Stoa
Vol. 16, no. 32, pp. 25-44
ISSN 2007-1868
DOI: https://doi.org/

CONCIENCIA E INTELIGENCIA ARTIFICIAL

Consciousness and Artificial Intelligence

EDUARDO RUIZ MAZÓN
Universidad Nacional Autónoma de México
Secretaría de Relaciones Exteriores, México
eruizmazon@hotmail.com

RESUMEN: La equiparación de la Inteligencia Artificial (IA) con la conciencia humana ha sido una de las principales motivaciones del desarrollo de la primera, si bien sus aplicaciones tecnológicas actuales le han dado una propia justificación y utilidad económica sin precedentes en las sociedades modernas y a nivel global. No obstante, sigue siendo pertinente analizar las similitudes y diferencias esenciales entre la IA y la conciencia, así como si existen bases teóricas para afirmar que la primera puede replicar a la segunda. Este trabajo busca ubicar la discusión del tema en el contexto del manejo filosófico de las categorías epistemológicas de sujeto y objeto, comparándolas con los fundamentos técnicos y teóricos de la computación, destacando principalmente los enfoques fenomenológico y conductista y sus respectivas posiciones en torno a esta tecnología. Con ello se intenta contribuir a una discusión fundamentada en el marco del creciente interés general por la IA.

PALABRAS CLAVE: Conciencia · inteligencia artificial · algoritmo · semántica · subjetividad · intencionalidad · cerebro · intersubjetividad · cultura.

ABSTRACT: The comparison of Artificial Intelligence (AI) with human consciousness has been one of the main motivations for the development of the former, although its current technological applications have given it its own justification and unprecedented economic utility in modern societies and globally. However, it is still pertinent to analyze the essential similarities and differences between AI and consciousness, as well as whether there are theoretical bases to affirm that the former can replicate the latter. This essay tries to approach the discussion of the matter in the context of the philosophical use of the epistemological categories of subject and object, comparing

Recibido el 4 de enero de 2025
Aceptado el 27 de mayo de 2025

them with the technical and theoretical foundations of computing, mainly highlighting behavioristic and phenomenological approaches and their respective positions on this technology. With this in view, the aim is to contribute to an informed discussion, in the framework of the general interest in AI.

KEYWORDS: Consciousness · AI · algorithms · Semantics · subjectivity · intentionality · brain · intersubjectivity · mind · culture.

1. Introducción

Es bien conocido que los pioneros de la Inteligencia Artificial (IA) establecieron como uno de los objetivos principales de esta tecnología la simulación de actividades o fenómenos propios de la conciencia humana. A más de sesenta años de la invención de la IA y la acuñación de su nombre, pareciera que el debate sobre la identificación u homologación entre la IA y la conciencia humana estuviera superado, dada una delimitación clara de ambos conceptos y sus respectivas características esenciales.

No obstante, la ubicuidad actual de las aplicaciones y programas de IA en la actividad económica y cotidiana de las sociedades contemporáneas, ha revivido la pertinencia de este debate. El lanzamiento reciente de ChatGPT y de otros robots conversacionales, como DeepSeek, Copilot o Gemini, la discusión sobre las ventajas o riesgos del aprovechamiento de la IA, los cuestionamientos sobre los aspectos éticos de la IA (Calvo 2024, p. 87), la creciente propaganda a favor de su uso y alimentación por parte de los usuarios en internet contribuyendo a la minería de datos y, finalmente, la regulación de distintas aplicaciones de IA como reconocedores de rostros en lugares públicos o calificadores de conducta social, incluyendo su restricción o incluso prohibición en algunos países, son ejemplos de procesos que han atraído la atención de amplios círculos de opinión a este debate.

Este trabajo tiene por objeto retomar dicha discusión desde una perspectiva filosófica y epistemológica, presentando de manera puntual los fundamentos técnicos e históricos de la IA y revisando las discusiones en torno a la pertinencia o no de establecer analogías o identidades entre las computadoras y el cerebro, o bien entre los programas (y el *software* en general) y la conciencia o la mente. Para ello se exponen aspectos clave de las teorías sobre la subjetividad en la filosofía y conceptos fundamentales como intencionalidad, autoconciencia e intersubjetividad, de acuerdo con diversos autores clásicos y especialistas contemporáneos, con objeto de mostrar la especificidad de la conceptualización filosófica de la conciencia y la relevancia y validez de su distinción respecto a la IA en contextos teóricos y argumentativos más extensos o menos especializados. El interés principal no es formular propuestas novedosas, sino articular y actualizar enfoques que se han visto desatendidos, como los de la fenomenología o el idealismo dialéctico, frente a la predominancia de otras perspectivas, como el funcionalismo o el conductismo, con lo que quizá podrían ser calificados como novedosos.

