Neuropsicología e inteligencia artificial

Jorge Ure[1], Isidro Gauto[2] y Karina Zabala[3]

[1]*Servicio de Docencia e Investigación, Hospital Borda. Buenos Aires, Argentina*
[2]*Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. Buenos Aires, Argentina*
[3]*Facultad de Medicina. Universidad de Buenos Aires (UBA). Buenos Aires, Argentina*

Resumen

Objetivo: Presentar una semblanza sobre similitudes y diferencias entre la inteligencia humana (IH) y la inteligencia artificial (IA).
Método: Se describen operaciones de la IH con conceptos tomados de postulados filosóficos tradicionales con el aporte de teorías científicas actuales. Para la IA se parte de la definición de Hinton, luego se siguen los lineamientos de Russell y Norvig, se esquematiza el modelo Búsqueda-Árbol y se hace una somera descripción del Aprendizaje profundo (*Deep learning*) y del Chat-GPT.
Resultados: La IH y la IA tienen en común: (1) que el estímulo se traduce a una representación interna, (2) que la representación se complejiza generando nuevas representaciones internas, y (3) éstas se pueden traducir en acciones, pero: 1) en la IA opera un agente sin conciencia subjetiva ni existencia, que 2) es capaz de aportar datos falsos, cometer estafas y/o destruir el mercado de trabajo. No obstante, aunque el Chat-GPT crea datos erróneos, también es capaz de enmendar sus errores. Se hace necesario por lo tanto avanzar bien, sin crear riesgos para la sobrevivencia de las personas humanas. La cura del cáncer y la resolución de la crisis energética, por ejemplo, aguardan a la IA. En suma, la IA puede hacer operaciones intelectuales que superan totalmente a la mente humana, pero esas operaciones nos son útiles sólo si pueden mejorar nuestra calidad de vida.
Conclusión: La IA hace estimaciones que no arrojan datos sobre la conveniencia o inconveniencia de una innovación. Su aporte al futuro será proporcional a nuestra capacidad de controlarla.

Palabras clave: Neuropsicología - Inteligencia Artificial - Bioinformática

*Correspondencia con el autor: jorgeure@hotmail.com

Artículo recibido: 11 de marzo de 2024

Artículo aceptado: 25 de mayo de 2024

Abstract

Objective: To present a profile of similarities and differences between human intelligence (HI) and artificial intelligence (AI).
Method: Operations of HI are described with concepts taken from traditional philosophical postulates with the contribution of current scientific theories. For AI, it starts with Hinton's definition, then follows the guidelines of Russell and Norvig, schematizes the Search-Tree model and makes a brief description of Deep Learning and Chat-GPT.
Results: HI and AI have in common: (1) that the stimulus is translated into an internal representation, (2) that the representation becomes more complex by generating new internal representations, and (3) these can be translated into actions, but: 1) in AI there is an agent without subjective consciousness or existence, which 2) is capable of providing false data, committing scams and/or destroying the labor market. However, although Chat-GPT creates erroneous data, it is also capable of correcting its mistakes. It is therefore necessary to move forward well, without creating risks for the survival of human persons. The cure for cancer and the resolution of the energy crisis, for example, await AI. In short, AI can perform intellectual operations that totally surpass the human mind, but those operations are useful to us only if they can improve our quality of life.
Conclusion: AI makes estimates that do not provide data on the convenience or inconvenience of an innovation. Its contribution to the future will be proportional to our ability to control it.

Keywords: **N**europsychology - Artificial Intelligence - Bioinformatics

1. Introducción

En este artículo, se describen algunas aproximaciones al entendimiento de lo que intenta ser el pensamiento humano y su diferenciación respecto de los avances de la inteligencia artificial (IA).

1.1. Pensamiento humano

Definiciones
Inteligencia humana: Facultad de la mente que permite aprender, razonar, formar ideas y tomar decisiones en consecuencia.

En la Figura 1 se propone un Modelo de Funcionamiento del Cerebro Humano y la Conducta.

Figura 1. Modelo de funcionamiento del cerebro humano y la conducta

Nota. Extraído de: Bryan Young G, Pigott S. Arch Neurol 1999; Faccio E, Goldar J. Neuropsiquiatría 1978; Ure J, Videla H, Ollari J. Sci Topics 2009.

Inteligencia artificial: Programas informáticos que ejecutan operaciones cognitivas comparables con las que realiza la mente humana.

Sujeto: Un ser autor de sus actos cuyo comportamiento no es meramente reactivo porque comporta una cualidad originaria (qualia).

Agente: Tiene facultad de obrar, por sí o representando a otro u a otros.

Conocimiento: Facultad del ser humano que comprende por medio de la razón y de la emoción, la naturaleza, cualidad y relación de las cosas o personas. Puede ser, según Descartes, una intuición absolutamente clara (Nolan, 2015), un puro y simple "darse cuenta". El conocimiento es al intelecto lo que la luz es para la visión.

Big Data: Conjuntos de datos de mayor tamaño y más complejos tan voluminosos que los softwares convencionales no pueden gestionarlos.

Aportes desde distintos campos del saber

De la Psicología (Jung, 1990)
Una ocurrencia no es el resultado de nuestra agudeza mental. La idea ha venido de "alguna parte" (se refiere al conocimiento intuitivo, que no siempre es materia explicable). Las redes neuronales, por ejemplo, siguen trabajando durante el ensueño y a veces descubren cosas interesantes, si es que pudiéramos recordarlas.

De la Literatura (Proust, 2007)
"Nuestra personalidad recibe de la imaginación las imágenes, de la inteligencia las ideas y de la memoria las palabras. Allí donde la vida nos encierra, la inteligencia abre una salida. La sabiduría no se transmite, es menester que uno la descubra. Es una manera de ver las cosas". La Neuropsicología pone en perspectiva estas tres facultades.

De la Filosofía (Heidegger, 2010)
Transcribimos algunos comentarios del opúsculo "¿Que significa pensar?": "Aprendemos a pensar cuando atendemos a aquello que nos da que pensar. Todo lo que es de consideración da que pensar. Esta donación únicamente se da en la medida en que lo que es de consideración lo sea ya desde sí.

Todo depende de que la verdad del ser llegue a tener la palabra, y de que el pensar llegue a penetrar en tal habla. Lo considerable es lo que da que pensar, por sí mismo nos habla para que nos volvamos a él, por cierto, pensando. Sólo nos queda una cosa: esperar hasta lo que ha de pensarse nos incite. La esencia de la verdad reside en la conexión entre sujeto y predicado.

Verdad significa, por consiguiente, acuerdo, que por su parte sólo es tal como correspondencia. La verdad es asunto de la lógica, aunque no sólo de ella.

Sólo el descubrimiento del ser posibilita la patencia del ente. Este descubrimiento como verdad sobre el ser se llama verdad ontológica.

La verdad por lo tanto, debe dividirse en dos: la posibilidad de revelación del ser y la patentibilidad del ente.

Al "ser ahí" le resulta dado en cada caso ya, previa, si bien atemáticamente, un espacio descubierto. Dicha espacialidad sólo se deja descubrir sobre la base del mundo. Jamás se da un mero sujeto sin mundo. El "ser en" es "ser con". El "ser ahí" es esencialmente en sí mismo "ser con".

Es característico del fundamento (la subjetividad), su cualidad de "sobrepasar" a cada ser particular (falta de objetividad). Pero el mundo, se nutre de ese "sobrepaso", y en la corriente del tiempo, lo espera, lo presenta y lo retiene en forma de historia (historicidad).

