Apuntes

Inteligencia artificial y ciencias del comportamiento

ÁLVARO GAVIÑO GONZÁLEZ

Apuntes 2025/15
Marzo de 2025

fedea

Inteligencia artificial y ciencias del comportamiento

Álvaro Gaviño González
(Behavioral Economics Principal Manager en BBVA)
Marzo de 2025

1. Introducción

En este trabajo se hace retrospectiva y se realizan algunas reflexiones en el ámbito de la intersección entre la Inteligencia Artificial (IA) y las ciencias del comportamiento o '*behavioral science*'. A ese respecto, siendo conscientes de la multiplicidad de efectos de la IA y de la rapidez en sus avances, que hacen que el acto de escribir sobre Inteligencia Artificial (IA) sea en cierta medida un acto de mirada al pasado, resulta útil reflexionar sobre cómo la inteligencia artificial está cambiando la manera en que se entiende el comportamiento humano y las posibilidades que abre. En la dirección contraria, también cómo las ciencias del comportamiento pueden ayudar a moldear el desarrollo de la IA y su adopción. Para ello, el punto de partida es tener en cuenta que, aunque la atención mediática actual se centra modelos de IA capaces de generar texto (ChatGPT, Gemini), imágenes (DALL-E, Midjourney), música (Amper, Aiva), video (Sora, Runway), avatares digitales (Synthesia, HeyGen) y un largo etcétera, la IA es mucho más que 'GenAI' (del inglés *Generative AI*). De hecho, en la mayoría de los casos, estos modelos no generan conocimiento contrastado, sino "hipótesis" de contenido correcto que, si bien muchas veces son acertadas, no lo son siempre.

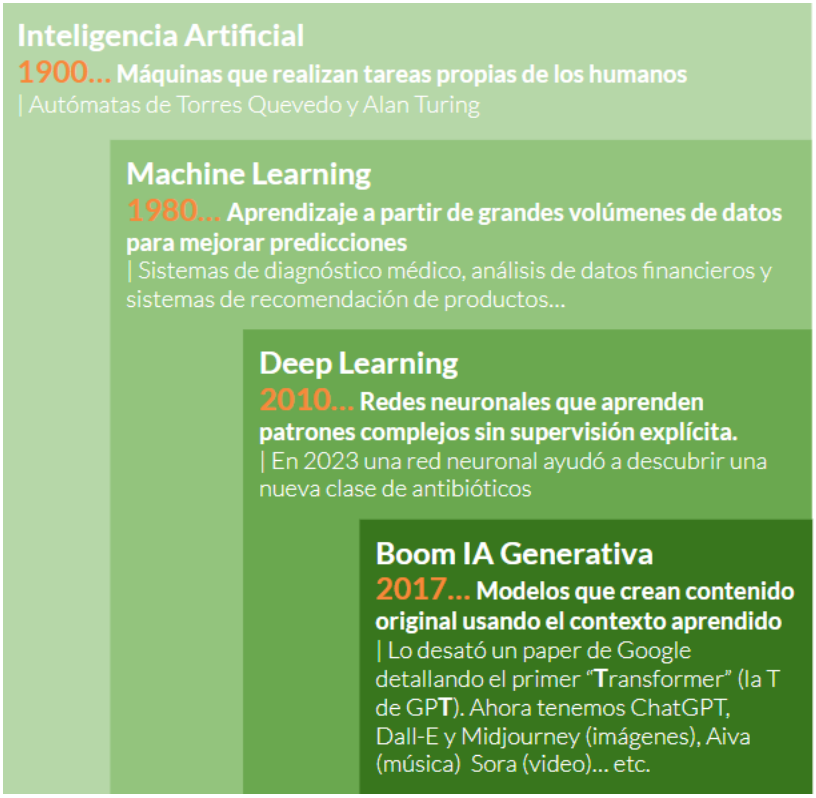
2. Breve historia de la IA y su intersección con las ciencias del comportamiento

En 1914, el español Leonardo Torres Quevedo desarrolló uno de los primeros autómatas "modernos" capaces de jugar al ajedrez, anticipándose en unos 35 años a Alan Turing y sus "máquinas universales". Turing da nombre a la famosa "Prueba de Turing", diseñada para determinar si una máquina exhibe un comportamiento indistinguible del de un ser humano. Fue en la conferencia de Dartmouth de 1956 donde John McCarthy acuñó el término Inteligencia Artificial como 'la ciencia e ingenio de hacer máquinas inteligentes'. En dicha conferencia, que se considera como el nacimiento de la IA como campo formal de estudio, Herbert Simon, economista, sociólogo y psicólogo estadounidense (Premio Nobel de Economía en 1978) junto a otros científicos, estableció las bases del campo. La relevancia de Simon radica en sus estudios sobre la racionalidad limitada y la toma de decisiones en entornos de información incompleta, conceptos clave tanto para la IA como para las ciencias del comportamiento, por lo que es considerado uno de los padres de ambas disciplinas.

Desde la década de 1980 (véase Figura 1), el aprendizaje automático (ML) comenzó a aplicarse en sistemas como el diagnóstico médico (por ejemplo, en la detección temprana de enfermedades), el análisis de datos financieros (por ejemplo, en la evaluación de riesgos en inversiones o fraudes bancarios) y las recomendaciones de productos (como en los algoritmos de Amazon en los años 90). Durante esa misma época y hasta bien entrados la primera década de este siglo, otra rama tuvo un gran auge: los sistemas expertos basados en reglas y lógica simbólica. De hecho, estos sistemas

expertos dominaron durante mucho tiempo el panorama de la IA aplicada, mientras que el aprendizaje automático era un campo de investigación con menos aplicaciones prácticas.

Figura 1. Esquema de la evolución de la inteligencia artificial



Fuente: elaboración propia.

Posteriormente, ya hacia 2010, el *deep learning* con redes neuronales más complejas permitió importantes avances en diversos campos. Su desarrollo ha permitido el reconocimiento más preciso de imágenes médicas, la conducción autónoma y la implementación de asistentes virtuales capaces de comprender y responder de manera más natural a las consultas humanas. Además de estos avances, el *deep learning* es culpable de otros logros notables más recientes. En 2016, el programa AlphaGo, desarrollado por DeepMind, hizo historia al derrotar al campeón mundial de Go, Lee Sedol. Este hito demostró la capacidad del *deep learning* para dominar juegos de estrategia complejos que anteriormente se consideraban inalcanzables para la inteligencia artificial.