2. Delimitación de conceptos

Si bien en un nivel psicológico especializado podría distinguirse entre conceptos como mente, conciencia, conocimiento, pensamiento o inteligencia, para efectos de este trabajo, entenderemos conciencia en un sentido amplio, es decir abarcando estos otros términos, pues dicho significado amplio es relevante para la delimitación que buscamos mostrar aquí entre la inteligencia humana y la inteligencia de las computadoras. El término "conciencia", aplicado a todos estos fenómenos mentales, cuenta en todo caso con una mayor tradición dentro de la filosofía y quizá de la psicología (Ryle 2009, pp. 138-145).

Por otra parte, independientemente de las actividades humanas físicas que han sido reemplazadas gradualmente con máquinas desde inicios de la industrialización y que pertenecerían más al campo de la robótica, como por ejemplo, el movimiento y ensamblaje de objetos, el trabajo manual y agrícola, los robots de servicio, etcétera, la IA en sentido estricto se asocia más con actividades "intelectuales" exclusivas del ser humano, como el razonamiento, el procesamiento complejo de información e ideas, la creación de obras de arte, la manifestación de emociones, la toma de decisiones, etcétera. En este último contexto es en el que cabría la pregunta de si la IA es capaz de reproducir o simular la conciencia o, más coloquialmente, si las computadoras son capaces de pensar.

La respuesta más sensata sería definir primero qué se entiende por "pensar": si pensamiento es la resolución de problemas matemáticos o técnicos de diversa índole específica, es claro que las computadoras piensan y en muchos casos, mejor que los seres humanos. Igualmente, frente a actividades como responder a preguntas proporcionando respuestas e información más o menos experta e incluso simulando la voz y el oído humanos, también cabría responder que las computadoras piensan. Incluso si se plantea la pregunta sobre si las computadoras "piensan que piensan", algunos responderían afirmativamente, destacando que precisamente la IA es una tecnología que permite que los algoritmos elaboren sus propios algoritmos y aprendan en el proceso, a partir de las experiencias y los datos recolectados por los sistemas del medio ambiente con sensores, por ejemplo, por medio del llamado aprendizaje profundo (*deep learning*). En este sentido existe actualmente diversidad de aplicaciones de la IA que llevan a cabo tareas y "acciones" con un nivel de sofisticación que hasta hace poco se consideraba netamente humano. Al respecto abundan los ejemplos: programas para ejecutar juegos de ajedrez, modelos extensos de lenguaje (los llamados LLMs), diagnósticos médicos, modelos climáticos, reconocimiento y procesamiento de imágenes, traductores, operaciones médicas, auto movilidad, operaciones militares, asesoría financiera, pronósticos económicos, etcétera.

En este punto sería pertinente distinguir lo que los especialistas definen como IA "débil" y IA "fuerte". En el primer caso hablaríamos de programas capaces de ejecutar actividades y resolver problemas *específicos* similares a los ejecutados y resueltos por seres humanos, mientras que en el segundo caso se hablaría de una capacidad *general* de reemplazar al intelecto humano en sus diversas actividades más complejas.

Para zanjar esta cuestión, hasta ahora más hipotética que real, se emplea con frecuencia como criterio la llamada "prueba de Turing". Si bien la formulación original del argumento de Turing era restringida a la de la ejecución de las reglas de un "juego de imitación" que planteó originalmente en su artículo de 1950 (Turing, 1950, p. 439), la prueba de Turing se entiende actualmente en círculos de discusión filosófica y computacional de manera más amplia y consiste en afirmar que si una persona que interactúa con una máquina y con otra persona, sin saber dicha diferencia, considera que las acciones y respuestas de la primera son indistinguibles de la segunda, puede atribuirse a la máquina en cuestión el carácter de consciente o de inteligente.

Como se puede apreciar, esta prueba recurre a un criterio netamente funcionalista que ha sido rebatido, argumentando que efectos similares no suponen una misma causa (Searle 1980, p. 6). No obstante, si bien aún no se han inventado sistemas de IA estrictamente "fuertes", la existencia de cada vez más sofisticadas e integradas aplicaciones de IA muestra la complejidad creciente de la pregunta sobre si la IA es consciente o no.

