¿Singularidad? Limitaciones, capacidades y diferencias de la inteligencia artificial frente a la inteligencia humana

CLARIDADES
REVISTA DE FILOSOFÍA
ISSN 1889-6855 · ISSN-E 1989-3787

Singularity? Limitations, possibilities, and differences of artificial intelligence compared to human intelligence

Pablo Carrera

Universidad Isabel I (España)

Fecha de envío: 09/04/2024
Fecha de aceptación: 13/09/2024

DOI: 10.24310/crf.16.2.2024.19654

Resumen

En este artículo nos planteamos la cuestión de si realmente la IA ha alcanzado el nivel de la inteligencia humana, algunas de las razones que nos llevan a este estado de opinión, así como varias de las diferencias fundamentales entre la IA y la inteligencia humana. Realizamos un breve recorrido del desarrollo histórico de la IA, para después revisar las capacidades reales e importantes limitaciones de las técnicas de aprendizaje profundo en las que se basan los avances recientes

en IA. Abordamos particularmente el argumento de que las capacidades cognitivas complejas son indisociables de un cuerpo biológico en interacción con un mundo físico y sociocultural, frente a una IA basada en un axioma dualista y cognitivista que ha sido señalado como incompleto o parcial. Finalizamos considerando los riesgos reales de la IA en la actualidad, así como algunas especulaciones sobre su futuro desarrollo.

Palabras claves

Inteligencia artificial; cognitivismo; neurociencia; conciencia; cognición corporizada.

Abstract

In this paper, we reflect on whether AI has reached the level of human intelligence, some of the reasons that lead us to that state of opinion, as well as on some of the fundamental differences between AI and human intelligence. We briefly review the historical development of AI and delineate the real capacities and important limitations of deep learning techniques, underlying most of the recent advancements in AI. We particularly explain the argument of how complex cognitive skills are inseparable from a biological body interacting with a physical and socio-cultural world. This view contradicts the dualist and cognitivist axiom on which IA is fundamentally based, which has been criticized as incomplete or partial. We finalize considering some current real risks of AI and some speculations on its future development.

Keywords

Artificial intelligence; cognitivism; neuroscience; consciousness; embodied cognition.

I. Introducción

La inteligencia artificial (IA) es uno de los temas de moda en la actualidad científica y tecnológica; ocupa portadas, se fundan empresas y *start-ups* relacionadas con ella, y todo discurso que pretenda actualizado hace referencia a ella (Leaver & Sdrarov, 2023). A la vez, la IA es un negocio boyante: desde el 2019 al 2021 la inversión privada en IA se ha duplicado a nivel mundial (de 60 a más 120 millones de dólares), y tanto las patentes como las publicaciones científicas relacionadas con la IA han proliferado en los últimos años (Giattino et al., 2023).

El crecimiento del negocio de la IA se debe al desarrollo de mejores modelos basados en el aprendizaje profundo, especialmente de modelos de IA generativos y del procesamiento del lenguaje masivos como Chat-GPT y similares (Blank, 2023). La (aparente) preocupación por el rápido desarrollo de la IA ha sido tal, que muchas de las grandes figuras del campo —incluidos los dirigentes de las principales empresas desarrolladoras— firmaron en 2023 una carta abierta en la que pedían una pausa en el desarrollo de modelos de IA más potentes que Chat GPT-4, con el fin de evitar posibles consecuencias catastróficas para la humanidad (Future of Life Institute, 2023; Leaver & Srdarov, 2023).

No es sorprendente, por tanto, que desde otros campos se vea esta situación con cierta preocupación, y que, desde la filosofía o la psicología nos planteemos preguntas como las que guían este monográfico: ¿Qué nos diferencia como humanos de la IA? O, ¿hay una diferencia cualitativa que impida que la labor humana sea suplantada por la labor no humana? Pero, cabría preguntarse: ¿realmente estamos en un punto en el que la IA haya alcanzado o superado la inteligencia humana? En este artículo intentamos argumentar por qué estamos todavía lejos de llegar a ese punto, así como algunas de las razones que nos llevan a pensar que sí podríamos estarlo.

II. Algunos apuntes sobre el desarrollo histórico de la IA: mitos, hipérboles e inviernos

La visión de la IA en la sociedad invariablemente se ha visto mezclada con elementos del cine, la literatura y la ciencia ficción, dando lugar a lo que se ha llamado el mito de la IA (Natale & Ballatore, 2020; Pilling & Coulton, 2019). De hecho, antes de los albores de la IA ya había películas como *Metrópolis* de Fritz Lang (1927), en la que aparecen robots antropomórficos inteligentes. En 1945 un ingeniero estadounidense llamado Vannevar Bush publicó un ensayo en el que llamaba a usar la tecnología para el manejo de grandes cantidades de información, como un complemento aumentativo de la memoria humana (Bush, 1945). Creó una maquina llamada Memex que permitía etiquetar información con códigos para recuperarla por asociación, lo que supone un antecedente del hipervínculo. La revista *Life* publicó ese ensayo con el subtítulo de «las máquinas empezarán a pensar», denotando la hipérbole y antropomorfización que ha caracterizado el tratamiento de los medios a los avances en computación e IA (Bush, 1945).

En 1950, Alan Turing publicó el influyente artículo «*Computing machinery and intelligence*», en el que se trataban ideas seminales como el procesamiento del lenguaje natural o el aprendizaje automático, así como la información como un elemento fundamental de la realidad. En este trabajo también se describió el famoso —y discutido por su excesivo funcionalismo— test de Turing, en el que se propone que se evalúe la inteligencia de una máquina por su habilidad de producir una conversación indistinguible de la de un humano (Turing, 1950). Empezó también a ganar fuerza en esta época, junto con el inicio de las teorías cognitivistas en psicología, la manida metáfora de la mente humana como una computadora, que discutiremos en más detalle en una sección posterior.

Es en 1956 cuando se acuñó el término inteligencia artificial, concretamente por John McCarthy y otros pioneros como Marving Minsky, en la conocida como conferencia de Dartmouth. Esta conferencia tenía como objetivo «encontrar la manera de que las máquinas usen el lenguaje humano, formen conceptos abstractos, solucionen problemas hasta ahora reservados a los humanos y se mejoren a sí mismas» (McCarthy et al., 2006: 12). Partiendo de la equiparación de la mente como computadora y de una comprensión exclusivamente lógica y racional de la inteligencia, estos autores pioneros de la IA proclamaron que «todo aspecto del aprendizaje o cualquier otro elemento de la inteligencia puede ser descrito de forma tan precisa que se puede hacer que una computadora lo simule» (McCarthy et al., 1006: 12).

Durante estos años y hasta finales de los años 60 se dio la época dorada de la IA: los medios y el público quedan fascinados por esta tecnología, sobre la que creció un desaforado optimismo y se volcaron una expectativas poco realistas. Este optimismo fue alimentado tanto por muchos de los expertos en IA como por el sensacionalismo de los medios de comunicación (Natale & Ballatore, 2020; Jiang et al., 2022). En EEUU, por ejemplo, hubo bastante interés en desarrollar la traducción automática del ruso al inglés, como potencial arma a utilizar en la Guerra Fría. IBM desarrolló una computadora, IBM 701, que tradujo exitosamente 60 frases del ruso al inglés en un evento público en 1954, lo que fue suficiente para que IBM sacara una nota de prensa diciendo que la IBM 701 era un «versátil cerebro electrónico» (IBM, 1954).

Los innegables avances en computación se confundían con los bastante discretos avances en IA, y servían para alimentar las optimistas predicciones de los propios expertos. En 1965, Herbert Simon en su libro «*The shape of automation for men and management*» predijo que «las máquinas serán capaces en 20 años de hacer cualquier trabajo que pueda hacer un humano» (Simon, 1956, citado en Mitchell, 2024: 1). Marvin Minsky del MIT, una de las principales figuras en el desarrollo temprano de la IA, afirmaba en la revista *Life* en 1970 que «de aquí a tres años tendremos una máquina con la inteligencia general de un ser humano promedio. Me refiero a una máquina que será capaz de leer a Shakespeare, darle cera a un auto, decir una broma o pelearse» (Minsky, 1970, citado en Mitchell, 2024: 1). No solo estas predicciones no se cumplieron, si no que el avance desde esa época en alcanzar una IA comparable a la inteligencia humana ha sido muy limitado.

