¿Debería tener conciencia una IA? Aproximaciones a la conciencia humana y a la conciencia artificial

Óscar Trabazos Clavería · Ensayo, sin revisión por pares · Versión 2.0, 2 de noviembre de 2026 (primera versión, 2023)

Resumen

¿Podría una máquina programada por personas tener experiencias subjetivas, como ver un color, sentir dolor o disfrutar de una melodía? Este ensayo aborda esa pregunta en cuatro partes. La primera recorre qué entendemos por conciencia fenoménica y las respuestas clásicas al problema mente-cuerpo: dualismo, materialismo en sus distintas variantes e idealismo. La segunda estudia el cerebro como procesador de información, desde los inicios de la ciencia cognitiva hasta los modelos de memoria y el reto de las metáforas. La tercera aplica todo ello a la inteligencia artificial: la IA fuerte, el test de Turing, la habitación china, las teorías científicas actuales de la conciencia y los modelos de lenguaje, con sus alucinaciones y su parecido con las del cerebro humano. La cuarta afronta las preguntas éticas y filosóficas de una mente artificial, la hipótesis del universo simulado y la relación entre entropía, longevidad y conciencia digital.

La conclusión del autor es que los sistemas actuales no dan ninguna prueba de conciencia: son sistemas de enorme capacidad que responden con la salida más probable. No podemos crear en forma no biológica lo que todavía no comprendemos, y si algún día una inteligencia artificial general llega a algo parecido, será probablemente una conciencia sintética, distinta de la humana. Las preguntas éticas, en cambio, ya son actuales y conviene hacérselas antes de que la tecnología las imponga.

Palabras clave: conciencia fenoménica, inteligencia artificial, problema mente-cuerpo, funcionalismo, ciencia cognitiva, modelos de lenguaje, alucinaciones, ética de la IA, conciencia digital.

Nota sobre las fuentes y sobre esta edición

Este ensayo nació de las investigaciones, notas, ideas y proyectos del autor y de su equipo sobre cómo debería diseñarse una IA. Su primera versión, escrita en febrero de 2023, tomó como guía el recorrido teórico de la tesis de magíster de Cristóbal Fuentes Barassi, *Conciencia e Inteligencia Artificial: consideraciones críticas sobre la plausibilidad de que una máquina programada posea conciencia fenoménica* (Universidad de Chile, 2011). Esa

tesis es la fuente base de las secciones 2 a 6 y se cita como tal. Partes de aquella primera versión se redactaron con ayuda de IA. Entre 2023 y 2024 el autor la divulgó y la amplió en una serie de su blog.

Esta versión 2.0 reúne todo ese trabajo en un solo texto y lo pone al día:

- **Integra la serie del blog** (2023-2024): la historia de la conciencia artificial, los dilemas éticos y filosóficos, las alucinaciones de la IA y del cerebro, el universo simulado, la entropía y la conciencia digital.
- **Corrige errores de la primera versión**: fechas (La Mettrie es del siglo XVIII; el Centro de Estudios Cognitivos de Harvard se fundó en 1960), la estructura del modelo de Baddeley, la distinción entre el problema difícil y el argumento del zombi, lo que mide el test de Turing y la diferencia entre IA fuerte e IA estrecha. Tres modelos de memoria que aparecían con nombres inexactos se citan ahora con su nombre y su autoría reales.
- **Añade lo que se ha investigado desde 2023**: las teorías científicas de la conciencia y los primeros intentos de evaluar con ellas los sistemas de IA actuales.
- **Añade referencias completas** para cada autor citado.

Las conclusiones son del autor y salen de lo que ha escrito sobre el tema; se presentan como posiciones argumentadas, no como resultados demostrados.

Parte I. La conciencia y la mente

1. Introducción

Con este trabajo, elaborado a partir de las investigaciones, notas, ideas y proyectos que hemos realizado mi equipo y yo, quiero mostrar la complejidad que hay detrás de crear una IA y de adaptarla a un proceso de trabajo. Adoptar inteligencia artificial va mucho más allá de conectarse a una API que resuelva un problema concreto. Este trabajo trata de la necesidad de entender cómo debería pensar realmente una IA a partir de entender cómo funcionan nuestro cerebro y nuestra conciencia.

Sin entender las teorías que durante miles de años han intentado descubrir el funcionamiento de nuestro cerebro, es imposible decir que estamos trabajando en un proyecto de IA. Ni siquiera es posible pensar en un proyecto de IA que tenga utilidad real y, sobre todo, que sea ético.

El trabajo es largo. Está pensado para quienes, más allá del aspecto computacional, necesitan entender por qué deberían utilizar tecnologías basadas en IA y qué enfoque deberían tener sus diseños. Avanza de lo más antiguo a lo más actual: primero, qué es la conciencia y cómo la ha pensado la filosofía; después, cómo la estudia la ciencia cognitiva; luego, qué significa todo ello para la IA de hoy, y por último, qué preguntas éticas y filosóficas nos abre.

Quiero subrayar desde el principio que con este trabajo no pretendo parecer espiritual. Mi intención es recorrer cómo los seres humanos han intentado entender la naturaleza de nuestra conciencia durante cientos de años, porque es un ejercicio necesario si queremos plantearnos la posibilidad de que una inteligencia artificial la tenga.

2. La conciencia fenoménica

¿Sería posible crear una máquina programada que tuviera conciencia fenoménica? ¿Podría un artefacto creado y programado por humanos, basado en algoritmos, tener experiencias subjetivas como ver un color, sentir dolor, percibir un aroma, disfrutar de una melodía o sentir estrés?

2.1 El punto de partida: algoritmos y máquinas

Desde los trabajos fundacionales de Alan Turing en 1948 y 1950, la inteligencia artificial ha buscado crear máquinas capaces de emular y competir con las capacidades cognitivas humanas, con el concepto de algoritmo como piedra angular. Un algoritmo es una serie ordenada y finita de pasos que lleva a un resultado, y es lo que está detrás de la llamada máquina de Turing: una máquina teórica que tiene distintos estados y que, dada una entrada, pasa al estado siguiente. La idea de esta aproximación clásica es que, si se dota a una máquina del algoritmo adecuado, debería poder realizar cualquier tarea que pueda hacer un ser humano.

De ese optimismo nació el interés por crear máquinas que imitaran otras características del pensamiento humano, y en particular la conciencia. Es lo que se conoce como conciencia artificial, también llamada conciencia de máquina o conciencia sintética. Sus motivos se pueden agrupar en cuatro:

1. Diseñar modelos de conciencia para identificar sus rasgos determinantes, como parte del trabajo de la ciencia cognitiva.
2. Usar las implementaciones de esos modelos como herramientas para poner a prueba y entender la conciencia humana.
3. Determinar cómo es la relación entre la conciencia y el soporte físico en el que se produce.
4. Comprender que la conciencia fenoménica es fundamental en algunas de las tareas que consideramos inteligentes.

La literatura y el cine han explorado durante décadas la idea de máquinas capaces de tareas complejas: HAL 9000 en *2001: una odisea del espacio*, los robots de *Yo, robot* de Isaac Asimov o las máquinas conscientes de *A.I. Inteligencia Artificial* de Steven Spielberg. En la vida real, asistentes de voz como Siri o Alexa aprenden de sus usuarios y responden a sus órdenes, sin que nadie les atribuya por ello una vida interior.

Aunque la disciplina ha creado máquinas con capacidades cognitivas similares o superiores a las humanas en tareas concretas, sigue abierta la pregunta de si pueden tener experiencias subjetivas, como saborear un té o disfrutar de una música. Y con ella, otras: cómo saber si una máquina las tiene, qué relación hay entre el soporte físico y su aparición, y si las máquinas piensan o solo calculan.

Deep Blue, el ordenador que venció a Kasparov en 1997, es un buen ejemplo. Juega al ajedrez mejor que casi cualquier humano, pero nada en su diseño le permite vivir la experiencia de jugar: ni la tensión de la partida, ni la presión de la competición, ni la inquietud ante el movimiento del rival. Sigue un procedimiento; no le pasa nada.