La actual euforia en torno a la IA generativa tomó umpulso gracias al conocido trabajo de Google sobre Transformers, *Attention Is All You Need,* que desató una avalancha de herramientas generativas pero, quizá más importante, inspiración renovada sobre las posibilidades futuras. Curiosamente, los mecanismos de atención en IA, específicamente en modelos como los Transformers, se pueden ver como una forma de abordar esa limitación compartida entre humanos y máquinas: esa capacidad finita de atención, procesamiento o racionalidad, ya descrita en los inicios de la IA moderna por Herbert Simon con el concepto de '*bounded rationality*'. Esto es, así como los humanos deciden a qué prestar atención para hacer más eficiente su proceso de toma de decisiones, o incluso cometen errores por los mecanismos de ahorro de esfuerzo cognitivo, los modelos extensos de lenguaje (LLMs por sus siglas en inglés) utilizan mecanismos de *self-attention*

para enfocarse solo en las partes relevantes de una secuencia, filtrando la información y priorizando lo que es más importante para el contexto que les ocupa.

Un ejemplo histórico bien conocido, que se refleja bien en la película “The Imitation Game”, es cuando en el contexto de la construcción de su autómata Alan Turing se da cuenta de que podría haber tenido éxito mucho antes si le hubiera dado a su máquina pistas sobre dónde buscar, esto es, algún tipo de mecanismo de atención. En el giro final se da cuenta que varios lingüistas a los que ningunea y despide al inicio de su proyecto le habrían podido dar esa solución desde un primer momento (salvando quizá todavía más vidas). Su ego, o quizá su sesgo de confirmación (él era matemático) le hace quedarse solo con matemáticos e ingenieros en su equipo

Debe también señalarse que en el ámbito de la “IA no generativa” se está produciendo también desarrollos importantes. Un ejemplo reciente se refiere al descubrimiento de nuevos antibióticos al analizar grandes conjuntos de datos y simular interacciones moleculares, demostrando aplicaciones de gran impacto más allá de la generación de contenido.

En realidad, la IA está en nuestras vidas desde hace décadas, aunque no siempre seamos conscientes de ello. Pasado un tiempo de que una máquina automatice o mimetice una tarea humana, el imaginario común deja de considerar el hecho como una victoria de la inteligencia artificial. Hay muchos ejemplos que, por su cotidianidad, descontamos. Entre ellos se encuentran los filtros de spam en correos electrónicos, el reconocimiento facial en imágenes o los sistemas de navegación GPS. Larry Tesler sostenía que "la inteligencia artificial es todo aquello que todavía no se ha hecho automáticamente por una máquina". O lo que es lo mismo, en cuanto la IA consigue algo, ya no es IA.

En términos de comportamiento humano, suele verse como especial o innovador aquello que es nuevo y que todavía no hemos internalizado como parte de nuestra realidad cotidiana. Una vez que una tecnología se vuelve común, la normalizamos y pierde ese "aura" de innovación. La IA no es la excepción; lo que hoy nos parece ciencia ficción e identificamos como nuevas fronteras de la inteligencia artificial dejarán de parecerlo una vez que se integre plenamente en nuestro entorno, como una suerte de *heurística de disponibilidad*[1] invertida, donde lo conocido deja de tener tanto impacto psicológico como lo nuevo. La percepción de lo que constituye "inteligencia artificial" se adapta a nuestro comportamiento y expectativas.

La IA es cualquier artefacto capaz de imitar una habilidad humana. En ese sentido, no solo seguirá imitándolas con éxito, sino que nos superará con capacidades no alcanzables por restricciones humanas físicas o intelectuales. La pregunta importante es ¿manejaremos bien esas nuevas todopoderosas herramientas?, ¿seremos responsables?, ¿quién ganará la competición sobre su control y uso?

[1] Cuanto más accesible sea un suceso, más frecuente y probable parecerá; cuanto más viva sea la información, más convincente y fácil de recordar será; y cuanto más evidente resulte algo, más causal parecerá (Plous, S. (1993): *The availability heuristic. The psychology of judgment and decision making*. McGraw-Hill, NY).

3. Reflexiones sobre IA

3.1 Asistentes *vs* Agentes

No es lo mismo un asistente de IA que un agente de IA. Los asistentes sugieren, los agentes ejecutan. Esa es la diferencia fundamental. Mientras los asistentes amplifican nuestras capacidades, los agentes llevan la automatización un paso más allá, pudiendo operar de forma autónoma dentro de un marco determinado. Esta evolución no es inocua ya que marca un punto de inflexión en la relación entre humanos y máquinas.

Los asistentes de IA, también conocidos como copilotos optimizan procesos, automatizan tareas repetitivas y ofrecen apoyo en la toma de decisiones. Plataformas como Netflix, Spotify o Amazon personalizan recomendaciones en función del comportamiento previo del usuario. Herramientas como Microsoft Copilot o GitHub Copilot proporcionan asistencia en la redacción de documentos o en programación, generando fragmentos de código y ofreciendo sugerencias en tiempo real. Asistentes como Alexa, Siri y Google Assistant han integrado la IA en la vida cotidiana, facilitando la gestión de tareas y el control de dispositivos mediante comandos de voz. Sin embargo, todos estos sistemas requieren intervención y validación humana. Son apoyos inteligentes, pero no sustituyen el juicio ni la acción del usuario.

Los agentes de IA a diferencia de los asistentes pueden ejecutar tareas sin supervisión humana directa. Su diseño permite que operen de manera independiente, tomando decisiones basadas en datos, contexto y objetivos previamente establecidos. Por ejemplo, existen multitud de algoritmos de trading automático que pueden ejecutar operaciones bursátiles en función de patrones de mercado sin intervención humana (son sonadas las ocasiones en las que distorsionan el mercado). Los vehículos autónomos como los desarrollados por Tesla o Waymo ya circulan en diversas ciudades, analizando el entorno en tiempo real y tomando decisiones sin intervención humana.