Las promesas no cumplidas y los escasos avances en IA frente a las expectativas creadas llevaron a finales de los sesenta y principios de los setenta a lo que se conoce como el primer invierno de la IA. La financiación a proyectos relacionados con la IA se paralizó y hubo una general desilusión con las posibilidades reales de esta tecnología (Jiang et al., 2022; Mitchell, 2021). En 1973 se emitió el informe Lighthill en Reino Unido, encargado por el gobierno británico, en el que se afirmaba que «en ningún área del campo [de la IA] los descubrimientos han producido el gran impacto que fue prometido» (Lighthill, 1973: 8). Esto provocó la paralización total de la financiación de proyectos relacionados con la IA en este país.

En EEUU, el invierno de la IA había comenzado antes, causado por los fracasos en el programa de traducción automática. En 1966 un informe de un comité de expertos en el procesamiento automático del lenguaje (ALPAC) al gobierno estadounidense refutaba que fuera posible la traducción automática en el corto o medio plazo. Como consecuencia, se suspendió la financiación en ese campo, y por tanto, en IA en general. En la cultura popular, sin embargo, el entusiasmo por la IA no decayó, con el lanzamiento de películas de ciencia ficción con una IA inquietante como *2001, una odisea en el espacio* —en la que Marvin Minsky sirvió como consultor— o simpática y antropomorfa, como en *La Guerra de las Galaxias*.

A principios de los años 80 resurgió el interés por la IA, esta vez aplicada principalmente al ámbito comercial e industrial. Se desarrollan en esta época los llamados sistemas de experto, basados en reglas de

tipo «si X–entonces Y» (lo que se conoce como un sistema *top-down* o arriba-abajo; Jiang et al., 2022). También se renovó el interés en una IA con objetivos más ambiciosos por parte de algunos estados: el Gobierno de Japón lanza en 1981 el proyecto *Fifth Generation computer*, con los objetivos de desarrollar ordenadores que pudieran tener conversaciones, interpretar imágenes o razonar como un ser humano. Estados Unidos también renueva la financiación en IA general con el programa *Strategic Computing Initiative*.

De nuevo, creció la inversión en IA, y de nuevo este crecimiento fue acompañado de un gran entusiasmo sobre las posibilidades de la IA basado en expectativas poco realistas y en una cobertura mediática sensacionalista. Sin embargo, a finales de los 80 colapsó el mercado de *hardware* especializado en procesar el lenguaje de programación LISP (el utilizado predominantemente para programar IA en EEUU en la época), que había florecido por la adopción de sistemas de IA experto por parte de empresas. El proyecto *Fifth Generation computer* acabó en 1993 sin alcanzar casi ninguno de sus objetivos iniciales (Mitchell, 2021).

Los sistemas de experto podían enfrentarse con éxito a tareas muy específicas e incorporaron avances como las redes neuronales. Pero las empresas abandonaron el uso de esta tecnología paulatinamente a principios de los 90 por sus limitaciones: dificultades para computar situaciones demasiado complejas, producción de errores crasos si se alimentaban de datos inusuales, dificultades para generalizar o enfrentarse a nuevas situaciones, o ser excesivamente cara de mantener y actualizar, entre otros problemas (Mitchell, 2021; Jiang et al., 2022).

El campo de la IA quedó desacreditado de nuevo, dando lugar al segundo invierno de la IA. Durante la década de los noventa y la primera década del siglo XXI se siguió avanzando en campos relacionados como el *machine learning* o aprendizaje automático, si bien los expertos en esta materia diferenciaban su disciplina de la entonces desacreditada IA. Algunos hitos importantes posteriores, sobre todo en cuanto la percepción social de la IA, son la primera vez que una IA (Deepblue) ganó en el ajedrez a un campeón mundial, Kasparov, en 1997. Más adelante, el programa AlphaGo, de Google Deepmind, pudo ganar a jugadores profesionales de este complejo juego asiático, incluso al aprender desde cero mediante aprendizaje por reforzamiento (Silver et al., 2018).

Al mito de la IA se han añadido elementos como la popularidad del libro publicado en 2005 *La singularidad está cerca*, de Ray Kurzweil. La singularidad es la idea de que es inevitable un crecimiento exponencial de las capacidades de la IA hasta un punto en el que aparezca una IA superinteligente, capaz de crear IA generales más inteligentes que los humanos. Según esta visión, esto puede tener consecuencias imprevisibles, incluyendo visiones más o menos apocalípticas que están más cerca de la ciencia ficción que de la ciencia. Aunque el libro ha sido criticado por su poca o nula base científica y por su naturaleza especulativa (por ejemplo, Modis, 2006), introdujo en el imaginario popular la idea de que el aumento en las capacidades de la IA se puede dar de forma exponencial y potencialmente catastrófica (recordemos la ya mencionada carta abierta; Future of Life Institute, 2023).

Virtualmente todos los avances de los últimos años en IA han sido posibles gracias al uso de técnicas de aprendizaje profundo o *deep learning*, de la mano del incremento en la capacidad de computación y la disponibilidad de cantidades masivas de datos para entrenar a los algoritmos (Jiang et al., 2022; López de Mantaras, 2020; Mitchell, 2021). De nuevo, y a pesar de los antecedentes, encontramos un optimismo desaforado en las posibilidades de desarrollo de la IA, unido a una burbuja de inversión y al consabido sensacionalismo mediático.

Las expectativas poco realistas y el optimismo sobre las posibilidades de desarrollo de la IA son, una vez más, alimentadas por los propios ejecutivos de las empresas líderes en el campo: Sam Altman, CEO de OpenAI, predijo en su blog que la IA «hará casi de todo, incluyendo nuevos descubrimientos científicos que expandirán nuestra concepción sobre todo» (Altman, 2021); Shane Legg, cofundadora de Deepmind de Google, predijo que la IA sobrepasará a la inteligencia de nivel humano a mitad de la década del 2020 (Despres, 2008); y Mark Zuckerberg, CEO de Meta, declaró en 2015 que el objetivo de Facebook para los siguientes cinco o diez años era sobrepasar al ser humano en todos los sentidos básicos: visión, lenguaje, oído o cognición general (McCraken, 2015).

Pero ¿cuánto hay de verdad en estas predicciones? ¿Estamos, por fin, ante el inicio de una IA general que llegue al nivel de la inteligencia humana, o incluso la sobrepase? ¿O se trata, una vez más, del sensacionalismo y el mito tecnológico que parecen indefectiblemente unidos a los avances en IA? Para

intentar arrojar algo de luz sobre estas cuestiones parece necesario abordar, aunque sea superficialmente, el funcionamiento de las técnicas de la IA en la actualidad, así como la diferencia entre la IA estrecha y la IA general.

III. ¿Estamos cerca de que la ia supere a la inteligencia humana? Aclaraciones sobre la IA general y específica y el aprendizaje profundo

¿A qué nos referimos con inteligencia? Definir qué es la inteligencia es una tarea compleja y que podría llevar a un extenso tratado. Para los objetivos de este trabajo definimos la inteligencia como la habilidad para aprender de la experiencia y realizar tareas complejas de manera flexible y adaptativa en diferentes entornos, a lo que podríamos añadir la capacidad de razonamiento abstracto y de comprensión de conceptos complejos (Gottfredson, 1997).