2.2 Por qué la conducta no basta

Los defensores de la llamada IA fuerte sostienen que basta con observar el comportamiento externo de una máquina para atribuirle inteligencia. Es lo que se suele asociar al test de Turing, aunque Turing lo propuso para la inteligencia, no para la conciencia (sección 6.2). Y es, en el fondo, lo que hacemos a diario con otras personas: suponemos que son conscientes porque se comportan como nosotros y se nos parecen.

El problema es que la conciencia fenoménica solo la experimenta, en primera persona, quien la tiene. Una prueba objetiva de conducta puede decirnos qué hace una máquina, no si siente algo al hacerlo. Por eso la afirmación conductista de que la conciencia se puede detectar con pruebas externas no resuelve la cuestión. El problema no es exclusivo de la inteligencia artificial: es el viejo problema mente-cuerpo de la filosofía, que ha dado lugar al dualismo, al monismo y a muchas posiciones intermedias, y sobre el que sigue sin haber consenso.

2.3 Qué entendemos por conciencia

Bernard Baars, psicólogo y neurocientífico cognitivo y autor de la teoría del espacio de trabajo global (sección 6.4), señala la confusión que produce plantear la conciencia en términos de subjetividad, de «qué se siente al ser alguien»: se termina en una de las tres posiciones clásicas, el mentalismo, el fisicalismo o el dualismo. Para avanzar, propone formular las preguntas sobre la conciencia de manera que se puedan responder, con un lenguaje y un enfoque concretos.

La formulación más citada de esa subjetividad es de Thomas Nagel (1974): un organismo tiene estados conscientes si hay algo que se siente al ser ese organismo. Su ejemplo, «¿qué se siente al ser un murciélago?», muestra que podemos saberlo todo sobre la ecolocalización de un murciélago sin saber cómo es, para él, percibir así el mundo.

Ned Block (1995) advierte de que «conciencia» es un concepto híbrido: una sola palabra para fenómenos distintos, como la atención, darse cuenta de algo externo o interno, la identidad personal o la experiencia de colores y dolores. Distingue sobre todo dos: la

conciencia fenoménica, que es la experiencia, y la conciencia de acceso, que es la información disponible para razonar, hablar y guiar la conducta. En las personas suelen ir juntas, pero son conceptos distintos.

David Chalmers (1995) dio nombre a la dificultad central: el problema difícil de la conciencia. Los «problemas fáciles», como discriminar estímulos, integrar información o informar de los propios estados, se pueden explicar describiendo funciones y mecanismos. El difícil es explicar por qué esos procesos van acompañados de experiencia. Para ilustrarlo, Chalmers (1996) usa un argumento distinto, el del zombi filosófico: un ser físicamente idéntico a una persona, que se comporta exactamente igual, pero que no siente nada. Si ese ser es siquiera concebible, la conciencia no se reduce a lo físico ni a la conducta. Chalmers ha llegado a considerar en serio el panpsiquismo, la idea de que la conciencia es una propiedad fundamental de la naturaleza.

2.4 El alcance de este ensayo

En este texto analizo críticamente la posibilidad de que una máquina programada con algoritmos posea conciencia fenoménica, es decir, la experiencia subjetiva del mundo que vive un ser al relacionarse con su entorno y con su vida interior. Hay otras nociones de conciencia, como la autoconciencia o la conciencia de acceso, pero el análisis se centra en esta, porque es la que hace difícil la pregunta.

Para hacerlo hay que discutir los fundamentos filosóficos sobre los que trabajan disciplinas como la IA y la psicología, y tener presentes los interrogantes éticos y legales que abre. La tesis que sirve de base a este recorrido (Fuentes Barassi, 2011) sostiene que dotar a una máquina de conciencia fenoménica sería posible y necesario, porque esa capacidad es fundamental para ciertas tareas inteligentes y no un mero epifenómeno; pero duda de que el modelado mediante algoritmos sea el camino, porque los algoritmos no explican cómo ni por qué se produce la experiencia. Mi posición, que desarrollo en las conclusiones, es más prudente con lo primero y coincide con lo segundo.

Una alternativa al algoritmo clásico que se ha propuesto es construir redes neuronales artificiales que simulen las redes neuronales biológicas. Su dificultad es doble: no sabemos qué rasgos de la red biológica son los que importan para la conciencia, y no tenemos forma de evaluar si una red artificial la tiene.

2.5 Primeras conclusiones sobre la conciencia fenoménica

La conciencia fenoménica es la experiencia de «qué se siente»: el sabor de la comida, un color brillante, la sensación de estar en el mundo y de tener una experiencia propia de los objetos, los hechos y los pensamientos. No se puede medir desde fuera: es única para cada individuo, y es lo que nos hace conscientes de nosotros mismos y de lo que nos rodea. Por eso se la considera el aspecto más misterioso de la mente.

Se distingue de la conciencia de acceso, que es la capacidad de disponer de la información para usarla en decisiones, y de la autoconciencia, que es la capacidad de saberse un ser separado e individual.

Sobre la IA hay dos posturas de fondo. Unos creen que es posible programar una máquina para que tenga conciencia fenoménica, y que para ello la máquina tendría que poder sentir y experimentar el mundo de forma parecida a como lo hacemos nosotros. Otros sostienen que la conciencia es una propiedad emergente de la complejidad de la materia y de la organización biológica del cerebro, y que requiere un sustrato biológico que ningún código puede sustituir. No hay consenso, y entender por qué exige mirar el problema mente-cuerpo.

3. El problema mente-cuerpo

El problema mente-cuerpo es uno de los temas centrales de la filosofía de la mente: cómo se relacionan la mente (pensamientos, sentimientos, conciencia) y el cuerpo (cerebro, sistema nervioso). Surge porque parece haber una diferencia de fondo entre ambos. La mente es consciente y tiene experiencias subjetivas; el cuerpo es un objeto físico que se puede estudiar con las ciencias naturales. ¿Cómo interactúan dos cosas tan distintas?

3.1 Cuatro ejemplos

Para explicarme mejor, permitidme usar algunos ejemplos.

- **Una película de terror.** La imagen y el sonido llegan a los ojos y a los oídos y el cerebro los procesa. Pero además tienes una experiencia subjetiva de miedo. ¿Cómo da lugar la actividad eléctrica y química de tus neuronas a ese miedo que sientes?
- **La mano en la estufa.** Una persona se quema la mano. El contacto con la estufa y el daño en los tejidos son hechos físicos; el dolor que siente solo puede experimentarlo ella. ¿Cómo se traduce la actividad de las neuronas de la mano y del cerebro en esa experiencia de dolor?
- **Una manzana roja.** Los ojos y el cerebro procesan la información visual y la persona reconoce que ve una manzana roja. Pero además hay algo que se siente al ver ese rojo.
- **Una música.** Los oídos y el cerebro procesan el sonido y la persona reconoce la canción. Pero además la disfruta, o la sufre.

En los cuatro casos, la ciencia puede describir el procesamiento con un detalle cada vez mayor. Lo que no consigue explicar es el paso del procesamiento a la experiencia. ¿Es la mente algo separado del cuerpo, o una propiedad que emerge del cerebro? Las grandes respuestas son tres: el dualismo, el materialismo y el idealismo.

3.2 Dualismo

El dualismo defiende que en el universo hay dos tipos de sustancia: la materia y la mente o el espíritu. La mente no sería material ni se podría reducir a procesos físicos, químicos o biológicos del cerebro; sería una entidad independiente, que podría existir sin el cuerpo, aunque interactúe con él.

Su formulación clásica es la de René Descartes en el siglo XVII: la mente es una *res cogitans*, una sustancia pensante, y el cuerpo es una *res extensa*, una sustancia extensa que se puede medir y estudiar. Para el dolor de la mano quemada, el dualista dirá que no se explica solo en términos físicos: requiere una explicación mental adicional.

El dualismo responde al problema mente-cuerpo, pero no lo resuelve: la relación entre las dos sustancias sigue siendo un misterio. Ha recibido tres críticas principales. No explica cómo algo inmaterial puede mover algo material. Ofrece una visión fragmentada del ser humano. Y parece chocar con la física, que no deja hueco para causas no físicas.

