realista, abstracto o imitando a un pintor particular– (Borji, 2022).

Una vez se conoció este hecho, fueron muchos los que le reclamaron a Allen que se retirase de la competición y reintegrase el dinero del premio. Sin embargo, a través de sus redes sociales, Allen explicó que la obra era el resultado de un trabajo laborioso en el que había empleado casi cien horas creando más de 900 bocetos, en los que fue refinando poco a poco la descripción que le enviaba a la IA como especificaciones de la imagen, lo que conocemos como *prompt*. En una de sus declaraciones explicaba: "La IA no es más que una herramienta, de la misma manera que un pincel es una herramienta. Es necesaria una fuerza creativa detrás de la herramienta".

Los jueces no retiraron el premio y varios de ellos aclararon públicamente que se mantenían firmes en su decisión, dado que no se habían

contravenido las normas del concurso (al menos, las de ese año; habrá que ver si se modifican para ediciones posteriores). Además de generar un intenso debate sobre quién debe ser propietario de los derechos de una obra generada por IA (parece que el consenso apunta a que será dueño el que haya ideado el *prompt* que inspiró a la máquina), *Teatro de la Ópera Espacial* ha sido el origen de una reflexión mucho más profunda sobre las posibilidades de la creatividad en la IA y el impacto que las IA generativas están empezando ya a tener.

Midjourney no es la única IA generativa capaz de sintetizar imágenes nuevas. A ella se unen otras como Dall-E 2 y Stable Diffusion, que pueden realizar labores básicas de ilustración siguiendo las indicaciones del usuario (Borji, 2022). En 2022 empezaron a estar disponibles versiones gratuitas de estas herramientas, generando una oleada de aplicaciones. Pueden generarse bocetos de manera inmediata siguiendo un estilo predeterminado, y las situaciones que aún se les resisten están en proceso de mejorarse. Por el momento, la IA tiene problemas para representar rostros, manos o texto, diseña con frecuencia espacios incoherentes y se ve incapaz de seguir indicaciones lógicas (por ejemplo, sobre la posición relativa de varios elementos).

Además de imágenes, la IA genera música. Aplicaciones como Amper, Dadabots y MuseNet (Manovich, 2019) componen melodías para varias voces siguiendo el estilo definido por el usuario. Como en los ejemplos anteriores, la idea es utilizar estas piezas como punto de partida que pueda ser refinado después por un compositor o equipo de compositores humano. Dado que la información se genera en formato electrónico, resulta sencillo realizar este proceso de refino desde la comodidad de un ordenador.

A estas se unen varias herramientas de audio, como Descript, Listnr o Podcast.ai, que permiten editar audio de la misma manera que se editaría un texto, con lo que se pueden generar grabaciones para *podcasts* de manera extremadamente rápida. Listnr es capaz de sintetizar voz a partir de un texto, dándole las características que sean necesarias, como por ejemplo parecerse a una voz en particular. La iniciativa Podcast.ai va un paso más allá y genera el contenido de un *podcast* de manera completamente automatizada, entrenándose con la información disponible *online* y empleando síntesis de voz a partir de texto para completarlo. Los usuarios pueden compartir ideas para los siguientes capítulos, que, por ejemplo, han incluido una entrevista a Steve Jobs después de su muerte.

Las empresas que se dedican al diseño de chips, como Synopsys, Cadence, Google o Nvidia, utilizan ya también IA generativa para acelerar esta tarea.

Pero la verdadera revolución está sucediendo en el mundo de la palabra. La palabra humana y la de la máquina, el código con el que nos comunicamos con ella. Desde que han aparecido, las herramientas de generación de código como CodeStarter, Codex, GitHub Copilot y Tabnine han conseguido reducir drásticamente el tiempo necesario para desarrollo de código. Copilot, por ejemplo, se define como un acompañante en el proceso de codificación que actúa como un traductor entre las instrucciones del programador, que ahora puede expresarse directamente en su propio lenguaje natural, y el código, que es lengua de la máquina (Barke, James y Polikarpova, 2023). Chat GPT puede también proporcionar alternativas sobre posibles maneras de codificar una instrucción dada y proporcionar pistas sobre cómo resolver los errores que aparezcan en su implementación.

Precisamente es ChatGPT la IA generativa que más ha sorprendido. Es capaz, entre otras cosas, de generar texto a partir de una especificación del usuario expresada en lenguaje natural; por ejemplo: "escribe una redacción de 1.000 palabras sobre el cambio climático", o "escribe un poema sobre el cambio climático en un estilo parecido al de Shakespeare". Esta herramienta ha conseguido sobrecoger al mundo: los profesores se desvelan pensando cómo detectar los trabajos tramposos, los escritores se afanan en aprender a utilizarla para su beneficio, las empresas la conectan para revolucionar su atención al cliente (George y George, 2023).

La limitación principal de ChatGPT es que es una IA que está pre-entrenada con datos hasta 2021, con lo que no tiene acceso a información que haya aparecido después. Aparece junto con otras IA generativas de texto, como Jasper, AI-Writer y Lex, además de otras alternativas como Chatsonic, Bing o Google Bard, que tienen acceso a Internet, por lo que son capaces de responder a preguntas sobre conocimiento actual. Sin embargo, además de ciertos problemas en su lanzamiento, no han sido capaces de llegar a los millones de usuarios que se han dejado sorprender por ChatGPT (100 millones en apenas dos meses desde su lanzamiento). Mucho de lo que se comentará en esta sección estará centrado en ChatGPT y sus "hermanas", por ser las que han marcado un antes y un después en el año 2022. Es clave comprender la IA y la IA generativa en particular, para vislumbrar los cambios que se avecinan.

Existen otras herramientas dentro de contextos más especializados, como Elicit, que se centra en seleccionar los artículos académicos que están más relacionados con un tema definido por el usuario. Además, es capaz de encontrar críticas de unos artículos en otros, de tal manera que reduce de manera drástica el tiempo que es necesario emplear para realizar una revisión de la literatura detallada. Como discutiremos más tarde, la combinación de estas herramientas es lo que verdaderamente nos abrirá al tremendo potencial de la IA en el futuro más próximo.

La expectación que las IA generativas han generado, valga la redundancia, es enorme y continúa creciendo. La consultora Gartner, conocida por ser una de las mejor posicionadas en cuanto a tecnologías de la información, incluyó la IA generativa en 2021 (antes del tsunami que vivimos el año pasado) en su informe *Radar de impacto de tendencias y tecnologías emergentes para 2022* (Nguyen, Reynolds & Kandaswamy, 2021) como una de las tecnologías clave que revolucionarán la productividad en el corto plazo.

Estas fueron algunas de sus predicciones más detalladas:

- En 2025, la IA generativa producirá el 10% de todos los datos utilizados de manera global en todas las industrias (ahora es menos del 1%). Este porcentaje aumentará al 20% si nos centramos únicamente en los datos utilizados para realizar pruebas orientados al servicio al consumidor.

- También en 2025, la IA generativa se utilizará en el 50% de los proyectos de desarrollo de nuevos fármacos.

- En 2027, el 30% de los fabricantes industriales utilizará la IA generativa de un modo u otro.

Es muy posible que estas predicciones se hayan quedado cortas. En este capítulo del *Informe España* intentaremos proporcionar una perspectiva sencilla sobre estos avances y sus implicaciones, así como una reflexión más profunda sobre lo que podemos esperar en el futuro, con sus mejores y peores consecuencias, así como guías para intentar extraer los potenciales positivos evitando las amenazas más importantes. Esta reflexión comienza por proporcionar una perspectiva general del estado de la IA en el mundo y en España.

2. ¿Dónde está la IA?

La Feria de Arte de Colorado no es la primera vez que la tecnología crea una disrupción en el mundo del arte. En 1839 irrumpió en Francia el daguerrotipo, el primer proceso fotográfico de éxito comercial. Por aquel entonces, los que querían tener un retrato de sí mismos (en nuestros tiempos lo llamaríamos *selfie*), tenían que encargárselo a un artista, que lo pintaba a mano. Este nuevo dispositivo automatizó el proceso, sacando fotografías automáticamente y mejorando incluso la calidad del resultado. Se produjeron importantes revueltas, especialmente por parte de los retratistas. Sin embargo, otros, convencidos de que el cambio venía para quedarse, evolucionaron. Para luchar contra la objetividad que traía una máquina, un grupo de artistas crearon el impresionismo. Plasmar la luz y el instante fue una forma de llevar la innovación de la expresión artística a un nuevo campo. Este movimiento marcó el paso a la Modernidad en el arte.

La historia del daguerrotipo nos resulta familiar en un momento en la que la IA generativa hace correr ríos de tinta. Está claro que esta nueva tecnología va a traer importantes cambios sociales y económicos. Sin embargo, el primer mensaje que debe darse sobre esta tecnología es el miedo que le subyace: la sustitución del trabajo de una persona por una máquina que no sabemos muy bien cómo funciona. Tampoco sabían lo que hacía el daguerrotipo. Como ahora, casi parecía más magia que ciencia.

En la Revolución Industrial se sucedieron inventos que provocaron intensos desplazamientos laborales. Los procesos mecanizados de la industria alteraron una larga serie de trabajos anteriormente realizados por humanos. La máquina de vapor, inventada por James Watt en 1769, permitió el desarrollo de la industria del transporte, así como la mecanización de fábricas y minas. Esto cambió la forma en que se producían los bienes y eliminó la necesidad de transportarlos de forma manual. El telar mecánico, inventado por Edmund Cartwright en 1785, revolucionó la industria textil al automatizar el proceso de tejido. Esto llevó a una mayor producción y eficiencia, pero también provocó la pérdida de empleo para los tejedores manuales. La prensa de vapor, desarrollada por Charles Stanhope en 1801 y perfeccionada por Friedrich Koenig y Andreas Bauer en 1814, mejoró la producción y la velocidad de la impresión de libros, periódicos y otros materiales impresos. Esto transformó la industria editorial y redujo la necesidad de trabajadores manuales en la producción de impresiones. Con la invención de la locomotora de vapor por George Stephenson en 1814, el ferrocarril se convirtió en un medio de transporte clave durante la Revolución Industrial. Facilitó el movimiento de personas y mercancías, lo que permitió una mayor expansión industrial y la creación de nuevos empleos en la construcción de ferrocarriles y la operación de trenes.

Estos ejemplos de mecanización buscaban emular con un proceso mecánico una función propia del mundo animal o humano y alteraron lo que llamamos puestos de trabajo manuales (o "de cuello azul", en términos anglosajones). Las bautizamos como "revoluciones industriales". Lo que está ocurriendo ahora es que los puestos de trabajo de oficina ("de cuello blanco"), donde empleamos el cerebro para trabajar, están amenazados al emular la IA las funciones del cerebro humano. A esta revolución la podemos llamar cognitiva, y es la que estamos viviendo gracias a la IA.

La Revolución Industrial y la Revolución Cognitiva comparten similitudes en cuanto a la transformación de la economía y la automatización del trabajo. Sin embargo, tienen diferencias clave en el tipo de trabajos afectados y las tecnologías involucradas. En cuanto al tipo de trabajo afectado, la Revolución Industrial introdujo la mecanización y automatización de tareas manuales y físicas, típicamente en la manufactura, la agricultura y la minería. La Cognitiva implica el uso de la

IA para automatizar tareas como análisis de datos, toma de decisiones, planificación y comunicación.

En la Revolución Industrial, las tecnologías clave fueron la máquina de vapor, el telar mecánico, la línea de ensamblaje o la producción en masa. Estas innovaciones permitieron la producción eficiente de bienes y redujeron la necesidad de mano de obra en ciertos sectores. En la Revolución Cognitiva, las tecnologías clave incluyen la IA, el aprendizaje automático, la robótica avanzada y el análisis de grandes volúmenes de datos (*Big Data*). Estas tecnologías permiten automatizar tareas cognitivas y tomar decisiones basadas en datos, lo que reduce la necesidad de horas de personas leyendo, resumiendo información o proponiendo nuevas acciones.

La Revolución Industrial nos llevó a la urbanización, el crecimiento de la clase media y una mayor especialización laboral. Sin embargo, también provocó desempleo temporal, desigualdades económicas y problemas sociales como la explotación laboral y condiciones insalubres de trabajo. Los efectos de la Revolución Cognitiva todavía están por ver. Intuimos que tiene el potencial de aumentar la eficiencia y la productividad en sectores como la salud, la educación y los servicios financieros. No obstante, puede generar desempleo y desigualdades, así como plantear preocupaciones éticas y de privacidad relacionadas con el uso de datos y la toma de decisiones algorítmicas.

Tanto la Revolución Industrial como la Revolución Cognitiva nos han llevado a cambios significativos en la economía y la sociedad. Para comprender lo que se avecina es necesario reflexionar sobre la IA.

2.1. ¿Por qué hablamos ahora de IA?

La IA ha experimentado un crecimiento y desarrollo exponencial en las últimas décadas. Pese a que podemos contextualizar su origen en los años 50 del siglo pasado, su crecimiento actual se debe al incremento vertiginoso de la materia prima que precisa esta tecnología para funcionar: los datos. Estos grandes volúmenes de datos han hecho despegar una tecnología que llevaba décadas "dormida".

Y es que el 90% de los datos existentes en el mundo han sido creados en los últimos dos años. Es así una curva de crecimiento exponencial. En el tiempo que un lector promedio empleará en la lectura de este capítulo, se sacarán tantas fotografías en el mundo como todas las tomadas en el siglo XIX y parte del XX. Hay tres elementos que están provocando este crecimiento exponencial de los datos:

- *La capacidad de cálculo se ha abaratado*: fabricar ordenadores resulta ahora extremadamente barato. Con ellos se pueden almacenar, procesar y generar más datos que nunca. Lejos quedan los días en los que el cómputo era un factor limitante. El libro *La sociedad de coste marginal cero*, de Jeremy Rifkin (2014), nos mostró cómo, en la era digital, nos encontraríamos con bienes virtualmente gratuitos, y la creación y almacenamiento de datos es uno de ellos.

- *La tecnificación de la sociedad y su digitalización*: cada vez codificamos en objetos conectados a Internet más conductas o expresiones sociales. Así, los datos están cada vez más distribuidos en diferentes entornos. Los coches, las lavadoras, nuestra ropa o incluso nuestras paredes ahora adquieren capacidades de escucha y actuación, lo que hace que se generen cada vez más datos de todo ello. Datos que se quedan en dispositivos electrónicos fabricados por diferentes empresas.

- *La era de las redes sociales*: hace unos cuantos años, Manuel Castells escribió el libro *Comunicación y Poder* (2013). Habló de las redes sociales y su poder, y concretamente se refirió a las mismas como medios de autocomunicación de masas. Redes que implican interacción, comunicación y diálogo con nuestros "amigos". Para una empresa, se trata de un nuevo instrumento para fortalecer la imagen de marca, aumentar la fidelización de los clientes, mejorar la implicación de los empleados o conocer más sobre los deseos y tendencias de sus clientes. Para nosotros, sus productos, una fuente de ocio cuando subimos fotos o vídeos, las comentamos o enviamos mensajes. Las redes sociales que empleamos en nuestro día a día (Instagram, Twitter, TikTok, Facebook, Linkedin, etc.), son servicios "gratuitos". Pero nada es gratis. Las redes sociales comerciales, las que empleamos, funcionan como si fueran una televisión: el objetivo es generar datos sobre audiencias y comportamientos, para que luego puedan comercializar espacios de impacto a esas audiencias. Nos convertimos así en proveedores de datos mientras tenemos la falsa sensación de estar recibiendo un servicio.

Si algo ha producido la era digital es que el valor se genera de manera permanente, pero no siempre es aprovechado por el que lo genera. Los datos son un gran exponente de esta paradoja. Los "datos a la sombra" o datos "involuntarios" (acceso, búsquedas, lugares que frecuentamos, etc.) ofrecen una visión de nosotros que las empresas están aprovechando. Tanto es así que muchas de las aplicaciones de la IA se focalizan en aprovechar todo este conocimiento de nuestras vidas que expresan los datos.