Esta definición de inteligencia es lo que se conoce como inteligencia general. Tal y como se desprende de la propia definición de IA y de las declaraciones de los líderes pasados y actuales del campo, lo que el programa de investigación en IA ha aspirado desde su inicio es a llegar a una inteligencia equiparable a la humana, denominada IA general (Fjelland, 2020; Korteling et al., 2021; López de Mantaras, 2017). Es importante diferenciar la IA general de lo que se ha llamado IA estrecha o específica (*narrow artificial intelligence*). La IA estrecha, también llamada IA débil, pueden desempeñarse bien o incluso a niveles superhumanos en un rango limitado de tareas predefinidas y estructuradas, pero son totalmente incapaces fuera de esas tareas predefinidas.

Todos los avances en IA en la actualidad, incluyendo los programas de IA generativa (Chat-GPT, DALL-E, etc.), los famosos programas DeepBlue (especializado en ajedrez) o AlphaGo (especializado en el juego asiático Go), o los programas de reconocimiento de imagen, son programas de IA específica. Incluso los vehículos autónomos son una combinación de diferentes IA específicas, especializadas en percibir el entorno, integrar la información en tiempo real, diseñar y modificar una ruta, tomar decisiones, etc. (Jiang et al., 2022). Las IA específicas son apropiadas para tareas muy bien predefinidas, con reglas explícitas y estructuradas, susceptibles de ser objeto de cómputo. En tareas de este tipo, en realidad, programas de IA superan a la inteligencia humana —en ciertos aspectos, y con limitaciones para generalizar—, ya que la inteligencia humana tiene unas limitaciones

en capacidad de procesamiento de datos y memoria que la actual capacidad de computación de una computadora supera (Korteling et al., 2021). Sin embargo, la IA específica no es adecuada para entornos o tareas con poca estructura, sin reglas consistentes y en las que suceden a menudo eventos imprevistos, raros o poco comunes, ya que la IA específica no tiene capacidad de razonar de manera general para adaptarse a diferentes situaciones. Por ejemplo, si bien el programa de IA DeepBlue fue capaz de derrotar al maestro del ajedrez Kasparov, era totalmente incapaz de hacer cualquier otra tarea que no sea jugar al ajedrez (de Saint Laurent, 2018).

Y ¿cómo funcionan las IA específicas actuales? Aunque hay diferentes tipos, como los modelos de experto mencionados anteriormente, en la actualidad los avances en IA se basan, casi en su totalidad, en el aprendizaje profundo (*deep learning*), un tipo de aprendizaje automático o supervisado (*machine learning*), a menudo implementado mediante lo que se ha llamado redes neuronales artificiales (Mitchell, 2021; de Saint Laurent, 2018; López de Mantaras, 2020). Las técnicas de aprendizaje automático y formas básicas de redes neuronales existen desde hace décadas, pero es aproximadamente desde la década del 2010 que el desarrollo de IA específicas mediante el uso de redes neuronales profundas se acelera, principalmente gracias al aumento en la capacidad de computación y a la disponibilidad de cantidades masivas de datos de entrenamiento (Jiang et al., 2022; Mitchell et al., 2021; de Saint Laurent, 2018).

De una forma simplificada y breve, el aprendizaje automático se refiere a procedimientos estadísticos para, dada una información determinada (variables independientes), predecir un resultado, normalmente la clasificación en una determinada categoría (variable dependiente). Se requiere, por tanto, de bases de datos de entrenamiento, en las que el resultado o variable dependiente es conocida y está etiquetado. Comparando los valores predichos con los valores observados como medida de acierto, el algoritmo se aplica de forma iterativa y se va perfeccionando para minimizar el error en su capacidad de predicción. Una vez entrenado, el algoritmo puede ser usado para predecir el resultado o la categoría deseada, y cuanta mayor cantidad de datos reales haya usado como entrenamiento, más generalizable o eficaz podrá ser (de Saint Laurent, 2018). Por ejemplo, para detectar qué correos electrónicos son spam, se puede entrenar un algoritmo con bases de datos en los que haya correos categorizados como

spam. Usando diferentes parámetros como datos de entrada (por ejemplo, inclusión de palabras como «oferta» o presencia de faltas de ortografía), el algoritmo perfeccionará su capacidad de predecir los correos que son spam en base a estas variables, y se podrá usar con esta función más allá de la base de datos de entrenamiento.

El aprendizaje profundo mediante el uso de redes neuronales artificiales básicamente complejiza este modelo básico utilizando diferentes capas combinadas, cada una capaz de procesar información a partir de unos datos de entrada y de dar una salida (con una mayor complejidad), que es procesada por la siguiente capa hasta llegar a un resultado final o datos de salida, normalmente la probabilidad de una clasificación. Esto hace que las redes neuronales artificiales sean eficaces en detectar patrones complejos en los datos, haciendo que hayan tenido relativo éxito en campos como el reconocimiento visual o el procesamiento del lenguaje natural. En la práctica, se trata de la aplicación sucesiva de modelos estadísticos relativamente sencillos (regresión lineal, logística o estadística Bayesiana), combinados de manera compleja, asignando pesos (algo parecido a los coeficientes de regresión) según la capacidad de cada característica procesada de los datos de entrada para predecir la clasificación correcta, lo que implica una inmensa cantidad de computación (López de Mantaras, 2020).

Además, los programadores deben matizar e ir ajustando diferentes hiperparámetros para obtener un modelo lo más eficaz posible, por lo que, en realidad, el aprendizaje más que automático es supervisado. Se denominan redes neuronales porque las neuronas inspiraron su arquitectura, con una entrada y una salida que está conectada a otra neurona, transmisión de datos de forma jerárquica y procesamiento en paralelo (en vez de un procesamiento central, como en el caso de las CPU), pero en realidad tienen más que ver con la estadística y la computación que con la neurociencia (de Saint Laurent, 2018). Para una explicación más detallada, pero adecuada para un público general, sobre los distintos tipos de aprendizaje profundo detrás de los avances en IA actuales, así como de sus limitaciones, ver el excelente resumen de López de Mantaras (2020), fundador y exdirector del Instituto de Investigación de Inteligencia Artificial del CSIC.

En definitiva, las actuales técnicas de aprendizaje profundo lo que básicamente hacen es, dada unas entradas de datos, extrapolar un resultado o salida estadísticamente probable basado en el análisis de patrones y

asociaciones estadísticas en bases de datos de entrenamiento (cuanto más numerosos, mejor). Este resultado o salida puede ser también lenguaje que parezca humano, como es el caso de los modelos de lenguaje masivos como Chat-GPT (Chomsky, Roberts & Watumull, 2023). En otros casos, como en el aprendizaje profundo por reforzamiento (el utilizado para los algoritmos de juegos de mesa como AlphaGo), la lógica es algo distinta ya que se trata de aprender a maximizar las acciones que sean reforzadas con el resultado deseado (López de Mantaras, 2020).

Hay bastantes limitaciones implícitas en el funcionamiento del aprendizaje profundo, como que —en general— no se sabe del todo la «ruta» o aprendizaje que ha tomado el algoritmo para llegar a la respuesta correcta. De hecho, los cálculos que hace el algoritmo para aumentar la corrección de sus predicciones son, en general, poco transparentes, dándose un efecto «caja negra» por el que a menudo no se sabe muy bien qué información se está teniendo en cuenta y si están aprendiendo lo que el programador intentaba. Esto es peligroso, ya que estos sistemas son proclives a los «atajos»: reglas de decisión que funcionan bien en determinados escenarios, aunque tengan poco que ver con el aprendizaje que se quería conseguir, por lo que generalizan mal a situaciones reales y dan lugar a errores (Geirhos et al., 2023).

Un ejemplo clásico (intencional) es el de un algoritmo que se quiso entrenar para distinguir correctamente entre perros de raza husky siberiano y lobos, una clasificación complicada de realizar. Sorprendentemente, el algoritmo consiguió una buena tasa de éxito en la clasificación, pero se debía a que, en las imágenes de entrenamiento, los lobos aparecían con un fondo de nieve; el algoritmo usaba la presencia de nieve o no para clasificar la imagen como lobo o como husky. Evidentemente, esto significa que arroja resultados correctos por razones equivocadas, y solo en esa base de datos, ya que ante fotos de lobos que no aparezcan con nieve, el algoritmo fallará y no reconocerá que es un lobo (Ribero, Singh & Guestrin, 2016).