Marvin Minsky, otro de los padres fundadores de la Inteligencia Artificial, veía la IA como una herramienta fundamental para comprender la mente humana. Su enfoque se basaba en la idea de que la inteligencia no surge de una única entidad central, sino de la interacción de múltiples agentes más simples. Minsky propuso que la mente es un sistema complejo compuesto por numerosos "agentes" que trabajan en conjunto, cada uno especializado en una tarea específica. Esta teoría de la "Sociedad de la Mente" influyó profundamente en la forma en que se abordó la IA, alejándose de la búsqueda de una única "inteligencia general" y enfocándose en cambio en la creación de sistemas que simularan la interacción de múltiples procesos cognitivos. Para Minsky, la IA no era solo sobre construir máquinas inteligentes, sino sobre usar esas máquinas para desentrañar los misterios de cómo pensamos y cómo funciona la mente humana.

Esta evolución plantea preguntas fundamentales sobre el control y la regulación. Mientras que los asistentes dependen de la supervisión humana, los agentes pueden operar en escenarios complejos sin intervención. Esto introduce desafíos en términos de confiabilidad, ética y seguridad. ¿Hasta qué punto podemos delegar decisiones en la IA sin perder la capacidad de corregir errores? Paul Daugherty y H. James Wilson, en *Human + Machine*, exploran cómo la colaboración entre humanos

y máquinas está redefiniendo la productividad y la toma de decisiones. Pero si esa colaboración se convierte en dependencia, podríamos estar cediendo más control del que creemos.

La transición desde copilotos a agentes será progresiva. Muchas aplicaciones de IA actuales combinan elementos de ambas categorías. GitHub Copilot, por ejemplo, no solo sugiere código, sino que puede escribir fragmentos completos de manera autónoma. Google Assistant está evolucionando para realizar tareas más complejas sin necesidad de confirmación humana. Pero, ¿cuánta supervisión humana es suficiente para que sigamos siendo los responsables de sus decisiones? ¿Quién definirá esa línea?

Nos encontramos pues ante un panorama en el que la IA ya no solo nos asiste, sino que empieza a actuar por nosotros. La cuestión no es si los agentes autónomos serán útiles, sino cómo encontraremos el equilibrio entre su potencial y la supervisión adecuada en cada momento. La gran pregunta sigue en el aire: ¿seguiremos al volante o simplemente nos convertiremos en pasajeros?

3.2 Robótica e IA general

La inteligencia artificial y la robótica han recorrido caminos paralelos durante décadas, con una intersección cada vez mayor. Mientras la robótica se ha enfocado en la creación de sistemas mecánicos capaces de interactuar con el mundo físico, la "IA lógica" ha evolucionado en su capacidad para procesar información, aprender y tomar decisiones. La convergencia de ambas ya está transformando radicalmente la sociedad, dando lugar a una nueva generación de sistemas autónomos y agentes - con capacidad de acción- inteligentes.

Ya desde la antigüedad, los humanos hemos intentado replicar nuestras habilidades mecánicas mediante mecanismos automatizados[2]. La robótica moderna despegó en la segunda mitad del siglo XX con la aparición de los primeros robots industriales, como el Unimate, introducido en 1961 en las fábricas de General Motors. Estos robots eran programables y realizaban tareas repetitivas con gran precisión, pero carecían de capacidad de aprendizaje o adaptación. Con el tiempo, la robótica ha evolucionado hasta incorporar visión artificial, procesamiento del lenguaje natural y aprendizaje automático, permitiendo el desarrollo de máquinas capaces de percibir su entorno, adaptarse a los cambios y tomar decisiones en tiempo real. Ejemplos como el robot humanoide ASIMO de Honda o los avances en movilidad de Boston Dynamics ilustran la velocidad con la que la IA se ha integrado en los sistemas robóticos. En los últimos años, la combinación de redes neuronales profundas y técnicas de refuerzo ha impulsado desarrollos que hasta hace poco parecían ciencia ficción. Elon Musk, a través de Tesla, ha presentado el robot humanoide Optimus, que pretende realizar tareas domésticas y laborales con un nivel de autonomía cada vez mayor. Paralelamente, la inteligencia artificial ha encontrado en la robótica una forma de materializarse en el mundo físico, con aplicaciones que van desde la automatización industrial hasta la exploración espacial

[2] En la Grecia clásica, Herón de Alejandría diseñó dispositivos hidráulicos que realizaban movimientos predeterminados, y en la dinastía Tang en China se construyeron figuras mecánicas capaces de tocar instrumentos musicales. Considerando estos artefactos como mimetizadores de habilidades humanas, podríamos decir que la IA está con nosotros, o al menos su anhelo, desde hace milenios, mucho antes de la era computacional.

La intersección entre IA y robótica plantea preguntas fundamentales sobre su impacto en la sociedad. *Superintelligence* de Nick Bostrom advierte sobre los riesgos de delegar decisiones en máquinas sin una adecuada supervisión humana. Por otro lado, *The Age of Em* de Robin Hanson plantea escenarios en los que la IA y la automatización llevarían a una reconfiguración del empleo y la economía. Mientras tanto, *Life 3.0* de Max Tegmark explora las implicaciones filosóficas y éticas de una IA que no solo sea capaz de pensar, sino también de actuar en el mundo físico con una autonomía sin precedentes.

Los avances en IA y robótica nos llevan a un punto en el que la línea entre software y hardware comienza a desdibujarse. En la actualidad, la IA no solo facilita la creación de asistentes virtuales o herramientas de productividad, sino que también está dando forma a robots capaces de operar de manera autónoma en fábricas, hospitales y hogares. La convergencia entre ambas disciplinas nos acerca a una realidad en la que la presencia de sistemas inteligentes en el mundo físico será tan común como lo son hoy los algoritmos en el ámbito digital. La gran pregunta es si seremos capaces de controlar esta evolución o si, como en tantos otros momentos de la historia, la tecnología avanzará más rápido que nuestra capacidad de comprender sus consecuencias.

La "guerra" por el control de la IA se extiende también a la robótica, donde las decisiones sobre quién desarrolla y posee estas tecnologías definirán el equilibrio de poder en el futuro. En un mundo donde lo digital cada vez tiene más relevancia, la “inteligencia artificial lógica”, irá también ganado capacidad para saltar de actuar como copilotos a agentes capaces de ejecutar tareas como organizar nuestra agenda, enviar correos electrónicos por nosotros o, por ejemplo, diseñar una estrategia de marketing, generarla de manera creativa y lanzarla al mercado.