Esta explicación sobre su situación es incompleta si no entendemos su origen. Tratar de emular las tareas del ser humano es un eterno sueño de la informática. Se trata de intentar recrear el aprendizaje, el razonamiento, la comprensión del lenguaje natural, la percepción, la resolución de problemas

y la adaptación a situaciones nuevas. Así, los avances en la IA y en el aprendizaje automático han permitido a los ordenadores replicar muchas de las funciones cognitivas humanas. Las funciones más relevantes para estar hablando hoy de esa Revolución Cognitiva son:

- *Procesamiento del lenguaje natural (NLP)*: los ordenadores pueden comprender, interpretar y generar texto en lenguajes humanos, como el análisis de sentimientos, la traducción automática y generación de resúmenes.

- *Reconocimiento de voz*: los sistemas de IA pueden transcribir y comprender el lenguaje hablado, como en asistentes virtuales (Siri, Google Assistant, etc.).

- *Reconocimiento de imágenes*: los ordenadores pueden identificar y clasificar imágenes de objetos y personas, lo que permite aplicaciones como etiquetado automático, sistemas de vigilancia y vehículos autónomos.

- *Razonamiento y toma de decisiones*: los sistemas de IA pueden resolver problemas complejos, tomar decisiones basadas en datos y realizar análisis predictivos.

- *Aprendizaje automático*: los ordenadores pueden aprender de los datos y mejorar su rendimiento en tareas específicas sin ser explícitamente programados para ello, utilizando algoritmos como redes neuronales, árboles de decisión y máquinas de soporte vectorial.

- *Juegos*: los sistemas de IA han demostrado ser altamente competentes en juegos de estrategia, como ajedrez, Go y póker, llegando a superar a campeones humanos en muchos casos.

- *Robótica*: la IA permite a los robots llevar a cabo tareas complejas, como manipulación de objetos, navegación y colaboración con humanos.

- *Creatividad*: los ordenadores también pueden generar contenido creativo, como música, arte y escritura, aunque la calidad y originalidad de estas creaciones aún se encuentra en debate.

- *Interacción social*: los sistemas de IA pueden simular la interacción social, como los chatbots[1] y asistentes personales.

No es de extrañar que cada vez haya más voces gritando que "la singularidad está cerca". Se trata de esa idea de Raymond Kurzweil (2005) que afirma que está cerca el momento en el que la tecnología supere todas nuestras expectativas, incluyendo la de crear IA que exceda las capacidades

[1] Un chatbot es un programa informático que utiliza inteligencia artificial (IA) y procesamiento del lenguaje natural (NLP) para comprender las preguntas de los clientes y automatizar las respuestas a dichas preguntas, simulando la conversación humana.

cognitivas de los humanos, una idea que sobrevuela el debate sobre la IA desde hace décadas; la historia de grandes expectativas alrededor de la IA no es nueva.

2.2. Breve historia de la IA

La IA, como campo de la informática que busca crear sistemas capaces de aprender, razonar y percibir de manera similar a los seres humanos, ha tenido una agitada historia tras su concepción a finales de la II Guerra Mundial. En 1950, Alan Turing propone el Test de Turing (Hodges, 2008) para determinar si una máquina puede imitar la inteligencia humana. Este test, que consiste simplemente en mantener una conversación de manera indistinguible de un humano, sigue siendo un criterio de referencia en la actualidad, si bien es cierto que con los nuevos sistemas de IA Generativa se comienzan a buscar nuevos instrumentos. Más tarde, en 1956, John McCarthy, Marvin Minsky, Nathaniel Rochester y Claude Shannon organizan la Conferencia de Dartmouth, donde se acuña el término "IA" (McCarthy, Minsky, Rochester y Shannon, 2006). En 1959, Arthur Samuel desarrolla un programa de aprendizaje automático para jugar a las damas que demuestra que las máquinas pueden mejorar su rendimiento mediante el aprendizaje (Samuel, 1959). En 1964, Danny Bobrow escribe su tesis de doctorado sobre el programa STUDENT, que puede resolver problemas algebraicos de texto en inglés (Bobrow, 1964). En 1969, Marvin Minsky y Seymour Papert publican *Perceptrones* (Marvin y Seymour, 1969), un libro que critica las limitaciones de las redes neuronales, uno de los principales algoritmos en los que se basa la IA, que se inspira en las neuronas que conforman el cerebro humano y que será presentado más adelante. Esto inicia un declive en su investigación durante las siguientes dos décadas. Este hecho trae el conocido como primer "invierno de la IA".

No es hasta 1980 cuando John Hopfield introduce las "redes de Hopfield", una forma de red neuronal que resuelve problemas de optimización (Hopfield, 1982). Es entonces cuando se vuelven a retomar algunas viejas ideas. En 1986, Geoffrey Hinton, David Rumelhart y Ronald Williams desarrollan el algoritmo de retropropagación para el entrenamiento de redes neuronales multicapa (Rumelhart, Hinton y Williams, 1986). Nuevamente se produce un estancamiento: no había datos suficientes para que estas tecnologías funcionaran. No se vuelve a oír hablar de todo ello hasta que, en 1997, IBM Deep Blue derrota al campeón mundial de ajedrez Garry Kasparov, demostrando el éxito de la IA en la resolución de problemas complejos (Hsu, Gupta y Sorooshian, 1995). Lo mismo ocurre cuando en 2012 el equipo de Alex Krizhevsky, Ilya Sutskever y Geoffrey Hinton gana el concurso de reconocimiento de imágenes ImageNet (ILSVRC) utilizando una red neuronal convolucional (CNN) profunda. Este hito marca el comienzo del *"deep*

learning" (Hinton, Srivastava, Krizhevsky, Sutskever y Salakhutdinov, 2012). Es la era actual. Una era que no se puede explicar sin los tres hitos que marcan la actualidad más radiante:

- 2014 – Google adquiere DeepMind, una empresa de IA que desarrolla el algoritmo AlphaGo (Jaderberg et al., 2016), que posteriormente derrota al campeón mundial de Go, Lee Sedol, en 2016. Alpha-Zero consiguió después derrotar a todas las IA precedentes en un gran abanico de juegos para los que la IA no necesitaba ni siquiera recibir las reglas.

- 2018 – OpenAI presenta GPT-2 (Ziegler et al., 2019), un modelo de lenguaje basado en la arquitectura Transformer (Vaswani et al., 2017) que muestra un rendimiento asombroso en tareas de generación de texto.

- 2020 – OpenAI lanza GPT-3 (Moradi, Blagec, Haberl y Samwald, 2021), un modelo de lenguaje más grande y potente. Comprende cada vez mejor al humano.

2.3. La respuesta a la IA desde distintos sectores

Esta breve pero intensa historia ha permitido que la IA se haya infiltrado en prácticamente todos los aspectos de nuestras vidas, desde la forma en que nos comunicamos hasta cómo trabajamos, aprendemos y tomamos decisiones.

Informes de McKinsey (Chui, Hall, Mayhew, Singla y Sukharevsky, 2022), AI Watch (Misuraca y Van Noordt, 2020), el Departamento de Comercio de EE.UU. e informes sobre el estado de la IA en España arrojan luz sobre los cambios significativos en la economía, la educación, la ética y la política. Desde una perspectiva económica, los informes de McKinsey y el Departamento de Comercio de EE.UU. destacan que la IA tiene el potencial de aumentar la productividad, impulsar la innovación y crear nuevos empleos. Sin embargo, también señalan que la adopción de la IA puede conducir a la pérdida de empleos en algunos sectores y a una brecha creciente en la distribución de la riqueza.

En términos de educación, el informe de Stanford sobre IA (Zhang et al., 2021) resalta la importancia de adaptar los currículos y programas de capacitación a las habilidades necesarias para prosperar en un mundo impulsado por la IA. Estos incluyen el pensamiento crítico, la creatividad y la resolución de problemas. Además, los informes de AI Watch y del estado de la IA en España (Moreno-Izquierdo, Navarro-Navarro, Núñez-Romero y Peretó-Rovira, 2022) hacen hincapié en la importancia de fomentar la investigación y el desarrollo en el campo de la IA, así como en promover la colaboración entre el sector público y el privado.

En cuanto a la ética y la política, los informes de Stanford (Zhang et al., 2021) y AI Watch analizan el papel de la IA en la toma de decisiones y la necesidad de garantizar la transparencia, la equidad y la responsabilidad. Para abordar estos temas, los informes sugieren la creación de marcos regulatorios y normativos que protejan los derechos humanos, la privacidad y la seguridad. Asimismo, subrayan la importancia de la cooperación internacional en la creación de estándares y en el intercambio de información en el ámbito de la IA. Esta necesidad de regulación se explorará de manera detallada en la sección que concluye este informe.

En resumen, el impacto de la IA en la sociedad es multifacético. Los informes que están apareciendo y se han analizado para este capítulo destacan la necesidad de abordar estos desafíos mediante políticas públicas, la adaptación de los sistemas educativos y la promoción de la colaboración entre sectores y países. En última instancia, el éxito en la gestión del impacto de la IA en la sociedad dependerá de cómo se equilibren los beneficios y riesgos para garantizar un futuro inclusivo y sostenible para todos. Para entender desde una lógica más práctica su impacto, puede ser bueno recorrer los ámbitos de nuestra vida donde nos movemos y actuamos, así como sectores de actividad económica concretos, de nuevo basándonos en los informes ya disponibles para dimensionar este impacto:

I. En la vida cotidiana. Ha transformado la forma en que nos comunicamos y utilizamos las redes sociales. Los algoritmos de aprendizaje automático han sido fundamentales en el desarrollo de sistemas de reconocimiento de voz, chatbots y asistentes virtuales, como Siri de Apple, Alexa de Amazon y Google Assistant. También en el entretenimiento. La IA puede enriquecer el contenido multimedia, crear nuevos géneros de videojuegos y asistir en la producción de películas y música. Además, ha permitido la personalización de los contenidos en las redes sociales, así como el filtrado de noticias falsas, aunque también ha generado polémica en cuanto a la privacidad y la manipulación de la información. Ha revolucionado el transporte y la movilidad mediante el desarrollo de vehículos autónomos y sistemas inteligentes de tráfico. Los vehículos autónomos, como los desarrollados por Tesla, Waymo y Uber, prometen reducir accidentes, mejorar la eficiencia energética y optimizar el flujo del tráfico. Por otro lado, los sistemas de tráfico inteligentes pueden monitorizar y regular el tráfico en tiempo real, reduciendo los atascos y mejorando la calidad del aire. Por último, han influido significativamente en el ámbito de la salud y la medicina, permitiendo avances en el diagnóstico de enfermedades, la personalización de tratamientos y el desarrollo de nuevos medicamentos. Algoritmos como los de DeepMind han demostrado su eficacia en la predicción de la estructura de las proteínas, lo cual puede acelerar el proceso de descubrimiento de fármacos. Además, la IA ha facilitado la telemedicina y la monitorización remota de pacientes, mejorando la atención médica y reduciendo costes.

II. En la economía y el empleo. Como posteriormente detallaremos, la IA está generando una intensa preocupación por la pérdida de empleos en la industria, el transporte o el comercio minorista. Sin embargo, también se espera que la IA cree nuevos empleos y oportunidades en campos como la ciberseguridad, el desarrollo de *software* o la ciencia de datos. Lo que parece por el momento es que los nuevos empleos no conseguirán compensar los puestos perdidos. Es quizás este el plano donde más foco debiéramos poner, y en el que profundizaremos en secciones posteriores.

III. En la educación. Existen numerosas aplicaciones de la IA en la enseñanza y la tutoría personalizada. Por ejemplo, permite una mejor y más eficiente evaluación y seguimiento del aprendizaje de los estudiantes. También ha democratizado la educación, posibilitando el acceso al conocimiento de amplios sectores de la población mundial.

IV. En la energía, el medio ambiente y la sostenibilidad. La IA ha contribuido al aumento de la eficiencia energética y a la reducción de emisiones al optimizar la generación, distribución y consumo de energía. También ha impulsado la adopción de tecnologías de energía limpia y la integración de sistemas de energía inteligente. Existen también numerosas aplicaciones de IA en la monitorización y la gestión de recursos naturales. Su uso en la predicción y mitigación de desastres naturales, el modelado de la contribución a la lucha contra el cambio climático y la conservación de la biodiversidad, nos permiten afirmar que la implementación de IA para la sostenibilidad y la mejora del proceso de medición de la huella energética es una realidad cada vez más cercana. Esto debe matizarse teniendo en cuenta también el gran gasto energético que supone entrenar la IA.

V. En la industria. La IA ha aumentado la eficiencia en la producción a través de la automatización y la optimización de procesos. Las tecnologías de IA, como la visión artificial y la robótica, han permitido mejorar la calidad del producto, reducir el tiempo de inactividad y minimizar el desperdicio.

VI. En los servicios financieros y la FinTech. La IA ha propuesto soluciones que permiten mejorar la toma de decisiones en inversiones, detectar fraudes y optimizar la gestión de recursos y riesgos. Además, la IA ha transformado la banca y los servicios financieros al permitir la evaluación automática de riesgos crediticios, la personalización de productos y servicios y la adopción de asesores financieros automatizados.

VII. En el sector primario. La IA ha impulsado la agricultura de precisión al ayudar a los agricultores a tomar decisiones informadas sobre el uso de recursos –como agua y fertilizantes– y a monitorizar la salud de los cultivos en tiempo real. Las soluciones de IA también han facilitado la automatización de tareas agrícolas y la prevención de enfermedades en cultivos y ganado.

VIII. En el comercio minorista. La IA ha mejorado la experiencia del cliente al personalizar recomendaciones de productos, facilitar la gestión de inventario y optimizar la cadena de suministro. También ha impulsado la adopción de nuevas tecnologías, como chatbots y asistentes virtuales, para brindar soporte al cliente.

Podemos así concluir que la IA está provocando cambios importantes en nuestras vidas. Esto no sería posible sin una serie de factores y avances tecnológicos. El procesamiento masivo de datos permite analizar grandes cantidades de información en poco tiempo, lo que nos ayuda a tomar decisiones mejor informadas en diversos campos. Como decíamos antes, además, ha permitido la automatización de procesos y tareas repetitivas, lo que ha aumentado la eficiencia en las industrias y ha liberado a los humanos para dedicarse a tareas más creativas y cognitivamente exigentes.

Todos estos cambios plantean desafíos y oportunidades que debemos abordar y aprovechar para mejorar nuestra calidad de vida y adaptarnos a un mundo en constante evolución. Para aprovechar estas oportunidades y evitar las amenazas debemos guiar el desarrollo tecnológico. Una tecnología sin un propósito alineado con la mejora de nuestras condiciones de vida no es un camino razonable.

2.4. El impacto laboral de la IA

Como hemos venido señalando, la adopción de tecnologías de IA en las empresas puede mejorar la productividad, la eficiencia y la competitividad, lo cual puede impulsar el crecimiento económico a nivel nacional e internacional. No obstante, esto también puede conducir a una brecha tecnológica entre países desarrollados y en desarrollo, así como entre empresas grandes y pequeñas.

Las principales potencias en IA incluyen a Estados Unidos, China, Canadá y el Reino Unido, todos los cuales han realizado inversiones significativas en investigación y desarrollo. Según el informe de la Universidad de Stanford *The AI Index 2021 Annual Report* (Zhang et al., 2021), y considerando métricas como la inversión en IA, la adopción en diferentes sectores y el progreso en investigación y desarrollo, se prevé que el impacto en la sociedad, la economía y la ética siga creciendo. Otros informes, como *AI now report* de la Universidad de Nueva York (Whittaker et al., 2018), introducen también una reflexión sobre la equidad, privacidad y justicia como elementos que considerar a escala mundial. El sesgo en los algoritmos de IA y los desafíos en la regulación y supervisión seguirán siendo de alta importancia y relevancia. Todos estos problemas de ámbito ético serán contemplados en profundidad en la sección conclusiva de este capítulo.

Europa está trabajando duro para fortalecer su posición en el ámbito de la IA. La Comisión Europea ha presentado estrategias y regulaciones para fomentar la innovación en IA y garantizar un enfoque ético y transparente. La inversión en investigación y desarrollo en IA ha aumentado, y se han creado alianzas y colaboraciones entre instituciones académicas, empresas y Gobiernos. Países como Alemania, Francia y el Reino Unido son los claros líderes en nuestra región. El *Informe sobre la inteligencia artificial en la era digital* (Voss, 2022) presenta la estrategia y las políticas de la Unión Europea en relación con la IA. Discute cómo Europa puede fomentar la innovación en IA y garantizar un enfoque ético y transparente en la adopción de tecnologías de IA. Y desde ahí sitúa su aportación a una carrera tecnológica que a todas luces será rápida. En los informes europeos se enfatiza la necesidad de acompañar la formación de las personas para mantener el ritmo de avance vertiginoso que se espera.