Si bien las capacidades de cálculo y de clasificación del aprendizaje profundo son impresionantes, resulta cuanto menos extravagante pretender que estas técnicas de IA están a punto de dar lugar a algo remotamente parecido a la conciencia humana, la intencionalidad, o la inteligencia compleja y general del ser humano. El lector quizás se pregunta que, si todos estos avances se han dado en IA específica o estrecha, ¿qué avances

ha habido respecto a alcanzar una IA general, comparable a la humana? La conclusión del informe del Consejo Nacional de Ciencia y Tecnología del Gobierno de Estados Unidos, de 2016, nos parece aun plenamente vigente en este sentido:

> Un amplio abismo parece separar la IA estrecha actual del desafío mucho más difícil de la IA general. Los intentos de alcanzar la IA general mediante la expansión de soluciones de IA estrecha han logrado pocos avances durante muchas décadas de investigación. El consenso actual de la comunidad de expertos del sector privado, con el que coincide el Comité de Tecnología del NSTC, es que la IA general no se logrará hasta dentro de al menos décadas (National Science and Technology Council, 2016: 7).

Con el añadido de que, en lugar de décadas, se puede tardar siglos en alcanzar una IA general. Algunos autores van más allá y defienden que este objetivo es imposible de alcanzar, debido a diferencias fundamentales entre la inteligencia humana y la IA. En la siguiente sección ahondamos en algunos de estos argumentos.

IV. El error de Descartes y el acierto de Dreyfus: Por qué la IA es cualitativamente diferente de la inteligencia humana

Lo primero que nos gustaría apuntar, particularmente siendo esta una revista de filosofía, es que no pretendemos hacer una crítica o análisis del corpus filosófico de Descartes, objetivo que sobrepasaría con creces nuestra capacidad. Con el error de Descartes nos referimos simplemente a la crítica de la tradición dualista que propone que la mente y el cuerpo son entidades separadas (*res cogita* y *res extensa*), es decir, que para entender las operaciones mentales y el pensamiento podemos prescindir del cuerpo. Un planteamiento que, de una manera u otra, ha sido axiomático en el desarrollo de la IA (Brödner, 2019; Gill, 2019; Mitchell, 2021).

Respecto a las diferencias entre la IA actual y la inteligencia humana, nos limitaremos a revisar algunas de las cuestiones más relevantes. En primer lugar, trataremos algunas de las diferencias en la manera en que los humanos y la IA perciben, razonan, aprenden o usan el lenguaje, y, en segundo lugar, nos centraremos en el argumento del cuerpo como base indisociable de la inteligencia general del ser humano.

IV. 1. Funcionalismo y las limitaciones del uso del lenguaje y el aprendizaje en la IA

Es importante diferenciar lo que entendemos cómo inteligencia de lo que es una simulación de inteligencia. Las propuestas clásicas de la IA proponen que esta será capaz de ser inteligente, o incluso de dar lugar a una conciencia, ya que desde esta perspectiva se entiende que es el mero procesamiento y computación de símbolos en base a reglas lo que da lugar a la inteligencia, como proponía la Hipótesis de Símbolos Físicos de Newell y Simon (1976). En la práctica, en el campo de la IA se ha adoptado una perspectiva funcionalista de la inteligencia: si una computadora es capaz de resolver un problema que requiere una determinada capacidad cognitiva en humanos, se considera que la IA muestra o tiene esa capacidad (Searle, 1980). Recordemos que el test de Turing proponía que una computadora se puede considerar inteligente si, en una conversación, responde de una forma que sea indistinguible a la de un humano.

Esta perspectiva funcionalista de la inteligencia fue criticada por John Searle con una conocida metáfora, la de la «habitación china», que resumió de la siguiente manera (Searle, citado en López de Mantaras, 2020: 58):

> Supongamos que un angloparlante que no tiene ni idea de chino se encierra en una habitación en la que dispone de un conjunto muy completo de reglas, escritas en inglés, sobre cómo manipular caracteres chinos y cómo generar otros a partir de tales manipulaciones. A continuación, desde el exterior se le proporcionan una serie de caracteres en ese idioma y él, aplicando las reglas mencionadas, procede a transformarlos en otros caracteres chinos que devuelve al exterior, de manera que estos resulten ser respuestas a los caracteres de entrada indistinguibles de las que daría alguien que habla chino con fluidez.

Aunque el *outpout* o la salida pueda ser (más o menos) indistinguible de la respuesta de alguien que entendiera bien el chino, nadie podría decir que ese hipotético sujeto entiende el idioma chino solo porque pueda manipular símbolos sintácticamente siguiendo instrucciones precisas. Es decir, no hay ninguna comprensión semántica (del significado) de esos símbolos. De igual manera, los algoritmos que procesan el lenguaje humano, desde Chat-GPT a traductores automáticos, no comprenden, en ninguno de los sentidos que le podamos dar a ese verbo, el texto que reciben. Tan solo detectan patrones y devuelven los patrones más probables, en base a miles o millones de asociaciones estadísticas en textos analizados, analogía que se puede aplicar

también a otros ámbitos de la IA como el reconocimiento de imágenes (Chomsky, Roberts & Watumull, 2023; Leaver & Srdavov, 2023).

Noam Chomsky y otros expertos en lingüística e IA estadounidenses escribieron una recomendable pieza en The New York Times, llamada «La falsa promesa de Chat-GPT», donde argumentan algunas de las principales diferencias entre el aprendizaje y uso del lenguaje por modelos de IA y humanos. El procesamiento de lenguaje en la IA se basa en el análisis sobre que palabras suelen ir asociadas a otras en grandes cantidades de texto (convertidas en números y vectores y computadas), por lo que no puede diferenciar lo posible de lo imposible, ni detectar relaciones de causa y efecto, solo asociaciones (Chomsky, Roberts & Watumull, 2023). Como señalan estos autores, esta forma de procesar y producir lenguaje es fundamentalmente diferente de como la procesa y produce un ser humano. Nuestra especie aprende una compleja gramática de forma implícita y natural a partir de una exposición mínima al lenguaje, pero apoyada en una interacción continuada en un entorno social con intencionalidad compartida y en un soporte biológico con millones de años de evolución.

Otro aspecto en el que se pone crudamente de manifiesto las limitaciones de la IA y sus diferencias respecto a la inteligencia humana es en la transferencia del aprendizaje. En el sentido que solemos entenderlo, tanto a nivel coloquial como científico, aprender algo implica que se pueda transferir o generalizar ese aprendizaje a nuevos contextos (Raviv et al., 2022). Sin embargo, esto es notablemente difícil para los algoritmos de aprendizaje profundo. De hecho, hay todo un subcampo del aprendizaje automático, denominado aprendizaje por transferencia, dedicado al problema de cómo conseguir que los algoritmos puedan transferir con éxito sus aprendizajes a nuevas situaciones (Mitchell, 2021; Ranaweera & Mahmoud, 2021).

Por ejemplo, el algoritmo AlphaZero, una versión extendida de AlphaGo, aprendió a jugar al ajedrez y al shogi (un juego de mesa japonés parecido al ajedrez) desde cero, además de al Go, mediante el aprendizaje por reforzamiento (Silver et al., 2018). Pero el algoritmo requería una red neuronal separada para cada juego, y tuvo que aprender desde cero cada uno. Es decir, lo aprendido sobre cómo ganar en un juego no era transferido en absoluto a su aprendizaje sobre cómo ganar en otro (López de Mantaras, 2020). Desde una visión funcionalista, podríamos

pensar ¿Qué importa, si el algoritmo ha sido capaz de alcanzar niveles superhumanos y vencer a jugadores humanos expertos en esas tareas?