3. Antecedentes históricos y tecnológicos

Históricamente, la IA se asocia con las tecnologías para la construcción de calculadoras y computadoras y las teorías de la llamada cibernética. Las aspiraciones de Leibniz de crear máquinas o lenguajes universales capaces de razonar a la manera del ser humano y el planteamiento de posibles sistemas numéricos distintos al decimal, como los binarios por Juan Caramuel (Pérez Martínez 2022, p. 110), son referencias importantes para la teorización y fundación de tecnologías que posibilitaron la automatización de la computación y diferentes formas de inferencia lógica basadas en ella. Norman Wiener, uno de los inventores del término "cibernética", entendía a ésta como una ciencia en la que no solo la computación, sino también la comunicación y el control formaban una teoría integral aplicable tanto a las máquinas como a los animales, entre cuyas características esenciales se incluía el concepto de "retroalimentación" (Wiener 1985, p. 11).

Con la cibernética y sus ramificaciones, así como logros tecnológicos, se hace patente la concepción filosófica y científica de la mente o el pensamiento como máquina, algoritmo o, en términos más abstractos, como función matemática, con su esquema tripartita esencial de "insumo (*input*)-estructura (o proceso)-resultado (*output*)", latente en las filosofías racionalistas vinculadas con el pensamiento formal y matemático desde el Racionalismo del S. XVII y convertidas en programa de investigación en el positivismo lógico y la filosofía analítica.

En su artículo seminal *A Symbolic Analysis of Relay and Switching Circuits*, Claude Shannon (1938, p. 471) mostró cómo era posible representar ecuaciones matemáticas y proposiciones e inferencias de la lógica formal booleana por medio de circuitos basados en relevadores electromagnéticos, que posteriormente serían reemplazados por bulbos y transistores, posibilitando así el escalamiento de memorias y procesa-

dores artificiales con la manufactura de chips conteniendo literalmente millones de millones de circuitos y transistores.

Si bien en sus etapas iniciales la programación de aplicaciones computacionales se conceptualizó en términos de lógica proposicional y de predicados deterministas, el gradualmente más complejo empleo de cálculos probabilistas ha resultado más eficiente en las aplicaciones de IA generativa y es actualmente el estándar de sus algoritmos. De acuerdo con ello, los programadores crean una base de conocimiento con múltiples datos recolectados de expertos y de la minería realizada a través de los insumos proporcionados por los millones de usuarios de las aplicaciones en internet. Estos datos son incorporados por los programadores para crear una base de reglas que opera mediante probabilidades condicionadas y distribución de máxima entropía, cuando no se cuenta con un conocimiento pleno de las variables a ser consideradas (Ertel 2017, pp. 136-145). La programación humana se constituye aquí en un sistema de inferencias semánticas, si bien la programación desencadenada por la propia IA se desarrolla de manera automática. Para entender la evolución de la IA, no obstante, es importante conocer la base electromecánica de la computación y su fundamento lógico-matemático.

4. Sintaxis vs. Semántica

El diseño de programas y algoritmos basados en las tecnologías de semiconductores, los sistemas numéricos binarios, la lógica formal booleana y la probabilidad condicionada, tiene como característica el procesamiento de información basado en el cálculo automatizado. Este tipo de mecanización del pensamiento, no obstante, tiene una limitación esencial, consistente en la representación y expresión de lenguajes de manera exclusivamente sintáctica. El nificado de las expresiones en dichos lenguajes, si es que cabe el término "significado", se transmite mediante la manipulación física (o sea electromagnética) de las "señales" o marcas hechas en la memoria y las unidades de procesamiento, fabricadas actualmente en diversos medios materiales, pero principalmente en semiconductores.

Por ejemplo, para que los transistores en un circuito conviertan cargas y corrientes eléctricas en símbolos o comandos, se requiere que sean configurados en lo que se conoce como autómatas finitos o "máquinas de estados finitos" (Hwang, 2005, p. 4). Para que un circuito convierta un insumo (*input*) de una secuencia de unos y ceros en un resultado (*output*) de un pixel, una letra o letras (palabras, oraciones, párrafos, reglas, inferencias, etcétera), se requiere una secuencia determinada (en serie o paralelo, dependiendo si se usan los conectores "y" u "o" del algebra booleana) de almacenamiento o liberación de una carga o corriente eléctrica en la base de silicio semiconductor del transistor. Así, para que una computadora reconozca la presión de las teclas correspondientes y escriba en la pantalla la palabra "hola", se requiere que el input asociado con "h", por decir 01100100110 (secuencia de símbolos de un lenguaje máquina determinado) sea reconocido por el autómata (o sea la secuencia de transistores o *flip-flops* específica) diseñado para dicho símbolo, y que acepte (y no