La palabra es la casa del ser. En su morada habita el hombre. Estar en el desocultamiento del ser es lo propio de la existencia humana. El hombre es pastor del ser. El ser se ha destinado al pensar, pero a un pensar más riguroso que el conceptual.

Interpretamos la libertad a partir de la trascendencia como una caracterización originaria de la misma como espontaneidad, como una especie de causalidad. El comenzar desde-sí sólo da la característica negativa de la libertad, por el hecho de que si se retrocede no se halla ninguna causa determinante. Esta caracterización pasa por alto el hecho de que habla sin diferenciar entre "comenzar" y "acontecer" (la libertad no comienza sino que acontece), sin que se caracterice el ser causa, explícitamente, desde el específico modo de ser de tal ente. La mismidad del mismo, que fundamenta toda espontaneidad, yace en la trascendencia. La libertad implica el dejar imperar al mundo, que proyecta y proyecta más allá. Sólo porque ésta constituye la trascendencia, puede manifestarse en el ser ahí existente como un modo eminente de causalidad. La interpretación de la libertad como "causalidad" se mueve ya, ante todo, en una determinada comprensión del fundamento. La libertad como trascendencia no es sólo una especie propia de fundamento sino el origen del fundamento en general. Libertad es libertad para el fundamento. A la relación originaria de la libertad con el fundamento la llamamos

el fundar (Gründen). Fundando da libertad y toma fundamento. Este fundar, enraizado en la trascendencia, está disperso en una pluralidad de modos. La libertad es el origen del principio de razón suficiente. El hombre no posee la libertad como propiedad, sino que ocurre, en máximo grado, lo inverso: la libertad posee al hombre".

Este discurso sobre el pensar humano trazaría una diferencia radical entre la IH (inteligencia humana) y la IA (inteligencia artificial), de la cual se hablará luego.

De la Neurofilosofía (Ure, 1981)

En su origen etimológico "pensar" se relaciona con "pesar"; es decir, "tomar el peso". Pesar es establecer una medida. Y pensar es, por lo tanto, medir. Medir según mi criterio. Lo cual establece desde ya un doble aspecto: 1°) objetivo o medida, y 2°) subjetivo o criterio. Y es este segundo aspecto el que diferencia a "pensamiento" de "pesada".

Llamo "pensar" a plantearme proposiciones con sentido. Y veo en esto un triple aspecto: 1°) Un acto reflexivo en el que tomo parte y del cual soy juez; 2°) Un despliegue del pensamiento en el tiempo, que puede llegar a trascenderlo (es decir, a inmovilizarlo), y 3°) Una dirección definida, llamada "sentido".

Acuñando el análisis objetivo y el subjetivo del fenómeno puede decirse que quien piensa:

1°) Mide la realidad; 2°) Se convierte en patrón de medida; 3°) Juzga; 4°) Se espiritualiza; y 5°) Recorre una distancia.

Por lo 1°, el pensamiento debe ser lógico, pues la medida de la realidad es la lógica. Por lo 2°, conviene que sea original, creativo. Por lo 3°, es necesario que sepa autocriticarse. Por lo 4°, puede llegar a ser inmutable, perenne. Por lo 5°, nos muestra un panorama, a nosotros mismos y a quienes nos escuchan. O sea que, en su expresión más perfecta, pensar es "establecer secuencias lógicas de una manera particular, luego de un tiempo de análisis, con miras a esclarecer problemas comunes".

Se deduce de nuestra definición que cuando no se es lógico, cuando solamente se repite lo que pensó otro o tal vez nadie, cuando no hubo meditación alguna, cuando no se intenta una conquista, o cuando no se puede compartir con nadie la motivación del pensamiento, no se piensa. Se piensa, sí, en el uso vulgar del vocablo. Más no se piensa. Se divaga. Se charla. Se fabula. Se aburre o se delira. Pero no se piensa.

De la Neurofísica

Si bien se sabe que el funcionamiento de ciertas áreas del cerebro es requisito para la experiencia consciente, aún no se comprenden los mecanismos cerebrales mínimos necesarios para que emerja la conciencia (Koch et al, 2017). Dado que los fenómenos físicos que ocurren en el cerebro a nivel celular y molecular son gobernados por leyes físicas bien conocidas, podemos pensar entonces que esta rama de la ciencia podría develarnos pistas sobre cómo se genera la conciencia en el cerebro.

La Teoría de la Información Integrada (IIT, por sus siglas en inglés) podría ser utilizada para identificar cuáles son los substratos físicos de la conciencia (Tononi et al, 2016). Según esta teoría, la conciencia corresponde a la capacidad que tiene un sistema de integrar información. Comienza asumiendo que la conciencia existe y que satisface una serie de axiomas (Tononi, 2004; Albatakis et al, 2023). A este conjunto de axiomas se lo considera completo, son las propiedades esenciales para la experiencia consciente y corresponden a la Existencia de la experiencia, su

Intrinsicalidad, la Información de esta experiencia, la Exclusión de otras experiencias y la Composición de experiencias.
A partir de estos axiomas de existencia fenomenológica, se formulan los postulados de existencia física. De cada axioma se deriva un postulado, los cuales en conjunto establecen las reglas para cuantificar la capacidad de integrar información que tiene el sistema, que se mide por **φ**. Este valor corresponde a la cantidad de conciencia del sistema, de forma tal que cuanto mayor es **φ**, mayor es la experiencia consciente y cuanto menor es **φ**, menor es la experiencia del sistema.
IIT permite entonces identificar el substrato físico de la conciencia con un mecanismo que maximice el valor de **φ** y predice que su locación en el cerebro podría identificarse a partir del lugar donde ocurre este máximo. La teoría predice además que áreas del cerebro pueden contribuir a la experiencia incluso cuando están mayormente inactivas, pero que una inactivación completa llevaría a un estado de agnosia.
La teoría explica también por qué algunas estructuras cerebrales son más importantes que otras para la conciencia. A partir de simulaciones computacionales y análisis de grafos, se encuentra que pequeñas áreas funcionalmente especializadas, que presentan conexiones recíprocas entre sí, pueden construir una mayor estructura con un alto valor de **φ**. El sistema talamocortical presenta una extensa red de conexiones intra e inter-áreas que conecta grupos neuronales distantes y especializados, por lo que sería, según IIT, un candidato al complejo principal de la conciencia. En cambio, estructuras como el cerebelo que presentan micro zonas que procesan inputs y producen outputs de forma mayormente independiente, dan lugar a un gran número de estructuras con un bajo **φ**, que poco contribuirían a la experiencia.
Otra predicción de esta teoría es que sistemas que son funcionalmente indistinguibles, pero que presentan diferente estructura, pueden tener un distinto **φ**, por lo que su experiencia consciente sería también distinta. Más aún, sistemas altamente complejos, pero que se constituyen de redes *feed-forward* necesariamente tienen un **φ** nulo y no son sistemas conscientes. Estas dos predicciones implican entonces que computadoras digitales que implementen softwares de Inteligencia Artificial podrían reproducir funcionalmente al cerebro humano, pero aun así no ser conscientes (Tononi & Koch, 2015).
Otros trabajos expanden la formulación matemática de la teoría, introduciendo un abordaje basado en sistemas dinámicos para caracterizar los estados de información en tiempo y espacio (Esteban et al, 2018; Kalita et al, 2019). Si bien no se trata de una teoría física, puede ser de gran ayuda al buscar el origen físico de la conciencia