En una situación de juego de mesa, en la que todas las reglas son conocidas, la tarea a realizar está muy bien definida y todos los posibles estados futuros son modelables y computerizables, los algoritmos de aprendizaje profundo pueden mostrar el éxito de AlphaZero, aun cuando no transfieran su aprendizaje de una situación a otra. Pero el mundo real es distinto: no se pueden computar todos los estados posibles ni futuros movimientos, se dan multitud de eventos inesperados, etc. (López de Mantaras, 2020). Sin una inteligencia flexible, capaz de aplicar lo aprendido a situaciones diferentes por medio de analogías, cualquier acción se vuelve prácticamente imposible en el mundo real. La falta de capacidad de generalización de la IA es un problema serio para su desarrollo y para alcanzar una hipotética IA general, que por ahora no tiene solución.

En realidad, muchos de las limitaciones de la IA se deben a que no tiene sentido común, es decir, no tiene el vasto conocimiento, en gran parte tácito, adquirido a lo largo de una vida en interacción con un mundo físico y social-cultural (Davis & Marcus, 2015). Con esto nos referimos al conocimiento básico sobre el mundo real: sus reglas físicas, lo que es posible y no, la causalidad, o cómo se comportan los objetos inanimados y los seres vivos, entre otros muchos aspectos. A lo que se une la capacidad de usar ese conocimiento para reconocer y hacer predicciones sobre las situaciones que nos encontramos, abstraer conceptos generales a partir de particulares, o hacer analogías considerando la experiencia previa. Capacidades que un bebé de un año va aprendiendo de manera natural, mediante la interacción repetida con el mundo, pero que hoy por hoy nadie sabe cómo incorporar a las técnicas de IA (Mitchell, 2021; Davis & Marcus, 2015). Lo nos lleva al siguiente punto: para adquirir sentido común es esencial estar en el mundo, y para estar en el mundo hace falta un cuerpo. Un cuerpo que, además, es tan indisociable de las capacidades cognitivas complejas que se han tratado de emular con la IA como una flor del tronco y las raíces de un árbol.

IV. ii. *El cuerpo como base de la conciencia y de las capacidades cognitivas complejas*

Los argumentos que critican el programa de investigación de la IA desde la posición de que el cuerpo es indisociable de las capacidades cognitivas complejas surgieron desde casi el principio de la IA. Entre estas voces críticas, sobresale el filósofo americano Hubert Dreyfus; ya en la época dorada de la IA, a mitad de los años 60, fue una de las voces escépticas que señalaron las limitaciones y dificultades para realizar una IA general frente al entusiasmo y optimismo generalizado.

En su trabajo *Alquimia e inteligencia artificial*, de 1965, y su ampliación en el libro *Lo que no pueden hacer las computadoras*, de 1972, avanzó uno de sus principales argumentos, que atañe a un axioma fundacional en el campo de la IA: según Dreyfus, el comportamiento inteligente humano no se puede reducir al pensamiento explícito y formal, susceptible de ser convertido en operaciones discretas lógico-matemáticas (y, por tanto, de ser llevadas a cabo por mera computación; Dreyfus, 1973). Revisando las reflexiones de Dreyfus de la época, sus predicciones han envejecido notablemente mejor que las de los exultantes pioneros de la IA como Minsky, que en general recibieron con desprecio las críticas del filósofo.

En desarrollos posteriores de sus críticas al programa de investigación de la IA, Dreyfus expandió sus argumentos para señalar que la inteligencia humana es indivisible de un cuerpo que está en el mundo, autodeterminado, con un sustrato metabólico y biológico y que responde al mundo de forma intuitiva, en vez de siguiendo reglas explícitas mediadas por una representación mental (Brödner, 2019; Gill, 2019). Aunque esta crítica atañe más bien a la IA simbólica del inicio del campo de la IA que a la IA conexionista basada en el aprendizaje profundo actual, la IA conexionista tampoco requiere de un cuerpo que interactúe con el mundo real.

Dreyfus y otros críticos atribuyen este error de base a una tradición de largo arraigo en el pensamiento occidental: el dualismo mente-cuerpo, la concepción de que hay una mente racional o inteligencia independiente del cuerpo y de elementos como los sentimientos. Si el papel del cuerpo se reduce a ser una fuente de *inputs* (percepción) y de *outputs* (conducta), y todo lo que llamamos conocimiento o inteligencia se traslada a reglas explícitas y conceptos o representaciones (lo que se ha denominado conocimiento proposicional), nada impediría que creáramos un modelo

computacional o algorítmico de dichas reglas y representaciones, ya sea mediante programación lógica o mediante aprendizaje profundo. Es decir, exactamente la visión de la IA que se proponía en el documento fundacional del campo de la IA (ver la sección sobre el desarrollo histórico de la IA), compartida o extendida en la Hipótesis de Símbolos Físicos de Newell y Simon (1976).

La concepción dualista se vio reflejada también en el cognitivismo imperante en la época dorada de la IA, con el auge de las metáforas de la mente como una computadora dedicada al procesamiento de información, en el que la mente sería el *software* y el cerebro el *hardware* (Brödner, 2019). Esta concepción dualista —parcial e incompleta—, sigue siendo relevante hoy en día, a juzgar por las absurdas ideas (llamadas platónico-cartesianas por algunos autores) de que podremos «subir nuestra mente» a una nube digital en apenas unas décadas (Gill, 2019; Mitchell, 2021; Woolaston, 2013).

Frente a esta concepción dualista, la perspectiva que aporta Dreyfus y otros autores es que la inteligencia se da desde y a partir de un cuerpo —en inglés se usa el término *embodied*— y de la experiencia inmediata y consciente en el mundo (Dreyfus, 1973; Brödner, 2019). El comportamiento inteligente humano responde en el mundo, en gran parte, de forma intuitiva, en base a un «saber cómo» tácito, en vez de a unas reglas explícitas (el más elaborado «saber qué»; Brödner, 2019). Las capacidades cognitivas complejas, que se han intentado emular de forma separada por la IA, funcionan de forma interrelacionada; cuando vemos algo en una imagen y lo reconocemos como parte de una categoría, no estamos analizando cada detalle de ese algo y comparándolo con millones de imágenes de lo mismo anteriores para decidir si pertenece a esa categoría, sino que lo relacionamos con otros elementos que no están en la imagen y hacemos analogías basadas en nuestra experiencia (López de Mantaras, 2020).

El campo de la cognición corporizada ha mostrado que las capacidades cognitivas complejas, como el pensamiento conceptual, dependen del cuerpo, al estar inextricablemente asociadas a las emociones, la percepción, o a representaciones neuronales de estados físicos del cuerpo (Foglia & Wilson, 2013). Por ejemplo, la investigación en neurociencia sugiere que las áreas cerebrales dedicadas a la cognición compleja están asociadas a las áreas dedicadas a sistemas sensoriales y motores, y que capacidades como el pensamiento abstracto usan representaciones neuronales que transmiten

información del cuerpo (Damasio, 1994). Como casi todo en la evolución, las funciones más complejas, como las capacidades cognitivas superiores, se desarrollan utilizando capacidades y funciones más básicas.

El neurocientífico Antonio Damasio ha sido uno de los más conocidos defensores del papel del cuerpo y los sentimientos como la base de la que emerge la cognición compleja, criticando también la concepción dualista de la mente y el cuerpo desde la neurociencia (Damasio, 1994; Damasio & Carvalho, 2013). En el trabajo de este autor, la piedra fundamental a partir de la que se desarrolló la inteligencia y la conciencia humanas fue la necesidad de mantener la homeostasis del metabolismo, es decir, mantener los parámetros biológicos del cuerpo en un rango compatible con la vida. Esto requiere una continua percepción y representación neuronal del estado del organismo, de las vísceras y el sistema musculoesquelético, lo que constituye el bloque más básico de la conciencia (Damasio & Meyer, 2009).