rechace) dicho *input*. Podría ser, por ejemplo, una serie de tres transistores que solo acepten el estado finito compuesto de una secuencia de dos unos seguidos, eliminando o rechazando otras secuencias que no incluyan dos unos seguidos, al no poder ocupar o almacenar en la misma transición los dos últimos transistores de la secuencia del circuito. El "1" significaría "conservar la carga eléctrica" en el transistor y el "0" "liberar la corriente eléctrica". Con el resto de las letras de la palabra sucedería algo similar, al igual que con procesos de computación más complejos y escalados a circuitos con dimensiones de millones de millones de transistores y de millones de millones de insumos, comandos y resultados en lenguaje máquina. Así es cómo "piensan" las computadoras, de una manera muy simplificada.

Si se emplea la lógica proposicional como lenguaje formal de programación, es decir como interface entre una persona y una computadora, esta barrera semántica se constata en la necesidad de definir la significación, o sea la relación entre un signo y su significado, como un cálculo automático de tablas de verdad, que puede tener solo dos valores de verdad o más, dependiendo de si se trata de lógica tradicional booleana, modal o *fuzzy logic*. La combinación correcta o válida de estos signos o valores, predefinida como tal, es el criterio semántico básico utilizado por la máquina, o sea una reducción a sus posibilidades sintácticas. Dos fórmulas o proposiciones son semánticamente equivalentes, si tienen los mismos valores de verdad en todas las interpretaciones (Ertel 2017, p. 25), pero la interpretación de las proposiciones aisladas es algo externo a la combinación, que es lo único que puede verificar la máquina al operar. Los valores de verdad asignados a las proposiciones son condición del cálculo y no su resultado. Para la interpretación, el lenguaje formal de dicho cálculo requiere el meta lenguaje natural del programador y del usuario de la computadora, que son los que conocen el significado de los signos más allá de su combinación y los que finalmente interpretan dichos signos, tanto como *inputs* o como *outputs*. Es decir, el proceso "consciente" se da entre las personas que interactúan o se comunican por medio de la computadora, no en la computadora misma.

Esta limitación no es superada tampoco empleando un cálculo probabilístico o una probabilidad condicionada basada en el Teorema de Bayes o en la llamada *fuzzy logic*, más amplia que la lógica proposicional o de predicados, que se fundamenta en el cálculo decidible y completo de las tablas de verdad con valores binarios. La capacidad de la IA de aprender y elaborar algoritmos de manera autónoma a partir de nuevos insumos y datos tampoco rebasa esta barrera semántica, aunque implique procesos combinatorios más complejos y una retroalimentación permanente, así como una comunicación continua con el medio ambiente de la máquina, por medio de sensores. Ello se constata en distintos tipos de IA generativa que procesan *inputs* de manera específica como, por ejemplo, redes neuronales contrapuestas, es decir generadoras y discriminadoras (las llamadas GANs); modelos transformadores a partir del reconocimiento de largas secuencias de signos (como ChatGPT); redes neuronales con reforzamiento de aprendizaje a partir del uso de pesos (*weights*) y rutinas de entrenamiento, y otros tipos de aplicaciones. Por otra parte, tampoco parecería ser superada dicha limitación mediante la utilización de computadoras cuánticas, en el caso

de que pudieran ser fabricadas de manera confiable y con capacidades de escalamiento en el futuro, pues la llamada "supremacía cuántica" se refiere más a una potenciación de la capacidad de cálculo (Roth 2024, p. 196) que a una diferenciación cualitativa del procesamiento de la información, aun considerando a los llamados "*qubits*" como espectros continuos de codificación de datos antes que pautas discretas de almacenamiento y procesamiento, como los transistores. Y es que el problema de fondo no es la dualidad de estados físicos de los circuitos en cero o en uno (o estados intermedios), sino la limitación de una máquina para interpretar semánticamente el significado de un signo.

5. Subjetividad

A partir de esta explicación resumida sobre la base material y el modo de operación de las computadoras y de los sistemas de procesamiento de información y datos, podemos establecer que la distinción esencial entre la conciencia y la IA se encuentra en lo que conocemos como *subjetividad*. Aquí es donde la pregunta no sería si las computadoras piensan, o si piensan que piensan, sino más bien si están conscientes de que piensan. Si bien el concepto de subjetividad tiene una larga tradición en la filosofía y particularmente en la epistemología, para efectos de este trabajo, bastará destacar dos características esenciales de la misma, relevantes para ilustrar la diferencia entre la IA y la conciencia: la intencionalidad y los llamados "qualia".