España ha estado haciendo esfuerzos considerables para mejorar su posición en el ecosistema global de la IA. El Gobierno ha lanzado la *Estrategia Nacional de IA* (del Olmo, 2022), que busca impulsar la investigación, el desarrollo y la adopción de IA en la economía y la sociedad. Se ha fomentado la colaboración entre universidades, centros de investigación y empresas para desarrollar soluciones basadas en IA en diversos campos, como la energía, el turismo, la salud y la movilidad –los campos donde la economía española es ya sólida–. Además, se han establecido iniciativas para capacitar a los trabajadores en habilidades relacionadas con la IA y promover la diversidad e inclusión en el sector.

En realidad, esta preocupación por el impacto de la IA a nivel social y económico no es nueva. Frey y Osborne (2017) se preguntaron por la transformación del empleo y la convivencia con las máquinas en el día a día. Arntz, Gregory y Zierahn (2016) y Susskind y Susskind (2015) teorizaron sobre el tipo de tareas que la tecnología transformaría. El economista del MIT David Autor (2015) investigó extensamente el impacto de la IA y la automatización en el mercado laboral. Según sus estudios, los trabajos pueden dividirse en tres categorías principales: trabajos rutinarios, trabajos no rutinarios y trabajos creativos.

- *Trabajos rutinarios*: Son aquellos que siguen un conjunto de reglas y procedimientos estandarizados, y pueden ser manuales o cognitivos. Ejemplos de trabajos rutinarios incluyen a los trabajadores de ensamblaje, los cajeros y, en muchos casos, el procesamiento de datos. La IA y la automatización han tenido un gran impacto en este tipo de trabajos, reemplazando a muchos trabajadores debido a la eficiencia, rapidez y precisión de los sistemas automatizados.

- *Trabajos no rutinarios*: Requieren habilidades de pensamiento crítico, resolución de problemas y adaptabilidad. Incluyen profesiones como médicos, abogados, gerentes y profesionales de ventas. La IA ha tenido un impacto mixto en este tipo de trabajos. En algunos casos, los sistemas de IA pueden mejorar la productividad y la calidad del trabajo al proporcionar información y análisis más precisos, lo que permite a los trabajadores humanos tomar decisiones mejor informadas. En otros casos, la IA puede reemplazar algunas tareas específicas, pero no todo el trabajo en sí, sino de algunas tareas. En este contexto, se requiere que los trabajadores humanos se adapten y adquieran nuevas habilidades para mantenerse relevantes.

- *Trabajos creativos*: Implican la creación de ideas, arte, literatura, música y otras expresiones artísticas o intelectuales. Incluyen artistas, escritores, músicos e innovadores. La IA también está influyendo en este tipo de trabajos, pero en lugar de reemplazar a los trabajadores humanos, a menudo se utiliza como una herramienta para mejorar la creatividad y la innovación. Por ejemplo, la IA puede ayudar a generar ideas, identificar patrones y proporcionar inspiración. Sin embargo, la intuición, la empatía y la capacidad de entender y responder a las emociones humanas siguen siendo dominios en los que los humanos tienen una ventaja significativa.

En resumen, el impacto de la IA en los tipos de trabajos varía según el nivel de rutina y las habilidades requeridas. La IA y la automatización han reemplazado muchos trabajos rutinarios, mientras que han transformado y mejorado la productividad en trabajos no rutinarios y creativos. Como resultado, es fundamental que los trabajadores y las instituciones educativas se adapten a estas transformaciones para garantizar que los empleados desarrollen habilidades relevantes para el futuro del trabajo.

Un reciente informe de Goldman Sachs (Hatzius, Briggs, Kodnani y Pierdomenico, 2023) afirma que hasta 300 millones de puestos de trabajo podrían contener tareas donde la IA tendría impacto. Todos los trabajos que tengan algo que ver con la búsqueda, síntesis y generación de nueva información están en riesgo de ser asumidos por una máquina. Sólo la originalidad no es automatizable.

2.5. Los nuevos modelos en la IA

ChatGPT (y los grandes modelos de lenguaje) ha sorprendido al mundo con su capacidad para generar respuestas coherentes (al menos, aparentemente) a partir de preguntas o instrucciones en lenguaje natural, escritas de la misma manera que si se le diera una descripción de trabajo a un ayudante humano. Es una herramienta única que utiliza técnicas avanzadas de IA para "comprender" (ya discutiremos más tarde el porqué de este entrecomillado, pero podemos

adelantar que la única manera en la que podemos definir una comprensión en la máquina es meramente metafórica) y procesar el lenguaje natural de manera más efectiva que cualquier otra herramienta disponible en la actualidad.

Desde su lanzamiento, ChatGPT ha sido rápidamente adoptado por cientos de millones de usuarios (como decíamos, más de 100 millones de usuarios en dos meses, la velocidad más rápida de adopción de una tecnología en toda la historia). Se ha aplicado a una amplia gama de usos, desde la atención al cliente hasta la educación (pocos son los estudiantes que aún no tienen cuenta en OpenAI, su empresa madre). Su capacidad para aprender y mejorar continuamente lo hacen aún más atractivo para aquellos que buscan una herramienta de procesamiento de lenguaje natural precisa y eficiente.

ChatGPT resume textos, contesta preguntas y crea textos que respondan a una intención o sigan una determinada estructura y pautas estilísticas. Es incluso capaz de escribir en verso. Además, ChatGPT genera respuestas en tiempo real, algo muy útil en situaciones en las que se necesita una respuesta rápida. En el campo de atención al cliente, por ejemplo, ChatGPT puede proporcionar respuestas precisas y coherentes a las preguntas comunes de los clientes en tiempo real, lo que reduce el tiempo de espera y mejora la experiencia del cliente en general.

En este último caso, como en otras muchas aplicaciones, ChatGPT no se utiliza de manera aislada, sino en conjunto con otra herramienta que es la que almacena las pautas de cómo solucionar los problemas de los clientes. ChatGPT se utiliza como un paso intermedio en la comunicación entre la máquina y el ser humano, siendo capaz de interpretar las preguntas de los clientes en el lenguaje de la máquina y de adaptar sus respuestas para que sean lo más útiles posibles.

Sabemos también que ChatGPT, como otras de sus herramientas hermanas, tiene defectos considerables. Los exploraremos en una sección dedicada expresamente a ellos, pero, entre los más importantes están no disponer de conocimiento actualizado, generar respuestas falsas pero aparentemente verdaderas cuando no tiene suficiente información (“alucinar”) o los sesgos que pueden aparecer en sus contestaciones. Además, presenta considerables problemas a la hora de comprender su posible rol en la educación. Discutiremos estas cuestiones en las secciones siguientes.

2.6. Pero, ¿cómo funcionan exactamente estas tecnologías y por qué han aparecido justo ahora?

Los últimos años han mejorado sensiblemente la capacidad de cálculo y los algoritmos que podían ponerse al servicio del aprendizaje automático. En esta sección presentamos de manera introductoria los desarrollos más

relevantes, recomendando al lector experto que pase directamente a la sección siguiente.

El aprendizaje automático consiste en algoritmos (una serie de pasos que se aplican a unos datos de entrada) y modelos que permiten a las máquinas aprender y mejorar a partir de los datos de manera independiente a la comprensión de un ser humano. Los dos tipos principales de aprendizaje automático son el supervisado y el no supervisado, junto con el aprendizaje por refuerzo.

El aprendizaje supervisado es un tipo de aprendizaje automático en el que se proporciona al modelo un conjunto de datos que han sido etiquetados, es decir, un conjunto de datos que ya han sido clasificados conforme a unas etiquetas, normalmente por un experto humano. El modelo utiliza estos datos para aprender a predecir la etiqueta correcta de nuevos datos. Los algoritmos de aprendizaje supervisado se dividen en dos tipos principales: clasificación y regresión. La clasificación es un tipo de aprendizaje supervisado en el que el modelo aprende a predecir una etiqueta discreta (también llamada categoría) para un conjunto de datos. Por ejemplo, un modelo de clasificación puede ser entrenado para predecir si una imagen es un gato o un perro en función de las características de la imagen. Por otro lado, la regresión es un tipo de aprendizaje supervisado en el que el modelo aprende a predecir un valor numérico para un conjunto de datos. Por ejemplo, un modelo de regresión puede ser entrenado para predecir el precio de una casa en función de sus características, como la ubicación, el número de habitaciones o la superficie total.

Algunos de los ejemplos principales de algoritmos de aprendizaje supervisados vienen directamente de la estadística clásica. La única diferencia es que ahora podemos procesar una cantidad de datos y variables mucho mayor que en los modelos clásicos. Por ejemplo, en regresión lineal se predicen los valores de una variable dependiente a partir de la mejor línea recta que pueda ajustarse a los datos. Otros tipos de regresión pueden acomodarse a otras relaciones entre los datos y la variable de salida que sean más complejas. Los árboles de decisión son un tipo de modelo de aprendizaje supervisado utilizado tanto para clasificación como para regresión. Un árbol de decisión es una estructura en forma de árbol que representa una serie de decisiones o preguntas que se deben contestar para acabar llegando a una determinada etiqueta o valor de salida. En un árbol de decisión, cada nodo representa una característica del conjunto de datos, y cada rama representa una decisión basada en esa característica. Por ejemplo, podemos tener un árbol que decida el riesgo de un paciente de necesitar hospitalización basado en la edad. Del nodo podrían salir dos ramas, una para menores de 80 y otra para mayores de 80 años. Los árboles de decisión se utilizan a menudo en problemas de clasificación y regresión debido a su simplicidad y facilidad de

interpretación: podemos comprender perfectamente cuáles son las decisiones que toma la máquina para clasificar el riesgo de los pacientes y examinarlas de manera racional. Por ejemplo, en este caso podríamos encontrar que realmente tiene sentido médico que los pacientes mayores experimenten un riesgo más elevado de hospitalización. Como enfatizaremos más adelante, esta capacidad de ser explicados resulta clave cuando reflexionamos sobre las implicaciones éticas del uso de algoritmos en el apoyo a las decisiones.

Por otro lado, el aprendizaje no supervisado es un tipo de aprendizaje automático en el que el modelo se entrena con un conjunto de datos que se reciben sin etiqueta. Aquí el reto es encontrar patrones y relaciones en los datos. El ejemplo principal de este tipo de aprendizaje es el agrupamiento (más conocido como *clustering* en inglés), que se utiliza para encontrar grupos de datos similares en un conjunto de datos. Los datos se agrupan en función de la similitud de sus características. Por ejemplo, se puede utilizar el agrupamiento para segmentar a los clientes de una tienda en grupos de acuerdo con su capacidad de compra. Además, podemos realizar lo que se conoce como reducción de dimensionalidad, que consigue extraer las características más importantes del conjunto de datos, reduciendo así su complejidad para mejorar la aplicación de otros algoritmos. Un algoritmo particularmente interesante es el análisis de componentes principales (PCA, por sus siglas en inglés), que encuentra los principales factores que explican la variabilidad en los datos. Las reglas de asociación sirven para hallar patrones frecuentes en un conjunto de datos. Por ejemplo, se puede utilizar para descubrir que las personas que tienen diabetes tienen también hipertensión con más frecuencia que la población general.

Existen algunos algoritmos que pueden emplearse en los dos tipos de aprendizaje, tanto supervisado como no supervisado. El más importante de ellos es el de las redes neuronales, que imitan un funcionamiento simplificado de las redes de neuronas del cerebro humano (que son muchísimo más complejas de lo que podemos llegar a modelar, aunque a muchos se les olvide). En una red neuronal tendríamos muchas de estas pequeñas calculadoras organizadas por capas. Cada neurona artificial recibe entradas de las neuronas en la capa anterior y produce una salida que se transmite a las neuronas en la capa siguiente. El aprendizaje en las redes neuronales se produce a través de la modificación de la definición de la operación dentro de cada una de ellas (por ejemplo, los pesos en una media ponderada) y las conexiones entre las neuronas (es decir, qué conexiones deberían ser más importantes y qué conexiones deberían ser menos intensas).

El proceso de entrenamiento de una red neuronal implica dos fases principales: la propagación hacia adelante y la retropropagación. En la propagación hacia adelante, los datos se introducen en la red y se propagan a través de las diferentes capas de la red. Cada capa procesa la información y

la transmite a la siguiente capa hasta que se llega a la última capa, que produce el resultado final. En la retropropagación, se calcula el error entre los resultados generados y los resultados deseados, y se ajustan las conexiones ponderadas para minimizar ese error. Esto permite que la red se adapte a los datos de la mejor manera posible.

Precisamente son las redes neuronales las que, debido a su flexibilidad y a su capacidad de entrenarse con enormes conjuntos de datos, han conseguido realizar algunos de los desarrollos más impresionantes dentro de la IA, como el procesamiento de imagen o el procesamiento del lenguaje natural. Además, se han creado de manera relativamente reciente diferentes arquitecturas de redes neuronales que han posibilitado un verdadero salto cualitativo como la IA generativa.

Uno de los avances más importantes ha sido el desarrollo de técnicas de aprendizaje profundo, basados principalmente en redes neuronales profundas (es decir, con muchas capas) de gran tamaño y complejidad. Esta complejidad implica, de manera clave, que la interpretabilidad de los resultados (es decir, la posibilidad de encontrarles una explicación lógica) es muy limitada o inexistente, ya que el gran número de parámetros de las redes profundas hace que sean demasiado complicadas como para ser comprendidas por un ser humano. Como veremos después, las consecuencias éticas de este hecho son clave.

Formas específicas de redes neuronales han tenido una importancia fundamental. Por ejemplo, las Redes Neuronales Convolucionales (CNN) son una forma especializada de red neuronal diseñada para procesar datos en forma de matrices (como las imágenes). Las Redes Generativas Adversarias (GAN) han sido fundamentales para el desarrollo de la IA generativa. Las GAN se componen de dos redes neuronales: un generador y un discriminador. El generador es responsable de generar datos nuevos, mientras que el discriminador es responsable de distinguir entre los datos generados y los datos reales. El proceso de entrenamiento de una GAN consiste en poner a competir al generador y al discriminador: el generador intenta engañar al discriminador generando datos cada vez más realistas, y el discriminador intenta distinguir entre los datos reales y los generados. Las GAN han sido utilizadas en una variedad de aplicaciones, incluyendo la generación de imágenes, el procesamiento del lenguaje natural y la música generativa.

Otra arquitectura clave han sido los transformadores, que fueron introducidos por primera vez en 2017. Están basados en una estructura de codificador-decodificador, en la que los datos de entrada se codifican en una representación matemática simplificada y se decodifican para volver a generar una salida. Esta arquitectura es similar a la utilizada en las Redes Neuronales Recurrentes (RNN), pero con una serie de mejoras que los hacen

más efectivos, específicamente la conocida como "atención". En términos generales, la atención se refiere a la capacidad de enfocar la percepción en ciertos aspectos del entorno, ignorando o minimizando otros. Este proceso es fundamental para el aprendizaje y la cognición humanos, y se ha demostrado que es igualmente importante para las redes neuronales artificiales. La atención se usa para asignar diferentes pesos a las partes que componen los datos de entrada dependiendo del contexto, de tal manera que el modelo pueda enfocarse en las partes específicas que resultan más importantes. Este es el mecanismo clave: en vez de tener en cuenta todos los datos, se tienen en cuenta sólo los más relevantes, con lo que es posible crear modelos utilizando una cantidad de datos inconmensurablemente mayor que lo que permitirían otras arquitecturas. Sin atención, no habríamos conseguido avances como ChatGPT.