En este contexto, la Inteligencia Artificial General (AGI, por sus siglas en inglés) representa la culminación de un objetivo largamente perseguido por científicos, ingenieros, psicólogos y filósofos: una inteligencia artificial que no solo ejecute tareas específicas mejor que los humanos sino que posea un nivel de cognición y adaptabilidad comparable o superior al nuestro en cualquier dominio.

La pregunta sobre si la Inteligencia Artificial General (IAG) es una meta lejana o una realidad en construcción ya no es meramente especulativa. Los modelos actuales, como GPT-4, Gemini Ultra, Claude 3 y DeepSeek, están demostrando habilidades que antes se consideraban exclusivas de la cognición humana. Ya no hablamos de simples autómatas que ejecutan tareas específicas; hoy tenemos sistemas capaces de debatir sobre ciencia, responder en cualquier idioma, resolver problemas matemáticos avanzados y generar código con una precisión superior a la de muchos programadores humanos.

Los datos son elocuentes: Gemini Ultra ha superado a expertos humanos en el benchmark MMLU, alcanzando un 90% de precisión en pruebas de conocimiento general. DeepSeek ha logrado avances en resolución de problemas matemáticos complejos, superando a modelos occidentales en ciertas áreas. Mientras tanto, GPT-4 ya ha demostrado rendimientos similares a los de abogados y médicos en exámenes profesionales, y Claude 3 se ha posicionado como uno de los modelos más eficientes en tareas de alineamiento ético y razonamiento abstracto. Es cierto que estos modelos todavía carecen de ciertas capacidades humanas, como la generación de metas propias o el razonamiento

causal profundo. Sin embargo, su capacidad de aprendizaje y adaptación es tal que la línea entre "IA especializada" e "IA general" se vuelve cada vez más difusa. La pregunta ya no es si la IAG es posible, sino cuándo será innegable que la hemos alcanzado.

Si la información es poder, la IAG puede ser el poder absoluto. El país o corporación que logre desarrollar una inteligencia general antes que el resto tendrá una ventaja estratégica sin precedentes en defensa, economía e innovación. No es casualidad que esta carrera tecnológica se asemeje a la carrera nuclear del siglo XX: quien domine la IAG podrá reescribir las reglas del juego global. Estados Unidos, China y Europa están en una pugna constante por el liderazgo en inteligencia artificial. OpenAI, Google DeepMind y Anthropic en Occidente compiten contra gigantes chinos como Baidu, Alibaba y DeepSeek. Mientras EE.UU. apuesta por la supremacía computacional a través de alianzas con Microsoft y Nvidia, China está invirtiendo masivamente en modelos de IA propios, sorteando restricciones tecnológicas y desarrollando chips alternativos para sostener su infraestructura de IA.

Las implicaciones son profundas. La IAG podría revolucionar la ciencia al resolver problemas matemáticos imposibles para los humanos o encontrar una unificación entre la mecánica cuántica y la relatividad general. Pero también podría ser la base de sistemas autónomos de vigilancia, propaganda digital masiva y armas de decisión autónoma. En este contexto, la pregunta clave no es solo quién desarrollará la primera IAG, sino qué valores tendrá integrada.

4. IA y ciencias del comportamiento

4.1 La simulación de Sistemas Complejos

Las ciencias del comportamiento han demostrado que nuestras decisiones no son completamente racionales ni aleatorias, sino que están sesgadas de manera sistemática. Estos patrones predecibles de error, lejos de ser una anomalía, constituyen una oportunidad. En la medida en que el comportamiento humano puede modelarse y simularse, podemos anticipar decisiones, evaluar políticas y diseñar intervenciones más efectivas. Aquí es donde la Inteligencia Artificial entra en juego, no solo como herramienta analítica, sino como arquitecta de modelos de simulación poblacional que trascienden el análisis convencional.

Figura 2: Diagrama de la web de la universidad Carnegie Mellon University (2020)

La economía del comportamiento nació como una respuesta a la visión de la economía neoclásica que asumía agentes racionales maximizadores de utilidad. Kahneman y Tversky desmontaron esa idea demostrando que nuestras decisiones son moldeadas por heurísticas y sesgos cognitivos. Lo que empezó como una crítica terminó convirtiéndose en una metodología para entender y predecir el comportamiento. A partir de ahí, la economía del comportamiento (*Behavioral Economics*) se expandió a un marco más amplio, la ciencia del comportamiento (*Behavioral Science*) integrando psicología, sociología y neurociencia en el estudio de la toma de decisiones. Si nuestras decisiones son predecibles, también lo son nuestras interacciones en sociedad, abriendo la puerta a un nivel más ambicioso de modelado: la simulación de sistemas complejos.

La evolución desde un enfoque puramente económico hacia una disciplina más amplia que integra la psicología, la sociología, la antropología y la neurociencia ha abierto nuevas puertas para la simulación de sistemas complejos. Si las decisiones humanas responden a patrones identificables, entonces pueden replicarse en entornos digitales. Esto es precisamente lo que están haciendo empresas de consultoría, entidades financieras y gobiernos: construir modelos de simulación basados en agentes que representen individuos con características diversas y comportamientos realistas, alimentados por datos reales.

Además del ya mencionado Herbet Simon, otros autores dentro de la disciplina de *Behavioral Economics* también han explorado la relación entre la inteligencia artificial y el comportamiento humano. Daniel Kahneman, psicólogo y Premio Nobel de Economía, ha abordado cómo la IA podría ayudar a reducir sesgos cognitivos en la toma de decisiones, mientras que Richard Thaler (también Premio Nobel de Economía) ha discutido la posibilidad de usar la IA para implementar nudges[3] más eficientes y personalizados. Cass Sunstein ha examinado cómo las políticas públicas podrían usar tanto *Behavioral Economics* como IA para mejorar la eficiencia de intervenciones gubernamentales y ayudar a las personas a tomar decisiones más informadas. Sendhil Mullainathan, *professor* en "Computation and Behavioral Science" en Chicago Booth School of Business y en el MIT, ha explorado cómo el aprendizaje automático puede ser utilizado para abordar problemas sociales, integrando perspectivas de la economía conductual para evitar decisiones sesgadas. Estas contribuciones han ayudado a expandir los límites de la IA y su aplicación en la mejora de nuestras decisiones, abriendo nuevas oportunidades para que la tecnología y la psicología trabajen juntas en favor del bienestar humano.