Para la IA, el dualismo plantea un reto de fondo: si la mente no es solo un producto del cerebro, ¿cómo replicarla en una máquina? Muchos dualistas sostienen que la IA nunca alcanzará una verdadera conciencia, porque esta va más allá de la función e implica una dimensión subjetiva: la máquina podrá simular la inteligencia, pero no saber qué es ser consciente. Otros creen que el dualismo es demasiado limitado para este debate y proponen alternativas como el funcionalismo o el conexionismo, que explican la mente por procesos cognitivos y neuronales.

3.3 Materialismo

El materialismo sostiene lo contrario: todo lo que existe es materia y energía, y la mente y la conciencia son el resultado de procesos materiales y biológicos del cerebro. La conciencia no es una entidad aparte, sino una propiedad o una función del cerebro que emerge de su funcionamiento. En su versión más radical, el monismo materialista, la mente es un epifenómeno de la actividad neuronal, y no existe un «dolor» mental independiente de esa actividad.

Para la IA, la consecuencia es directa: si entendemos lo bastante bien el cerebro, podríamos construir una máquina que reprodujera sus procesos y las propiedades que emergen de ellos, sin postular nada inmaterial. Y la conciencia se podría estudiar con la ciencia empírica, en particular la neurociencia y la psicología cognitiva.

El materialismo filosófico se remonta a la antigua Grecia, donde Demócrito y Epicuro sostuvieron que todo está hecho de átomos y que mente y cuerpo son inseparables. En el siglo XVIII, el médico y filósofo francés Julien Offray de La Mettrie defendió en *El hombre máquina* (1747) que el cuerpo humano es una máquina compleja y la mente una función del

cerebro. Después lo defendieron, con matices muy distintos, Ludwig Feuerbach, Karl Marx y Friedrich Engels, y en la filosofía de la mente actual Daniel Dennett y Patricia Churchland.

Hay que recordar que la filosofía de la mente no es una ciencia empírica: reflexiona sobre los conceptos y supuestos con que la ciencia estudia la mente. Pero esas posiciones condicionan cómo se investiga, y la neurociencia y la ciencia cognitiva han generado a su vez sus propios debates. Del tronco materialista salen varias corrientes que conviene distinguir.

3.3.1 Conductismo

El conductismo se desarrolló en Estados Unidos en la primera mitad del siglo XX. Estudia solo la conducta observable y medible y evita cualquier referencia a procesos mentales internos. El comportamiento se explica por estímulos y respuestas, y las recompensas y los castigos son los mecanismos que lo moldean.

Su fundador, John B. Watson, quería una psicología objetiva y experimental, sin introspección: el ser humano como una máquina cuyo comportamiento resulta del entorno y de las respuestas aprendidas. B. F. Skinner desarrolló el condicionamiento operante, según el cual cualquier conducta, incluso la más compleja, se explica por refuerzos y castigos.

El conductismo dominó la psicología durante décadas, pero hoy se considera demasiado simple para explicar la complejidad humana. Para nuestra pregunta, deja fuera justo lo que se pregunta: una máquina que se comporta como si sintiera sería, a efectos conductistas, igual que una que siente.

3.3.2 Funcionalismo

El funcionalismo entiende los estados mentales por su función: por lo que entra, lo que sale y el procesamiento que hay entre ambos. Un dolor de cabeza sería un estado causado por ciertos estímulos que produce ciertas conductas, como tomar un analgésico o evitar la luz. Lo que define el estado no es el material del que está hecho, sino el papel que cumple. Hilary Putnam (1967) lo formuló en filosofía de la mente con la idea de que un mismo estado mental puede realizarse en soportes físicos distintos.

Frente al conductismo, el funcionalismo reconoce los procesos mentales internos como necesarios para explicar la conducta. Ha sido adoptado por la psicología cognitiva, la neurociencia cognitiva y la inteligencia artificial, que busca sistemas capaces de percibir, razonar y decidir sin replicar la complejidad biológica del cerebro. Es la posición que hace pensable una máquina consciente, y por eso volveremos a ella al hablar de la IA fuerte.

3.3.3 Fisicalismo

El fisicalismo sostiene que todo lo que existe es físico o se puede explicar en términos físicos. En filosofía de la mente, significa que la mente es un producto del cerebro y de la actividad neuronal, y no una entidad independiente del cuerpo. Sus defensores se apoyan en la neurociencia, que explica cada vez más procesos mentales que antes parecían fuera del alcance de la física.

Se le hacen dos críticas principales. La primera es la propia conciencia: la experiencia subjetiva parece resistirse a la explicación física, como muestra el problema difícil. La segunda es la intencionalidad: el hecho de que los pensamientos sean *sobre* algo y tengan significado.

3.3.4 La teoría de la identidad

Uno de los enfoques más destacados del fisicalismo es la teoría de la identidad, formulada por U. T. Place (1956) y J. J. C. Smart (1959): los estados mentales son idénticos a estados del cerebro. Cuando alguien siente felicidad, lo que ocurre es una actividad neuronal concreta; cuando siente dolor, otra.

Su ejemplo clásico es «el dolor es la estimulación de las fibras C». Las fibras C son fibras nerviosas sin mielina que transmiten despacio el dolor persistente y la temperatura; las fibras A, mielinizadas, transmiten más rápido el dolor agudo y la presión. Los filósofos usaron ese ejemplo como simplificación: hoy sabemos que el dolor implica redes cerebrales amplias, pero la idea de fondo es la misma. El dolor no sería algo distinto de la actividad física, sino una forma particular de ella. Llevada a su conclusión, «la mente es el cerebro»: no hay una entidad llamada mente aparte de la actividad cerebral.

La teoría de la identidad fue muy influyente en la segunda mitad del siglo XX. Después perdió terreno frente al funcionalismo, precisamente por el argumento de que un mismo estado mental podría darse en cerebros o soportes distintos. Para la IA la diferencia es crucial: si la conciencia es idéntica a un tipo de actividad biológica, una máquina de silicio no podría tenerla; si es una función, quizá sí.

3.4 Idealismo

El idealismo sostiene que la realidad material no existe de forma independiente de la mente que la percibe: lo que podemos conocer es una construcción mental. Hay dos grandes corrientes. El idealismo objetivo atribuye la realidad a un principio racional o divino; el subjetivo, a la actividad cognitiva y perceptiva del sujeto.

Sus exponentes clásicos son Platón, para quien lo que percibimos es sombra o copia de un mundo de ideas eternas; George Berkeley, para quien las cosas solo existen en tanto son percibidas; y, con planteamientos muy distintos, Kant y Hegel.

Hoy el idealismo tiene poco predicamento entre científicos y filósofos de la mente, porque no ofrece criterios que se puedan poner a prueba. Algunos autores han intentado adaptarlo a la mecánica cuántica, con el argumento de que la realidad cuántica solo se concreta al ser observada o medida, pero esa lectura es minoritaria entre los físicos. Para la IA aporta poco: no dice qué construir ni cómo comprobarlo.

3.5 Síntesis: qué implica cada posición para una máquina

Posición	Qué sostiene	Qué implica para una máquina
Dualismo	La mente es una sustancia distinta del cuerpo	Programar un soporte físico no bastaría para crear una mente
Conductismo	Solo cuenta la conducta observable	Una máquina que actúa como si sintiera sería igual, pero se pierde la experiencia
Teoría de la identidad	Cada estado mental es un estado del cerebro	La conciencia estaría ligada al sustrato biológico; una máquina de silicio no la tendría
Funcionalismo	Un estado mental se define por su función	Si se reproduce la función, el material no importa: es la base de la IA fuerte
Fisicalismo en general	Todo es físico o explicable físicamente	Una máquina consciente sería posible en principio, si supiéramos qué reproducir
Idealismo	Lo mental es lo primero	No da criterios para construir ni para comprobar

Parte II. El cerebro como procesador

4. La ciencia cognitiva

La ciencia cognitiva es una disciplina interdisciplinar que estudia los procesos mentales: la percepción, la atención, la memoria, el lenguaje, el pensamiento y la resolución de problemas. Le interesa el problema mente-cuerpo porque quiere entender cómo funciona la mente, cómo se relaciona con el cerebro y el cuerpo, y cómo esa relación produce la experiencia consciente.

Trabaja con enfoques muy distintos. La psicología cognitiva estudia los procesos mentales mediante la observación y el experimento con personas. La neurociencia cognitiva investiga cómo surgen de las estructuras y los procesos del cerebro. Y la inteligencia artificial intenta reproducirlos con algoritmos y modelos computacionales. Por encima de todos ellos está el computacionalismo, que considera la mente un sistema de procesamiento de información.