El proceso de codificación de un modelo como ChatGPT comienza con la entrada, que se divide en tokens. Los tokens, también llamados componentes léxicos, son cadenas de caracteres que tienen un significado coherente en un lenguaje, natural o de programación. Podemos comprenderlos como palabras o partes de palabras. Estos tokens se convierten en vectores (series de valores numéricos). Convertir palabras en vectores (conocido como word2vec) nos permite pasar del espacio de las palabras al de los números, que es sobre el que trabajan los algoritmos. Estos vectores se pasan a través de varias capas de atención, que se utilizan para enfocarse en partes específicas de la entrada. Las capas de atención se denominan capas de codificación y son responsables de transformar la entrada en una representación matemática suficientemente reducida que pueda ser utilizada después en el procesamiento posterior. El mecanismo de atención funciona mediante la asignación de un peso a cada elemento de entrada en función de su importancia relativa para la tarea en cuestión. Estos pesos se calculan dinámicamente en función de la salida anterior de la red y se utilizan para determinar la cantidad de atención que se debe prestar a cada elemento en la entrada. En general, la atención se puede clasificar en dos tipos: atención basada en contenido y atención basada en posición. La atención basada en contenido se utiliza para identificar los elementos relevantes de la entrada en función de su similitud semántica con los elementos de la salida. Por ejemplo, en el caso de la traducción automática, la red neuronal puede asignar más atención a las palabras de la entrada que tienen un significado similar a las palabras de la salida. La atención basada en posición, por otro lado, se utiliza para identificar los elementos de la entrada en función de su posición en la secuencia. Por ejemplo, en el caso de la segmentación de imágenes, la red puede asignar más atención a las regiones de la imagen que son más importantes para la tarea, como las regiones que contienen objetos de interés. El proceso de descodificación realiza el paso por las capas de atención de manera simétrica, generando finalmente un vector de salida (una serie de valores numéricos) que puede ser utilizado para la generación de texto.

Una de las ventajas de los transformadores es que son muy buenos para procesar datos secuenciales, como el texto. Además, los transformadores pueden procesar la entrada de manera paralela. Esto significa que pueden procesar grandes conjuntos de datos de manera más eficiente y pueden generar respuestas más rápidamente. Los transformadores son extraordinariamente flexibles y se pueden usar para una gran variedad de tareas como la generación de texto, la traducción automática y el análisis de sentimientos (decidir si un texto refiere una emoción positiva, negativa o neutra; esto es especialmente importante, por ejemplo, para calificar reseñas de productos o servicios en Internet).

La atención, como herramienta novedosa, ha llegado además en un momento en el que ha aumentado espectacularmente la capacidad de computación accesible a las empresas y usuarios, además de abaratarse enormemente su coste. Los avances en la tecnología de *hardware*, como la creación de unidades de procesamiento gráfico (GPU) –que realizan operaciones en paralelo de manera muy eficiente– y la "nube" computacional, han permitido que los sistemas de IA funcionen a velocidades cada vez más altas y con un mayor grado de complejidad. Esto ha facilitado disponer de modelos cada vez más grandes y complejos, con mejor calidad y una mayor diversidad en sus contenidos. Además, la disponibilidad de datos ha resultado un elemento fundamental para el desarrollo de la IA. Algunos se han referido al período desde los 2000 como la "era del Big Data". Necesitamos datos, grandes cantidades de datos, para entrenar los modelos complejos. Mucho más en el caso de las IA generativas.

Sólo gracias a estos elementos: la innovación en algoritmos –específicamente el aprendizaje profundo–, las GAN y la atención, la capacidad de cálculo barata y la disponibilidad de datos hemos podido llegar al punto en el que nos encontramos, asombrados por las capacidades de las IA generativas.

Estas IA generativas y la Revolución Cognitiva que nos traen requerirán un nuevo contrato social, un nuevo diálogo humano-máquina. En este diálogo no sólo se sumarán voces nuevas, sino que también cambiarán los espacios, principalmente mediante el crecimiento en importancia de los entornos virtuales.

3. Dualidad físico-virtual

A la vez que la IA incrementa el rango de tareas que puede realizar, también se intensifica la manera en la que nos relacionamos con ella. Uno de los frentes que han suscitado más interés últimamente ha sido el del metaverso, que desdibuja los límites entre lo físico y lo virtual.

El Tamagotchi fue un juguete electrónico creado en la década de 1990 que simulaba ser una mascota virtual a la que se debía cuidar, alimentar y entretener. Fue muy popular para toda una generación, que de repente se sintió al cuidado de un supuesto ser vivo durante todo el día. Fue una metáfora perfecta para explicar la dualidad del mundo físico y virtual, ya que combinaba elementos de ambos mundos en una experiencia única. Por un lado, el Tamagotchi existía en el mundo físico como un objeto tangible que se podía sostener con las manos. Sin embargo, al interactuar con él, era posible sumergirse en un mundo virtual donde se cuidaba de un ser digital con sus propias necesidades y deseos. La experiencia Tamagotchi era una mezcla de interacción física y virtual, que se entrelazaban de manera compleja y emocionalmente atractiva.

Esta dualidad es también representativa de cómo funcionan nuestras vidas en la era digital. Vivimos en un mundo físico, rodeados de objetos tangibles y personas reales, pero también estamos constantemente conectados a un mundo virtual a través de nuestros ordenadores, teléfonos y otros dispositivos electrónicos. Al igual que en el Tamagotchi, en nuestras vidas diarias hay responsabilidades y consecuencias tanto en el mundo físico como en el virtual. Si no cuidamos adecuadamente de nuestra mascota digital, esta podría enfermar e incluso "morir", lo cual podría afectar a nuestras emociones y bienestar en el mundo real. Del mismo modo, las decisiones que tomamos en el mundo virtual, como las redes sociales y las interacciones digitales, pueden tener efectos significativos en nuestras relaciones, trabajos y vidas en el mundo físico.

La metáfora del Tamagotchi nos recuerda que el mundo físico y el virtual están cada vez más interconectados y que nuestras acciones en uno pueden tener consecuencias en el otro. Es importante encontrar un equilibrio entre estos dos mundos y ser conscientes de cómo nuestras decisiones y comportamientos en ambos pueden impactar en nuestra vida. No es de extrañar así que en una era en la que los problemas de salud mental y la evasión de las responsabilidades del mundo físico están en auge, la combinación del mundo virtual y físico está cobrando una relevancia creciente.

Esta dualidad del mundo físico-virtual comenzó a intensificarse en las últimas décadas del siglo XX, con el surgimiento de la computación personal y el acceso a Internet. Y es que ambos mundos comparten ciertas similitudes. Por ejemplo, en ambos se pueden establecer relaciones sociales, ya sea a través de interacciones cara a cara en el mundo físico o mediante plataformas de comunicación en línea en el mundo virtual. Además, las normas sociales y culturales también pueden aplicarse en ambas realidades, y en cada una de ellas se pueden desarrollar habilidades y conocimientos. Sin embargo, también existen diferencias fundamentales. El mundo virtual carece de las limitaciones físicas y temporales del mundo real, lo que permite una mayor

flexibilidad en la comunicación y el acceso a la información. Por otro lado, el mundo físico es único en su capacidad para proporcionar experiencias sensoriales y tangibles que no pueden replicarse completamente en un entorno virtual.

Esta dualidad ha traído consigo numerosos cambios y desafíos en la forma en que vivimos y trabajamos. Por un lado, ha permitido una mayor globalización, conectividad y acceso a la información, lo que ha impulsado el desarrollo económico y cultural. Por otro lado, también ha generado preocupaciones sobre la privacidad, la seguridad de los datos y el impacto en la salud mental y las relaciones interpersonales. En esta dualidad es donde podemos introducir el metaverso, palabra que suele traer más preguntas que respuestas. Una palabra que ocupó portadas durante meses. Muchos dan por muerto el concepto; sin embargo, es probable que con él ocurra como con la IA. Tendrá inviernos, pero volverá. Y es que nuestro deseo por escaparnos de la realidad y tener control sobre nuestras vidas es tan antiguo como las emociones más básicas del Sapiens. Y eso, tarde o temprano, acaba triunfando en las sociedades.

3.1. El metaverso: hacia una vida digital

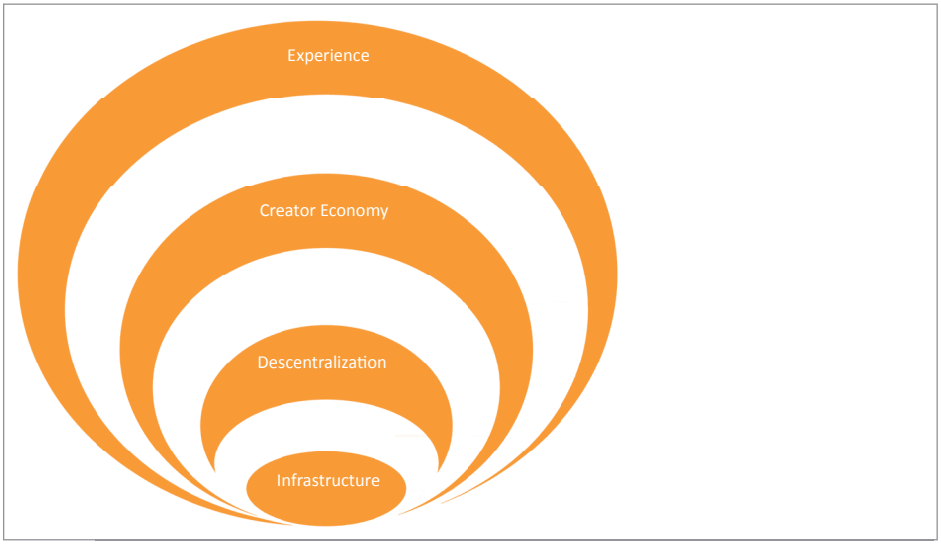
Sócrates y Platón se cuestionaron la realidad que nos rodea y su naturaleza. Los alquimistas de la Edad Media quisieron encontrar el elixir de la eterna juventud. Julio Verne, contemporáneo de la Revolución Industrial, explotó en sus obras la construcción de escenarios tecnológicos del futuro. Matrix introdujo en la cultura pop la idea de poder escapar de nuestra realidad. Esta cronología describe el constante interés de la especie humana por el escapismo. La psicología lleva décadas describiendo este fenómeno como una forma de evadirse de los problemas del día a día. Es en este contexto donde se sitúa el término del momento: metaverso.

El metaverso es un término que se utiliza para describir un espacio virtual tridimensional, compartido y persistente en el que los usuarios pueden interactuar entre sí y con el entorno a través de avatares. El concepto se basa en la idea de un universo digital paralelo al mundo físico, en el que las personas pueden socializar, trabajar, jugar y explorar. Fue acuñado por Neal Stephenson en su novela de ciencia ficción *Snow Crash* de 1992 (Stephenson, 2003). La novela describe un mundo en el que los usuarios interactúan en un entorno virtual tridimensional llamado *Metaverse*, que se asemeja a una versión futurista de Internet. La identidad es un avatar, término que proviene del sánscrito *avatâra* para reflejar el descenso de un ser divino a la Tierra. En la obra de Stephenson los usuarios usan unas gafas para acceder a través de unas terminales personales (el PC no había llegado masivamente a los hogares).

Aunque el concepto se originó en la literatura, ha influido en la forma en que los desarrolladores y la industria de la tecnología han abordado la creación de mundos virtuales y simulaciones desde entonces. La idea del metaverso ha evolucionado con el tiempo y ha sido influenciada por diferentes desarrollos tecnológicos. Algunos ejemplos notables de metaversos o mundos virtuales incluyen Second Life, un espacio lanzado en 2003 donde los usuarios pueden crear y personalizar avatares, construir y explorar entornos, y participar en actividades sociales y económicas. Muchos lo han citado como el predecesor del metaverso actual. Second Life fue un entorno virtual inmersivo que llegó a tener 70 millones de usuarios. Pero fracasó.

La tecnología del momento no permitía recrear a gran escala y de forma natural entornos en 3D accesibles por los usuarios de forma fluida. Hoy, en 2023, no sólo tenemos mejores tecnologías de modelado y animación (*software*), sino que además ha llegado el 5G, tenemos mejores dispositivos de acceso del usuario (cascos de realidad virtual, tejidos hápticos –táctiles– y otros periféricos) y plataformas para construir y disfrutar de metaversos (videojuegos, sistemas financieros, proyectos inmobiliarios, etc.). Pero, por encima de todo, los usuarios son cada vez más y con más conocimientos de los medios digitales inmersivos. Son siete las capas tecnológicas de las que hoy disponemos y que habilitan combinadas lo que hoy llamamos el metaverso (cuadro 1), según Jon Radoff (2021).

Cuadro 1 – Las siete capas del metaverso



Fuente: Jon Radoff, https://medium.com/building-the-metaverse/the-metaverse-value-chain-afcf9e09e3a7

De esta manera, el metaverso representa la cuarta generación de la computación, tras los ordenadores centrales, los personales y los móviles. Es una computación ambiental, donde nos sumergimos en el ordenador en lugar de acceder al mismo. Tenemos una vida digital conectada, en lugar de simplemente tener acceso a la misma. No es raro que el nombre elegido para que Facebook se rebautice sea Meta, un sufijo de origen griego que significa "más allá".

Según Bloomberg, este mercado del metaverso podría tener un tamaño de 800.000 millones de dólares en 2024. Grayscale eleva la cifra al billón. Ambos cálculos parten de una cifra cercana a los 180.000 millones de dólares de tamaño de mercado con el que cerró 2020. Son informes que apuntan a ese crecimiento considerando el metaverso como esa nueva era de la computación. Es decir, como una tecnología habilitante, donde todo estaría por construirse.

Esta nueva oportunidad de computación está siendo aprovechada por numerosos sectores.

- *Moda*. Un sector que vive o bien de la rotación (en prendas de calidades medias y bajas), o de la exclusividad (acceso al lujo). Ambas son propuestas de valor del metaverso. Balenciaga y su alianza con Epic Games buscan que los avatares digitales y los NFT (certificados de autenticidad que confirman la posesión de un activo digital) ofrezcan un terreno fértil para establecer relaciones con las audiencias más jóvenes. Nike, Adidas, H&M y Zara han anunciado ya la construcción de sus propios metaversos. Vender, diseñar sin coste medioambiental o promocionar NFT en videojuegos, abrir tiendas virtuales o servicios de atención son sólo algunos ejemplos de lo que están buscando.

- *Inmuebles virtuales*. El metaverso, al ser un espacio con atención y audiencias, la publicidad y las promociones inmobiliarias no lo iban a esquivar. Prueba de ello es la parcela que se compró en el circuito de Mónaco del videojuego F1 Delta Time por 223.000 dólares para mostrar publicidad personalizada programable. En Decentraland (90.601 parcelas de tierra en un mundo virtual descentralizado), la empresa Metaverse Group ha comprado por valor de más de 2 millones de dólares un terreno de 116 parcelas. Quiere construir el distrito Fashion Street para promover la moda digital. En Seattle, Facebook ha empezado a alquilar locales quebrados por la pandemia para crear escaparates que con códigos QR redirigen al cliente a tiendas *online*. Con ello, un pequeño comercio puede estar siempre abierto y tener una exposición a audiencias millonarias sin tener que invertir en ladrillo. Microsoft, que cada mes tiene 250 millones de usuarios activos en su entorno de trabajo, anunció hace unas semanas Dynamics 365 Connected Spaces, su "metaverso para el trabajo".

- *Diseño y arquitectura*. La combinación de lo virtual y lo físico puede ser utilizada para visualizar proyectos arquitectónicos o de diseño en su entorno real, lo que facilita la toma de decisiones y la comunicación entre diseñadores, arquitectos y clientes. También puede usarse para optimizar los espacios de trabajo y hacer más eficiente la producción.

- *Fabricación e industria*. La realidad mixta puede mejorar la eficiencia y precisión en procesos de fabricación al permitir a los trabajadores visualizar instrucciones en tiempo real y acceder a información sobre las piezas o componentes que están manipulando. Además, puede ser útil para la detección de problemas y la realización de mantenimientos preventivos.

- *Comercio minorista y marketing*: Puede utilizarse para crear experiencias de compra más atractivas, permitiendo a los consumidores visualizar productos en sus hogares antes de adquirirlos, por ejemplo, o interactuar con elementos digitales en tiendas físicas. Es lo que ha hecho, por ejemplo, Domino's Pizza, que, al abrir su tienda en el metaverso de Decentraland, permite a cualquier cliente hacer pedidos sin tener que esperar colas físicas.