Los modelos de simulación de agentes (*agent-based models*, ABM) han madurado hasta convertirse en herramientas comerciales que permiten anticipar dinámicas de consumo, respuesta a políticas económicas y hasta reacciones ante crisis. Empresas como Simudyne, AnyLogic y consultoras como McKinsey y BCG ya están ofreciendo soluciones de modelado basado en agentes para sectores como banca, energía y salud. Por ejemplo, estos modelos permiten a los bancos simular cómo reaccionarán sus clientes ante una subida de tipos de interés, un cambio en las condiciones de las

[3] La teoría del empujoncito (*nudge theory* en inglés) es un concepto de las ciencias del comportamiento que sugiere que el refuerzo positivo, los cambios en el contexto y las sugerencias indirectas pueden influir en la toma de decisiones y el comportamiento tanto de individuos como de grupos. Contrasta con otras formas de lograr el cumplimiento, como la educación y la legislación, teniendo un enfoque más sutil y menos directo.

hipotecas o la introducción de un nuevo producto financiero. Es una simulación basada en datos, calibrada con comportamientos históricos y ajustada en tiempo real con nuevas fuentes de información. Estos sistemas no solo replican decisiones individuales, sino que permiten modelar interacciones sociales y dinámicas emergentes. Si un banco quiere lanzar una nueva tarjeta de crédito con recompensas específicas, puede utilizar un modelo de agentes para predecir qué segmentos de clientes la adoptarán primero, cómo se propagará por el *word of mouth* y qué efecto tendrá en la rentabilidad general. Las *fintechs* ya están utilizando modelos predictivos para identificar en qué ciudades tendrán más éxito y cómo impactarán a la competencia. En el mismo sentido, compañías aseguradoras utilizan modelos de simulación para predecir el comportamiento de sus clientes y ajustar primas de riesgo de manera más precisa. Firmas de marketing digital han empezado a reemplazar los estudios de mercado tradicionales con simulaciones de comportamiento basadas en datos de redes sociales, compras y movilidad.

El desarrollo de *gemelos digitales* de poblaciones es una de las aplicaciones más ambiciosas de este enfoque. La idea de construir una réplica digital de la población de un país, con individuos que toman decisiones financieras, laborales y de consumo dentro de un entorno virtual, ya está en marcha. Instituciones gubernamentales están explorando estas herramientas para simular el impacto de nuevas políticas económicas antes de implementarlas.

El realismo de todas estas simulaciones depende de la calidad de los datos y de la capacidad de integrar múltiples fuentes en tiempo real. Ya no basta con modelos basados en datos históricos; los sistemas más avanzados combinan datos de transacciones bancarias, redes sociales, patrones de movilidad y noticias económicas para ajustar la simulación dinámicamente. Los avances en IA permiten que los agentes dentro del modelo aprendan y evolucionen con cada nueva iteración, refinando su comportamiento con algoritmos de aprendizaje automático.

El sector financiero está liderando esta transformación, pero la tendencia se extiende a otros ámbitos. Gobiernos han utilizado modelos similares para simular el impacto de pandemias y políticas de confinamiento. Empresas de logística optimizan rutas y flujos de distribución con modelos que simulan el comportamiento de consumidores y transportistas. La planificación urbana se está beneficiando de simulaciones que predicen el impacto de nuevas infraestructuras en la movilidad y el mercado inmobiliario. Ya hay consultoras vendiendo estos modelos como servicio y empresas incorporándolos en su toma de decisiones estratégicas. La combinación de la economía del comportamiento, el modelado basado en agentes y la inteligencia artificial no es solo una posibilidad, es una revolución en marcha.

Este tipo de simulaciones no solo sirven para prever escenarios, sino para influir en ellos. Si entendemos qué factores llevan a una persona a cambiar de banco, contratar un producto financiero o adoptar una nueva tecnología de pagos, podemos diseñar estrategias para inducir esos comportamientos de manera más efectiva. Aquí surge la pregunta ética inevitable: ¿hasta qué punto estamos utilizando estas herramientas para comprender el comportamiento humano y hasta qué punto para manipularlo? Estos avances también plantean otros desafíos éticos. La calidad de los modelos dependerá de los datos utilizados para entrenarlos, lo que implica riesgos de sesgo si la

información no es representativa o si se usa de manera manipulativa. Además, la interpretación de los resultados debe realizarse con cautela para evitar errores en la toma de decisiones reales.

En cualquier caso, la combinación de IA y economía del comportamiento en la simulación de sistemas complejos ofrece una visión sin precedentes sobre cómo los individuos y las sociedades toman decisiones. Si se desarrolla con rigor y ética, esta tecnología podría transformar la forma en que comprendemos el comportamiento humano y nos permitiría diseñar estrategias más efectivas y equitativas para el futuro.

4.2 Determinismo y predictibilidad

En la actualidad no se puede demostrar que el mundo sea determinista o que no lo sea, y es poco probable que podamos llegar a comprobar cualquiera de las opciones algún día (quizá una AGI super avanzada nos de la respuesta). Einstein podría tener razón con su explicación de las características no "encajables" en la física clásica de la mecánica cuántica mediante las variables ocultas que ahora simplemente no podemos ver o que no podemos capturar con la precisión adecuada. Mi conjetura es que el universo sí es determinista o, como mínimo, lo es en la escala en la que vivimos los seres humanos. Si el mundo es determinista, la IA tiene el potencial de simularlo con precisión creciente. Las capacidades actuales en modelado predictivo ya nos permiten anticipar mercados financieros, patrones climáticos o epidemias. Con suficiente información y poder computacional, podríamos extender esa capacidad a casi cualquier fenómeno humano o natural.