Los mecanismos homeostáticos básicos (por ejemplo, retirarse ante algo que quema o hiere), compartidos por casi todo ser vivo, se ven expandidos en animales con un sistema nervioso más complejo por los sentimientos, que llevan aparejados una valencia (positiva o negativa). Esta valencia sirve para indicar desviaciones homeostáticas y señalar la naturaleza ventajosa o desventajosa de una situación fisiológica, facilitando el aprendizaje y funcionando como guías para un comportamiento adaptativo (Damasio & Carvalho, 2013). Los sentimientos emergen a partir de representaciones de estados internos del cuerpo, y los procesos neuronales asociados a sentimientos se dan en regiones relacionadas con la interocepción (percepción y representación de estados corporales). La aparición de los sentimientos en la evolución —que implican darse cuenta de que algo es positivo o negativo, en cierta manera ser consciente, aunque solo sea de ese aspecto— fue uno de los hitos en el camino a la conciencia y a las capacidades cognitivas complejas (Damasio & Carvalho, 2013; Damasio & Meyer, 2009).

La conciencia humana es una extensión más de los mecanismos homeostáticos, al permitir una mayor flexibilidad y planificación en entornos impredecibles: al saber de nuestra propia existencia y de la existencia de objetos y eventos fuera, al saber de nuestro pasado y poder hacer predicciones sobre el futuro, podemos planear y responder de manera flexible para evitar lo que sea dañino y acercarnos a lo que sea

beneficioso (Damasio & Meyer, 2009). Por tanto, la conciencia implica tanto la percepción del estado interno como de objetos y eventos externos en relación con uno mismo (y de los posibles cambios en el estado interno debidos a objetos y eventos externos). La maquinaria neuronal relacionada más estrechamente con la conciencia es contigua y está interconectada, en el tronco encefálico, con las estructuras que gobiernan la atención y las emociones y las que regulan los estados del cuerpo. Lógico, ya que todas estas funciones están al cargo, en último término, del proceso fundamental de mantener la homeostasis del cuerpo, es decir, de la supervivencia.

En definitiva, un abismo separa la inteligencia humana de las técnicas de IA actualmente viables, que adolecen de este complejísimo soporte biológico del que emergen las funciones cognitivas superiores y la conciencia. Debido a ésta y otras diferencias fundamentales ya revisadas, varios autores defienden que no se podrá alcanzar una IA general desde el paradigma del aprendizaje automático. Si esto ocurrirá o no es solamente especulación, ya que es falso que los avances actuales en la IA estrecha nos acerquen a una IA general. Podemos respirar tranquilos; la singularidad, el momento en que la IA superará la inteligencia humana, solo está cerca para quien valore más la ciencia ficción que la ciencia o para quien se deje llevar incautamente por la hipérbole sobre la IA en los medios. ¿Qué nos debería preocupar de la IA, entonces? Y, ¿qué direcciones apuntan al desarrollo de una IA general, de forma realista? En el siguiente apartado abordamos, someramente, estas cuestiones.

V. Riesgos reales y especulaciones sobre la IA en la actualidad

Los riesgos en el desarrollo de la IA actual, por tanto, no están en que la IA se mejore a sí mismo exponencialmente hasta llegar a una IA superinteligente. ¿A qué se debe, entonces, declaraciones como la alarmante carta abierta que pedía paralizar temporalmente el desarrollo de la IA por ser demasiado potente (Future of Life Institute, 2023)? Ese discurso distópico e hiperbólico consigue dos importantes beneficios para las principales empresas desarrolladoras de IA: en primer lugar, trasladan el mensaje implícito de que las técnicas de IA actuales son más potentes de lo que realmente son y de que están en una posición de crecimiento exponencial. Para un inversor, el crecimiento exponencial de la potencia de la IA se traslada a un crecimiento

también exponencial de las empresas desarrolladoras de IA, y en último término, de los beneficios por invertir en estas empresas. De hecho, la inversión en IA se ha multiplicado en los últimos años, convirtiéndose en un boyante negocio, o más bien una burbuja, en la que muchas empresas y *startups* incluyen el término IA solo para atraer inversión (Cooban, 2023).

El segundo objetivo que consigue es que el debate sobre los riesgos de la IA se centre en los supuestos peligros de una IA descontrolada y excesivamente poderosa —como hemos querido explicar, totalmente irreales—, en vez de en los riesgos o aspectos polémicos muy reales que presenta esta tecnología. Uno de ellos es la amplificación de los sesgos y discriminaciones existentes en la sociedad por parte de los algoritmos; al fin y al cabo, un algoritmo de aprendizaje profundo será tan certero y libre de sesgos como certeros y libres de sesgos sean los datos (etiquetados y producidos por humanos) en los que se ha entrenado. Con el problema añadido de la opacidad en saber qué información ha tenido en cuenta un algoritmo de aprendizaje profundo para llevar a cabo una clasificación y su tendencia a utilizar «atajos» si sirven para llegar a la solución para la que se le ha programado.

Por ejemplo, en el sistema judicial, un informe creado por IA para un juez o una jueza sobre el riesgo cuantitativo de reincidencia de un delincuente puede aparentar ser una información objetiva, rigurosa y neutra: al fin y al cabo, lo ha realizado una IA en base a análisis estadísticos complejos de cientos o miles de casos (y tenemos la tendencia a pensar que, si algo implica un análisis cuantitativo, es riguroso e imparcial). Sin embargo, en realidad estos algoritmos introducen los sesgos sistemáticos presentes en el sistema de justicia —por ejemplo, racistas—, implícitos en los datos en los que se entrena el algoritmo. De hecho, se ha demostrado que este sesgo ya ha ocurrido en casos en los que un algoritmo sobreestimaba el riesgo real de delinquir de personas negras frente a personas blancas en el sistema de justicia de EEUU (Okidegbe, 2022).

El otro riesgo que queremos mencionar está relacionado con la sobreestimación de las capacidades reales de la IA (LaGrandeur, 2023). Las técnicas de IA son en general frágiles, en el sentido de que son vulnerables a errores catastróficos por pequeñas desviaciones que entran en conflicto con sus datos de entrenamiento, o a funcionar mal frente a intentos adversariales (en los que se quiere confundir al algoritmo a propósito). La hipérbole y sobreestimación de sus capacidades reales puede dar lugar a que pongamos

demasiada confianza en una IA que en realidad no comprende nada y que es bastante estúpida, tal y como señala López de Mantaras (2020). Esto puede dar lugar a situaciones más o menos cómicas, como el caso del robot de seguridad que iba a patrullar el metro de Nueva York, anunciado por el alcalde y recibido con una gran cobertura mediática. Tuvo que ser finalmente retirado unos meses después, ya que no servía para realizar las tareas de vigilancia que se pretendía. De hecho, debía ser el robot el vigilado por dos policías humanos para que no fuera vandalizado por los usuarios del metro (Rubinstein & Meko, 2024). Si tenemos en cuenta el potencial uso de la IA en otras áreas como las armas letales, sin embargo, la sonrisa se nos puede congelar en la boca. No porque la IA decida rebelarse de forma autónoma contra nosotros, en una suerte de Terminator, sino porque cometa errores de generalización, sea frágil y en general más incapaz de diferenciar y tomar decisiones fiables de lo que sus programadores o usuarios estimen.

Hay otros aspectos no particularmente cómodos para las empresas desarrolladoras de IA, como los relacionados con la privacidad y confidencialidad de datos o con la propiedad intelectual, que por ser más del dominio público no trataremos. Pero, más allá de sus riesgos y limitaciones, es evidente que la IA estrecha actual ofrece posibilidades y oportunidades. Tiene y tendrá un gran impacto en el mercado laboral, la educación, y otros muchos aspectos de nuestra vida, aunque seguramente en menor medida de lo que se nos quiere hacer ver. Estas posibilidades son ampliamente difundidas en los medios y seguramente otros trabajos de este monográfico abordarán algunas de ellas.