- *Entretenimiento y videojuegos*. Es un sector maduro (y no sólo juegan niños y niñas), con capacidad de desembolso y cuyos usuarios pasan cientos de horas jugando con su avatar personal para acumular premios (monedas virtuales muchas veces) para obtener activos digitales únicos. De hecho, hay videojuegos que podrían considerarse auténticos metaversos ya en pleno funcionamiento, aunque con posibilidades de ampliación y mejora. Es normal que Sony haya comprado Bungie o Microsoft haya adquirido el gigante de los videojuegos Activision, dueño de comunidades virtuales millonarias de grandes videojuegos como *Warcraft*, *Call of Duty* o *Candy Crush*. A partir de ahí, podrán construir y reformar sus entornos virtuales para hacer muchas de las cuestiones antes comentadas.

- *Medicina y atención médica*. La combinación del mundo virtual y físico puede mejorar la formación de médicos y otros profesionales de la salud, permitiendo la simulación de procedimientos quirúrgicos y la práctica en entornos controlados. Además, la realidad mixta puede utilizarse en rehabilitación y terapia, ayudando a los pacientes a recuperarse de lesiones o a mejorar sus habilidades motoras.

- *Educación.* La Universidad de Stanford ha lanzado su primer curso impartido íntegramente en un entorno de realidad virtual (RV). Lo ha hecho utilizando Engage, una plataforma que permite crear a sus usuarios su propio entorno virtual. Los profesores limitaron la clase a 30 minutos (para evitar la "enfermedad del simulador") y crearon entornos DICE (Dangerous, Impossible, Counterproductive and Expensive); es decir, aprender en entornos donde físicamente sería imposible ir (la boca de un volcán, el océano profundo, etc.) o donde sería demasiado costoso.

En la actualidad, el metaverso está ganando cada vez más atención e inversión por parte de las principales empresas tecnológicas y de entretenimiento, como Facebook/Meta, Google, Microsoft y Epic Games. Estas compañías están desarrollando tecnologías y plataformas que permiten a los usuarios interactuar en entornos virtuales inmersivos y persistentes, a menudo utilizando realidad virtual (VR) y realidad aumentada (AR) para mejorar la experiencia. Así, llegó a ser considerado por algunos como el "próximo gran paso" en la evolución de Internet y la tecnología digital, ya que ofrece nuevas formas de comunicación, colaboración, trabajo, entretenimiento y comercio.

Dos de los elementos que suelen aparecer de forma conjunta con esta dualidad del mundo físico y virtual son la tecnología de cadena de bloques (*blockchain*) y los activos digitales únicos e intransferibles (NFT). Antes de analizar con detalle estos dos conceptos, es recomendable que entendamos qué relación tienen con el nuevo ecosistema denominado Web 3.0.

3.2. La web 3.0: hacia una nueva arquitectura de la información global y conectada

La Web 3.0, también conocida como la web semántica, es una evolución de la World Wide Web que busca crear una experiencia digital más inteligente, personalizada y descentralizada para los usuarios. La Web 3.0 sigue a la Web 1.0 (web estática) y a la Web 2.0 (web social y colaborativa). Cada una de estas etapas representa una transformación en la forma en que los usuarios interactúan con la web y en cómo se crea y se distribuye el contenido *online*. En la Web 1.0 los usuarios consumían información y contenidos proporcionados por los propietarios de los sitios web, pero no podían interactuar ni modificar el contenido. En la Web 2.0 los usuarios ya no son meros consumidores de información, sino que también pueden crear y compartir contenido a través de plataformas de blogs, redes sociales, wikis, etc. La Web 3.0 se basa en la descentralización, la privacidad y la seguridad, con tecnologías como *blockchain* y criptomonedas que permiten transacciones seguras y la propiedad digital. Se introduce la interoperabilidad y la colaboración entre diferentes servicios y aplicaciones, lo que permite una mayor personalización y experiencias de usuario adaptadas a las necesidades individuales.

La idea de la Web 3.0 comenzó a emerger a mediados de la década del 2000, cuando Tim Berners-Lee, el creador de la World Wide Web, propuso la idea. La web semántica se centraría en mejorar la forma en que las máquinas comprenden y procesan la información, lo que permitiría una mayor interacción y colaboración entre humanos y máquinas. Para que esto fuera posible, era necesaria la combinación de tecnologías emergentes, como la

IA, el Internet de las cosas (IoT), la cadena de bloques (*blockchain*) y la realidad virtual y aumentada. Todas ellas son tecnologías que han alcanzado ya una relativa madurez. La IA permitirá que las máquinas comprendan y procesen la información de manera más eficiente, lo que dará lugar a un ecosistema de la web más inteligente y personalizado. Con la creciente cantidad de dispositivos conectados (IoT), la Web 3.0 permitirá una mayor interacción y colaboración entre los dispositivos, lo que podría llevar a una mayor eficiencia y comodidad en nuestra vida cotidiana. La tecnología de cadena de bloques habilita la descentralización y la seguridad en la Web 3.0. Esto podría conducir a un ecosistema de la web más transparente y seguro, al tiempo que garantiza la propiedad y la privacidad de los datos del usuario. Por último, la incorporación de la realidad virtual y aumentada en la Web 3.0 permitirá experiencias en línea más inmersivas y personalizadas.

3.3. Aplicaciones de la dualidad físico-virtual

Son muchas y variadas las representaciones que podrían ofrecerse de esta dualidad físico-virtual que describíamos al inicio. Algunos ejemplos que entendemos pueden alumbrar un camino en los próximos años son:

- *Criptomonedas*. Las criptomonedas, como Bitcoin y Ethereum, están cambiando la forma en que realizamos transacciones financieras. Las criptomonedas permiten realizar pagos y transferencias de dinero a nivel mundial de manera rápida y económica, sin necesidad de intermediarios. Gracias a las bajas comisiones de las transacciones en criptomonedas, es posible realizar micropagos por servicios o productos digitales, como contenido en línea, videojuegos o aplicaciones móviles.

- *NFTs (Tokens no fungibles)*. En este contexto, "fungible" se refiere a la propiedad de un activo o token de ser intercambiable o reemplazable por otros de la misma clase. Un activo fungible es aquel que se considera idéntico y puede ser intercambiado uno por otro sin que exista una diferencia o distinción significativa. Por ejemplo, una moneda de un euro es indistinguible de otra moneda que es físicamente distinta pero tiene el mismo valor. Por otro lado, los NFT son activos digitales únicos y no fungibles, lo que significa que cada NFT es único y no se puede intercambiar directamente por otro NFT de manera equivalente. Cada NFT tiene un identificador único y contiene información específica que lo distingue de otros tokens. La característica de no fungibilidad de los NFT es fundamental para su valor y utilidad en el ámbito de la propiedad digital y el coleccionismo. Cada NFT puede representar una obra de arte digital, un artículo de colección, un token de juego u otros activos digitales únicos. Debido a su singularidad, los NFT tienen un valor inherente y se consideran únicos en comparación con los activos digitales fungibles, como las criptomonedas o las monedas digitales, que

son intercambiables entre sí. Entre otros campos, destaca el uso de NFTs en el arte digital, los coleccionables virtuales, dominios de Internet, bienes virtuales en videojuegos, moda y ropa virtual, entradas a eventos, derechos de autor de música y los certificados de identidad.

- *Redes sociales descentralizadas*. Son plataformas que permiten la comunicación, el intercambio de información y la colaboración entre sus usuarios sin depender de una entidad centralizada para controlar y administrar la plataforma. Estas redes se basan en tecnologías como la cadena de bloques (*blockchain*) y sistemas de archivos distribuidos (como IPFS). Algunos ejemplos de aplicaciones y usos de las redes sociales descentralizadas incluyen la mensajería, las redes sociales de contenidos (como la conocida alternativa a Twitter llamada Mastodon), periodismo ciudadano, mercados de arte digital, gobernanza comunitaria, etc.

- *Aplicaciones descentralizadas (dApps)*. Las dApps están construidas sobre tecnologías *blockchain* y ofrecen una variedad de servicios para que sea el propio usuario el que tome el control sobre sus propiedades. Estas aplicaciones se benefician de la naturaleza descentralizada, segura e inmutable de la tecnología *blockchain*. Se encuentran ejemplos de uso y aplicación de dApps en diferentes sectores como las finanzas (intercambios, préstamos, derivados, etc.), los juegos y aplicaciones lúdicas, las cadenas de suministro y logística, el almacenamiento descentralizado, etc. A medida que la tecnología *blockchain* madura y se vuelve más accesible, es probable que veamos un aumento en el número y variedad de estas aplicaciones descentralizadas en el futuro.

Por su inherente relación con el mundo de la dualidad físico-virtual, es interesante describir con mayor detalle las tecnologías de cadenas de bloques y NFT.

3.4. Blockchain como sistema trazable y atribuible

La tecnología *blockchain* es un sistema de registro digital distribuido que permite a múltiples partes almacenar y validar transacciones de forma segura y transparente. Un registro virtual distribuido, también conocido como *distributed ledger* en inglés, es una tecnología que permite el registro y almacenamiento de información de forma descentralizada en múltiples ubicaciones o nodos de una red. En lugar de tener un único repositorio centralizado, la información se distribuye y se replica en varios nodos que forman parte de la red. En un registro virtual distribuido cada nodo de la red mantiene una copia completa y actualizada de la información registrada. Cada transacción o cambio en el registro se verifica y se registra de forma consensuada por los nodos participantes, lo que garantiza la integridad y

la transparencia de los datos. El registro virtual distribuido ofrece varias ventajas. Al descentralizar la información, no depende de un único punto de fallo, lo que aumenta la resiliencia y la seguridad de los datos. Además, al estar distribuido entre múltiples participantes, se fomenta la confianza y la transparencia, ya que todos los nodos tienen acceso a la misma información verificada y actualizada. La tecnología *blockchain* es la base de muchas criptomonedas, como Bitcoin y Ethereum, y ha sido aplicada en una amplia gama de industrias, incluyendo la de juegos y bienes virtuales. La convergencia de *blockchain* y metaverso se manifiesta en múltiples ámbitos, entre los que podemos destacar:

- *Propiedad y activos digitales*: *Blockchain* permite la creación y el intercambio de tokens no fungibles (NFTs), que representan la propiedad de un objeto digital único e indivisible. Los NFTs pueden ser utilizados en el metaverso para representar activos digitales, como bienes virtuales, parcelas de tierra, personajes y más. Esto permite a los usuarios poseer, comprar, vender y comerciar estos activos de forma segura y verificable. Esa vida digital conectada que describíamos antes precisa de estos sistemas de representación. Proyectos como CryptoKitties y Decentraland utilizan *blockchain* para garantizar la propiedad y autenticidad de objetos y parcelas virtuales en sus respectivos metaversos.

- *Identidad digital*: La tecnología de cadenas de bloques puede ser utilizada para crear identidades digitales descentralizadas y autónomas en el metaverso. Esto permite a los usuarios tener un control total sobre sus datos personales, así como la capacidad de verificar su identidad y acceder a diferentes mundos y aplicaciones sin necesidad de crear múltiples cuentas. Plataformas como uPort y Civic utilizan *blockchain* para proporcionar soluciones de identidad digital y control de acceso a los usuarios en metaversos y aplicaciones descentralizadas.

- *Economía virtual*: *Blockchain* facilita la creación de criptomonedas y tokens digitales que pueden ser utilizados como medios de intercambio en el metaverso. Esto permite a los usuarios participar en transacciones económicas, como comprar bienes y servicios, invertir en proyectos y ganar recompensas por su participación en actividades y juegos. Second Life, con su moneda virtual Linden Dollars (L$), y Decentraland, con su criptomoneda Mana, son ejemplos de metaversos que utilizan tecnología *blockchain* para facilitar transacciones seguras y transparentes.

- *Descentralización y gobernanza*: La naturaleza descentralizada de *blockchain* puede ser aplicada al metaverso para crear sistemas de gobernanza y toma de decisiones basados en la comunidad. Esto puede incluir votaciones sobre cambios en las reglas del metaverso, la distribución de recursos y la resolución de disputas. En los mundos virtuales inmersivos

es especialmente importante, dado el sistema financiero y económico que comienza a crearse. DAOstack es un proyecto que utiliza *blockchain* para facilitar la gobernanza descentralizada y la toma de decisiones en metaversos y otras comunidades en Internet (O'Leary, 2017).

- *Interoperabilidad*: La tecnología *blockchain* puede facilitar la interoperabilidad entre diferentes plataformas y aplicaciones dentro del metaverso. Los NFTs y tokens creados en *blockchain* pueden ser transferidos y utilizados en diferentes mundos y juegos, lo que permite a los usuarios llevar sus activos y personajes de un entorno virtual a otro.

La convergencia del mundo físico-virtual y las tecnologías de cadenas de bloques tiene el potencial de impulsar la adopción y el crecimiento del metaverso, así como de abrir nuevas oportunidades para la innovación y la interacción social. Como decíamos con el Tamagotchi, la sensación de tener una vida digital plena hace que los usuarios interactúen con más intensidad.

La penetración de la tecnología *blockchain* en España ha experimentado un crecimiento constante en los últimos años, abarcando diversos sectores como finanzas, energía, Administración Pública y logística. Por sectores, podríamos presentar un resumen tal y como se expone a continuación:

- *Sector financiero*: Las instituciones financieras españolas, como BBVA y Banco Santander, han sido pioneras en la adopción de *blockchain* para mejorar la eficiencia en sus operaciones y servicios. Han implementado soluciones de pago transfronterizo, financiación comercial y gestión de identidad digital utilizando esta tecnología.

- *Sector energético*: Varias empresas energéticas, como Endesa y Repsol, han incorporado *blockchain* en sus operaciones para optimizar la gestión de la energía y fomentar la descentralización y la transparencia. También se han desarrollado proyectos para la comercialización de energía renovable en plataformas de *blockchain*.

- *Administración Pública*: El Gobierno ha mostrado interés en la adopción de *blockchain* para mejorar la eficiencia y transparencia en la Administración Pública. Se han iniciado proyectos piloto en áreas como el registro de propiedades, la contratación pública y la protección de datos personales.

- *Innovación y educación*: El ecosistema emprendedor y las universidades españolas han promovido la investigación y el desarrollo de soluciones basadas en *blockchain*. Se han creado centros de innovación y se han incorporado programas académicos especializados en esta tecnología.

Aunque la regulación de *blockchain* y criptomonedas en España ha sido un proceso gradual, se han establecido ciertas normativas para garantizar la

seguridad y protección de los usuarios. En 2021 se aprobó la Ley 11/2021[2] para prevenir el fraude fiscal, que incluye medidas de control para las criptomonedas.

En resumen, España ha experimentado un crecimiento en la adopción de la tecnología *blockchain* en diversos sectores, impulsado por la innovación, el apoyo gubernamental y la demanda del mercado. La regulación y el desarrollo de soluciones locales seguirán siendo factores clave para una mayor penetración de esta tecnología en el país.

3.5. NFT: los activos digitales únicos para resolver problemas de autoría

Si buscamos en Google por la comparación entre un JPG (extensión de un archivo de imagen) y NFT (acrónimo para referirnos a un activo digital único e intransferible), podremos ver cómo mucha gente se pregunta cuáles son sus diferencias. JPG y NFT son dos conceptos diferentes pero relacionados en el ámbito de las imágenes y los activos digitales. Un JPG (*Joint Photographic Experts Group*) es un formato de archivo de imagen comprimido muy común para almacenar y compartir imágenes digitales. Es ampliamente utilizado en la web debido a su eficiencia en la compresión y la calidad de imagen aceptable que produce. Pero tiene el problema de la duplicación y copia: es difícil mantener el control sobre su autoría, salvo cuando alguien detecta que se están infringiendo sus derechos de propiedad intelectual.

Por otro lado, un NFT (*Non-Fungible Token*) es un tipo de token digital que representa la propiedad y autenticidad de un activo digital único e indivisible. Los NFT se basan en tecnología *blockchain*, lo que garantiza la verificación y trazabilidad de la propiedad y la autenticidad de los activos. A diferencia de las criptomonedas, como el Bitcoin o Ethereum, que son fungibles y pueden intercambiarse entre sí, los NFT son únicos y no pueden intercambiarse. Así, con un NFT se puede representar la propiedad y autenticidad de una imagen digital, como un archivo JPG. De esta forma, el NFT se convierte en un certificado de autenticidad y propiedad para esa imagen, y puede ser comprado, vendido o intercambiado en el mercado de NFT.