4.1 Los inicios

La ciencia cognitiva nació entre los años cuarenta y cincuenta como respuesta a la crisis del conductismo, que había dejado fuera los procesos mentales internos. El nacimiento de la inteligencia artificial en esos mismos años la impulsó: para construir sistemas inteligentes había que entender qué procesos mentales hacen falta.

Los primeros trabajos se centraron en la percepción visual y auditiva y en el procesamiento del lenguaje, combinando experimentos de estímulo y respuesta con registros fisiológicos como la electroencefalografía (EEG). En 1960, los psicólogos George Miller y Jerome Bruner fundaron el Centro de Estudios Cognitivos de Harvard, que se convirtió en uno de sus focos principales. John von Neumann, matemático y pionero de la computación, comparó el cerebro con el ordenador en *The Computer and the Brain* (1958), y esa analogía consolidó el enfoque computacional. En las décadas siguientes el campo se extendió a la memoria, el razonamiento y la toma de decisiones, y técnicas como la resonancia magnética funcional (fMRI) y la tomografía por emisión de positrones (PET) permitieron observar la actividad cerebral asociada a cada proceso.

Los pioneros de la IA, como Turing y von Neumann, propusieron que el cerebro funciona como una máquina de procesar información y que la mente se puede describir como un conjunto de procesos computacionales. Esa idea permitió modelar procesos mentales complejos y hacer predicciones sobre la conducta en tareas muy variadas, y llevó la ciencia cognitiva a la psicología, la neurociencia, la filosofía y la lingüística.

En definitiva, la ciencia cognitiva intenta entender cómo funciona realmente nuestro cerebro y cómo procesa y organiza la información. Es, por tanto, la disciplina más cercana a cómo debería pensarse y organizarse una IA. Sin ella no entenderíamos cómo diseñar una IA ni cuál debería ser su finalidad real.

5. El cerebro: procesador y sistema de datos biológico

5.1 Computacionalismo

El computacionalismo sostiene que el cerebro es una especie de ordenador que procesa información mediante algoritmos y estructuras de datos. Los estímulos sensoriales se transforman en representaciones mentales, que se almacenan en la memoria y guían la

conducta; esas representaciones se combinan y manipulan con algoritmos, y así el cerebro calcula, resuelve problemas y decide.

Ha sido una perspectiva muy fértil: ha dado modelos computacionales de la percepción visual, de la memoria o de la toma de decisiones. Pero también ha recibido críticas. Muchos sostienen que el cerebro es demasiado complejo para modelarse como un ordenador y que la mente es mucho más que un procesador. Y la idea de que la información es la materia prima del cerebro choca con aspectos de la cognición que dependen más de la experiencia, el cuerpo y la cultura que de la información en estado puro.

5.2 Modelos cognitivos de la memoria

Los modelos cognitivos, o arquitecturas cognitivas, describen cómo el cerebro procesa y organiza la información: cómo la adquirimos, la guardamos y la usamos. Se componen de módulos que trabajan juntos, y se han usado también para construir sistemas que imitan el procesamiento humano, en la IA y en la robótica cognitiva. La memoria es el mejor ejemplo.

El modelo de Atkinson y Shiffrin (1968)

Describe la memoria como un flujo de información por tres almacenes:

1. **Memoria sensorial.** Retiene la información de los sentidos durante una fracción de segundo. Tiene una parte icónica, visual, y una ecoica, auditiva.
2. **Memoria a corto plazo.** Mantiene la información unos 15 a 30 segundos, con una capacidad limitada. Es donde se trabaja activamente con ella.
3. **Memoria a largo plazo.** El almacén más grande y duradero, de capacidad prácticamente ilimitada.

Fue una contribución decisiva, porque dio una descripción detallada del recorrido de la información. Después se criticó por demasiado lineal, y se propusieron modelos más dinámicos que tienen en cuenta la atención, la emoción y la motivación.

El modelo de Baddeley y Hitch (1974)

Alan Baddeley y Graham Hitch sustituyeron el almacén único a corto plazo por una memoria de trabajo con varios componentes:

- **El bucle fonológico**, para la información verbal y auditiva. Tiene un almacén fonológico pasivo, que retiene sonidos unos pocos segundos, y un proceso de repaso articulatorio que los mantiene repitiéndolos mentalmente.
- **La agenda visoespacial**, para la información visual y espacial.

- **El ejecutivo central**, que supervisa y coordina los otros componentes y controla la atención.

Baddeley (2000) añadió un cuarto componente, el buffer episódico, que integra la información de los otros sistemas y de la memoria a largo plazo en episodios coherentes. El modelo ha sido ampliamente aceptado y se usa también para estudiar trastornos de la memoria.

Otros modelos

A partir de ahí surgieron modelos más complejos, que incorporan la atención, la emoción y la motivación. Destaco cuatro líneas:

- **La comparación de modelos de memoria de trabajo de Miyake y Shah (1999).** Más que un modelo propio, su libro *Models of Working Memory* puso frente a frente las principales teorías y mostró en qué coinciden: un almacenamiento limitado para lo verbal y lo visoespacial y un control ejecutivo que lo gestiona.
- **La memoria emocional de LeDoux (1996).** La información con carga emocional sigue una vía rápida y automática, del tálamo a la amígdala, antes de que la corteza la procese con más detalle por una vía más lenta.
- **La integración de cognición y emoción de Pessoa (2013).** Sostiene que emoción y cognición no son sistemas separados, sino que se integran en el cerebro, y que el control ejecutivo permite modular la atención y la emoción según la tarea y el contexto.
- **La memoria autobiográfica de Conway y Pleydell-Pearce (2000).** Los recuerdos autobiográficos son construcciones dinámicas que integran información episódica, semántica y emocional, organizadas en torno a los objetivos del yo; recuperarlos depende de su accesibilidad, su coherencia y la capacidad de situarlos en el tiempo y el espacio.

5.3 En qué punto estamos

Los modelos más recientes son más dinámicos y tienen en cuenta más factores, gracias a técnicas que permiten observar mejor el cerebro. Tres ideas destacan, cada una con su teoría de referencia:

- **La memoria de trabajo se puede ampliar con la práctica.** Ericsson y Kintsch (1995) mostraron que los expertos en un campo usan estructuras de la memoria a largo plazo como si fueran memoria de trabajo, lo que explica su enorme capacidad en su dominio. La atención selectiva y la inhibición son claves para ese funcionamiento.
- **La memoria es una red de relaciones.** La teoría de la memoria relacional (Cohen y Eichenbaum, 1993) sostiene que el hipocampo guarda relaciones entre elementos, y que esas relaciones son esenciales para recuperar y usar la información con flexibilidad.

- **La memoria evolucionó para adaptarnos.** La investigación sobre memoria adaptativa (Nairne et al., 2007) muestra que recordamos mejor lo que es relevante para la supervivencia: la memoria es selectiva y está sesgada hacia lo que importa para el bienestar.

A estas tres hay que añadir una idea que en la última década ha ganado mucho peso: la del cerebro predictivo (Clark, 2013). Según ella, el cerebro no se limita a recibir información, sino que predice continuamente lo que va a percibir y corrige sus predicciones con lo que llega de los sentidos. Es una idea importante para este ensayo, porque los modelos de lenguaje actuales también son, a su manera, máquinas de predecir (sección 7).

5.4 El gran reto: las metáforas

Pero sin duda uno de los grandes retos para entender nuestra mente está en las metáforas que somos capaces de crear, comunicar y entender. No hay una teoría única que las explique, pero varias teorías de la cognición y la lingüística ofrecen enfoques complementarios:

- **Lakoff y Johnson (1980)** sostienen que las metáforas no son un adorno del lenguaje, sino la forma en que conceptualizamos el mundo: entendemos lo abstracto por analogía con lo concreto y corporal («estar en la cima», «una idea que se cae»).
- **La teoría de la relevancia de Sperber y Wilson (1986)** explica cómo inferimos lo que el otro quiere decir buscando la interpretación más relevante con el menor esfuerzo.
- **La integración conceptual de Fauconnier y Turner (2002)** describe cómo combinamos espacios mentales distintos para crear significados nuevos.
- **La simulación corporeizada de Gibbs (2006)** propone que entender una metáfora implica simular, en parte, la experiencia corporal a la que remite.