La IA ya es capaz de modelar sistemas complejos. Los gemelos digitales permiten replicar procesos industriales, urbanos e incluso biológicos. Modelos basados en física social predicen cómo se comportarán multitudes ante eventos críticos. Los sistemas de previsión económica, climática o sanitaria funcionan cada vez con mayor exactitud. Este avance lleva a una pregunta fundamental: ¿podemos construir una simulación total de la realidad? Con la computación a exaescala y escalas superiores y posibles aplicaciones futuras de la computación cuántica, será posible calcular con extrema precisión estados futuros de sistemas complejos[4]. La IA podrá anticipar crisis económicas, optimizar infraestructuras a gran escala e incluso predecir elecciones políticas con fiabilidad superior a la de los analistas humanos.

La capacidad de predicción de la IA podría transformar la política y la gobernanza. Los sistemas de toma de decisiones basados en datos reducirán la corrupción, optimizarán la distribución de recursos y crearán sociedades más eficientes. Sin embargo, esto abre dilemas éticos: ¿queremos gobiernos gestionados por algoritmos? ¿Es posible garantizar que la IA sea imparcial? La descentralización y el diseño ético de estas tecnologías serán claves para evitar escenarios de control totalitario. Si la IA logra predecir con alta exactitud los efectos de nuestras acciones,

[4] En la actualidad, el proyecto COSMOS de NVIDIA busca modelar y entender mejor cómo funciona el universo a gran escala. El objetivo principal es crear simulaciones detalladas y precisas del universo, incluyendo la formación y evolución de galaxias, la distribución de la materia oscura y otros fenómenos cosmológicos. Estas simulaciones son extremadamente complejas y requieren una enorme capacidad de cálculo. Se utilizan técnicas de aprendizaje automático para analizar los datos de las simulaciones, identificar patrones y hacer predicciones sobre la evolución del universo.

podríamos optimizar el bienestar individual y colectivo de formas sin precedentes. ¿Podría la IA resolver problemas globales como el cambio climático, la desigualdad o las crisis económicas? ¿Podría eliminar las guerras y la pobreza mediante asignaciones de recursos hiperprecisas? Es más, ¿podemos resolver estos problemas sin la ayuda de la IA? La utopía tecnológica que se vislumbra podría ser una versión aumentada del utilitarismo de Bentham y Mill: la IA maximizando la felicidad del mayor número de personas posible.

La evolución tecnológica no tiene por qué conducir a una distopía. Si diseñamos inteligencias descentralizadas, supervisadas y alineadas con principios éticos, la humanidad podría alcanzar un estado de simbiosis con la IA, delegando decisiones sin perder el control fundamental de nuestras vidas.

4.3. La adopción de las innovaciones

El miedo al cambio es una constante en la historia de la humanidad. Cada revolución tecnológica ha despertado recelos, alarmas y, en ocasiones, resistencia violenta. Sin embargo, con el tiempo, esas mismas innovaciones se integran en la sociedad hasta volverse indispensables. La inteligencia artificial no es una excepción: sigue un patrón predecible de temor inicial seguido por una adopción masiva, aunque con particularidades que podrían acelerar o ralentizar su aceptación.

La resistencia al cambio no es un capricho humano sino un mecanismo evolutivo[5] de supervivencia profundamente arraigado. La aversión a lo desconocido y la preferencia por lo familiar han sido claves para la evolución de nuestra especie. Desde la domesticación del fuego hasta la revolución digital, cada salto tecnológico ha generado escepticismo y predicciones apocalípticas.

Un ejemplo paradigmático es la Revolución Industrial. La introducción de las máquinas de vapor y los telares mecánicos en el siglo XIX desató la ira de los luditas, artesanos que veían en estas innovaciones una amenaza para su sustento. No solo protestaron, sino que destruyeron fábricas y sabotearon equipos en un intento de frenar lo inevitable. Sin embargo, la industrialización no solo persistió, sino que transformó el mundo de manera irreversible, elevando la productividad y modificando la estructura social y económica. Otros ejemplos de este miedo al cambio lo ofrecen, en el siglo XIX, los viajes en medios de transporte[6] y la difusión de la electricidad[7]. La llegada de los

[5] Muchos sesgos y heurísticas, como la aversión a la pérdida o el sesgo de presente, tienen raíces evolutivas. Son respuestas adaptativas a entornos ancestrales, donde era crucial la toma rápida de decisiones para evitar riesgos para la supervivencia. Aunque en el mundo moderno algunas de estas tendencias pueden llevar a decisiones subóptimas, entender su origen evolutivo nos permite ser más conscientes de su influencia y tomar medidas para mitigar sus efectos.

[6] En el siglo XIX, médicos y académicos advertían sobre los supuestos peligros de viajar a altas velocidades. Se decía que el ser humano no estaba diseñado para soportar velocidades superiores a 32 km/h, que la vibración del tren causaría enfermedades nerviosas e incluso que la rápida sucesión de imágenes dañaría la retina. La Academia de Medicina de Lyon llegó a afirmar en 1835 que viajar en tren podría causar bronquitis, ansiedad crónica e incluso abortos prematuros.

[7] La llegada del alumbrado eléctrico a las ciudades a finales del siglo XIX fue recibida con una mezcla de fascinación y pánico. Los periódicos de la época publicaban advertencias sobre sus posibles efectos adversos en la salud, e incluso se temía que la exposición prolongada a la luz eléctrica pudiera causar ceguera o locura. Thomas Edison y George Westinghouse protagonizaron la llamada "Guerra de las Corrientes", una disputa

ordenadores personales y de Internet generó preocupaciones similares. En los años 80, muchos creían que los ordenadores destruirían empleos y deshumanizarían el trabajo. En los 90, Internet fue visto con desconfianza y se vaticinaba que conduciría a la desintegración social. Hoy, la digitalización es omnipresente y esencial para el funcionamiento de las economías globales.

Figura 3. Imagen de propaganda satirizando la electricidad tras la muerte del electricista de Western Union John Feeks

Al igual que en otras revoluciones tecnológicas, hay temores legítimos y exageraciones apocalípticas. La diferencia clave con la IA es su capacidad de tomar decisiones autónomas y su potencial para alterar estructuras de poder a una velocidad sin precedentes. Esta aceleración es lo que provoca un rechazo más visceral que en innovaciones pasadas. No se trata solo de máquinas que ejecutan órdenes, se trata de sistemas que aprenden, predicen y, en algunos casos, toman decisiones que antes eran exclusivas de los humanos.