En su famoso experimento mental de "el cuarto chino", el filósofo estadounidense Searle procura demostrar que la IA no puede llevar a cabo procesos conscientes. Para ello elabora una argumentación con la presentación de un contraejemplo en el que, en sentido contrario a la argumentación tradicional de la cibernética, un ser humano puede imitar a una computadora, realizando actividades mecánicas aparentemente conscientes, pero que en el fondo son una mera combinación sintáctica de signos cuyo significado o dimensión semántica, al no ser conocidos en absoluto por dicho ser humano, son irrelevantes.

Imaginemos a una persona encerrada en un cuarto que no tiene el menor conocimiento del idioma chino y a la que se le proporciona un documento escrito en dicho idioma, junto con un conjunto de reglas en su propio idioma para correlacionar los signos del primer documento con estas reglas. Posteriormente se le proporcionan otras instrucciones por escrito en su propio idioma en las que se le indica qué signos chinos proporcionar en respuesta a estas instrucciones. Quienes proporcionan estos documentos llaman al primer documento "texto", al segundo documento "programa" y al tercer grupo "preguntas". El documento que la persona entrega después de correlacionar los signos del primero con el segundo y de éste con el tercero es llamado "respuestas" (Searle 1980, p. 3). Si se proporcionan las preguntas y las respuestas a un hablante de chino, dirá que la persona que respondió dichas preguntas entiende dicho idioma. No obstante, es claro, como afirma Searle, que lo único que hizo la primera persona fue manipular símbolos formales sin ninguna interpretación, o sea sin ningún conocimiento del significado de los caracteres chinos.

Searle afirma que, como en el caso de la persona en el cuarto chino, la máquina, alimentada por un programa o un algoritmo, produce resultados, pero es incapaz de crear estados mentales o significados pues carece de intencionalidad. Esta intencionalidad la explica como el efecto específico y distintivo de la actividad del cerebro que es un órgano constituido de manera fundamentalmente distinta a la constitución de una computadora, por más sofisticada que ésta sea. Searle concluye que tratar de simular la conciencia no es replicarla, pese a que la computación y el razonamiento humano produzcan aparentemente resultados similares por ser ideales o simbólicos. El argumento del cuarto chino de Searle ha sido ampliamente discutido, tanto en círculos de filósofos, como de especialistas en computación, y sus conclusiones son respaldadas por diversos autores (Bishop 2013, p.28).

Por otra parte, el concepto de intencionalidad tiene una larga tradición en la filosofía, desde el pensamiento escolástico y racionalista hasta la psicología experimental, pasando por la filosofía fenomenológica de Husserl y el existencialismo de Heidegger y Sartre. En el caso de la filosofía de Heidegger, por ejemplo, la intencionalidad se vincula con su intento por crear una teoría metafísica en la que las categorías tradicionales de sujeto y objeto sean superadas para establecer una relación primigenia entre la existencia humana individual ("Dasein") y el "ser" mediante lo que él denomina "suceso" (*Ereignis*) (Heidegger 1989, p. 11).

En el contexto de la Fenomenología de Husserl, la intencionalidad es "acto" de la conciencia y al mismo tiempo "vivencia" (*Erlebnis*) (Husserl 1913, p. 344). La intencionalidad no es una capacidad de la conciencia entre otras, sino su característica fundamental, que consiste en su correlación esencial con un objeto. La conciencia siempre es conciencia de algo y solo a partir de esta correlación es que puede hablarse de *evidencia*, tanto en un sentido empírico como ideal (Husserl, 1998, p. 24). Se trata de una especie de tensión entre dos polos, mismos que son inseparables, pero, al mismo tiempo, irreducibles entre sí. Este proceso, por medio del cual *aparecen* a la conciencia significados, forma parte de la especificidad de la subjetividad y de la dificultad para modelarla o reproducirla en procesos como los de la IA.

Esta evidencia de significados muestra también la irreductibilidad de la intencionalidad a la posesión de estados mentales, entendidos como meros substratos neuronales o electromagnéticos. En términos semióticos, para regresar al planteamiento sobre la dimensión semántica de la conciencia, dicha irreductibilidad correspondería a la imposibilidad de reducir un significado a un signo, entendido éste como un objeto exclusivamente empírico. Aquí cabría referirse a lo que Umberto Eco denominaba "infinitud" de la circularidad semiótica, es decir, el proceso por el cual signo y significado intercambian sus papeles continuamente, en el contexto de la actividad humana y cultural de la interpretación (Eco 1994, p. 85).