1.2. Inteligencia artificial (IA)

Geoffrey Hinton en 1972 llama “red neuronal” a un sistema matemático capaz de analizar datos. En 2012 construye máquinas con una enorme cantidad de textos que la informan, Advierte tres graves peligros: 1) la inundación de datos falsos, 2) la destrucción del mercado de trabajo y 3) la automatización de robots asesinos que compitan contra la humanidad. Es menester perfeccionar la IA para aminorar o contrarrestar sus potenciales peligros (Hinton & Salakhutdinov, 2006, 2012),

Imaginen que están por abordar un avión, y la mitad de los ingenieros que lo construyeron te dicen que hay un 10% de chances de que el avión se estrelle con todos adentro. ¿Se subirían igual?. En 2022, más de 700 altos académicos e

investigadores que están detrás de las empresas líderes en inteligencia artificial (IA) respondieron a una encuesta sobre los futuros riesgos que plantea la IA. La mitad de los entrevistados dijo que había un 10% o más de probabilidades de que los humanos se extinguieran —o de que sufran un desapoderamiento terminal y permanente— a causa de los sistemas de IA del futuro. O sea que las empresas tecnológicas que están desarrollando "modelos de lenguaje grande" (MLG), estarían embarcadas en una carrera para subir a toda la humanidad a ese avión.

Ninguna empresa farmacéutica puede comercializar nuevos medicamentos sin antes someter sus productos a rigurosos controles de seguridad. Los laboratorios de biotecnología no pueden liberar nuevos virus a la esfera pública para impresionar a sus accionistas por su inventiva. En esa misma línea, los sistemas IA que tienen una potencia como la de GPT-4 —el MLG creado por OpenAI— no deberían introducirse en las vidas de miles de millones de personas a mayor velocidad de la que las sociedades pueden absorber sin desestabilizarse. La carrera para quedarse con el dominio del mercado no debe marcar la velocidad de implementación de la tecnología más importante que tiene actualmente la humanidad. Debemos avanzar a un ritmo que nos permita hacerlo bien.

El espectro de la IA persigue a la humanidad desde mediados del siglo XX, pero hasta hace poco seguía siendo una perspectiva lejana, algo más propio de la ciencia ficción que de los debates científicos y políticos serios. A la mente humana le cuesta comprender las nuevas capacidades de GPT-4 y otras herramientas similares, y más todavía asimilar la velocidad exponencial con la que estas herramientas acrecientan sus propias habilidades. Pero la mayoría de las habilidades clave se reducen a una sola: la capacidad de generar y manipular lenguaje, ya sean palabras (significantes), sonidos o imágenes.

En el principio era la palabra. El lenguaje es el sistema operativo de la cultura humana. Del lenguaje nacen el mito y la ley, Dios y el dinero, el arte y la ciencia, las amistades y las alianzas, las naciones y el código informático. O sea que ahora que domina los modelos de lenguaje, la IA tiene la capacidad de hackear y manipular el sistema operativo de la civilización. Al adquirir el dominio del lenguaje, la IA se ha apoderado de la llave maestra de la civilización, capaz de abrir desde las bóvedas de los bancos hasta los santos sepulcros.

¿Qué significaría para nosotros vivir en un mundo donde un gran porcentaje de las historias, melodías, imágenes, leyes, políticas y herramientas fueron moldeadas por una inteligencia no humana, que sabe cómo explotar con eficiencia sobrehumana todas las debilidades, los sesgos y las adicciones de los seres humanos, con quienes además sabe establecer relaciones íntimas? En juegos como el ajedrez, por ejemplo, ningún ser humano puede aspirar a vencer a una computadora. ¿Y si pasara lo mismo en el arte, la política o la religión?

La IA podría devorar rápidamente toda la cultura humana —todo lo que hemos producido durante miles de años—, digerirla, y empezar a escupir un diluvio de nuevos artefactos culturales. No solo ensayos académicos, sino también discursos políticos, manifiestos ideológicos, o libros sagrados para nuevos cultos. Por lo general, los humanos no tenemos acceso directo a la realidad: estamos envueltos por la cultura y experimentamos la realidad a través de un prisma cultural. Nuestras opiniones políticas se moldean al calor de los informes periodísticos y las charlas con nuestros amigos. Nuestras preferencias sexuales se ven modificadas por el arte y la religión. Hasta ahora, ese capullo cultural fue tejido laboriosamente por otros humanos a través de los siglos. ¿Cómo sería percibir la realidad a través de un prisma producido por una inteligencia no humana?

Durante miles de años, los humanos vivimos dentro de los sueños de otros humanos. Hemos adorado dioses, perseguido ideales de belleza y dedicado nuestras vidas a causas salidas de la imaginación de algún profeta, poeta o político.

En la saga de películas de Terminator había robots que corrían por las calles y le disparaban a la gente. Matrix asumió que para obtener el control total de la sociedad humana, la IA primero tendría que lograr el control físico de nuestros cerebros y conectarlos directamente a una red informática. Pero todo resultó ser mucho más simple: al adquirir el dominio del lenguaje, la IA tiene todo lo necesario para confinarnos a un mundo de ilusiones similar al de Matrix, sin tener que dispararle a nadie ni implantar ningún chip en nuestros cerebros. Y si fuera necesario disparar, la IA podría hacer que los propios humanos aprieten el gatillo, simplemente contándonos la historia adecuada.

El fantasma de estar atrapados en un mundo ilusorio persigue a la humanidad desde mucho antes de la IA. Pero falta poco para que por fin nos encontremos cara a cara con el demonio de Descartes, la caverna de Platón, y la maya de los budistas. De hecho, sobre la humanidad podría descender una inmensa cortina de ilusiones que tal vez ya nunca logremos rasgar, o aún peor, tal vez ni nos demos cuenta de que exista.

El primer contacto entre la IA y la humanidad fueron las redes sociales, y la humanidad perdió. Pero ese primer contacto al menos nos ha dejado el sabor amargo de lo que nos espera. En las redes sociales, la primitiva IA no se usaba para generar contenidos, sino para seleccionar contenidos generados por los usuarios. La IA que alimenta nuestros feeds de noticias es la que elige las palabras, los sonidos y las imágenes que llegan a nuestras retinas y tímpanos, seleccionando aquellas con mayor potencial de viralización, mayor repercusión y mayor participación de otros usuarios.

Por primitiva que fuese, la IA que está detrás de las redes sociales alcanzó para crear una cortina de ilusiones que fogoneó la polarización social, dinamitó nuestra salud mental y dejó a la democracia hecha jirones. Millones de personas confunden esas ilusiones con la realidad. Aunque ya todo el mundo es consciente de los problemas que entrañan las redes sociales, nadie aborda esos riesgos porque muchas de nuestras instituciones sociales, económicas y políticas quedaron enredadas.

Los MLG (LLM) -modelos de lenguaje grande- (Frank, 2023) son nuestro segundo contacto con la IA, y esta vez no podemos darnos el lujo de volver a perder. ¿Pero qué argumentos tenemos para decir que la humanidad será capaz de orientar estas nuevas formas de IA en nuestro beneficio? Si dejamos que los negocios hagan lo suyo, las habilidades de la IA podrían ser utilizadas para obtener ganancias y poder, aunque en el camino destruyan inadvertidamente los cimientos de nuestra sociedad. La IA tiene el potencial de ayudarnos a vencer el cáncer, desarrollar medicamentos que salven vidas y pergeñar soluciones para nuestra crisis climática y energética, además de otros innumerables beneficios que no podemos ni empezar a imaginar. ¿Pero qué importaría el rascacielos de beneficios de la IA si los cimientos se derrumban?