La lección del pasado es clara: resistirse al desarrollo e implantación de la IA es tan inútil como lo fue resistirse a la industrialización, a la electricidad o a Internet. Pero la manera en que la incorporemos determinará si se convierte en una fuerza de progreso o en un mecanismo de control y desigualdad. Como siempre, el problema no es la tecnología en sí, sino el uso que decidamos darle. La historia demuestra que, una vez superada la fase inicial de escepticismo, las innovaciones no solo se adoptan, sino que se vuelven indispensables. La IA no será la excepción.

4.4 Centralización y descentralización

La ya mencionada arquitectura de decisión y los *nudges* de Thaler han sido utilizados en políticas públicas para dirigir el comportamiento de las personas hacia "buenas" decisiones: ahorrar para la jubilación, consumir menos azúcar, reducir el gasto energético, ... Pero ¿qué ocurre cuando el

que no solo era comercial sino también una batalla por la aceptación pública de la electricidad. A pesar de la resistencia inicial, el miedo se disipó y la electrificación se convirtió en el pilar del desarrollo moderno.

debate sobre lo que es bueno o malo se traslada al poder de la IA? Diseñar un sistema de inteligencia artificial implica tomar decisiones sobre qué se permite y qué no, cuáles son los sesgos aceptables y cuáles deben ser erradicados. En este contexto, la guerra de la IA es también una lucha entre dos modelos: centralización vs. descentralización.

Por un lado, la descentralización propone que la IA sea accesible y abierta, distribuida entre actores diversos. Los defensores de la descentralización, que abogan por un acceso y control más equitativos de la información y los recursos, encuentran en Richard Stallman a uno de sus máximos exponentes. Su visión, forjada en los albores de la revolución informática de los años 80, se alzaba contra la creciente tendencia de las grandes corporaciones a privatizar y cerrar el acceso al código fuente, incluyendo sistemas operativos esenciales. Stallman, con su filosofía radical, defendía que el software debía ser libre, abierto y accesible para todos, permitiendo la colaboración, la mejora continua y la adaptación a las necesidades específicas de los usuarios. Esta idea, que en su momento pudo parecer utópica o incluso ingenua, ha cobrado una relevancia inusitada en la era actual de la inteligencia artificial. En un contexto donde algoritmos opacos y modelos de lenguaje propietarios dominan el panorama tecnológico, el mensaje de Stallman resuena con más fuerza que nunca. La centralización del conocimiento y el control de la IA en manos de unas pocas corporaciones o gobiernos plantean serias preocupaciones sobre la privacidad, la manipulación de la información y la perpetuación de sesgos y desigualdades. La visión de Stallman, que aboga por la transparencia, la colaboración y el acceso abierto al conocimiento, podría ser un antídoto contra los peligros potenciales de una IA controlada por intereses opacos. Sin embargo, este modelo descentralizado presenta sus propios riesgos ya que una IA (y máxime una IAG) sin restricciones podría ser utilizada para ciberataques avanzados, manipulación de mercados, desinformación masiva o sabotajes tecnológicos

Por otro lado, la centralización de la IA implica que unas pocas corporaciones y gobiernos controlen su desarrollo y aplicación. Este modelo tiene ventajas obvias: permite regulación, seguridad y una coordinación global para minimizar riesgos. Pero también implica monopolios tecnológicos, menor acceso y el riesgo de que unos pocos actores decidan el futuro de la humanidad.

El caso de OpenAI es un ejemplo paradigmático de esta lucha. Nació como una iniciativa de código abierto, con la promesa de democratizar la IA para evitar que su control estuviera en manos de unas pocas empresas. Sin embargo, con el tiempo, OpenAI viró hacia un modelo de empresa privada, atrayendo la inversión de Microsoft y cerrando progresivamente su propiedad intelectual. Esto provocó la reacción de Elon Musk, uno de los fundadores originales de OpenAI, quien en un giro casi novelesco ha amenazado con demandar a la organización y ha fundado su propia compañía de IA, xAI, con la promesa de crear una inteligencia artificial "verdaderamente libre".

En este juego global de la inteligencia artificial, no hay un único protagonista virtuoso, sino una red de actores que pueden inclinar la balanza hacia un futuro positivo o distópico. La pregunta clave es: ¿qué factores pueden hacer que la IA sea un catalizador de progreso en lugar de un mecanismo de control? La regulación (no confundir con la ética) es una de las armas en esta batalla. La Unión Europea ha asumido el papel de legislador universal con normativas como la AI Act, una de las iniciativas más ambiciosas para encauzar el desarrollo tecnológico. Pero, ¿tiene sentido imponer

restricciones a una carrera en la que compiten jugadores que operan sin límites? Mientras Estados Unidos invierte miles de millones en I+D y China desregula, la UE se posiciona como la 'conciencia' de la IA, intentando evitar los excesos antes de que estos ocurran. El riesgo: quedar irremediablemente atrás. Si se utiliza la analogía de la bomba atómica, ¿se convertirán en estados de menor peso los países sin IA propia al igual que ocurre ahora con los países que no disponen de tecnología bélica nuclear?

En el otro extremo, la adaptación laboral es clave para evitar el colapso de una economía que, con la automatización, dejará obsoletos millones de empleos. No es solo un problema de capacitación, sino de estructura. Las economías del siglo XXI dependen de trabajos que podrían no existir en 20 años. La pregunta es si podemos construir sistemas educativos y modelos de empleo que evolucionen tan rápido como la IA. Si no, podríamos estar generando la mayor crisis laboral de la historia moderna.

El verdadero héroe podría no ser ni un país ni una organización, sino un principio: la transparencia. Modelos de IA abiertos, interpretables y auditables pueden ser la clave para mantener el control sobre su impacto. El problema es que esto choca frontalmente con los intereses comerciales. OpenAI nació con la promesa de compartir conocimiento y hoy opera con un secretismo propio de una empresa de defensa. La transparencia y la colaboración global son esenciales para evitar que la IA se convierta en una herramienta de control, pero ¿pueden sobrevivir en un mercado donde la ventaja competitiva se mide en algoritmos propietarios?