¿Cuáles son, entonces, las posibilidades de desarrollar una IA general? Como hemos reiterado, los avances en IA estrecha no implican que nos estemos acercando a una IA general. Algunos autores sugieren algunas direcciones por las que se podría avanzar hacia en esta dirección, que apenas pasan de ser especulaciones. Parece que hay cierto consenso en la necesidad de que la IA sea corpórea e interaccione con el mundo para que se pueda desarrollar una IA general, lo que nos lleva al campo de la robótica (López de Mantaras, 2020; Nilsson, 2006; Mitchell, 2021). En la robótica actual, la percepción e interacción con parámetros físicos de los robots es limitada y, en los casos en los que debe mantenerse en ciertos parámetros, el robot simplemente aplica un rango programado externamente a partir de la información detectada por sus sensores.

Una sugerente propuesta en este sentido, por el propio Antonio Damasio, es la robótica blanda, en la que se equiparía al robot con tejidos blandos en su superficie. Aunque implicaría una mayor vulnerabilidad, también daría lugar a una mayor capacidad de interacción con el medio, y, con los sensores adecuados, permitiría al robot obtener información de cuando algún objeto o evento externo está afectando esos tejidos. Dotándole de representaciones con valencia negativa o positiva sobre los estados de viabilidad de su propia supervivencia, se podría dar lugar a un equivalente artificial de los sentimientos e, hipotéticamente, a que el robot fuera responsable de su propia homeostasis. Estos autores proponen también la integración inter-modal de representaciones de los estados internos y sus desviaciones homeostáticas con la percepción del mundo y los objetos externos. Según estos autores, el interés en la propia supervivencia y el significado intrínseco de las percepciones de su estado interno, en relación con el mundo y los objetos externos, son los cimientos a partir de los cuales se podría llegar a algo parecido a la conciencia y la inteligencia general (Man & Damasio, 2019). Ni que decir tiene que esto es una perspectiva meramente especulativa en la actualidad, y no está nada claro cómo se podría dotar a la IA de esas capacidades mediante el aprendizaje profundo.

Otra interesante propuesta fue avanzada por el propio Turing: la máquina o robot niño. Se trataría de dotar a un robot no de las capacidades cognitivas de un adulto, sino de las de un bebé o un niño, y que, mediante la interacción repetida con el mundo físico y social, aprenda de la misma manera que un niño o niña lo hace (Turing, 1950; Nilsson, 2006). De igual manera, aunque teóricamente sugerente, nadie sabe cómo podríamos llevar a cabo tal empresa. En definitiva, lo que parece claro es que, de una manera u otra, la humanidad seguirá pensando en poder crear un día una IA general que se pueda comparar a la inteligencia humana. Al margen de hipérboles y sensacionalismo, tenemos poco más que especulaciones, pero solo el tiempo dirá si se puede avanzar en este interesante reto o hay impedimentos insalvables, tal y como algunos autores como Dreyfus sugieren.

VI. Conclusiones

En primer lugar hemos querido señalar que algunas de las razones por las que se da un estado de opinión en el que nos preguntamos si la IA está o estará de forma inminente al nivel de la inteligencia humana son la

hipérbole y el sensacionalismo que han acompañado a esta tecnología desde sus inicios (Natale & Ballatore, 2020; Pilling & Coulton, 2019). Hipérbole que ha sido alimentada por parte de sus desarrolladores y los medios de comunicación y que, combinada con el uso de metáforas poco exactas pero populares (por ejemplo, la de la mente como computadora), nos lleva a sobreestimar y entender mal las capacidades reales de la IA continuamente.

Como hemos tratado de argumentar, la inteligencia general humana no se limita a unas reglas explícitas y un conocimiento formal codificable y susceptible de ser computado, lo que invalida la hipótesis de partida del campo de la IA. Tampoco puede existir de forma independiente a un cuerpo que actúa e interacciona con el mundo, con un metabolismo biológico del que las capacidades cognitivas complejas emergen como mecanismos adaptativos para mantener la homeostasis. Capacidades cognitivas complejas —incluida la conciencia— que funcionan de forma interrelacionada entre ellas y con las emociones y estados corporales, ya que desde el punto de vista evolutivo se desarrollaron a partir de estos mecanismos de regulación y adaptación más básicos. El comportamiento inteligente humano consiste en la integración —en gran parte no consciente— de percepción holística, razonamiento abstracto, emociones, planificación y otras capacidades, al servicio en último término de la supervivencia. Una integración que es exquisitamente afinada tanto por una vida de experiencia acumulada de interacción con en el mundo (desde el punto de vista ontogenético) como por la evolución durante cientos de miles de años para optimizar la adaptación a entornos sociales complejos y cambiantes (desde el punto de vista filogenético).

Las técnicas y algoritmos de IA logran, mediante la programación en base a reglas explícitas lógico-matemáticas o el entrenamiento computando miles o millones de casos, emular algunos comportamientos inteligentes como reconocer imágenes, usar el lenguaje o ganar a juegos de mesa. Comportamientos que solo desde una perspectiva funcionalista podemos entender como realmente inteligentes; generalmente, presentan importantes limitaciones como la falta de generalización del aprendizaje o errores catastróficos ante pequeñas distorsiones. Y, evidentemente, el proceso por el que una IA puede llegar a un resultado exitoso en un problema que requiera inteligencia es fundamentalmente distinto al del ser humano. En realidad, la inteligencia solo está en los datos que el algoritmo

manipula y en el programador que prepara y entrena iterativamente un algoritmo que de ninguna manera comprende el significado de los símbolos que manipula.

Por lo tanto, es mucho lo que nos define y separa como humanos frente a las técnicas de la IA en la actualidad, y por ahora no está nada claro si va a disminuir realmente ese abismo, más allá de sobredimensionados avances en la IA estrecha. No son pocos los autores que proponen que nunca se podrá alcanzar una IA general, al menos desde el paradigma del aprendizaje profundo imperante (Chomsky, Roberts & Watumull, 2023; Fjelland, 2020). Parece más realista dedicar nuestros esfuerzos a saber cómo utilizar bien las grandes posibilidades de la IA como complemento a la inteligencia humana, usando las capacidades de computación y automatización que proporciona la IA en tareas bien definidas y siempre bajo nuestra supervisión. El otro esfuerzo que merece la pena hacer es, como con todo, educarnos y así debatir y valorar críticamente avances tecnológicos como la IA actual, para lo que monográficos como éste suponen una excelente iniciativa.

Referencias bibliográficas

Altman, S. [en línea]: «Moore's law for everything», en *Sam Altman* (2021). https://moores.samaltman.com [Consultado: 02/04/2024].

Blank, I. A. (2023): «What are large language models supposed to model? », *Trends in Cognitive Sciences*, 27(11), pp. 987–989. doi: 10.1016/j.tics.2023.08.006.

Brödner, P. (2019): «Coping with Descartes' error in information systems», *AI and Society*, 34(2), pp. 203–213. doi: 10.1007/s00146-018-0798-8.

Bush, V. [en línea]: «As we may think», en *Life* (1945). https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/ [Consultado: 02/04/2024].

Chomsky, N., Roberts, I., & Watumull, J. [en línea]: «The false promise of Chat-GPT» en *The New York Times* (2023). https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html [Consultado: 02/04/2024].

Cooban, A. [en línea]: «AI investment is booming. How much is hype?» en *CNN* (2023). https://edition.cnn.com/2023/07/23/business/ai-vc-investment-dot-com-bubble/index.html [Consultado: 26/03/2024].

Damasio, A. (1994): *Descartes' error: Emotion, reason, and the human brain.* New York City: Putnam.

Damasio, A., & Carvalho, G. B. (2013): «The nature of feelings: Evolutionary and neurobiological origins», *Nature Reviews Neuroscience*, 14(2), pp. 143–152. doi: 10.1038/nrn3403.

Damasio, A., & Meyer, K. (2009): «Consciousness: An overview of the phenomenon and of its possible neural basis», en Laureys, S. y Tononi, G. (eds) *The neurology of consciousness*. San Diego: Elsevier, pp. 1–14. doi: 10.1016/B978-0-12-374168-4.00001-0.