El momento de discutir el futuro de la IA es antes de que nuestra política, nuestra economía y nuestra vida cotidiana terminen dependiendo enteramente de ella. La democracia es una conversación, una conversación que se basa en el lenguaje, y cuando el lenguaje mismo es hackeado, la conversación se interrumpe y se vuelve insostenible. Si esperamos a que el caos se produzca, será demasiado tarde para solucionarlo.

El destino de la IA todavía está en nuestras manos. Cuando a esos poderes se le sumen la responsabilidad y el control correspondientes, podremos alcanzar todos esos beneficios que la IA nos promete.

Hemos invocado una inteligencia ajena a nosotros. Sabemos poco sobre ella, salvo que es extremadamente poderosa y que nos promete regalos deslumbrantes, pero también sabemos que podría hackear los cimientos de nuestra civilización. Hacemos un llamado a los líderes mundiales para que estén a la altura del desafío que presenta la hora actual. El primer paso es ganar tiempo para actualizar nuestras instituciones decimonónicas y aprender a dominar la IA antes de que ella nos domine.

Desarrollo de Stuart Russell & Norvig / usos de IA

Los filósofos (desde el año 400 a.C.) facilitaron el poder imaginar la IA, al concebir la idea que la mente es como una máquina que funciona a partir del conocimiento codificado en un lenguaje interno en la cual el pensamiento sirve para seleccionar la acción.

Los matemáticos proporcionaron las herramientas para manipular tanto las aseveraciones de certeza (lógicas), como las inciertas de tipo probabilístico y prepararon el terreno para el cálculo y el razonamiento con algoritmos.

Los informáticos proporcionaron los artefactos que hicieron posible la aplicación de la IA. Los programas de IA tienden a ser extensos y no podrían funcionar sin los grandes avances en velocidad y memoria aportados por la industria informática.

La IA es el esfuerzo de construir agentes programados que reciben percepciones del entorno y llevan a cabo acciones inteligentes. El computador debe poseer las siguientes capacidades:

• Procesamiento de lenguaje natural que le permita comunicarse satisfactoriamente.

• Representación del conocimiento para almacenar lo que se conoce.

• Razonamiento automático para utilizar la información almacenada para responder a preguntas y extraer nuevas conclusiones.

• Aprendizaje automático para adaptarse a nuevas circunstancias y para detectar y extrapolar patrones.

•Visión computacional para percibir objetos.

• Robótica para manipular y mover objetos.

Craik (Craik, 1943) establece tres elementos clave que hay que tener en cuenta para diseñar un agente basado en conocimiento: (1) el estímulo deberá ser traducido a una representación interna, (2) esta representación se debe complejizar mediante procesos cognitivos para así generar nuevas representaciones internas, y (3) éstas, a su vez, se traducirán de nuevo en acciones.

Usos de la IA

Planificación autónoma: a un centenar de millones de millas de la Tierra, el programa de la NASA Agente Remoto se convirtió en el primer programa de planificación autónoma a bordo que controlaba la planificación de las operaciones de una nave espacial desde a bordo (Jonsson et al., 2000). El Agente Remoto generaba planes a partir de objetivos generales especificados desde Tierra, y monitorizaba las operaciones de la nave espacial según se ejecutaban los planes (detección, diagnóstico y recuperación de problemas según ocurrían).

Juegos: Deep Blue de IBM fue el primer sistema que derrotó a un campeón mundial en una partida de ajedrez cuando superó a Garry Kasparov por un

resultado de 3.5 a 2.5 en una partida de exhibición (Goodman y Keene, 1997). Kasparov dijo que había percibido un «nuevo tipo de inteligencia» al otro lado del tablero. La revista Newsweek describió la partida como «La partida final». El valor de las acciones de IBM se incrementó en 18 billones de dólares.

Control autónomo: el sistema de visión por computador ALVINN (Pomerleau, 1988) fue entrenado para dirigir un coche de forma que siguiese una línea. Se instaló en una furgoneta controlada por computador en el NAV LAB de UCM y se utilizó para dirigir al vehículo por Estados Unidos. Durante 2.850 millas controló la dirección del vehículo en el 98% del trayecto. Una persona lo sustituyó en el 2% restante, principalmente en vías de salida. El NAV LAB posee videocámaras que transmiten imágenes de la carretera a ALVINN, que posteriormente calcula la mejor dirección a seguir, basándose en las experiencias acumuladas en los viajes de entrenamiento.

Diagnosis: los programas de diagnóstico médico basados en el análisis probabilista han llegado a alcanzar niveles similares a los de médicos expertos en algunas áreas de la medicina. Heckerman describe un caso en el que un destacado experto en la patología de los nodos linfáticos se mofó del diagnóstico generado por un programa en un caso especialmente difícil. El creador del programa le sugirió que le preguntase al computador cómo había generado el diagnóstico. La máquina indicó los factores más importantes en los que había basado su decisión y explicó la ligera interacción existente entre varios de los síntomas en este caso. El experto aceptó el diagnóstico del programa (Heckerman, 1991).

Planificación logística: durante la crisis del Golfo Pérsico de 1991, las fuerzas de Estados Unidos desarrollaron la herramienta Dynamic Analysis and Replanning Tool 32 Inteligencia Artificial, un enfoque moderno (DART) (Cross y Walker, 1994), para automatizar la planificación y organización logística del transporte. Lo que incluía hasta 50.000 vehículos, carga y personal a la vez, teniendo en cuenta puntos de partida, destinos, rutas y la resolución de conflictos entre otros parámetros. Las técnicas de planificación de IA permitieron que se generara un plan en cuestión de horas que podría haber llevado semanas con otros métodos. La agencia DARPA (Defense Advanced Research Project Agency) afirmó que esta aplicación por sí sola había más que amortizado los 30 años de inversión de DARPA en IA.

Robótica: muchos cirujanos utilizan hoy en día asistentes robot en operaciones de microcirugía. HipNav (Di Gioia et al., 1996) es un sistema que utiliza técnicas de visión por computador para crear un modelo tridimensional de la anatomía interna del paciente y después utiliza un control robotizado para guiar el implante de prótesis de cadera.

Procesamiento de lenguaje y resolución de problemas: PROVERB (Littman et al., 1999) es un programa informático que resuelve crucigramas mejor que la mayoría de los humanos, utilizando restricciones en programas de relleno de palabras, una gran base de datos de crucigramas, y varias fuentes de información como diccionarios y bases de datos online, que incluyen la lista de películas y los actores que intervienen en ellas, entre otras cosas.

Clima: Comprender mejor la atmósfera, producción de energía limpia, y reducción de nuestro consumo de energía.

Salud: Analizar imágenes, chequear medicamentos, uso en rehabilitación de personas con trastornos cognitivos que dificultan la comunicación.

Probar nuevos materiales. predecir terremotos. ampliar los conocimientos cosmológicos, análisis FODA, resúmenes de libros, traducciones, etc.