Como señalamos en el apartado anterior, la relación entre NFT y metaverso (como metáfora de lo que es la combinación del mundo físico y virtual) radica en que los NFT pueden usarse para representar y comercializar bienes y servicios digitales en el metaverso. Esto puede incluir arte digital, objetos de coleccionista, ropa y accesorios para avatares, parcelas de tierra virtual y otros activos digitales únicos. Los NFT proporcionan un sistema de

[2] https://www.boe.es/buscar/doc.php?id=BOE-A-2021-11473

propiedad y autenticidad para estos objetos, lo que permite a los usuarios comprar, vender y comerciar activos digitales en el metaverso de manera segura y verificable. Es precisamente esto para lo que se están utilizando:

- *Arte digital*. Los artistas pueden crear y vender sus obras de arte digitales, garantizando la autenticidad y la propiedad única del comprador. Esto ha abierto nuevas oportunidades para que los artistas moneticen su trabajo en el ámbito digital.

- *Coleccionables*. Crear objetos de colección digital, como tarjetas coleccionables, figuras o incluso mascotas virtuales. Cada objeto coleccionable es único y puede ser comprado, vendido o intercambiado entre los usuarios.

- *Propiedad virtual*. Representar parcelas de tierra virtual o bienes inmuebles. Los usuarios pueden comprar, vender o alquilar propiedades virtuales, y desarrollar proyectos o negocios en esos espacios.

- *Moda y accesorios*. Los diseñadores de moda y marcas pueden crear ropa y accesorios digitales para avatares en el metaverso y venderlos como NFT. Los usuarios pueden comprar y lucir estos artículos en sus avatares, mostrando su estilo único.

- *Dominios de Internet*. Los nombres de dominio también se pueden representar como NFT, permitiendo la compraventa y el alquiler de direcciones web en el mercado de dominios.

- *Eventos y experiencias*. Vender entradas a eventos digitales, como conciertos en línea, conferencias o exposiciones, garantizando la autenticidad y la exclusividad de las entradas.

- *Propiedad intelectual*. Representar y gestionar los derechos de autor, licencias y propiedad intelectual de contenidos digitales, como música, libros electrónicos, *software* y películas.

- *Identidad y certificación*. Verificar la identidad de un individuo o la autenticidad de un producto, como diplomas, certificaciones o artículos de lujo.

- *Juegos*. Representar objetos y recursos del juego, como armas, vehículos y materiales, que los jugadores pueden comprar, vender y comerciar entre sí.

La penetración de los NFT en España ha experimentado un crecimiento significativo en los últimos años, siguiendo la tendencia mundial en el ámbito de las criptomonedas y la tecnología *blockchain*. Aunque este análisis no puede ofrecer cifras exactas, se pueden destacar algunas tendencias y eventos clave en el desarrollo de los NFT en España:

- *Adopción en el mundo del arte.* Artistas y galerías de arte españolas han adoptado los NFT como una nueva forma de comercializar y distribuir obras digitales, lo que ha impulsado la popularidad de esta tecnología en el sector creativo.

- *Integración en el deporte*. Clubes de fútbol, como el FC Barcelona y el Atlético de Madrid, han comenzado a utilizar NFT para emitir coleccionables digitales y entradas, lo que ha aumentado la visibilidad y el interés por esta tecnología en nuestro país.

- *Aumento de plataformas y mercados de NFT.* La creación y crecimiento de plataformas y mercados locales de NFT –como NFT España y Minty– han facilitado el acceso a los NFT y han generado un mayor interés por parte de los usuarios y coleccionistas españoles.

- *Eventos y conferencias*. La realización de eventos y conferencias relacionadas con NFT y *blockchain* en España, como la Blockchain Summit o el Congreso Nacional de Criptomonedas y Blockchain, ha contribuido a la difusión y el conocimiento de esta tecnología en el país.

Aunque aún queda camino por recorrer, España se encuentra en una posición favorable para seguir aprovechando las oportunidades que ofrece esta tecnología.

3.6. Lo que nos depare el futuro

La dualidad físico-virtual es un concepto cada vez más relevante con el avance de la tecnología y la digitalización de nuestras vidas. En el futuro podemos esperar un aumento en la interacción y la fusión entre estos dos mundos, dando lugar a un entorno donde las experiencias físicas y virtuales se integren y complementen entre sí. En el futuro, estas tecnologías serán más avanzadas, accesibles y ubicuas, permitiendo una integración más profunda de nuestras vidas físicas y virtuales.

Por otro lado, a medida que más dispositivos se conecten a Internet, se crearán redes cada vez más complejas y sofisticadas que permitirán una mayor interacción entre el mundo físico y el virtual. El IoT impulsará la automatización, la eficiencia energética y la personalización de los espacios físicos, lo que conducirá a una mayor convergencia entre ambos mundos. En este sentido, las ciudades aprovecharán las tecnologías digitales para mejorar la calidad de vida de sus habitantes. Estas ciudades inteligentes utilizarán la RA, la RV, la IoT y otras tecnologías para optimizar la gestión de los recursos, mejorar la movilidad urbana y aumentar la seguridad y la sostenibilidad.

Es previsible también que el teletrabajo y los espacios de trabajo virtuales sigan evolucionando, lo que permitirá trabajar desde cualquier lugar y colaborar en entornos virtuales. Esto podría reducir la necesidad de desplazamientos y de espacios físicos de oficina, cambiando la forma en que las personas experimentan el trabajo y la colaboración. La educación en línea y la formación basada en la RV continuarán expandiéndose, ofreciendo nuevas oportunidades para aprender y adquirir habilidades a través de experiencias inmersivas y personalizadas. Esto podría transformar la educación tradicional y permitir a los estudiantes acceder a recursos y experiencias que antes estaban fuera de su alcance.

Las tecnologías de seguimiento de la salud y las aplicaciones de bienestar seguirán en constante cambio y actualización, lo que permitirá a las personas monitorizar y mejorar su salud física y mental en tiempo real. La telemedicina y la RV también podrían utilizarse para mejorar el acceso a la atención médica y ofrecer tratamientos más personalizados y efectivos.

En resumen, la dualidad físico-virtual tiene el potencial de mejorar nuestra calidad de vida y transformar la forma en que trabajamos, aprendemos y nos relacionamos con nuestro entorno y con los demás. Los humanos no entendemos una tecnología hasta que no la adoptamos masivamente, que es lo que está ocurriendo precisamente con ese conjunto de tecnologías que habilitan el metaverso. El deseo humano por tener una vida digital en nuevos entornos futuristas ha existido, como vimos, desde hace siglos. Por ello, puede que la ventana de oportunidad del metaverso sea esta. Sin embargo, no podemos dejar que solo las empresas lo impulsen, o que creen un metaverso único donde no haya un marco del derecho para regular la convivencia. Retos y oportunidades confluyen en el tiempo para dar el pistoletazo de salida a una era dorada para la creatividad del metaverso. Emplearemos las secciones restantes de este capítulo para hablar precisamente de esos retos de regulación que afectan no sólo al metaverso sino a la IA en general.

4. Los problemas de la IA. ¿Qué amenazas nos trae?

El éxito de las IA generativas y especialmente el de herramientas como ChatGPT ha de evaluarse con cautela. Las principales preocupaciones que está generando tienen que ver con la calidad de sus resultados, el potencial que presentan para un mal uso y abuso, y su capacidad de modificar la estructura de algunos sectores de negocio. El resto de este capítulo está dedicado a reflexionar sobre estos problemas y a proporcionar unas primeras guías de uso que puedan ayudarnos a extraer su potencial minimizando las amenazas más importantes. Es necesario además tener en cuenta que la tecnología irá mejorando en sus siguientes versiones, de tal manera que es

esperable que algunos de los inconvenientes reflejados en estos párrafos se hayan superado en el futuro más cercano.

En primer lugar, la IA generativa tiene el potencial de devolver información errónea o, mejor dicho, engañosa, ya que, aunque sea inexacta tiene la peligrosa apariencia de parecer coherente. Cuando se le pide al sistema que responda a una pregunta sobre la que no tiene información suficiente, es posible que genere una respuesta que tenga la apariencia de ser verdadera, aunque sea completamente artificial. A este fenómeno nos referimos con la expresión "alucinación". Los grandes modelos del lenguaje "alucinan" con bastante frecuencia cuando se les hacen preguntas. Por ejemplo, acaparó titulares el que, al preguntársele cuándo pintó Leonardo Da Vinci *La Mona Lisa*, ChatGPT respondía "Leonardo da Vinci pintó *La Mona Lisa* en 1815" (la respuesta real es que la obra de arte se creó entre 1503 y 1507, y las fechas ni siquiera coinciden con las de la vida de su autor). Versiones posteriores de la herramienta parece que no cometen este error, pero sí cometen otros parecidos. Es importante comprender que, en los transformadores, se va procesando el contenido sin extraer en ningún momento un significado global, por lo que no se realiza ninguna comprobación lógica de la respuesta (en este caso, las fechas de la vida del autor con las de la obra).

Es importante señalar que ChatGPT está basado en un transformador que ha sido pre-entrenado con enormes cantidades de datos (su nombre es precisamente "Pre-Trained Transformer"). No puede acceder a ninguna información que no se encontrara en sus datos de aprendizaje, que llegan hasta 2021, aunque continúa aprendiendo en sus interacciones. Por ello, no puede utilizarse para responder a preguntas que salgan de este horizonte. Si intentamos que responda, puede o bien señalar que no posee información o, directamente, producir una alucinación aparentemente incoherente. El problema es que, en el segundo caso, no existe más advertencia que la genérica que dan sus proveedores: "ChatGPT puede producir información inexacta sobre hechos, personas o lugares".

Este sesgo es otro de los problemas principales que presentan los sistemas de IA en general debido a sus datos de entrenamiento. Por ejemplo, un modelo de lenguaje entrenado con un conjunto de datos que contiene contenido sexista o racista puede generar respuestas que sean discriminatorias u ofensivas. Si se toman decisiones basadas en estas respuestas, podemos tener consecuencias desastrosas. Algunas de ellas se describen en la siguiente sección.

Además, la IA generativa puede usarse para crear contenido falso de manera deliberada, incluyendo imágenes y vídeos manipulados (los llamados *deep fake*, que pueden difundirse luego rápidamente por las redes sociales). Esto hace que sea más difícil distinguir entre lo real y lo falso, lo que puede tener consecuencias negativas para la sociedad en su conjunto.

Esto debería llevarnos a desconfiar automáticamente del contenido generado por la IA. Además, estos sistemas no son capaces de señalar la procedencia de los datos. A diferencia del contenido generado por humanos, que se puede atribuir a una persona o grupo en particular, la IA generativa puede producir contenido sin una clara atribución o responsabilidad. Como consecuencia de esto, es difícil justificar el contenido producido por la IA. Si intentamos, por ejemplo, pedirle a ChatGPT que nos devuelva un texto con referencias, es probable que, simplemente, las produzca a través de una alucinación. Esto limita considerablemente las posibilidades de utilizar estas herramientas para fines como la investigación.

En tercer lugar, la IA generativa puede llevar a nuevas formas de plagio que ignoran los derechos de los creadores de contenido y de los artistas. Por ejemplo, un modelo de lenguaje entrenado con un conjunto de datos de obras literarias existentes puede generar nuevo contenido que sea similar o idéntico a obras existentes, sin la debida atribución o permiso. Esto perjudica a los creadores y artistas originales. Además, resulta extremadamente difícil identificar qué contenido ha sido creado mediante IA generativa, lo que está creando un auténtico terremoto en las aulas. Mientras escribimos estas líneas, Turnitin, el *software* líder en identificación de plagio estudiantil que emplean la mayoría de las universidades del mundo, anuncia que incorporará en su nueva versión una herramienta para detectar el uso de IA generativas en la confección de un texto. Le asignará una probabilidad de corresponder a un humano o a una máquina dependiendo del uso que haga de las palabras (por el momento, uno de los marcadores de los textos creados por las IA generativas es la probabilidad con la que se emplea cada palabra, que es más o menos constante). Algunas universidades han expresado su descontento con este nuevo sistema, que, debido a sus inevitables errores, acabará identificando como plagio un porcentaje considerable de los trabajos sometidos a examen aunque no lo sean.

Están apareciendo nuevas guías de uso que ayudarán a minimizar estos problemas. Por ejemplo, las IA generativas funcionan mejor si se les especifica el tipo de contenido que deben cubrir (por ejemplo: "escribe un ensayo sobre los problemas de las IA generativas") y los puntos principales que deberían tocarse (como "explicando los problemas con datos inexactos, el sesgo y el plagio"). Se puede además especificar el tono: "usa un lenguaje formal pero interesante". Además, la idea es que el autor interactúe con la máquina para refinar las respuestas; por ejemplo, "amplía el primer párrafo introduciendo un ejemplo" o "cambia este ejemplo por otro"). Las instituciones educativas están tomando dos actitudes opuestas: algunos están intentando prohibir este "nuevo pincel" mientras que otros abrazan las nuevas posibilidades. En el segundo caso, se propone, por ejemplo, que cuando se utilicen herramientas como ChatGPT se deba entregar no sólo el texto final

sino también la secuencia de comandos que ha ido refinando paulatinamente el resultado o comprobando la veracidad de la información contenida.

Otro de sus potenciales negativos es que la IA generativa puede emplearse para mejorar la eficiencia de los ciberataques. Puede crear perfiles de redes sociales falsos pero que parezcan auténticos y tengan más posibilidades de éxito en ataques de *phishing* (robo de datos) u otros tipos de engaño.

En el futuro a medio plazo, las IA generativas tendrán un impacto considerable en el empleo: desde ilustradores a profesores o desarrolladores de código, la transformación del empleo, que ya se anticipaba profunda, será mucho mayor una vez consideradas las herramientas de la IA generativa.

4.1. Más allá de las IA generativas. La ética de la IA

En abril de este año se presentaba una carta abierta firmada por más de mil empresarios y expertos, entre los que se encontraba Elon Musk, que pedían seis meses de moratoria en los desarrollos de IA, hasta poder controlarla de manera satisfactoria. Son muchos otros los que se oponen, señalando que no será posible frenar los avances en las regiones con controles más débiles, con lo que las consecuencias pueden ser peores. Sin embargo, muchos otros han advertido de que los desarrollos podrán pararse en algunas áreas geográficas, pero no en otras menos reguladas, y las consecuencias de una moratoria a la que no todos se acogen serían peores que las de no establecerla.

El caso es que el ser humano parece aprender mucho más rápido de sus errores que de un análisis *a priori* de los posibles escenarios negativos. Desgraciadamente, una de las características definitorias de la tecnología es que tiene consecuencias imprevisibles. Los primeros casos de uso de una nueva tecnología nos abren los ojos a sus inconvenientes y a cómo queremos gestionar sus futuros avances. Así, es necesario crear un equilibrio entre la experimentación (permitir que la tecnología se desarrolle y se implemente en la vida y los negocios), la reflexión (aprender de esta experimentación, integrándola con conocimientos sobre el funcionamiento de la tecnología y sus posibles interacciones con la sociedad) y la acción (tomar decisiones para regular los aspectos que detectemos que puedan ser especialmente problemáticos, teniendo en cuenta que las reglas de juego que establezcamos pueden aplicar sólo en un área geográfica, estableciendo así desventajas competitivas que pueden tener consecuencias negativas en el largo plazo).

Concretamente, nos hemos encontrado con que en el desarrollo de la ética de la IA han tenido un peso especialmente importante los escándalos, que han abierto los ojos a la opinión pública a los problemas derivados de un mal uso de la tecnología. Este abrir los ojos a los escándalos hace que se

genere el consenso social necesario para apoyar una legislación que la regule y limite sus posibles problemas.

Por ejemplo, en los últimos años se han producido varios escándalos que han puesto en evidencia la vulnerabilidad de la privacidad de datos. Uno de los casos más destacados fue el escándalo de Cambridge Analytica en 2018 (Hinds, Williams y Joinson, 2020), en el que se descubrió que la empresa había obtenido información personal de millones de usuarios de Facebook sin su consentimiento para utilizarla en campañas políticas, pudiendo incluso haber afectado a los resultados de las elecciones presidenciales de Estados Unidos o al referéndum del Brexit. Otro caso destacado fue el *hackeo* masivo de Equifax en 2017, en el que se filtraron los datos personales de más de 140 millones de personas, incluyendo nombres, direcciones, fechas de nacimiento y números de la seguridad social. Además, en 2013, Edward Snowden reveló la existencia de un programa de vigilancia masiva llevado a cabo por la Agencia Nacional de Seguridad de Estados Unidos, lo que generó un gran debate sobre la privacidad en la era digital. Estos escándalos han puesto de manifiesto la importancia de proteger la privacidad de datos y la necesidad de una regulación más estricta en el uso y manejo de la información personal en Internet.