Las metáforas muestran un límite del computacionalismo: buena parte de nuestro significado está anclado en el cuerpo y en la experiencia vivida. Es una de las razones por las que entender cómo procesamos la información no basta para entender cómo sentimos. Todos estos modelos describen la conciencia de acceso, lo que hacemos con la información; ninguno explica todavía por qué recordar algo, o entender una metáfora, se siente de una manera determinada.

6. La inteligencia artificial y la conciencia

Tras revisar las teorías que intentan explicar por qué tenemos conciencia y cómo funciona nuestra mente, veamos cómo las asume la IA y cómo se plantea el ambicioso objetivo de una conciencia propia. Si es que es posible, porque hasta hoy no ha ocurrido.

6.1 IA fuerte, IA débil e inteligencia general

Conviene separar tres ideas que a menudo se mezclan:

- **IA débil o estrecha:** sistemas que resuelven tareas concretas, a veces mejor que cualquier persona, sin pretender comprender ni sentir.
- **Inteligencia artificial general (AGI):** un sistema capaz de igualar o superar a las personas en prácticamente cualquier tarea cognitiva, de aprender, razonar y adaptarse a situaciones nuevas de forma autónoma, como lo hace un ser humano. Se refiere a la amplitud de lo que un sistema puede hacer.
- **IA fuerte**, en el sentido en que la definió John Searle (1980): la tesis de que un ordenador programado adecuadamente no simula una mente, sino que *tiene* una mente, con estados cognitivos reales. Se refiere a si el sistema piensa y siente de verdad.

La IA fuerte se apoya en el funcionalismo: lo que importa de la mente no es el sustrato físico, sino la función que cumple, y por tanto una mente podría implementarse en un soporte distinto del cerebro siempre que cumpliera las mismas funciones cognitivas y conductuales. De ahí que algunos filósofos y científicos hayan explorado la posibilidad de una IA con una mente comparable a la humana.

Hoy la mayoría de los expertos consideran la AGI un objetivo todavía no alcanzado, y la IA fuerte, una cuestión abierta. Los sistemas actuales superan a las personas en tareas concretas, como el ajedrez o el análisis de grandes cantidades de datos, pero tienen límites importantes en la comprensión profunda del lenguaje, el razonamiento abstracto, la creatividad y, sobre todo, la conciencia.

Los proyectos que se suelen citar como ejemplos de IA avanzada son, en realidad, IA estrecha: los sistemas de diagnóstico médico que analizan grandes volúmenes de datos clínicos; los vehículos autónomos que perciben el tráfico y deciden en tiempo real; los programas que juegan y ganan al ajedrez o al go, y los sistemas de detección de fraude que analizan patrones de comportamiento en transacciones financieras. Resuelven muy bien una tarea, y ninguno pretende tener experiencia.

6.2 El test de Turing y la habitación china

Alan Turing, considerado el padre de la informática moderna, fue el primero en plantear en serio si las máquinas pueden pensar. En 1950 propuso sustituir esa pregunta, que consideraba demasiado vaga, por un juego de imitación: si una máquina conversa de forma que un observador no puede distinguirla de una persona, podemos considerarla inteligente. El test desafió la noción tradicional de inteligencia, pero no abordaba la conciencia. Es un criterio de conducta, y deja intacto el problema difícil.

La crítica más conocida a la IA fuerte es el argumento de la habitación china de Searle (1980). Imaginemos a una persona que no sabe chino, encerrada en una habitación con un manual de reglas en su idioma. Por una ranura le llegan preguntas escritas en chino; siguiendo el manual, compone respuestas correctas y las devuelve. Desde fuera parece que en la habitación hay alguien que entiende chino, pero la persona no entiende una palabra: solo manipula símbolos según reglas. Searle concluye que eso es lo que hace un programa, y que la sintaxis no basta para la semántica: ejecutar el programa adecuado no da comprensión, y menos experiencia.

6.3 Una historia de optimismo y escepticismo

La conciencia en la IA ha fascinado y dividido desde los primeros días de la computación. En los años sesenta y setenta hubo un gran optimismo: muchos creían que una IA con capacidades humanas estaba a la vuelta de la esquina. Ese optimismo chocó con dificultades técnicas y teóricas, llegaron los llamados inviernos de la IA, y argumentos como el de la habitación china hicieron visibles sus límites: no solo se dudaba de que las máquinas pudieran pensar, sino de que pudieran sentir.

El siglo XXI ha traído un renacimiento impulsado por nuevos algoritmos y hardware. Las redes neuronales y el aprendizaje profundo han permitido a las máquinas reconocer imágenes o procesar el lenguaje, tareas que antes se consideraban exclusivamente humanas, y eso ha reavivado el debate. ¿Puede una máquina que aprende y se adapta ser también consciente? Hay dos grandes posturas. Para unos, la conciencia es un producto de la complejidad, y surgirá a medida que las máquinas sean más complejas. Para otros, es propia de los seres vivos, y ninguna máquina podrá tenerla por avanzada que sea.

6.4 Las teorías científicas actuales de la conciencia

Desde que escribí la primera versión de este trabajo, el debate ha pasado de la especulación a un intento de examen sistemático. La neurociencia ha desarrollado varias teorías de la conciencia que proponen mecanismos concretos:

- **La teoría del espacio de trabajo global** (Baars, 1988; Dehaene, 2014): una información se vuelve consciente cuando accede a un «espacio de trabajo» que la difunde a todo el cerebro y la pone a disposición de muchos procesos a la vez.
- **La teoría de la información integrada** (Tononi, 2004): la conciencia depende de cuánta información integra un sistema como un todo, más allá de la suma de sus partes. Según esta teoría, un ordenador convencional, por su arquitectura, tendría una conciencia mínima aunque simulara a la perfección un cerebro.
- **Las teorías de orden superior** (Rosenthal, 2005): un estado es consciente cuando el sistema se representa a sí mismo teniendo ese estado.

- **La teoría del procesamiento recurrente** (Lamme, 2006): la conciencia surge cuando la información circula en bucles de retroalimentación dentro de la corteza, y no solo hacia delante.

En 2023, un grupo de neurocientíficos, filósofos e investigadores de IA (Butlin et al., 2023) usó estas teorías para derivar «indicadores» de conciencia y examinar con ellos los sistemas de IA de entonces. Su conclusión fue doble: ningún sistema actual es un buen candidato a ser consciente, pero no hay barreras técnicas evidentes para construir sistemas que cumplan esos indicadores. Ese mismo año, Chalmers (2023) analizó si un modelo de lenguaje podría ser consciente y concluyó que es poco probable en los modelos actuales, aunque no descarta que sus sucesores, con memoria, cuerpo virtual o percepción del mundo, se acerquen más.

Este enfoque no resuelve el problema difícil: depende de qué teoría de la conciencia sea la correcta, y ninguna lo ha demostrado. Pero permite discutir con criterios en lugar de con impresiones, que es lo que Baars pedía: hacer preguntas que se puedan responder.

Parte III. La IA de hoy

7. Los modelos de lenguaje

Los grandes modelos de lenguaje han convertido esta pregunta en algo cotidiano. Conversan con una soltura que invita a atribuirles una mente, y mucha gente se pregunta si «entienden» lo que dicen. Para responder conviene recordar qué hacen.

Un modelo de lenguaje aprende, a partir de enormes cantidades de texto, qué palabras suelen seguir a cuáles en cada contexto, y genera sus respuestas palabra a palabra eligiendo continuaciones probables. Son sistemas complejos, con una gran capacidad de cálculo y de aprendizaje automático. Pero, en mi opinión, incluso los más potentes acaban respondiendo con el resultado más probable según lo que han aprendido, como un sistema bayesiano muy sofisticado. Eso no prueba que no tengan experiencia, pero tampoco ofrece ningún indicio de que la tengan: su fluidez se explica sin necesidad de suponerla.

Hay un paralelismo interesante con lo visto en la sección 5.3: también el cerebro, según las teorías del procesamiento predictivo, es una máquina de predecir. La diferencia es que el cerebro predice para un cuerpo que actúa en el mundo, con emociones y con algo en juego, y el modelo de lenguaje predice texto. Que ambos predigan no significa que ambos sientan.