Por otra parte, los enfoques de tipo conductista, particularmente en su variante funcionalista, explican a la intencionalidad como la capacidad de tener "actitudes proposicionales", es decir disposiciones mentales sobre ciertas proposiciones (en sentido lógico) acerca del mundo, en las que el sujeto pensante, al creer en una cierta proposición (es decir, al suponer que es cierta) podría actuar de cierta manera (Chalmers

1996, p. 19). La mente, de alguna forma, procesa un determinado insumo, y produce alguna conducta (aún en potencia), de la misma forma en que una máquina o un algoritmo realizan una función o producen un resultado. No obstante, dicha explicación en realidad no agota el elemento semántico de la cuestión. Por ejemplo, las actitudes proposicionales frente a conceptos como "Apocalipsis" o bien "hoyo negro" serían muy probablemente las de una conducta de huida (potencial, desde luego). Sin embargo, las definiciones, explicaciones o interpretaciones de ambos conceptos o creencias son muy distintas y no pueden ser reducidas a unas mismas o similares actitudes proposicionales.

De igual manera, otros filósofos funcionalistas como Dennett, explican dicha intencionalidad en términos de disposiciones psicológicas o "comportamientos verbales" (Dennett 1991, p. 77) con propósitos conductuales específicos, pero restringidos a sujetos observables a los que se pueda "atribuir" de manera estrictamente objetiva dicha capacidad. Es decir, la introspección, herramienta importante para explorar los contenidos de la propia subjetividad y compararlos con los de otras subjetividades, no sería necesaria para caracterizar a la intencionalidad, como Denett la entiende. No obstante, en este enfoque el punto central relativo a los contenidos semánticos mentales y sobre todo su infinidad de variedades y significados -sean vistos de manera subjetiva u objetiva-, resultaría también soslayado.

Como una extensión de este tipo de discusiones rebasa el enfoque de nuestra presente aproximación, por intencionalidad entenderemos, de manera amplia y no estrictamente en el sentido de Husserl, la capacidad de la conciencia de estar dirigida a objetos o de representarlos. Para no caer en una discusión de carácter ontológico sobre la referencia o el sentido extra mental de dichos objetos, o bien en una discusión de carácter psicológico sobre la intencionalidad como motivación de las acciones y deseos, bastará que digamos aquí que intencionalidad es la capacidad de captar o crear significados. Ello además va más en conformidad con la actual discusión sobre la significación sintáctica o semántica a nivel de la IA y la conciencia, sin necesidad de adentrarse en una consideración de la totalidad de la experiencia consciente o de los correlatos ontológicos de la conciencia.

El otro elemento esencial y distintivo de la conciencia a destacar son los llamados "qualia", que se definen como el carácter subjetivo de las experiencias. ¿Cómo puede por ejemplo describirse el sabor del chocolate o el aroma de la vainilla? Podría responderse enunciando y comparando ciertas características como la dulzura, amargura, irritabilidad, etcétera, o bien se podrían explicar los componentes y estructuras de las moléculas de los objetos que provocan dichas sensaciones. No obstante, la vivencia subjetiva que cada individuo tiene al respecto es fundamentalmente intransmisible e inefable.

De hecho, el carácter subjetivo de las experiencias no puede ser explicado en términos de otras experiencias, pues ello supondría adscribirles una realidad objetiva, que, por definición no tienen. Esta subjetividad extrema puede ser aludida un tanto metafóricamente como un punto de vista único. "Si el carácter subjetivo de la experiencia es completamente comprensible solo desde un punto de vista, entonces cualquier

cambio hacia una mayor objetividad —o sea menor vinculación a un punto de vista específico— no nos acerca a la verdadera naturaleza del fenómeno: nos aleja aún más" (Nagel 1974, p. 445).

El concepto de "qualia" ha sido cuestionado por diversas corrientes de las filosofías conductistas y funcionalistas, argumentando que dicho carácter subjetivo de las experiencias no puede ser investigado por las ciencias experimentales y basadas en criterios objetivos (Dennett 1991, p. 70). Ante la imposibilidad de eliminar estas experiencias subjetivas inherentes a toda conciencia, que no solo podrían llamarse intuitivas, sino incluso evidentes, se han planteado como alternativas tanto en las ciencias cognitivas como en la filosofía de la mente las llamadas Teorías de la Mente (TM). Una TM postula en esencia que la conciencia solo puede ser atribuida a un sujeto observado a partir de sus conductas observables: El observador desarrolla su teoría a partir de estas observaciones sin asumir conceptos sobre la subjetividad del observado; por otra parte, se postula también que el observado tiene su propia TM sobre otros sujetos, para poder interactuar con ellos (también a través de sus mutuas observaciones).