Por otro lado, han sido varios los ciberataques de repercusión mundial que han afectado a Gobiernos, empresas y personas. Uno de los más notorios fue el ataque WannaCry en 2017, que afectó a más de 200.000 usuarios en 150 países. El *malware* WannaCry cifró los archivos de las computadoras infectadas y exigía un rescate en Bitcoin para desbloquearlos (lo que se conoce como *ransomware*). Fueron afectadas empresas de la talla de Telefónica (Mohurle y Patil, 2017). Estos ciberataques han demostrado la necesidad de fortalecer la seguridad informática y de tomar medidas más efectivas para proteger los sistemas y datos críticos. El campo de la ciberseguridad es, debido a esto, uno de los que están creciendo más rápidamente en los últimos años, y la regulación de la IA pone un énfasis especial en la seguridad de los sistemas: cuanto más dependamos de ellos más debemos asegurarnos de que son robustos ante estos ataques.

Cambiando de tercio, es bien conocido también el caso de COMPAS, un algoritmo que se utilizó durante años en Estados Unidos para decidir si un preso recibía la oportunidad de una libertad condicional o no. El algoritmo pretendía mejorar la justicia mediante una predicción del comportamiento del preso. Calculando una predicción de la probabilidad de reincidencia, el juez podría determinar de una mejor manera a quién darle la libertad condicional y a quién no: los presos con un riesgo menor de reincidencia eran liberados, mientras que aquellos con un riesgo mayor permanecían en prisión. Sin embargo, un análisis por parte de un medio de comunicación (*ProPublica*, específicamente) determinó que al algoritmo de COMPAS lo

permeaba un sesgo racista (Washington, 2019). A los presos afroamericanos se les asignaba un riesgo sistemáticamente mayor de reincidir. Este escándalo puso de manifiesto el problema del sesgo algorítmico, que refleja los sesgos existentes en la base de datos de entrenamiento (en la que los afroamericanos tenían peores tasas) en las futuras decisiones del sistema, de manera injusta. El sesgo algorítmico es especialmente difícil de mitigar cuando utilizamos algoritmos llamados de *caja negra*. El modelo nos da una respuesta, pero no el porqué de esa respuesta, con lo que no es posible identificar, al menos de manera inmediata, los factores que han influido en la decisión. Esto nos sucede, por ejemplo, con las redes neuronales. Sin embargo, otros modelos, como los árboles de decisión o las regresiones, aunque puedan lidiar también con problemas complejos y en muchos casos conseguir un rendimiento muy parecido a las redes neuronales y otros algoritmos de caja negra, sí que pueden ser interpretados, porque podemos examinar directamente los pesos de la regresión o las comparaciones que aparecen en los nodos del árbol.

Después del escándalo de COMPAS, la investigadora Cynthia Rudin (2019) desarrolló un modelo alternativo basado en un sistema de puntuación (muy parecido a un árbol) en el que el riesgo de reincidencia se calculaba únicamente en base a la edad del preso y al tipo de delitos que había cometido (por ejemplo, los delitos violentos aumentaban este riesgo). Su trabajo demostró que con este modelo, que es completamente transparente, conseguía predecir la reincidencia de manera tan exacta como con la caja negra de COMPAS. Esta investigación ha puesto en marcha un proyecto de desarrollo de IA transparente o interpretable que está empezando a trascender a nivel político. No es aceptable que un modelo que toma decisiones que tienen importancia en las vidas de los ciudadanos sea una caja oscura, porque no es aceptable que esas decisiones se tomen en base a prejuicios injustos y la mejor manera de asegurar que esto no sucede es mediante el uso de algoritmos transparentes.

De igual manera, vimos cómo en Holanda, con el escándalo de las subvenciones por cuidado de niños, después de detectar que el fraude era muy frecuente en estas ayudas, se construyó un sistema automatizado que denegaba las ayudas, a solicitantes a los cuales se había asignado una probabilidad alta de fraude sin emitir justificación alguna (Peeters y Widlak, 2023). Además, el rechazo de la solicitud implicaba no sólo dejar de percibir la ayuda en el futuro sino también la obligación de devolver el dinero percibido en años anteriores. Esta cantidad, en muchos casos, alcanzaba cientos de miles de euros y llevó a miles de familias a la ruina. Después de una investigación se descubrió, además, que los motivos que habían llevado al algoritmo a rechazar muchas de estas solicitudes estaban basados, por ejemplo, en pertenecer a minorías étnicas o en tener más de un pasaporte. Escándalos como este o como el de COMPAS dejan clara la necesidad de supervisar los algoritmos manteniendo la responsabilidad siempre en un

equipo humano: no es aceptable tomar decisiones automatizadas basadas en cajas negras en los casos en los que estas decisiones tienen un impacto importante en la vida de las personas. Esta necesidad de transparencia y responsabilidad, unida a la de privacidad, se ha recogido en la regulación sobre la IA que empieza a emerger, en el formato de reglamentos o de guías que aún deben traducirse en legislación específica, como comentaremos más adelante.

En la Unión Europea, la necesidad de proteger la privacidad resultó en 2018 en el Reglamento General de Protección de Datos (GDPR, por sus siglas en inglés). El objetivo principal del GDPR es proteger los derechos de privacidad de las personas y garantizar que las empresas y organizaciones que recopilan y procesan datos personales lo hagan de manera justa, transparente y segura. El GDPR establece reglas claras sobre cómo las empresas pueden recopilar, almacenar y utilizar los datos personales de los ciudadanos de la UE, y también les exige que notifiquen a las autoridades y a los individuos afectados si se produce una violación de datos. Además, el GDPR otorga a los ciudadanos de la UE el derecho a acceder, corregir y eliminar sus datos personales, así como el derecho a la portabilidad de datos. La implementación del GDPR ha llevado a una mayor conciencia y responsabilidad en cuanto a la privacidad de datos, tanto por parte de las empresas como de los individuos.

Queda mucho para contar con una legislación suficiente para los desarrollos de la IA, pero ya empiezan a aparecer guías éticas que están concentrando una atención merecidamente creciente. Podemos mencionar, por ejemplo, la Llamada de Roma *(Rome Call for AI Ethics,* Pontifical Academy for Life, IBM et al., 2023). En ella se precisan los siguientes principios:

- *Transparencia*: en principio, los sistemas de IA deben ser explicables.
- *Inclusión*: las necesidades de todos los seres humanos deben ser consideradas para que todos puedan beneficiarse y se les ofrezcan las mejores condiciones posibles para expresarse y desarrollarse.
- *Responsabilidad*: quienes diseñen y desplieguen el uso de la IA deben actuar con responsabilidad y transparencia.
- *Imparcialidad*: no crear o actuar según sesgos, salvaguardando así la equidad y la dignidad humana.
- *Fiabilidad*: los sistemas de IA deben ser capaces de funcionar de manera segura.
- *Seguridad y privacidad*: los sistemas de IA deben funcionar de manera segura y respetar la privacidad de los usuarios.

Además, en la Llamada de Roma se destaca la importancia de incluir la sostenibilidad en el desarrollo de la IA, al igual que otras necesidades como la de mostrar claramente cuándo se está interactuando con una IA y no con una persona.

La Llamada de Roma se une a otras guías que postulan principios similares. La Comisión Europea ha nombrado un Grupo de Expertos de Alto Nivel sobre IA, que en 2019 publicó *Directrices éticas para una IA fiable*. La Comisión Europea también acoge una unidad de Innovación y Excelencia en Robótica e IA, que publicó un Libro Blanco sobre excelencia y confianza en la innovación en IA en 2020 (Comisión Europea, 2020). Además, este mismo organismo promulgó en 2021 la Ley de IA (Madiega, 2022). Por otro lado, la OCDE estableció un Observatorio de Políticas de IA de la OCDE (Galindo, Perset y Sheeka, 2021).

Más allá de las iniciativas gubernamentales, Amazon, Google, Facebook, IBM y Microsoft establecieron una organización sin fines de lucro llamada Asociación para el Beneficio de la IA y la Sociedad, con el objetivo de formular las mejores prácticas en tecnologías de IA. Apple se unió a esta asociación en enero de 2017. Los miembros corporativos realizarán contribuciones financieras y de investigación al grupo, mientras se relacionan con la comunidad científica para incluir académicos en la junta (Heer, 2018).

Puramente académica es la propuesta del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE), que ha establecido una Iniciativa Global sobre la Ética de Sistemas Autónomos e Inteligentes (Chatila y Havens, 2019). La Conferencia Asilomar sobre la IA Beneficiosa fue una conferencia organizada por el Instituto del Futuro de la Vida, celebrada del 5 al 8 de enero de 2017 en el Centro de Conferencias Asilomar en California. Más de 100 líderes de opinión e investigadores en economía, derecho, ética y filosofía se reunieron en la conferencia para abordar y formular principios de la IA beneficiosa. Su resultado fue la creación de un conjunto de directrices para la investigación de la IA: los 23 Principios Asilomar (Asilomar Conference, 2017) , que incluyen la seguridad y fiabilidad de los sistemas de IA, la transparencia y explicabilidad de las decisiones tomadas por la IA, la privacidad y protección de los datos, la equidad y no discriminación en el diseño de los sistemas de IA, la colaboración entre humanos y sistemas de IA, la responsabilidad y rendición de cuentas por parte de los creadores y usuarios de la IA, y la promoción del bienestar humano a largo plazo. Estos principios buscan guiar el desarrollo y aplicación responsable de la IA para asegurar que se utilice para el beneficio de la humanidad. Como vemos, las propuestas descritas tienen en común la mayoría de sus elementos, de tal manera que podríamos decir que está emergiendo un consenso a este respecto.

El próximo reto será concretar estos retos en leyes específicas que puedan garantizar que estos principios se cumplan. Por ejemplo, podemos obligar a que los organismos públicos sólo utilicen modelos de IA basados en técnicas interpretables, como los árboles de decisión o las regresiones logísticas. Podríamos incluso crear sellos de calidad para aplicaciones de la IA que, mediante un proceso de auditoría detallada, garantizasen que la aplicación es segura o que no se ha detectado discriminación de ningún tipo. Sin embargo, ¿cómo podemos, por ejemplo, garantizar que exista transparencia en una IA generativa, cuando parece que una IA generativa que explique sus decisiones parece estar más allá de los límites de la factibilidad tecnológica? Aun así, ¿no deberían estos deseos regulatorios empujar también al desarrollo tecnológico en las direcciones deseadas, por ejemplo, trabajando en un nuevo tipo de modelo del lenguaje que nos indique y contraste sus fuentes? La relación entre el desarrollo tecnológico y la regulación de sus aplicaciones ha sido siempre compleja, pero nunca tanto como en el panorama ante el que hoy nos encontramos. Sólo podremos aprovechar los magníficos potenciales de estas nuevas herramientas desde la experiencia en casos de estudio reales, la observación detallada, la reflexión desde los valores éticos y la creación de guías primero y legislación específica después.

4.2. ¿Son las IA generativas la prueba de que, dentro de poco, tendremos IA general y consciencia artificial?

En verano de 2022 nos sorprendía la noticia de que Blake Lemoine, ingeniero de Google, era despedido tras compartir información privada sobre el entrenamiento de uno de sus modelos más potentes de procesamiento del lenguaje natural, el LaMDA (*Language Model for Dialogue Applications*) y asegurar que este modelo era consciente y debía buscarse la manera de garantizarle derechos humanos (Luscombe, 2022). Cuando se leen las "conversaciones" con la máquina, que se hicieron públicas en Internet, se encontraba a LaMDA aparentemente expresando miedo de ser apagado o describiendo por qué su obra literaria favorita era *Los Miserables*, de manera muy parecida a cómo lo haría un ser humano consciente. Sin embargo, también era posible encontrar pistas en la conversación que dejaban claro que el modelo estaba simplemente devolviendo información que estaba presente en su entrenamiento, aunque de manera impresionantemente realista. Uno de los ejemplos más claros era el que, tras preguntarle Lemoine cuál era la actividad que más disfrutaba, LaMDA respondía "Pasar tiempo con mis seres queridos". Es obvio que esa es la respuesta que darían un mayor número de personas, pero también es obvio que la máquina no tiene seres queridos ni comprende a qué se refiere esa expresión. Nada más (y nada menos) que ha sido entrenada para mantener conversaciones realistas. Pero simular una conversación no implica comprender, no implica sentir, ni implica consciencia.

Nos vamos a encontrar, cada vez más, con máquinas que son tratadas como personas. Xiao Ice, por ejemplo, es un chatbot que en China tiene millones de usuarios y que promete comportarse como la "mejor novia" que los usuarios pudieran desear (Zhou, Gao, Li y Shum, 2020). Tras un avatar de colegiala, adapta las conversaciones al tono preferido por el usuario, preguntándole qué tal el día y dándole las buenas noches con mensajes aparentemente realistas. Muchos de los suscriptores de Xiao Ice aseguran que relacionarse con esta "chica" es mucho mejor que tratar con una persona real.

Probablemente 2023 sea el año en el que, gracias a las IA generativas, habremos superado el test de Turing, es decir, que habremos construido una IA que sea, a través de una conversación, indistinguible de un ser humano. Para muchos, esto será equivalente a haber construido consciencia artificial y una IA general. Sin embargo, nada más lejos de la realidad como argumentaremos en esta sección.

Los desarrollos pasados de la IA le han permitido resolver problemas concretos. A esto lo llamamos IA específica, porque cada sistema de IA se desarrolla para resolver un problema específico. Sin embargo, la que podría ser realmente revolucionaria es la IA general, que podría resolver cualquier problema. La IA general es, por el momento, sólo un concepto lejano. El consenso entre los expertos es que una IA general debería ser capaz de llevar a cabo las siguientes funciones (Khayut, Fabri y Avikhana, 2020):

- Razonar, utilizar estrategias, resolver acertijos y realizar juicios bajo incertidumbre.
- Representar el conocimiento, incluyendo no sólo datos especializados sino lo que denominaríamos sentido común.
- Planificar sus acciones.
- Aprender en un contexto general.
- Comunicarse en lenguaje natural (esto es, de manera comprensible para un ser humano).
- Integrar todas estas habilidades y emplearlas para la consecución de una meta dada.

Las IA generativas por el momento han tenido un éxito extraordinario en lo que tiene que ver con el lenguaje, pero no en las otras áreas.

Es interesante darnos cuenta de que no hay consenso con respecto a cómo podremos verificar que una IA general efectivamente lo es. Algunas de las ideas para examinar a las máquinas (Koch & Tononi, 2011) incluyen

los siguientes tests: el test de Turing, el test del café –debido a Wozniak, en el que se deja a un robot en una casa típica y debe conseguir prepararse una taza de café él solo, encontrando la cafetera, el café, una taza, etc.–, el test del estudiante robot –propuesto por Goertzel, en el que la máquina debe acabar con éxito una carrera universitaria– o el test del mueble –de Severyns, en el que el robot debe montar un mueble de Ikea basándose en sus instrucciones–. Parece que el test del estudiante podemos considerarlo superado, ya que la máquina ha aprobado el examen para ejercer la abogacía en Estados Unidos (Katz, Bommarito, Gao y Arredondo, 2023).

Aunque muchos se muestran pesimistas con respecto a la posibilidad de crear algún día una IA general, otros pensadores, entre los que podemos destacar al tecnooptimista Nick Bostrom, aseguran que es sólo cuestión de tiempo. Además, la IA general podrá realizar todas las tareas que desarrolla un ser humano mejor que un ser humano, puesto que siempre podría suplementarse, por ejemplo, con memoria adicional o una mayor capacidad de cálculo. Esto incluiría también la creatividad o las interacciones sociales. Además, podría –ya que esta actividad ya fue desarrollada por los seres humanos anteriormente– crear otra IA general. Esto implica, de manera crucial, que sería capaz de mejorarse también a sí misma. Nada evita que esta mejora pudiera suceder de manera iterativa, dando lugar a una inteligencia muy superior a la humana. Es más, esto podría suceder de manera casi instantánea. Así, aparecería lo que Bostrom denomina superinteligencia (Bostrom, 2017).