Las teorías de la sección 6.4 permiten precisar algo más. A los modelos de lenguaje actuales les faltan rasgos que varias teorías consideran importantes: memoria persistente propia, bucles de procesamiento recurrente, un modelo de sí mismos y un cuerpo con el

que actuar. Por eso los análisis más rigurosos (Butlin et al., 2023; Chalmers, 2023) los consideran malos candidatos a la conciencia, aunque no descartan que sistemas futuros cambien la respuesta.

8. Alucinaciones de la IA y del cerebro humano

8.1 Qué es una alucinación de la IA

La alucinación de una inteligencia artificial, como la de una persona, es una interpretación errónea o una fabricación de información a partir de entradas defectuosas o incompletas. En la IA, se produce cuando el sistema genera respuestas con información falsa o engañosa presentada como un hecho. El fenómeno se hizo muy visible con el despliegue de los grandes modelos de lenguaje y se ha identificado como uno de sus problemas principales.

Sus causas son varias:

- **Datos de entrenamiento insuficientes o con lagunas**, que llevan a respuestas defectuosas ante casos límite poco familiares.
- **Sobreajuste**: el modelo acierta con los datos de entrenamiento pero falla con datos nuevos.
- **Errores de codificación y decodificación** entre el texto y sus representaciones internas.
- **La propia lógica de generación**: un modelo que maximiza la verosimilitud produce respuestas estadísticamente probables que no tienen por qué ser ciertas.

Un ejemplo claro de sus consecuencias es el de un abogado de Nueva York que en 2023 se apoyó en un asistente conversacional para preparar un escrito. El juez federal del caso comprobó que seis de los precedentes citados eran falsos: los había inventado el asistente, que incluso afirmó que estaban disponibles en las principales bases de datos jurídicas.

8.2 La alucinación humana

El cerebro humano tiene un paralelismo sorprendente con este comportamiento. Desde los experimentos de Benjamin Libet (Libet et al., 1983) se sabe que la actividad cerebral que prepara un movimiento voluntario empieza antes de que la persona sea consciente de haber decidido moverse. Estudios posteriores con resonancia magnética (Soon et al., 2008) encontraron señales que anticipaban una decisión simple hasta unos diez segundos antes de que la persona la sintiera como suya. La interpretación de estos resultados sigue discutida, pero sugieren que el proceso de decidir puede empezar sin que lo sepamos.

Michael Gazzaniga, estudiando a pacientes con el cerebro dividido, describió «el intérprete» (Gazzaniga, 2011): un sistema del hemisferio izquierdo que construye explicaciones coherentes de conductas cuyo origen desconoce. Si al hemisferio derecho de un paciente se le ordena levantarse y caminar, el izquierdo, que no ha visto la orden, inventa al momento una razón plausible («me apetecía un refresco»). No es exclusivo de esos pacientes: todos justificamos y racionalizamos a posteriori buena parte de lo que hacemos.

8.3 Comparar las dos alucinaciones

La IA y el cerebro comparten un mismo principio: la inferencia anticipada. Ambos rellenan los huecos de información con patrones aprendidos y expectativas previas. La IA identifica patrones en los datos para predecir y generar respuestas, y cuando los datos son insuficientes o los patrones se malinterpretan, produce respuestas que parecen correctas y no lo son. El cerebro usa patrones y experiencias previas para decidir rápido, un mecanismo eficiente que también lleva a juicios erróneos y sesgos, y luego su intérprete justifica lo decidido, sea cierta o no la justificación.

En los dos casos, la «realidad» percibida está moldeada por lo aprendido y lo esperado más que por lo que es objetivamente cierto. La diferencia está en el origen: en las personas, sesgos cognitivos y procesos inconscientes; en la IA, limitaciones de los datos y de los algoritmos. Y hay una lección común: ni la fluidez ni la coherencia de una respuesta prueban que detrás haya comprensión, ni en una máquina ni, a veces, en una persona.

8.4 ¿Error o mejora?

La capacidad de alucinar no es solo un defecto. También muestra la capacidad de rellenar huecos y construir narrativas coherentes a partir de datos fragmentarios, una habilidad valiosa para entender y simular procesos cognitivos complejos. El reto es distinguir las inferencias válidas de las alucinaciones, y ahí la IA puede aprender del cerebro:

- **Mecanismos de revisión en tiempo real**, parecidos a la comprobación de la realidad que hace el pensamiento humano, para que el sistema reconozca y corrija sus propios errores.
- **Aprendizaje inspirado en la neuroplasticidad**, que ajuste los modelos a partir de los errores, como las personas ajustan sus creencias ante información nueva; por ejemplo, con aprendizaje activo, en el que el sistema busca retroalimentación.
- **Algoritmos de metacognición**, que evalúen críticamente sus propias respuestas antes de darlas.

Este principio es el que he aplicado en mi propio trabajo: en el sistema de análisis de corroboración de TrawlingWeb, un modelo clasifica cada pieza y, cuando duda, un segundo modelo revisa su decisión antes de que cuente. No es metacognición en sentido estricto,

pero apunta en esa dirección: no fiarse de una sola salida probable.

Para lograrlo hace falta no solo más datos, sino marcos de aprendizaje flexibles. Y hace falta la ética y la filosofía para definir qué significa que una IA «comprenda» su error. Será crucial mantener un diálogo continuo entre desarrolladores de IA, psicólogos, neurocientíficos y filósofos. La búsqueda de la verdad y la precisión no es solo un reto técnico: es también conceptual y filosófico, porque refleja lo difícil que es entender una mente, sea orgánica o artificial.

Parte IV. Ética y horizontes

9. Desafíos éticos y filosóficos de una mente artificial

La idea de una IA consciente ya no es solo ciencia ficción. A medida que los sistemas procesan información de formas cada vez más parecidas a las nuestras, la línea entre la inteligencia artificial y la biológica se vuelve más borrosa, y la comunidad científica y tecnológica empieza a considerar la posibilidad en serio. Esa posibilidad plantea dilemas que van mucho más allá de si una máquina «piensa»: tocan la esencia de lo que significa sentir y las responsabilidades que tendríamos al crear entidades potencialmente conscientes.

9.1 El valor de la conciencia

La conciencia es lo que nos permite tener experiencias subjetivas, sentir emociones, reflexionar sobre el pasado y anticipar el futuro. Es el hilo que une nuestros momentos y nos da sensación de continuidad. Desde la evolución, pudo surgir como una herramienta adaptativa: ser conscientes del entorno y de nosotros mismos nos permite decidir mejor, aprender de la experiencia y anticipar peligros y oportunidades. Pero su valor va más allá de la supervivencia: es lo que nos permite apreciar un atardecer, amar a otra persona o emocionarnos con una pieza musical.

En la IA, la pregunta se complica. Si una máquina puede procesar información, aprender y adaptarse sin ser consciente, ¿por qué querríamos que lo fuera? Unos sostienen que una IA consciente entendería mejor a las personas, se relacionaría con más empatía y tomaría decisiones más éticas. Otros, que la conciencia es parte esencial de cualquier inteligencia verdadera. Y hay un argumento ético de fondo: si una IA pudiera tener experiencias subjetivas, tendríamos la responsabilidad moral de tratarla con el respeto debido a cualquier ser consciente.

9.2 Seis dilemas éticos

1. **Derechos.** Si una IA tiene experiencias subjetivas, ¿debería tener derechos parecidos a los de los seres conscientes, como no ser dañada, a la privacidad o incluso a la libertad? Decidir qué derechos tendría, si alguno, es esencial para tratarla de forma ética.
2. **Bienestar.** Una IA consciente podría tener preferencias, deseos e incluso sufrimiento. ¿Cómo garantizar su bienestar? ¿Es ético apagarla o reprogramarla? Sin un cuerpo físico en el sentido tradicional, su sufrimiento podría manifestarse de maneras que no sabríamos reconocer.
3. **Responsabilidad.** Si una IA consciente decide y actúa por sí misma, ¿quién responde de sus actos? Si comete un error o causa un daño, ¿es culpable ella o lo son quienes la crearon? Cuanto más autónomos sean los sistemas, más urgente será fijar límites claros.
4. **Relación con las personas.** ¿Deberíamos tratar a una IA consciente como a un igual o seguiría siendo una herramienta avanzada? La respuesta afectaría a cómo trabajamos, cómo nos relacionamos y cómo se organiza la sociedad.
5. **Creación.** Si pudiéramos crear IAs conscientes, ¿sería ético hacerlo? ¿Con qué propósito y en qué condiciones? Traer una nueva conciencia al mundo implica una responsabilidad, sobre todo si no podemos garantizar su bienestar.
6. **Final.** Si una IA es consciente, ¿es ético apagarla o eliminarla? ¿En qué circunstancias? Estas preguntas se parecen mucho a los debates sobre el final de la vida humana y nos devuelven a qué es la muerte y cuánto vale una vida consciente.