Davis, E., & Marcus, G. (2015): «Commonsense reasoning and commonsense knowledge in artificial intelligence», *Communications of the ACM*, 58(9), pp. 92–103. doi: 10.1145/2701413.

de Saint Laurent, C. (2018): «In defence of machine learning: Debunking the myths of artificial intelligence», *Europe's Journal of Psychology*, 14(4), pp. 734–747. doi: 10.5964/ejop.v14i4.1823.

Despres, J. [en línea]: «Scenario: Shane Legg» en *Future* (2008). https://tinyurl.com/hwzna364 [Consultado: 02/04/2024]

Dreyfus, H. (1973): «Crítica de la razón artificial», *Diálogos: Artes, Letras, Ciencias Humanas*, 9(1), pp. 11–18.

Dreyfus, H. (2007): «Why Heideggerian AI failed and how fixing it would require making it more Heideggerian», *Philosophical Psychology*, 20(2), pp. 247–268. doi: 10.1080/09515080701239510.

Fjelland, R. (2020): «Why general artificial intelligence will not be realized», *Humanities and Social Sciences Communications*, 7(1), pp. 1–9. doi: 10.1057/s41599-020-0494-4.

Foglia, L., & Wilson, R. A. (2013): «Embodied cognition», *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(3), pp. 319–325. doi: 10.1002/wcs.1226.

Future of Life Institute [en línea]: *Pause Giant AI Experiments: An Open Letter*. Future of Life Institute (2023). https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [Consultado: 26/03/2024].

Geirhos, R. et al. (2020): «Shortcut learning in deep neural networks», *Nature Machine Intelligence*, 2(11), pp. 665–673. doi: 10.1038/s42256-020-00257-z.

Giattino C., Mathieu E., Samborska, V. y Roser, M. [en línea]: «Artificial Intelligence», en *Our World in Data* (2023). https://ourworldindata.org/artificial-intelligence [Consultado: 26/03/2024].

Gill, K. S. (2019): «From judgment to calculation: the phenomenology of embodied skill: Celebrating memories of Hubert Dreyfus and Joseph Weizenbaum», *AI and Society*, 34(2), pp. 165–175. doi: 10.1007/s00146-019-00884-0.

Gottfredson, L. S. (1997): «Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography», *Intelligence*, 24(1), pp. 13–23. doi: 10.1016/S0160-2896(97)90011-8.

IBM [en línea]: «701 Translator», en *IBM Press Release* (1954). https://tinyurl.com/2vbkr22w [Consultado: 28/03/2024].

Jiang, Y. et al. (2022): «Quo vadis artificial intelligence?», *Discover Artificial Intelligence*, 2(1). doi: 10.1007/s44163-022-00022-8.

Korteling, J. E. y Hans. et al. (2021): «Human- versus Artificial Intelligence», *Frontiers in Artificial Intelligence*, 4, pp. 1–13. doi: 10.3389/frai.2021.622364.

LaGrandeur, K. (2023): «The consequences of AI hype», *AI and Ethics*, (0123456789), pp. 1–4. doi: 10.1007/s43681-023-00352-y.

Leaver, T., & Srdarov, S. (2023): «ChatGPT Isn't Magic», *M/C Journal*, 26(5), pp. 1–6. doi: 10.5204/mcj.3004.

Lighthill J. (1973): «Artificial intelligence: a general survey», en *Artificial Intelligence: a paper symposium*. Brooklyn: Science Research Council.

López de Mántaras, R. (2017): «Algunas reflexiones sobre el presente y el futuro de la Inteligencia Artificial», *Revista de Occidente*, 436, pp. 57–72.

López de Mántaras, R. (2020): «El traje nuevo de la inteligencia artificial», *Investigación y ciencia*, 526, pp. 52–59.

Man, K., & Damasio, A. (2019): «Homeostasis and soft robotics in the design of feeling machines», *Nature Machine Intelligence*, 1(10), pp. 446–452. doi: 10.1038/s42256-019-0103-7.

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), p. 12. doi: 10.1609/aimag.v27i4.1904.

McCracken, H. [en línea]: «Inside Mark Zuckerberg's bold plan for the future of Facebook», en *Fast Company* (2015). www.fastcompany.com/3052885/mark-zuckerberg-facebook [Consultado: 02/04/2024].

Mitchell, M. (2021): «Why AI is harder than we think» *ArXiv*. doi: 10.1145/3449639.3465421.

Mitchell, M. (2024): «Debates on the nature of artificial general intelligence», *Science*, 383(6689): doi: 10.1126/science.ado7069.

Modis, T. (2006): «The singularity myth», *Technological Forecasting and Social Change*, 73(2), pp. 104–112. doi: 10.1016/j.techfore.2005.12.004.

Natale, S. y Ballatore, A. (2020): «Imagining the thinking machine: Technological myths and the rise of artificial intelligence», *Convergence*, 26(1), pp. 3–18. doi: 10.1177/1354856517715164.

National Science and Technology Council [en línea]: *Preparing for the future of artificial intelligence.* Executive Office of the President. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf [Consultado: 01/04/2024].

Newell, A., & Simon, H. A. (1976): «Computer science as empirical inquiry», *Communications of the ACM*, 19(3), pp. 113–126. doi: 10.1145/360018.360022.

Nilsson, N. J. (2006): «Human-level artificial intelligence? Be serious!», *AI Magazine*, 26(4), pp. 68–75.

Okidegbe, N. (2022): «The democratizing potential of algorithms?», *Connecticut Law Review*, 53, pp. 1–47.

Pilling, F., & Coulton, P. (2019): «Forget the singularity, its mundane artificial intelligence that should be our immediate concern», *Design Journal*, 22(sup1), pp. 1135–1146. doi: 10.1080/14606925.2019.1594979.

Ranaweera, M. y Mahmoud, Q. H. (2021): «Virtual to real-world transfer learning: A systematic review», *Electronics*, 10(12). doi: 10.3390/electronics10121491.

Raviv, L., Lupyan, G., & Green, S. C. (2022): «How variability shapes learning and generalization», *Trends in Cognitive Sciences*, 26(6), pp. 462–483. doi: 10.1016/j.tics.2022.03.007.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016): «"Why should I trust you?" explaining the predictions of any classifier», *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pp. 97–101. doi: 10.18653/v1/n16-3020.

Searle, J. R. (1980): «Minds, brains, and programs», *Behavioral and Brain Sciences*, 3(3), pp. 417–424. doi: 10.1017/S0140525X00005756.

Silver, D. et al. (2018): «A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play», *Science*, 362(6419), pp. 1140–1144. doi: 10.1126/science.aar6404.

Turing, A. (1950): «Computing machinery and intelligence», *Mind, LIX*(236), pp. 433–460. doi: 10.1093/mind/LIX.236.433.

Woolaston, V. [en línea]: «We'll be uploading our entire minds to computers by 2045 and our bodies will be replaced by machines within 90 years, Google expert claims», en *Daily Mail* (2013). https://tinyurl.com/ht44uxzv [Consultado: 02/04/2024].

Pablo Carrera: Docente e investigador en la Facultad de Ciencias de la Salud de la Universidad Isabel I.

Líneas de investigación:

– Desarrollo socioemocional y cognitivo en la infancia, adversidad y protección a la infancia.

Publicaciones recientes:

– Carrera, P., Zablah, F. M., de la Rosa, Y., & Benito-Gomez, M., (2024). «Scaling up Attachment and Biobehavioral Catch-up with Latine families: Implementation processes and effectiveness». *Infant Mental Health Journal*. doi: 10.1002/imhj.22141

– Carrera, P., Ferrari, L., Cáceres, I., Ranieri, S., Rosnati, R., Palacios, J., & Román, M. (2024). «Birth country identification and exploration in adolescents internationally adopted from Russia». *Developmental Child Welfare*. doi: 10.1177/25161032241296108

– Jiménez-Morago, J. M., Carrera, P., & León, E., (2024). «La intervención ante el maltrato infantil». En. L. Jiménez y V. Hidalgo (coords.), *Intervención familiar. Necesidades y apoyos*. Editorial Universidad de Sevilla.

Correo-e: pablotomas.carrera@ui1.es