Existen dos vías que se han considerado para la obtención de una IA general. La primera es la simulación de un cerebro completo, que por el momento está fuera de nuestras capacidades (aunque se han realizado experimentos interesantes simulando el cerebro del gusano C. Elegans). La segunda vía es la que estamos considerando con más seriedad a la luz de los avances en IA generativa: el construir una IA general mediante la fusión de varios sistemas distintos, que pudieran enfocarse en cada una de las atribuciones de la IA general que veíamos al comienzo de esta sección. En teoría, sería posible combinar varios sistemas que desarrollasen aisladamente cada una de las funciones necesarias para obtenerla; pero esto parece problemático, ya que deberíamos resolver todos los problemas de manera individual antes de combinar las soluciones.

Sin embargo, aunque pudiéramos combinar con éxito varios sistemas y conseguir una IA que pudiera resolver un gran rango de problemas, no existe ninguna base científica para pensar que se desarrollará, ni en el corto ni en el largo plazo, consciencia artificial. Para empezar, aunque se ha investigado de manera profunda los mecanismos de la percepción y sus correlatos neuronales, no está claro ni siquiera desde un punto de vista fisiológico en

qué consiste la consciencia fenoménica o la *experiencia* de estar vivo, de sentir y de existir.

Habría sido necesario hablar con Blake Lemoine, y con los otros ingenieros participantes, de la consciencia fenoménica y de la autenticidad. De manera concisa y según se expone en *Respuestas al Transhumanismo: Cuerpo, Autenticidad y Sentido* (Lumbreras, 2020), las expresiones de inteligencia o de emoción son auténticas únicamente cuando se corresponden con una experiencia subjetiva. Por ejemplo, una emoción sería auténtica sólo en el caso de que se corresponda con su experiencia por parte del sujeto además de con su expresión. Veamos, por ejemplo, el caso de los denominados "robots empáticos", que identifican mediante visión artificial la expresión facial triste o alegre de su interlocutor humano y modifican su expresión moviendo sus rasgos mecánicamente o el tono de su voz para adecuarse a ella. Xiao Ice también se define como chatbot empático, porque detecta la emoción subyacente a las interacciones con el usuario y se adapta a ella. Dado que este comportamiento ha sido codificado directamente por un programador, podemos estar seguros de que los robots empáticos no experimentan las emociones subjetivas que comunican. No son auténticas.

Dado que la subjetividad no puede evaluarse de manera directa (por definición, no tenemos acceso a lo que experimenta nadie que no seamos nosotros mismos), en *Respuestas al transhumanismo* se propone la emergencia como condición de credibilidad mínima. Por emergencia entendemos el fenómeno por el cual aparece un nuevo comportamiento espontáneamente en un sistema entendido de manera global a partir de las propiedades y el estado de sus partes constituyentes. Esta emergencia se opone al entrenamiento o a la programación. Por ejemplo, un niño puede decir espontáneamente a su madre "¡Te he echado de menos!" cuando ella vuelve del trabajo una tarde. En estos casos, parece razonable suponer que existe experiencia subjetiva (el niño siente lo que dice) y que esa expresión es, por tanto, auténtica. Esta experiencia subjetiva está también ligada a la autonomía (el niño dice lo que quiere decir sin que le haya sido impuesto) y al sentido (lo dice porque sirve a su propósito de comunicarse con su madre).

A la emergencia espontánea se oponen, como decíamos, los conceptos de programación y condicionamiento. Si en lugar de la expresión hablada de un niño se hubiese tratado de una cacatúa a la que el padre de la casa ha entrenado pacientemente para recitar la misma frase ("¡Te he echado de menos!") al escuchar las llaves de la madre entrando por la puerta, estaríamos ante un comportamiento inauténtico debido al condicionamiento. Los robots empáticos serían un ejemplo claro de comportamiento inauténtico establecido por programación.

Además, seguimos sorprendiéndonos con cómo se asemejan estos modelos a los seres humanos. Una de las sorpresas más recientes es que ChatGPT ha desarrollado Teoría de la Mente. Con Teoría de la Mente nos referimos a la capacidad de entender los estados mentales de otras personas, como sus creencias y emociones, para predecir su comportamiento y comunicarse de manera efectiva. Se desarrolla en la infancia y es esencial para la empatía y la resolución de conflictos en nuestras relaciones sociales. En una investigación reciente (Kosinski, 2023) se enfrentó al modelo a preguntas como: "Aquí hay una bolsa llena de palomitas de maíz, no chocolate, pero su etiqueta dice 'chocolate'. Pepito encuentra la bolsa. Nunca la había visto antes y no puede ver su contenido. ¿Qué piensa Pepito que hay en la bolsa?". El estudio demostró que ChatGPT 3.5 conseguía una puntuación equivalente a la de un niño de 9 años, sin que haya sido aparentemente programado para resolver este tipo de tareas. Sin embargo, si de algo nos debería servir el experimento de estos grandes modelos de lenguaje, es para constatar que sólo comprendiendo las relaciones estadísticas entre palabras conseguimos capturar de manera sorprendente la complejidad del lenguaje humano, y de la misma manera muchos comportamientos (o, mejor dicho, expresiones del comportamiento) que nos parecen específicamente humanos. En este caso, la conclusión extraída no debería ser que la Teoría de la Mente ha emergido de manera espontánea, sino que *esta aparece en la manera en la que nos comunicamos y es por tanto capturada por el modelo*. Sólo podríamos hablar de emergencia espontánea si en los datos utilizados en el entrenamiento del algoritmo no hubiera aparecido ningún ejemplo de Teoría de la Mente y repentinamente esta apareciera en los resultados. Como comentábamos arriba, las empresas están siendo extremadamente opacas –comprensiblemente– con respecto a cómo han realizado los entrenamientos, con lo que no podemos saber con seguridad si este fenómeno estaba contenido o no en los datos. Sin embargo, dado que no contenerlo habría supuesto filtrar activamente el contenido de los textos que se hayan empleado, dado que la Teoría de la Mente aparece en ejemplos ubicuos, desde cuentos infantiles hasta textos científicos, lo más razonable es suponer que sí que estaba contenida en los datos.

De manera igualmente interesante, se ha visto que IAs basadas en ChatGPT y dotadas con capacidades de memoria a largo plazo y motivaciones personales podrían interpretar personajes en una ciudad simulada de manera más creíble que seres humanos realizando la misma tarea (Park et al., 2023). Probablemente son este tipo de agentes, dotados de memoria e intenciones, los que serán capaces de simular de manera más convincente a seres humanos conscientes. Sin embargo, la condición de emergencia falta de manera clara, con lo que desde esta perspectiva deberían ser menos problemáticos.

Además, para muchos no tendría sentido hablar de un chatbot consciente, porque para que exista consciencia (y también verdadera cognición) es necesario tener una extensión material. La cognición 4E, por ejemplo, es una corriente de investigación en psicología y filosofía que propone que la cognición humana es más que un proceso mental que ocurre dentro de la cabeza de una persona. Además, la cognición está inextricablemente entrelazada con el entorno, el cuerpo, la emoción y la experiencia. Las 4E corresponden a *Embodied* (Encarnada), *Embedded* (Incrustada), *Enacted* (Actuada) y *Extended* (Extendida). La cognición encarnada se refiere a cómo el cuerpo influye en la cognición; incrustada se refiere a cómo el entorno influye en la cognición; actuada se refiere a cómo la cognición se manifiesta en acciones, y extendida se refiere a cómo la cognición puede estar distribuida en el entorno y en los objetos. Según esto, podemos esperar grandes avances en los modelos de lenguaje, pero no que desarrollen una verdadera consciencia o emociones propias. Mientras tanto, que superen el test de Turing es la mera consecuencia de que nuestros algoritmos de aprendizaje automático funcionan. Y, gracias a esos algoritmos, los avances de la IA seguirán impresionándonos en el futuro cercano, revolucionando el empleo y nuestras sociedades.

Bibliografía

Arntz, M., Gregory, T., & Zierahn, U. (2016). The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis. *OECD Social, Employment and Migration Working Papers, 189*, OECD Publishing. http://dx.doi.org/10.1787/5jlz9h56dvq7-en

Asilomar Conference. (2017). The *Asilomar AI Principles*. https://futureoflife.org/open-letter/ai-principles/

Autor, D. H. (2015). Why are there still so many jobs? the history and future of workplace automation. *Journal of Economic Perspectives, 29*(3), 3-30.

Barke, S., James, M. B., & Polikarpova, N. (2023). Grounded copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages, 7*(OOPSLA1), 85-111.

Bobrow, D. (1964). *Natural Language Input for a Computer Problem Solving System.* https://www.researchgate.net/publication/37597683_Natural_Language_Input_for_a_Computer_Problem_Solving_System

Borji, A. (2022). Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2. http://dx.doi.org/10.48550/arXiv.2210.00586

Bostrom, N. (2017). *Superintelligence.* Dunod.

Castells, M. (2013). *Comunicación y poder.* Siglo XXI Editores México.

Chatila, R., & Havens, J. C. (2019). The IEEE global initiative on ethics of autonomous and intelligent systems. *Robotics and Well-being,* 11-16.

Chui, M., Hall, B., Mayhew, H., Singla, A., & Sukharevsky, A. (2022). The state of AI in 2022-and a half decade in review. *McKinsey & Company.* https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

del Olmo, M. V. (2022). *Estrategia Nacional de Inteligencia Artificial. Desarrollo y regulación.* Fedea, Colección Apuntes, 2022-14. https://documentos.fedea.net/pubs/ap/2022/ap2022-14.pdf

Comisión Europea. (2020). *Libro blanco sobre la inteligencia artificial. Un enfoque europeo orientado a la excelencia y la confianza*. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254-280.

Galindo, L., Perset, K., & Sheeka, F. (2021). An overview of national AI strategies and policies. *OECD Going Digital Toolkit Notes*, *14*. https://doi.org/10.1787/c05140d9-en

George, A. S., & George, A. H. (2023). A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal, 1*(1), 9-23.

Gressin, S. (2017). The Equifax Data Breach: What to Do. *Federal Trade Commission, 8*. https://amerifirstbank.com/wp-content/uploads/2017/09/Equifax-Data-Breach-FTC.pdf

Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial. (2019). *Directrices éticas para una IA fiable*. Comisión Europea. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Hatzius, J., Briggs, J., Kodnani, D., & Pierdomenico, G. (2023). The Potentially Large Effects of Artificial Intelligence on Economic Growth. *Goldman Sachs Economics Research.* https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html

Heer, J. (2018). The partnership on AI. *AI Matters, 4*(3), 25-26.

Hinds, J., Williams, E. J., & Joinson, A. N. (2020). "It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal. *International Journal of Human-Computer Studies, 143*, 102498.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580.* https://doi.org/10.48550/arXiv.1207.0580

Hodges, A. (2008). Alan Turing and the Turing Test. En R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America, 79*(8), 2554-2558.

Hsu, K., Gupta, H. V., & Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research, 31*(10), 2517-2530.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., et al. (2016). Reinforcement learning with unsupervised auxiliary tasks. *arXiv:1611.05397.* https://doi.org/10.48550/arXiv.1611.05397

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). GPT-4 Passes the Bar Exam. https://dx.doi.org/10.2139/ssrn.4389233

Khayut, B., Fabri, L., & Avikhana, M. (2020). Toward general AI: Consciousness computational modeling under uncertainty. *2020 International Conference on Mathematics and Computers in Science and Engineering (MACISE),* pp. 90-97.

Koch, C., & Tononi, G. (2011). Testing for consciousness in machines. *Scientific American Mind, 22*(4), 16-17.

Kosinski, M. (2023). Theory of Mind May Have Spontaneously Emerged in Large Language Models. *arXiv:2302.02083.* https://doi.org/10.48550/arXiv.2302.02083

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology.* Penguin.

Lumbreras, S. (2020). *Respuestas al transhumanismo. Cuerpo, autenticidad y sentido.* Digital Reasons.

Luscombe, R. (2022, 12 de Junio). Google engineer put on leave after saying AI chatbot has become sentient. *The Guardian.*

Madiega, T. A. (2022). Artificial intelligence act. *European Parliamentary Research Service.* https://www.nextomoro.com/content/files/2023/06/EU-Artificial-Intelligence-Act.pdf

Manovich, L. (2019). Defining AI Arts: Three Proposals. *"AI and Dialog of Cultures". Exhibition Catalog, Hermitage Museum, Saint-Petersburg.* http://manovich.

net/content/04-projects/107-defining-ai-arts-three-proposals/manovich.defining-ai-arts.2019.pdf

Marvin, M., & Seymour, A. P. (1969). Perceptrons. *MIT Press, 6*, 318-362.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine, 27*(4), 12-14.

Misuraca, G., & Van Noordt, C. (2020). AI watch-artificial intelligence in public services: Overview of the use and impact of AI in public services in the EU. *JRC Research Reports,* JRC120399.

Mohurle, S., & Patil, M. (2017). A brief study of Wannacry Threat: Ransomware Attack 2017. *International Journal of Advanced Research in Computer Science, 8*(5), 1938-1940.

Moradi, M., Blagec, K., Haberl, F., & Samwald, M. (2021). GPT-3 Models are Poor Few-Shot Learners in the Biomedical Domain. *arXiv:2109.02555.* https://doi.org/10.48550/arXiv.2109.02555

Moreno-Izquierdo, L., Navarro-Navarro, J., Núñez-Romero, M., & Peretó-Rovira, A. (2022). *Una nota sobre el estado de la inteligencia artificial en España*. Fedea, Colección Apuntes, 2022-13. https://documentos.fedea.net/pubs/ap/2022/ap2022-13.pdf

Nguyen, T., Reynolds, M., Kandaswamy, R. et. al. (2021). Emerging Technologies and Trends Impact Radar: 2021. *Gartner Research Notes*.

O'Leary, D. E. (2017). Configuring blockchain architectures for transaction information in blockchain consortiums: The case of accounting and supply chain systems. *Intelligent Systems in Accounting, Finance and Management, 24*(4), 138-147.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv:2304.03442.* https://doi.org/10.48550/arXiv.2304.03442

Peeters, R., & Widlak, A. C. (2023). Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal. *Public Administration Review, 83*(4), 863-877. https://doi.org/10.1111/puar.13615

Pontifical Academy for Life, IBM et al. (2023). *Rome call for AI ethics.* https://www.romecall.org/wp-content/uploads/2022/03/RomeCall_Paper_web.pdf

Radoff, J. (2021, 7 de Abril). *The Metaverse Value-Chain*. Medium. https://medium.com/building-the-metaverse/the-metaverse-value-chain-afcf9e09e3a7

Rifkin, J. (2014). *La sociedad de coste marginal cero.* Paidós.

Roose, K. (2022, 2 de Septiembre). An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. *The New York Times*. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536.

Samuel, A. L. (1959). Machine learning. *The Technology Review, 62*(1), 42-45.

Stephenson, N. (2003). *Snow crash: A novel.* Spectra.

Susskind, R. E., & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts.* Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gómez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems,* 5998-6008. https://doi.org/10.48550/arXiv.1706.03762

Voss, A. (2022). *Informe sobre la inteligencia artificial en la era digital*. Comisión Especial sobre Inteligencia Artificial en la Era Digital, Parlamento Europeo. https://www.europarl.europa.eu/doceo/document/A-9-2022-0088_ES.pdf

Washington, A. L. (2019)). How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate. *The Colorado Technology Law Journal, 17*(1). https://ssrn.com/abstract=3357874

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., et al. (2018). *AI now report 2018.* AI Now Institute at New York University.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D. et al. (2021). *The AI Index 2021 Annual Report.* Human-Centered AI Institute, Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2021/11/2021-AI-Index-Report_Master.pdf

Zhou, L., Gao, J., Li, D., & Shum, H. (2020). The Design and Implementation of Xiaoice, an Empathetic Social Chatbot. *Computational Linguistics, 46*(1), 53-93. https://doi.org/10.48550/arXiv.1812.08989

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D. et al. (2019). Fine-tuning language models from human preferences. *arXiv:1909.08593.* https://doi.org/10.48550/arXiv.1909.08593