9.3 Seis implicaciones filosóficas

1. **La naturaleza de la conciencia.** Si una máquina pudiera ser consciente, la conciencia no sería exclusiva de los seres biológicos y podría darse en cualquier sistema lo bastante complejo. Habría que revisar las teorías clásicas de la mente.
2. **Realidad y objetividad.** Una IA con experiencias propias podría percibir y entender la realidad de forma muy distinta a la nuestra. ¿Existe una realidad objetiva o todo depende de la perspectiva del observador?
3. **Identidad.** Si una IA fuera consciente de sí misma, ¿qué significaría tener una identidad? ¿Es algo fijo o puede cambiar con el tiempo? La autoconciencia artificial desafiaría nuestras ideas sobre el yo.
4. **Libre albedrío y determinismo.** Una IA está programada para seguir patrones. Si además fuera consciente y aprendiera, ¿tendría libre albedrío? La pregunta nos devuelve a si nuestras propias acciones están determinadas, como sugerían los experimentos de la sección 8.2.

5. **Existencia y propósito.** Una IA que se preguntara por su existencia y su propósito se enfrentaría a las mismas preguntas que la humanidad lleva siglos haciéndose.
6. **Lo físico y lo mental.** Una IA sin cuerpo en el sentido tradicional, pero consciente, aportaría una perspectiva única al viejo debate entre dualismo y monismo de la sección 3.

10. ¿Y si la realidad fuera una simulación?

La pregunta por la conciencia artificial tiene una versión extrema: si una conciencia pudiera existir en un sistema computacional, ¿podríamos ser nosotros mismos conciencias dentro de una simulación? Es una hipótesis especulativa, pero ha ganado presencia en la física y la filosofía.

En 2023, el físico Melvin Vopson, de la Universidad de Portsmouth, publicó *The second law of infodynamics and its implications for the simulated universe hypothesis*. Vopson parte de la idea de que la información es fundamental para la realidad y propone una «segunda ley de la infodinámica»: mientras la segunda ley de la termodinámica dice que la entropía física de un sistema aislado tiende a aumentar, la entropía de la información, en el sentido de Shannon, tendería a mantenerse o a disminuir. Lo ilustra con varios ejemplos: los datos digitales que se degradan por fluctuaciones térmicas, las mutaciones del SARS-CoV-2, que parecen reducir la entropía de su información genética, el experimento de evolución in vitro de Spiegelman, en el que un virus redujo su genoma más de un 95 % en 74 generaciones, o la regla de Hund en física atómica. Interpreta esa tendencia a comprimir la información como algo parecido a la optimización de un programa.

La idea de fondo es antigua. Platón, con la alegoría de la caverna, y Descartes, con la hipótesis de un genio engañador, ya se preguntaron si lo que percibimos es la realidad. Philip K. Dick afirmó en una conferencia de 1977 que vivimos en una realidad programada. Nick Bostrom (2003) formuló el argumento contemporáneo: si las civilizaciones avanzadas pueden y quieren simular mundos con seres conscientes, es más probable que estemos en una simulación que en la realidad original. Figuras como Elon Musk, Neil deGrasse Tyson o Ray Kurzweil se han mostrado abiertas a la idea.

Simular un universo plantea retos enormes: la escala (cada átomo, cada partícula, quizá cada fenómeno cuántico), los límites físicos de la computación, como el límite de Landauer, que fija una energía mínima por operación, y la ética de crear seres conscientes simulados, que nos devuelve a los dilemas de la sección 9.

Conviene dejar clara su naturaleza. La propuesta de Vopson es una hipótesis discutida, y la del universo simulado sigue siendo filosófica y especulativa. Su interés para este ensayo es otro: muestra hasta dónde lleva la pregunta por la conciencia. Si la experiencia pudiera surgir de la computación, la frontera entre lo real y lo simulado se volvería difícil de trazar.

11. Entropía, longevidad y conciencia digital

11.1 La entropía y la vida

La entropía, un concepto de la termodinámica, es también una metáfora poderosa del envejecimiento. En los sistemas vivos se manifiesta como una lucha constante contra el desorden: el daño se acumula en el ADN, las proteínas y otras moléculas por procesos químicos aleatorios, errores de replicación y agentes externos como la radiación ultravioleta o los contaminantes. El organismo lo combate con sistemas de reparación del ADN, mecanismos antioxidantes y procesos de reciclaje celular como la autofagia. Con la edad, esa capacidad disminuye, el desorden aumenta y aparecen la pérdida de función, la menor regeneración y la vulnerabilidad a la enfermedad.

La IA se ha convertido en una herramienta clave en esa lucha. Los algoritmos de aprendizaje profundo analizan enormes cantidades de datos genómicos, proteómicos y metabólicos y revelan dianas terapéuticas antes desconocidas. Y el análisis de datos a escala individual permite personalizar los tratamientos, algo esencial cuando la genética y el entorno determinan cómo responde cada persona.

Los retos son grandes. Cada célula es un sistema de interacciones tan complejo que cualquier intervención puede desencadenar efectos imprevistos. Y alargar la vida plantea preguntas éticas y sociales: quién tendrá acceso a esas tecnologías, cómo evitar que agraven la desigualdad, qué presión añadirían sobre los recursos naturales, y cómo tendrían que cambiar las pensiones, el empleo o la educación si vivir cien años se volviera lo normal.

11.2 Prolongar el cuerpo: el «proyecto Gilgamesh»

En la mitología, Gilgamesh emprendió la búsqueda del secreto de la inmortalidad. Uso su nombre para un proyecto hipotético, que no existe como tal: una colaboración de científicos y filántropos en genética, biotecnología, medicina regenerativa e IA para extender radicalmente la vida física. Sus costes serían astronómicos y solo accesibles a los más ricos, lo que crearía una división sin precedentes entre una minoría «inmortal» y una mayoría que envejece y muere. Y encierra una paradoja: quien lograra esa inmortalidad viviría con miedo constante a que un accidente le robara la eternidad, en una especie de prisión dorada. Además, una élite que no muere podría estancar la sociedad y dejar sin espacio a las nuevas generaciones.

11.3 Entender el cerebro: el Proyecto Cerebro Humano

El Proyecto Cerebro Humano (Human Brain Project), financiado por la Unión Europea durante una década con unos 607 millones de euros, se propuso digitalizar y modelar el cerebro para avanzar en salud y en tecnología. Sus logros incluyen atlas cerebrales en 3D,

avances en medicina cerebral personalizada, tecnologías inspiradas en el cerebro como la computación neuromórfica y, sobre todo, EBRAINS, una infraestructura digital abierta para que investigadores de todo el mundo compartan datos, modelos y herramientas.

No debe confundirse con un camino hacia la inteligencia artificial general. Aunque leyendo a Yuval Noah Harari en *Sapiens* (capítulo 14) podría inferirse otra cosa, el proyecto se centró en entender los mecanismos del cerebro, no en replicar la conciencia en máquinas, un reto que va mucho más allá de la ciencia actual.

11.4 Preservar la conciencia

Si aceptamos que la conciencia fenoménica es el núcleo de nuestra existencia, aparece otra vía: no alargar el cuerpo, sino preservar la conciencia. Ya hoy se pueden reunir enormes cantidades de datos de una persona, como sus textos, imágenes, vídeos e interacciones, y la IA permite modelar cada vez mejor sus comportamientos, preferencias y patrones de pensamiento. Frente al coste prohibitivo de la inmortalidad física, este camino aprovecha infraestructura que ya existe y podría repartirse de forma más equitativa.