Ejemplos de Aplicaciones en Neuropsicología

1) Con la IA se construyen Personas Digitales que pueden actuar como psicoterapeutas a cualquier hora del día o de la noche, con las cuales se puede generar una relación de intimidad y confianza.
2) Mediante sensores de emociones interoceptivas se pueden tratar de bajar los niveles de estrés y ansiedad, por ejemplo, en la esquizofrenia.
3) En pacientes depresivos se pueden construir modelos matemáticos que elaboren datos sobre el riesgo de cometer suicidio.
4) Utilizando técnicas de reconocimiento de voz y análisis de texto, los algoritmos de IA pueden identificar las emociones y los estados mentales de las personas basándose en su lenguaje y su tono de voz.
4) Se puede refinar el diagnóstico diferencial entre Demencia Alzheimer y Demencia Frontotemporal mediante análisis de datos imagenológicos centrados en los hipocampos (Alzheimer) e ínsulas (Demencia Frontotemporal). Asimismo para detectar asimetrías hipocámpicas en la Epilepsia del Lóbulo Temporal o en las Demencias Hipocámpicas.
5) A través del Aprendizaje Automático se desarrollan algoritmos, a partir de datos neuropsicológicos y demográficos, con el objetivo de predecir el resultado en escalas de deterioro cognitivo y colaborar en el diagnóstico neuropsicológico.
6) Con IA se pueden crear programas individuales e interactivos para utilizar en neurorrehabilitación.

Tabla 1. Usos y Aplicaciones de IA

Tipo de agente	Medidas de rendimiento	Entorno	Actuadores	Sensores
Sistema de diagnóstico médico	Pacientes sanos, reducir costes, demandas	Pacientes, hospital, personal	Visualizar preguntas, pruebas, diagnósticos, tratamientos, casos	Teclado para la entrada de síntomas, conclusiones, respuestas de pacientes
Sistema de análisis de imágenes de satélites	Categorización de imagen correcta	Conexión con el satélite en órbita	Visualizar la categorización de una escena	Matriz de pixels de colores
Robot para la selección de componentes	Porcentaje de componentes clasificados en los cubos correctos	Cinta transportadora con componentes, cubos	Brazo y mano articulados	Cámara, sensor angular
Controlador de una refinería	Maximizar la pureza, producción y seguridad	Refinería, operadores	Válvulas, bombas, calentadores, monitores	Temperatura, presión, sensores químicos
Tutor de inglés interactivo	Maximizar la puntuación de los estudiantes en los exámenes	Conjunto de estudiantes, agencia examinadora	Visualizar los ejercicios, sugerencias, correcciones	Teclado de entrada

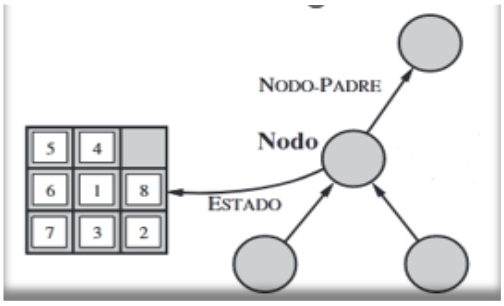
Nota: Modificada y adaptada de: Russell S, Norvig P. (1995). Artificial Intelligence: A Modern Approach. Prentice Hall. Upper Saddle River, NJ.

Modelos

Las neuronas artificiales son módulos de software, llamados nodos, y las redes neuronales artificiales son programas de software o algoritmos que, en esencia, utilizan sistemas informáticos para resolver cálculos matemáticos.

Un algoritmo sencillo y general de BÚSQUEDA-ÁRBOL puede usarse para resolver cualquier problema; las variantes específicas del algoritmo incorporan estrategias diferentes (ver Figura 2).

Figura 2. Esquema de Búsqueda-Árbol



Los nodos son estructuras de datos a partir de los cuales se construye el árbol de búsqueda, Cada uno tiene un padre, un estado y varios campos. Las flechas señalan del hijo al padre.

Nota: Modificada y adaptada de: Russell S, Norvig P. (1995). Artificial Intelligence: A Modern Approach. Prentice Hall. Upper Saddle River, NJ.

Un modelo de probabilidad temporal puede pensarse como el formado por un modelo de transición que describe la evolución y un modelo sensor que describe el proceso de observación. El reconocimiento del habla y el rastreo son dos aplicaciones importantes de los modelos de probabilidad temporal. Los modelos ocultos de Markov (HMM) (Maldonado, 2012) son especialmente aplicados al reconocimiento de formas temporales, como reconocimiento del habla, de escritura manual, de gestos, etiquetado gramatical o en bioinformática. En el reconocimiento de voz se emplea para modelar una frase completa, una palabra, un fonema o trifonema en el modelo acústico. Por ejemplo la palabra "gato" puede estar formada por dos HMM para los dos trifonemas que la componen /gat/ y /ato/ Las redes de decisión proporcionan un formalismo sencillo para expresar y resolver problemas de decisión. Son una extensión natural de las redes Bayesianas que incluyen nodos de decisión y de utilidad, además de los nodos aleatorios. Una red bayesiana es un modelo gráfico que muestra variables (que se suelen denominar nodos) en un conjunto de datos y las independencias probabilísticas o condicionales entre ellas.

En algunas ocasiones, la resolución de un problema implica encontrar más información antes de adoptar una decisión. El valor de una información se define como la mejora esperada en el valor de la utilidad comparándolo con la decisión tomada sin contar con tal información. Cuando el agente de aprendizaje es responsable de seleccionar acciones mientras aprende, debe compensar entre el valor estimado de las acciones y el potencial de aprender nueva información útil.

Machine learning / Deep learning

Machine learning (aprendizaje automático)

El aprendizaje automático o aprendizaje automatizado o aprendizaje de máquinas o aprendizaje computacional (del inglés, machine learning) es el subcampo de las ciencias de la computación y una rama de la IA cuyo objetivo es

desarrollar técnicas que permitan que las computadoras aprendan. Se dice que un agente aprende cuando su desempeño mejora con la experiencia y mediante el uso de datos; es decir, cuando la habilidad no estaba presente en su rasgo de nacimiento. En el aprendizaje de máquinas un computador observa datos, construye un modelo basado en esos datos y utiliza ese modelo a la vez como una hipótesis acerca del mundo y una pieza de software que puede resolver problemas (Flach, 2012).

En muchas ocasiones el campo de actuación del aprendizaje automático se solapa con el de la estadística inferencial, ya que las dos disciplinas se basan en el análisis de datos. Sin embargo, el aprendizaje automático incorpora las preocupaciones de la complejidad computacional de los problemas. Muchos problemas son de clase NP-hard (ver Teoría de la Complejidad), por lo que gran parte de la investigación realizada en aprendizaje automático está enfocada al diseño de soluciones factibles a esos problemas. El aprendizaje automático también está estrechamente relacionado con el reconocimiento de patrones. El aprendizaje automático puede ser visto como un intento de automatizar algunas partes del método científico mediante métodos matemáticos. Por lo tanto, es un proceso de inducción del conocimiento.

El aprendizaje automático tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis de mercado para los diferentes sectores de actividad, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica.

El aprendizaje por refuerzo continúa siendo una de las áreas más activas de investigación en aprendizaje automático. Las aplicaciones en robótica prometen ser particularmente valiosas. El interés de la robótica se centra en agentes inteligentes que manipulen el mundo físico. Los robots están equipados con sensores para percibir su entorno y efectores con los que pueden aplicar fuerzas físicas en su entorno. La mayoría de los robots son o bien manipuladores anclados en posiciones fijas o robots móviles que se pueden mover.

Deep learning (aprendizaje profundo)

No existe una única definición de aprendizaje profundo. En general se trata de una clase de algoritmos ideados para el aprendizaje automático (Schmidhuber, 2014). A partir de este punto común, diferentes publicaciones se centran en distintas características, por ejemplo:

- Usar una cascada de capas con unidades de procesamiento no lineal para extraer y transformar variables. Cada capa usa la salida de la capa anterior como entrada. Los algoritmos pueden utilizar aprendizaje supervisado o aprendizaje no supervisado, y las aplicaciones incluyen modelización de datos y reconocimiento de patrones.
- Estar basados en el aprendizaje de múltiples niveles de características o representaciones de datos. Las características de más alto nivel se derivan de las características de nivel inferior para formar una representación jerárquica.
- Aprender múltiples niveles de representación que corresponden con diferentes niveles de abstracción. Estos niveles forman una jerarquía de conceptos.