Con ello, no obstante, estaríamos frente a un dilema metodológico, relacionado con lo que en la psicología social se conoce como el problema de la "atribución": no involucrarse con los observados y, al mismo tiempo, involucrarse con los mismos (Malle, 2022, p.93). Por evitar el concepto de subjetividad, esta variante del conductismo caería en la inconsistencia de suponer que el observador es ajeno al objeto de estudio (o sea que es "objetivo"), pero que al mismo tiempo es subjetivo, en tanto sus observaciones, como interacciones, son parte de lo observado (que lo afectan como sujeto observador y sujeto interactuante, es decir eventualmente observado). En todo caso, atribuir al observador observado que tiene teorías y no certezas supondría tener la certeza de que no tiene certezas, lo cual sería una contradicción con la misma definición de "teoría".

6. Autoconciencia

La intencionalidad y los qualia están estrechamente vinculados con lo que sí podríamos llamar "certeza" subjetiva. Dicha certeza nos permite integrar los estados mentales en un principio unitario que conocemos como el "yo".[1] Este yo, que Kant identificó como el sujeto que acompaña todo juicio sintético a priori o "apercepción" (Kant 1998, p. 222), muestra también una capacidad de pensarse o referirse a sí mismo. De esta manera, la capacidad de captar significados implica también la posibilidad de captar la propia subjetividad, no solo como simple introspección sino como un objeto intencional. Por ello la certeza subjetiva es indisociable de la intencionalidad. Esta autorreferencia o "autointencionalidad" es parte esencial del pensamiento consciente, pero cabría preguntar si también puede ser parte de los procesos de IA.

[1] El "super yo" y el "ello" de las teorías psicoanalíticas serían comprendidos en el término "yo", pues para efectos de la presente discusión, se referirían finalmente a la mente o a lo que denominamos aquí subjetividad en sentido amplio.

Entendida como algoritmo de algoritmos o como metalenguaje de lenguajes, la IA puede ser también vista como una especie de autorreferencia. No obstante en este aspecto se distingue también fundamentalmente de la conciencia. Ello puede ser ilustrado mediante el famoso teorema de la completud de la lógica de predicados de primer orden de Gödel, así como con el aún más famoso teorema de la incompletud de las matemáticas. Sin entrar en detalles técnicos sobre ambos teoremas, baste decir que en el primer caso la jerarquización y separación clara de niveles semántico y sintáctico permite transmitir o "derivar" (Ertel 2017, p. 28) la consistencia y completud del nivel semántico al nivel sintáctico. Si una inferencia es válida o verdadera en términos semánticos, esta validez también será aplicable en términos sintácticos, que la máquina puede procesar, una vez definidos de manera correcta las reglas de inferencia y los valores de verdad. El algoritmo recibe un insumo y el cálculo de predicados con sus reglas de inferencia y definiciones produce de manera estrictamente sintáctica un resultado o una inferencia válida.

No sucede lo mismo en el segundo caso. En su famoso teorema de 1931, Gödel (1931, pp. 181-193) demuestra que tanto los sistemas axiomáticos de la Lógica matemática, más allá de la de predicados de primer orden, como los de la Teoría de Conjuntos y, por extensión cualquier sistema axiomático matemático, son incompletos e indecidibles. Esta conclusión se encuentra también vinculada con la formalización sintáctica de los lenguajes que se emplea para programar las aplicaciones de IA. Dicha formalización, a nivel de lenguaje, puede ser también llevada a cabo a nivel de metalenguaje, es decir, extendida del nivel sintáctico al nivel semántico. De esta manera se puede formalizar, por ejemplo, que se diga "esta oración es falsa", en referencia a esta misma oración. Si la oración es falsa, entonces es, por exclusión, verdadera; y si es verdadera, entonces (por su propio contenido definido) es falsa. La consecuencia de formalizar el contenido semántico es una contradicción explícita, que mina todo el sistema axiomático del que se deriva este tipo de proposiciones. Ello ha provocado el abandono por parte de matemáticos y filósofos de los intentos por identificar o reducir cualquier tipo de demostración semántica a una demostración sintáctica por medio de la aplicación mecánica o automática de un algoritmo.