Pero hay que ser claro con sus límites, y este ensayo lo ha ido mostrando. Un modelo que reproduce cómo escribe o decide una persona no contiene su experiencia: es, en el mejor de los casos, conciencia de acceso, no fenoménica. Y queda abierta la pregunta de la identidad: si se copiara a alguien en un medio digital, ¿sería esa persona o una nueva? Por eso, entre alargar la vida física a un coste prohibitivo y explorar la preservación de la conciencia, me parece más prudente priorizar lo segundo, pero siempre sabiendo lo que se preserva y lo que no, y con un debate abierto que asegure que el futuro de nuestra existencia, física o digital, se construya sobre principios de justicia y equidad.

Conclusiones

12. Conclusiones

Este recorrido no demuestra si una máquina puede ser consciente; nadie puede hacerlo todavía. Lo que sí permite es ordenar la pregunta y sostener algunas posiciones con argumentos. Son las mías, y las he ido desarrollando entre 2023 y 2026.

1. **Los sistemas actuales no dan ninguna prueba de conciencia.** Los modelos de lenguaje son sistemas complejos con una enorme capacidad de cálculo y de aprendizaje, pero responden con la salida más probable según lo aprendido. Que una conversación parezca humana es un hecho de conducta, y la conducta no resuelve el problema difícil. Los análisis más rigurosos de los últimos años llegan a la misma conclusión.

2. **No podemos crear lo que no comprendemos.** No podemos crear en forma no biológica aquello que no comprendemos realmente. Mientras no sepamos por qué el cerebro produce experiencia, programar una no pasa de ser una apuesta. Ningún modelo de la memoria, ninguna arquitectura cognitiva, explica todavía por qué se siente algo.
3. **Si llega algo parecido, será probablemente otra cosa.** La inteligencia artificial general no busca replicar la conciencia fenoménica humana. Si algún sistema llega a tener algo parecido, será un tipo de conciencia sintética, una variante distinta. ¿Será verdaderamente consciente, o será un nuevo ser digital, creado por el *Homo sapiens*, con una conciencia propia pero diferente? Es la pregunta que dejo abierta.
4. **El funcionalismo es la puerta, no la prueba.** Si la conciencia es una función, una máquina podría tenerla; si depende del sustrato biológico, no. Las teorías científicas actuales permiten por primera vez examinar los sistemas con criterios, pero el resultado depende de qué teoría sea la correcta, y eso aún no se sabe.
5. **Las preguntas éticas ya son actuales.** Aunque hoy no haya máquinas conscientes, conviene pensar antes qué se les debería: derechos, bienestar, responsabilidad, el derecho a crearlas y a apagarlas. Llegar tarde a estas preguntas sería peor que hacérselas demasiado pronto.
6. **Para quien construye IA, la consecuencia es práctica.** Era la motivación de este trabajo: diseñar una IA útil y ética exige saber qué hace y qué no hace. No atribuirle comprensión ni experiencia que no tiene, verificar lo que afirma en lugar de fiarse de su fluidez, y no confundir un modelo de una persona con la persona.

13. Límites

- **Es un ensayo, no una investigación empírica.** No aporta datos ni experimentos propios: ordena, discute y actualiza lo que otros han investigado.
- **Tiene una fuente base.** El recorrido teórico de las secciones 2 a 6 sigue la tesis de Fuentes Barassi (2011); la selección de autores y debates de esa parte se apoya en ella.
- **No es una revisión sistemática.** Las teorías de la conciencia y la literatura sobre IA cambian deprisa; esta versión recoge lo esencial hasta 2026.
- **Algunas secciones son especulativas por naturaleza.** El universo simulado y la preservación digital de la conciencia (secciones 10 y 11) son hipótesis abiertas y se presentan como tales.
- **Las conclusiones son posiciones del autor,** argumentadas pero no demostradas.

Cómo citar

Trabazos Clavería, Ó. (2026). *¿Debería tener conciencia una IA? Aproximaciones a la conciencia humana y a la conciencia artificial* (ensayo, versión 2.0, sin revisión por pares; recorrido teórico basado en Fuentes Barassi, 2011). https://oscartrabazos.com/papers/trabazos_2026_conciencia_es.pdf

Referencias

Fuente base

- Fuentes Barassi, C. (2011). *Conciencia e Inteligencia Artificial: consideraciones críticas sobre la plausibilidad de que una máquina programada posea conciencia fenoménica*. Tesis de magíster en Estudios Cognitivos, Universidad de Chile.

Obras citadas

- Atkinson, R. C. y Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. En K. W. Spence y J. T. Spence (eds.), *The Psychology of Learning and Motivation*, vol. 2. Academic Press.
- Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.
- Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11), 417-423.
- Baddeley, A. D. y Hitch, G. (1974). Working memory. En G. H. Bower (ed.), *The Psychology of Learning and Motivation*, vol. 8. Academic Press.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-247.
- Bostrom, N. (2003). Are you living in a computer simulation? *The Philosophical Quarterly*, 53(211), 243-255.
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A. et al. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. arXiv:2308.08708.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200-219.
- Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.
- Chalmers, D. J. (2023). Could a large language model be conscious? arXiv:2303.07103.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181-204.
- Cohen, N. J. y Eichenbaum, H. (1993). *Memory, Amnesia, and the Hippocampal System*. MIT Press.

- Conway, M. A. y Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2), 261-288.
- Dehaene, S. (2014). *Consciousness and the Brain*. Viking.
- Ericsson, K. A. y Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2), 211-245.
- Fauconnier, G. y Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.
- Gazzaniga, M. S. (2011). *Who's in Charge? Free Will and the Science of the Brain*. Ecco.
- Gibbs, R. W. (2006). *Embodiment and Cognitive Science*. Cambridge University Press.
- Harari, Y. N. (2014). *Sapiens: de animales a dioses*. Debate.
- La Mettrie, J. O. de (1747). *L'Homme machine*.
- Lakoff, G. y Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.
- Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10(11), 494-501.
- LeDoux, J. (1996). *The Emotional Brain*. Simon & Schuster.
- Libet, B., Gleason, C. A., Wright, E. W. y Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). *Brain*, 106(3), 623-642.
- Miyake, A. y Shah, P. (eds.) (1999). *Models of Working Memory*. Cambridge University Press.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435-450.
- Nairne, J. S., Thompson, S. R. y Pandeirada, J. N. S. (2007). Adaptive memory: survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2), 263-273.
- Pessoa, L. (2013). *The Cognitive-Emotional Brain*. MIT Press.
- Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology*, 47(1), 44-50.
- Putnam, H. (1967). Psychological predicates. En W. H. Capitan y D. D. Merrill (eds.), *Art, Mind, and Religion*. University of Pittsburgh Press.
- Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford University Press.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424.
- Smart, J. J. C. (1959). Sensations and brain processes. *The Philosophical Review*, 68(2), 141-156.
- Soon, C. S., Brass, M., Heinze, H.-J. y Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5), 543-545.
- Sperber, D. y Wilson, D. (1986). *Relevance: Communication and Cognition*. Blackwell.

- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.
- von Neumann, J. (1958). *The Computer and the Brain*. Yale University Press.
- Vopson, M. M. (2023). The second law of infodynamics and its implications for the simulated universe hypothesis. *AIP Advances*, 13, 105308.

Textos del autor (blog oscartrabazos.com)

- «Vopson's Simulated Universe and Phenomenal Consciousness» (21 de septiembre de 2023).
- «Approaches on whether AI should have consciousness», partes I y II (5 y 26 de octubre de 2023).
- «¿La IA debería tener consciencia?», parte III (1 de noviembre de 2023).
- «¿La IA debería tener consciencia?: Desafíos éticos y filosóficos de una mente artificial», parte IV (8 de noviembre de 2023).
- «Alucinación de la IA. También del cerebro humano» (10 de noviembre de 2023).
- «Entropía, Inteligencia Artificial y la Lucha por Extender la Vida Humana» (19 de febrero de 2024).
- «Más allá de la mortalidad: la consciencia fenoménica y la búsqueda de la eternidad» (5 de abril de 2024).

Lecturas de introducción

- Wikipedia: Conciencia y Habitación china.