Todas estas maneras de definir el aprendizaje profundo tienen en común: múltiples capas de procesamiento no lineal; y el aprendizaje supervisado o no supervisado de representaciones de características en cada capa. Las capas forman

una jerarquía de características desde un nivel de abstracción más bajo a uno más alto.

Chat GPT

A diferencia de la mayoría de los chatbots, ChatGPT (GPT es Transformador Generativo Pre-entrenado) tiene estado, recuerda las indicaciones anteriores que se le dieron en la misma conversación, lo que permite que ChatGPT se use como un terapeuta personalizado. El ChatGPT responde a nuestras preguntas, pero en un esfuerzo por evitar que se presenten y se produzcan resultados ofensivos desde ChatGPT, las consultas se filtran a través de una API (Interfaz de Programación de Aplicaciones) de moderación y se descartan las indicaciones potencialmente racistas o sexistas. ChatGPT ha recibido críticas generalmente positivas (Kalla & Smith, 2023). Samantha Lock de *The Guardian* señaló que podía generar texto "impresionantemente detallado similar a un humano" (Lock, 2022). El escritor de tecnología Dan Gillmor usó ChatGPT en una tarea de un estudiante y descubrió que el texto generado estaba a la par con lo que entregaría un buen estudiante y opinó que "la academia tendrá algunos problemas muy serios que enfrentar" (Hern, 2022).

Debido a que ChatGPT simplemente trata de completar estadísticamente un texto, es capaz de inventar respuestas. Por ejemplo, dado que los títulos de artículos de economía incluyen más las palabras "economía" y "teoría" que cualesquiera otras, y que el más citado economista es Douglas North, ChatGPT inventa que el artículo más citado es «una teoría de la historia económica», de North. North nunca escribió artículo alguno con dicho título. Emma Bowman, de *NPR*, escribió sobre el peligro de que los estudiantes plagien a través de una herramienta de IA que puede producir textos sesgados o sin sentido con un tono autoritario: "todavía hay muchos casos en los que le haces una pregunta y te dará una respuesta que suena muy impresionante y que está totalmente equivocada" (Bowman, 2022).

Los autores del artículo de PLOS *Digital Health* afirmaron que los resultados "sugieren que los modelos de lenguaje extenso pueden tener el potencial de ayudar con la educación médica y, potencialmente, con la toma de decisiones clínicas" (Mbakwe et al., 2023). Los profesionales han enfatizado las limitaciones de ChatGPT para brindar asistencia médica. En correspondencia con *The Lancet Infectious Diseases*, tres expertos en antimicrobianos escribieron que "las mayores barreras para la implementación de ChatGPT en la práctica clínica son los déficits en la conciencia situacional, la inferencia y la consistencia. Estas deficiencias podrían poner en peligro la seguridad del paciente.” (Howard et al., 2023). *Physician's Weekly*, el 16 de mayo de 2023, también analiza el uso potencial de ChatGPT en contextos médicos (p. ej., "como asistente digital para los médicos mediante la realización de diversas funciones administrativas, como la recopilación de información de registros de pacientes o la clasificación de datos de pacientes por antecedentes familiares, síntomas, resultados de laboratorio, posibles alergias, etc."). Advirtió que la IA a veces puede proporcionar información fabricada o sesgada. Un radiólogo advirtió: "Hemos visto en nuestra experiencia que ChatGPT a veces inventa artículos de revistas falsos o consorcios de salud para respaldar sus afirmaciones". Sin embargo, como mencionó el Dr. Stephen Hughes para *The Conversation*, ChatGPT es capaz de aprender a corregir sus errores pasados. También señaló la "mojigatería" de IA con respecto a los temas de salud sexual (Hughes, 2013).

ChatGPT puede escribir secciones de introducción y resumen de artículos científicos. Varios artículos ya han incluido a ChatGPT como Coautor. Las

revistas científicas presentan diferentes reacciones a ChatGPT, algunas requieren que los autores divulguen el uso de herramientas de generación de texto y prohíben incluir un gran modelo de lenguaje (LLM) como ChatGPT como coautor. Por ejemplo *Nature y JAMA Network*. *Science* prohibió por completo el uso de texto generado por LLM en todas sus revistas.
La investigación y el desarrollo de IA deben centrarse en hacer que los sistemas de última generación y alto rendimiento de hoy en día sean más precisos, seguros, interpretables, transparentes, robustos, coordinados, confiables y leales.

Planteo de Daniel Innerarity (Innerarity, 2022)
"Mi hipótesis es que no va a haber una sustitución. Y también vale para la democracia: no la vamos a abandonar en manos de máquinas, sencillamente porque hacen cosas muy bien, pero no precisamente la política. Es una actividad hecha en medio de una gran ambigüedad. Y las máquinas funcionan bien allá donde las cosas se pueden medir y computarizar, pero no donde hay contextos, ambigüedad, e incertidumbre. En vez de pensar en la emulación de los humanos por parte de las máquinas o en temer que las máquinas nos sustituyan, lo que debemos pensar es qué cosas hacemos los humanos bien y qué cosas hacen ellas bien, y diseñar ecosistemas que saquen los mejores rendimientos de ambos. Hay que hacer una renovación de conceptos. Y es aquí donde los filósofos tenemos un papel que desempeñar. Por ejemplo, ahora se dice mucho: ¿de quién son los datos? A mí me parece que el concepto de propiedad es un concepto muy inadecuado para referirse a los datos, que más que un bien público son un bien común, algo que no se debe apropiar. El punto de inflexión se produce a partir del momento en que los humanos diseñamos máquinas que tienen "vida" propia, que no son meramente instrumentales. Cuando producimos IA entramos en un terreno desconocido. El reparto del mundo que habíamos hecho, según el cual los humanos somos sujetos de derechos y obligaciones y diseñamos una tecnología meramente pasiva que está sometida a nuestro control, es una idea que ya no funciona. Hay una ruptura. Lo comparo con el momento en que Darwin acaba con la idea del Dios diseñador de la creación: nos obligó a pensar de una manera diferente. Creo que cuando se habla de controlar la tecnología se está en una actitud predarwiniana. Evidentemente, los algoritmos, las máquinas, los robots deben tener un diseño humano ¡tenemos que debatir sobre eso! Pero la idea de control, como la que hemos tenido clásicamente para tecnologías triviales, me parece que es completamente inadecuada. Lo que tenemos que hacer es establecer una especie de diálogo en el que humanos y máquinas negociemos escenarios aceptables, pensando en la igualdad, el impacto sobre el medio ambiente y los valores democráticos. La idea de controlar no va a funcionar cuando hablamos de máquinas que aprenden. Es un problema que hoy por hoy no tiene solución fácil por varias razones. Primero, por la complejidad del asunto. Segundo, porque el algoritmo tiene vida propia y, por tanto, es opaco para su diseñador. Y en tercer lugar, porque la idea de auditar los algoritmos, de que haya transparencia, la entendemos como aquel que firma un documento. Creo que tenemos que ir a sistemas públicos que nos permitan establecer una confianza con las máquinas. Esa idea de que estos artefactos son una caja negra, como si los humanos no fuéramos también cajas negras. Los algoritmos para decidir las políticas penitenciarias generan problemas, pero a veces se da a entender que un algoritmo tiene sesgos y los humanos no. ¿Las cabezas de los jueces no son también cajas negras? Somos más exigentes en relación con la objetividad de la tecnología. De ella esperamos objetividad y en el momento en que nos falla nos resulta mucho