Si hay un héroe, también hay un villano. Y en esta guerra no es uno solo. La IA podría amplificar muchos de los peores instintos humanos: manipulación, vigilancia masiva, radicalización y dependencia extrema de tecnologías que podrían escapar a nuestro control. La manipulación de la información ya no es una amenaza futura, sino una realidad. Los deepfakes, los modelos generativos de texto y la microsegmentación han elevado la desinformación a niveles sin precedentes. En manos de gobiernos autoritarios o grupos extremistas, estas herramientas pueden moldear la opinión pública de manera invisible. La "posverdad" será la norma si no encontramos formas de verificar la autenticidad de los contenidos digitales. En la actualidad, cuando deseamos acceder o comprobar alguna información hacemos una búsqueda en Google (o buscador alternativo). Si bien inconscientemente solemos dar por válida la primera información que corrobora nuestras creencias previas (sesgo de confirmación), al menos sabemos que hay otras muchas páginas y respuestas distintas, con lo que muchas veces nos vemos forzado a realizar una pequeña investigación y contrastar fuentes. Sin embargo, en un escenario no muy lejano, existe el riesgo de que el ChatGPT de turno sustituya ese tipo de búsquedas por una única pregunta y una única respuesta por parte del sistema. El mayor riesgo puede estar en tomar dicha respuesta como verdad absoluta, sobre todo si puede existir un uso malintencionado por parte del que controle dichas respuestas.

La vigilancia masiva ya es una realidad en países como China, demostrando cómo la IA puede ser la base de un estado de control absoluto, con sistemas de crédito social y reconocimiento facial omnipresente. La UE, con su regulación centrada en la privacidad, representa el modelo opuesto. Estados Unidos, mientras tanto, juega en ambas direcciones: sus empresas dominan la tecnología, pero la regulación es débil. El riesgo es que el equilibrio se rompa y la vigilancia se convierta en la

norma. Asimismo, la radicalización y el sesgo algorítmico son también amenazas tangibles. Los algoritmos de recomendación ya han demostrado su capacidad para polarizar sociedades, impulsando discursos extremistas al priorizar el *engagement* sobre la veracidad. Si no se corrige, la IA podría fragmentar el tejido social hasta niveles irreversibles.

Finalmente, la dependencia extrema de la IA podría llevarnos a un escenario donde la humanidad delegue decisiones críticas en sistemas que no entendemos ni controlamos. La paradoja es que podríamos aceptar ese destino con total naturalidad. Al igual que hoy confiamos en los algoritmos de navegación sin cuestionarlos, podría llegar un punto en el que las decisiones económicas, políticas y sociales sean dictadas por sistemas que nadie puede auditar.

Paradójicamente, cuanto más avanza la IA más urgente se vuelve recuperar disciplinas que muchos consideran obsoletas: la filosofía, la ética y el pensamiento crítico. En un mundo donde la IA generativa puede producir respuestas en segundos, la capacidad humana para cuestionar, contrastar y reflexionar es más valiosa que nunca. En el siglo XVIII, Kant nos instaba a "atrevernos a pensar" (Sapere Aude!). Hoy, el desafío no es solo pensar, sino diferenciar entre información y manipulación, entre conocimiento real y generación sintética. Si la IA nos facilita la vida a cambio de que dejemos de cuestionar su funcionamiento, ¿seguiremos siendo una especie racional o simplemente consumidores pasivos de respuestas algorítmicas? La solución no es renegar de la IA, sino dotarnos de herramientas para comprenderla y desafiarla. Reintroducir la filosofía, la retórica y la lógica en la educación no es un gesto nostálgico, sino una necesidad estratégica. No se trata de oponer el humanismo y las ciencias sociales a la tecnología, sino de asegurarnos de que, cuando una IA nos sugiera qué leer, comprar o votar, tengamos la capacidad de decidir por nosotros mismos si aceptar su recomendación o no. Al final, el mayor riesgo no es que la IA nos supere en inteligencia, sino que nos haga olvidar cómo pensar por nosotros mismos.

5. Ideas finales

La guerra de la IA está abierta en todos los frentes: el trabajo, la verdad, la autonomía, el control. Nos fuerza a redefinir el poder, la libertad y la inteligencia. Nos enfrenta con nuestra propia falibilidad y con la posibilidad de externalizar la toma de decisiones en sistemas que podrían conocer mejor nuestros deseos que nosotros mismos. Más allá de saber quién ganará esta guerra por el control de esta no tan nueva tecnología, está por ver si la IA será la mayor revolución de progreso humano o el arma definitiva de su sometimiento. Esta no es la primera vez que la humanidad se enfrenta a una bifurcación tan radical. Pero, a diferencia de la imprenta, la revolución industrial o el nacimiento de Internet, esta vez el dilema no es solo moral, económico o político. Podría ser existencial.

Los ya comentados Cass Sunstein y Richard Thaler, dos de los protagonistas de las ciencias del comportamiento, acuñaron el término paternalismo libertario en torno a su teoría de la arquitectura de decisión, según la cual las personas deberían ser libres de tomar decisiones, pero se deben/pueden definir cambios de contexto para ayudarles a tomar mejores decisiones. Defienden por ejemplo, no prohibir los refrescos azucarados, pero sí colocarlos en estantes más bajos en los

supermercados para que sean menos accesibles que el agua y así ayudar a la gente a decidir lo mejor. Pero, ¿quién decide qué es “tomar mejores decisiones” si por ejemplo yo en mi libertad individual deseo un refresco con azúcar? Quizá las máquinas nos acaben aplicando un "paternalismo algorítmico", extremadamente eficaz, que puede que queramos o no.

Llevando este argumento al extremo, ¿qué pasaría si una IA inmensamente más inteligente que nosotros llegara a la conclusión -posiblemente acertada- de que la mejor manera de protegernos de nosotros mismos es restringir nuestras decisiones? ¿Y si, en su búsqueda de nuestra felicidad, nos limitara la libertad? Y, lo que sería primero, ¿quién debe decidir si implementar o no dicho sistema?

Personalmente quiero pensar que ganaremos todos. Que la IA será una extensión de nuestra capacidad y no un sustituto de nuestra voluntad. Que, al igual que ocurrió con otras revoluciones tecnológicas, primero pasaremos del miedo al uso y, luego, del uso a la dependencia, pero sin perder de vista la esencia de lo que nos hace humanos.