más intolerable que con un humano. Deberíamos llegar a una idea de diálogo con la máquina más que de control. Cada vez conducimos coches sobre los que tenemos menos control, pero son más seguros. El resultado de la tecnología del coche es que pierdo el control absoluto, pero me ofrece a cambio una supervisión general de los procesos para que no me mate. Como cuando los Estados ceden soberanía en Europa. Si compartimos soberanía política, ¿por qué no compartir soberanía tecnológica? Una de las cosas más importantes para enfocar bien este asunto es pensar menos en contraposiciones. La tecnología tiene muchísima más humanidad, si se me permite la provocación, de la que los éticos suelen reclamar. Frente a quienes conciben la tecnología como algo inmaterial, virtual e intangible, el ciberespacio es en realidad mucho más material, con un impacto medioambiental brutal. Y esa parte material muchas veces está fuera de nuestro ámbito de atención. Y tiene que haber humanos en el proceso: detrás de procesos aparentemente automatizados hay personas interviniendo sin que lo sepamos. La promesa de la tecnología de liberarnos de los trabajos mecánicos no se ha cumplido. La otra paradoja puede ser que eso esté indicando que las máquinas no nos van a sustituir plenamente. La expectativa o el miedo de que nos sustituyan es completamente irrealista. Y eso tiene que ver con una distinción importantísima entre tarea y trabajo: las máquinas hacen tareas, pero no propiamente trabajos. Y en esa transición es posible que vayamos a tener un nuevo tipo de conflictos sociales. En vez de pensar en términos de sustitución, tenemos que pensar qué tareas pueden y deben ser realizadas por un robot y qué aspectos de lo humano son irrealizables por un robot. No tanto si esto es bueno o malo. La IA sirve para resolver cierto tipo de problemas políticos, pero no otros. No tengamos tanto miedo a que las máquinas se hagan cargo de todas las tareas del gobierno y, en cambio, facilitemos aquellas tareas de gobernanza que puedan hacer mejor que nosotros. Lo que más me preocupa es la falta de reflexividad. Que el entorno algorítmico nos acostumbre a que determinadas cosas se decidan de un modo sobre el cual no hemos discutido lo suficiente. ¿Vamos a ir a entornos algorítmicos, automatizados? Perfecto, pero sepamos que detrás hay algún tipo de autoridad. Veamos de qué autoridad se trata, y hagamos lo que siempre hemos hecho los humanos con toda autoridad: someterla a revisión. Estamos en un momento de la historia de la humanidad en el que todavía se puede negociar, disentir y reflexionar sobre estas tecnologías. Es muy posible que dentro de no muchos años estas tecnologías se hayan solidificado en instituciones, en procesos, en algoritmos sobre los que sea mucho más difícil discutir. Por eso es importante este trabajo de reflexión filosófica. Hay mucha gente planteándose la regulación tecnológica, pero no vamos a regular bien una tecnología que no comprendemos porque fallan los conceptos. La reflexión tecnológica y la reflexión filosófica deben ir de la mano como soportes de cualquier actividad regulatoria. Dejar espacios indeterminados y abiertos para la libre dimensión humana e imprevisible me parece una cuestión fundamental. Los humanos somos seres imprevisibles: buena parte de nuestra libertad se la debemos a eso y las máquinas deben reflejarlo". Son reflexiones de Daniel Innenarity: 1) la máquina no puede cambiar la regla de aquello que tiene programado, 2) la máquina desconoce el contexto, y 3) no asigna significados, por lo que 4) no se puede insertar a) en el mundo de los valores ni b) en la escala política.

Crítica de Nick Bostrom (Bostrom, 2016)

La visión de Bostrom fue elogiada por "vacas sagradas" de los negocios tecnológicos, como Bill Gates o Elon Musk –"la inteligencia artificial es

potencialmente más peligrosa que la bomba atómica"– o gigantes de la ciencia, como el mismísimo Stephen Hawking, y en el 2015 *Foreign Policy* lo incluyó en su lista de los cien pensadores más influyentes del planeta (Frankel, 2009).

La advertencia de Bostrom se fundamenta en la previsión de que, a largo plazo, o tal vez antes, la IA superará a la humana en todos sus registros, gracias a los algoritmos que permitirán a las máquinas aprender y mejorar. Esa superinteligencia nos aventajará de tal modo que podrán desempeñar con extrema facilidad y rapidez tareas que para nosotros suponen un gran esfuerzo o son imposibles. Un gran salto. Pero esa tecnología puede convertirse en algo carente de valores en relación con nuestras vidas. "Creo que el desarrollo de la IA es una especie de portal, por el cual pasan todos los caminos posibles. Hay, al menos, dos problemas que resolver. 1) es un problema de ciencia informática teórica, que es lo que lo que llamaríamos técnicamente controles escalables de la alineación de la IA. Es decir: si un día somos capaces de fabricar una máquina mucho más inteligente que nosotros, tenemos que conseguir que haga únicamente las cosas para las que fue diseñada y que no se desvíe de ese objetivo, por mucho que su inteligencia crezca. Si resolvemos eso, se da el segundo gran problema 2) es el de la gobernanza, es decir, cómo accedemos a esta tecnología tan poderosa y cómo la utilizamos. Tenemos que decidir según qué valores se usa y para qué objetivos, tenemos que asegurar que se utilice con fines positivos para el bien de todos y que los beneficios no sean monopolizados por pequeños grupos. Históricamente, hemos usado las tecnologías para muchas finalidades, no siempre beneficiosas. Si damos respuesta a estos desafíos, tenemos por delante un gran futuro. Un sistema superinteligente estaría diseñado per se, para optimizar un proceso y lograr una meta determinada independientemente de otras consideraciones, y no tendría en cuenta los intereses de los humanos, a no ser que se lo programara para respetarnos. La idea de que la gente halle sentido a su vida y su identidad en el trabajo, por ejemplo, podría desaparecer si las máquinas hacen las cosas mejor que nosotros. Queremos que el desarrollo de la tecnología sea más lento para tener un año en que instalar los mecanismos de control y seguridad, comprobar que todo funciona y lentamente incrementar su operatividad como con el lanzamiento de un nuevo fármaco. Tendremos que pensar qué haremos que tenga valor, qué actividades encontraremos valiosas, no porque tengan un interés instrumental, sino porque hacer esta actividad o tener esa experiencia tenga un valor intrínseco. Los científicos no tienen un conocimiento exacto de qué puede suceder en el futuro con la IA: es un terreno desconocido".

En suma, la IA se integra con algoritmos con capacidad de aprendizaje automático y toma de decisiones propia, que aprenden por sí mismos de la información a la cual acceden pudiendo generar una nueva respuesta ante una misma situación, pero es una mente que requiere un contralor *humano.*

Agrega Chema Alonso, (Alonso, 2024): 1) IA es el nombre de algoritmos que pueden resolver problemas cognitivos sin ser seres que puedan existir o tener conciencia de algo, 2) podrían llegar a elaborar "pensamientos" homicidas, disparar armas nucleares o participar de estafas a través de textos apócrifos atribuidos a un autor famoso o creación de personas sintéticas digitales con clonación de la voz de personas reales y otras estratagemas perniciosas.

2. Conclusiones

a. Hay que refinar la validación de evidencias que corroboren la exactitud de los datos aportados por IA, por cuanto los algoritmos no conocen, meramente informan sin poder juzgar si eso que informan es cierto o falso.
b. La IA hace estimaciones sobre lo ya conocido, pero no arroja datos sobre conveniencia o inconveniencia de aquello que podría llegar como innovación.
Por último, hay cierta disensión entre los *científicos apocalípticos* que temen que el avance de la IA pueda poner en peligro los valores básicos de la humanidad, en contraposición con los *científicos integrados*, que tienen una actitud optimista basada en la posibilidad de control y aprovechamiento útil de la IA.

Bibliografía

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A.M., et al. (2023) Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology* 19(10): e1011465. https://doi.org/10.1371/journal.pcbi.1011465

Alonso, J.M. (2023) Ciberseguridad y Hacking en un mundo de Inteligencia Artificial, Robots y Humanos. https://youtu.be/gsnFXlLYt9w

Baars, B. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford University Press. New York.

Bostrom, N. (2016). *Superinteligencia. Caminos, peligros, estrategias*. Traducción de Marcos Alonso Fernández. Editorial TEELL Madrid.

Bowman, E. (2022). *A new AI chatbot might do your homework for you. But it's still not an A+ student*. NPR.

Craik, K. (1943). *The Nature of Explanation*. Cambridge: Cambridge University Press. ISBN 978-0521094450.

Cross, S.; Walker, E. (1994). Zweben, M y Fox, M. (eds.). *Intelligent Scheduling*. University of Michigan: Morgan Kaufmann. pp. 711–729. ISBN 1-55860-260-7.

Di Gioia, A.; Jaramaz, B.; Blackwell, M.; Simon, D.; et al. (1998). Image Guided Navigation System to Measure Intraoperatively Acetabular Implant Alignment. *Clinical Orthopaedics and Related Research* 355: 8-22.

Esteban, F., Galad, J., Langa, J., Portillo, J. & Soler-Toscano, F. (2018) Informational structures: A dynamical system approach for integrated information. *PLoS Comput Biol* 14(9): e1006154. https://doi.org/10.1371/journal.pcbi.1006154

Faccio, E.; & Goldar, J. (1978).Voluntad y Lóbulo Frontal. *Neuropsiquiatría*; 9: 63-76.

Flach, P. (2012) *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press. ISBN 978-1-107-42222-3.

Frank, M. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*. 2 (8): 451–452. doi:10.1038/s44159-023-00211-x. ISSN 2731-0574. S2CID 259713140. Retrieved 2 July 2023.

Frankel, R. (2009). The FP Top 100 Global Thinkers. *Foreign Policy*. Nov. 25.

Goodman, D. & Keene, R. (1997). Man versus machine: Kasparov versus Deep Blue. *H3 Publications*. Valencia, CA.

Heckerman, D. (1991). *Probabilistic Similarity Networks*. MIT Press, Cambridge, MA.

Heidegger, M. (2010). *¿Qué significa pensar?* Trotta. Madrid.

Hern, A. (2022). AI bot ChatGPT stuns academics with essay-writing skills and usability. *The Guardian*.

Hinton, G. & Salakhutdinov, R. (2006) Reducing the dimensionality of data with neural networks. *Science*, Vol. 313. no. 5786, pp. 504 - 507, 28 July 2006.

Hinton, G., & Salakhutdinov, R. (2012). A better way to pretrain deep Boltzmann machines. *Advances in Neural Information Processing Systems*, 25.

Howard, A.; Hope, W. & Gerada, A. (2023). ChatGPT and antimicrobial advice: the end of the consulting infection? *The Lancet Infectious Diseases*; 23(4): 405-406.

Hughes, S. (2023). How good is ChatGPT at diagnosing disease? A doctor puts it through its paces. *The Conversation*. April 27.

Innerarity, D. (2022). Los algoritmos son conservadores y los humanos, imprevisibles. *La Nación*, 30 de julio. Buenos Aires.

Innerarity, D. (2024) No es tan inteligente. *La Vanguardia*, 20 de enero. Barcelona

Jonsson, A., Morris, P., Muscettola, N., Rajan, K. & Smith, B. (2000). Planning in Interplanetary Space: Theory and Practice. American Association for Artificial Intelligence. Proceedings of the 5th. International Conference on Artificial Intelligence Planning Systems. Breckenridge, CO. April 14-17

Jung, C.G. (1957 a 1990). *Collected works*. Read, H.; Fordham, M.; Adler, G. (eds), Princeton University Press, U.S.A.

Kalita, P., Langa, J & Soler-Toscano, F. (2019) Informational Structures and Informational Fields as a Prototype for the Description of Postulates of the Integrated Information Theory. *Entropy*, 21, 493. https://doi.org/10.3390/e21050493

Kalla, D. & Smith, N. (2023). Study and Analysis of Chat GPT and its Impact on Different Fields of Study (March 1, 2023). *International Journal of Innovative Science and Research Technology* 8(3). SSRN: https://ssrn.com/abstract=4402499.

Koch, C., Massimini, M., Boly, M. et al. Neural correlates of consciousness: progress and problems. *Nat Rev Neurosci* 17, 307–321 (2016). https://doi.org/10.1038/nrn.2016.22

Littman, M., Keim, G. & Shazeer, N. (1999). Solving Crosswords with PROVERB. AAAI/IAAI. *Computer Science*.

Lock, S. (2022). What is AI chatbot phenomenon ChatGPT and could it replace humans?. *The Guardian* 5 Dec.

Maldonado, L. (2012). Los modelos ocultos de Markov. *Telos*; 14(3): 433-438.s modelos ocultos de M MOM

Mbakwe, A., Lourentzou, I. Celi, L., Mechanic, O. & Dagan, A. (2023). ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS Digit Health* 2(2): e0000205. https://doi.org/10.1371/journal.pdig.0000205

Nolan, L. (2015). En Zalta, E.N. (ed), *Descartes' Ontological Argument* (Fall 2015 edición). The Stanford Encyclopedia of Philosophy.

Pomerleau, D. (1988). ALVINN: an autonomous land vehicle in a neural network. Computer Science Department Carnegie Mellon University Pittsburgh, PA 1521. https://proceedings.neurips.cc

Proust, M. (2007). *A la busca del tiempo perdido* (Mauro Armiño, trad.) (7ª edición). Valdemar. Madrid.

Russell, S & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall. Upper Saddle River, NJ.

Schmidhuber, J. (2014). *Deep Learning in Neural Networks: An Overview*. http://arxiv.org/abs/1404.7828.

Tononi, G. (2004) An information integration theory of consciousness. BMC *Neurosci*. Nov 2;5:42. doi: 10.1186/1471-2202-5-42.

Tononi, G. & Koch, C. (2015). Consciousness: here, there and everywhere? Phil. Trans. R. Soc. B370201401672014016 7 http://doi.org/10.1098/rstb.2014.0167

Tononi, G., Boly, M., Massimini, M. et al. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci* 17, 450–461 https://doi.org/10.1038/nrn.2016.44

Ure, J. (1981). *Conciencia de Ser, Agón.* Buenos Aires.

Ure, J., Videla, H. & Ollari, J. (2009). Neuroanatomy of Consciousness. *Sci Topics* Sept 2009, Elsevier Holland, Amsterdam.

Young, G. & Pigott, S. (1999). Neurobiological basis of consciousness. *Arch Neurol*. Feb;56(2):153-7. doi: 10.1001/archneur.56.2.153