Toda inferencia verdadera en sentido semántico es válida en sentido sintáctico, pero no toda inferencia valida en sentido sintáctico implica una inferencia verdadera en sentido semántico. La validez semántica es más amplia que la sintáctica y el esquema tripartita insumo-proceso-resultado resulta insuficiente en un razonamiento en el que buscamos determinar una "autointencionalidad" de carácter semántico. Para efectos de la instrumentación tecnológica, esta indecidibilidad e incompletud teórica o lógica es poco relevante pues el sistema de IA realiza tareas específicas o resuelve problemas de manera concreta y genera algoritmos, sin tener que verificar la completud o incompletud de todo su sistema o si llegará a un momento en que concluya el último de sus algoritmos. No obstante, para efectos de discernir si el algoritmo es consciente, es decir acompañado de una certeza subjetiva e intencional, capaz de comprender o generar la dimensión semántica del proceso, esta limitación es relevante.

La certeza subjetiva puede ser también entendida como conciencia de sí misma, es decir como autoconciencia, pero no en un sentido extensionalista como recursión, sino en un sentido de significación semántica. El concepto de autoconciencia tiene igualmente una tradición en la filosofía. Con Hegel, por ejemplo, dicha autoconciencia se explica como una relación, pero no algorítmica. En su *Fenomenología del Espíritu*, la autoconciencia se deriva de la conciencia en su proceso de continua oposición y asimilación respecto al objeto, cuyo dinamismo o "historia" abarca desde la certeza sensible, la percepción de una cosa concreta, el entendimiento de una fuerza detrás de un fenómeno y la inversión del intelecto, donde lo que aparenta ser un objeto resulta ser una ley, una generalidad teórica o una constante de la naturaleza, que a su vez se vuelve un objeto perceptible, por ejemplo con la revolución copernicana, donde el movimiento aparente del Sol y los astros que percibimos desde la Tierra se convierte en un modelo explicativo abstracto del pensamiento y al final éste en un objeto observable (por ejemplo por un astronauta, si es hipotéticamente capaz de ver al sistema solar desde la suficiente distancia . . .).

En esta continua dinámica entre el sujeto y el objeto, donde los papeles de uno y otro se intercambian, Hegel (1988) señala que "la conciencia de un otro, de un objeto en general, es necesariamente autoconciencia, ser reflejado en sí, conciencia de sí misma, en su ser otro" (p. 118). La intencionalidad de la conciencia es bajo esta perspectiva, indisociable de la certeza subjetiva y la representación de un objeto es esencialmente una facultad de la mente, pero al mismo tiempo un fenómeno de la misma. Se trata de una relación dialéctica donde el objeto deja de serlo para ser el sujeto y el sujeto a su vez se objetiva como fenómeno o como significado.

Esta dinámica de oposiciones, donde el principio de no contradicción de la lógica clásica binaria no es válido, es una relación inmediata. La certeza subjetiva es en este contexto intuitiva, pero al mismo tiempo intencional y no puede ser modelada por un esquema algorítmico tripartita del tipo insumo-proceso-resultado. Mientras en los algoritmos y las funciones matemáticas la autorreferencia se puede explicar como *recursividad*, o sea un proceso que se tiene a sí mismo como parte o insumo, el cual en principio se puede extender *ad infinitum*, en el caso de la autoconciencia la inmediatez de la relación de la conciencia consigo misma es una *infinitud cualitativa o actual*, es decir una distinción de lo indistinto o una identidad diferenciada (Hegel 1986, pp. 156-166).

7. Intersubjetividad y cultura

Hemos visto anteriormente que, de acuerdo con Searle, la intencionalidad de la conciencia es causada por el cerebro. La configuración biológica y la función de este órgano y del sistema nervioso por extensión, que hasta ahora no han sido comprendidas cabalmente y mucho menos identificadas con la configuración material o hardware de las computadoras, permite que haya "poderes causales" del cerebro sobre la conciencia.

Pero entonces, siguiendo a Searle, cabría la pregunta: ¿cómo es posible que la conciencia, que es una capacidad producida por un cerebro individual, se comunique o entienda con otras conciencias individuales? La respuesta obvia sería que dicha correspondencia se explica por la similitud biológica de los distintos cerebros de los sujetos que se comunican. Pero esta similitud es una propiedad abstracta, no solo de uno o dos cerebros tomados individualmente, por lo que se necesitaría un medio concreto para posibilitar esa comunicación. En este sentido, la intencionalidad, aunque pueda ser vista como sustentada en el cerebro a nivel individual, tiene que estar vinculada con la capacidad semiótica de la conciencia, es decir la capacidad para crear sistemas de signos y de significados, o sea lenguajes.

El lenguaje, y no los cerebros, serían el medio objetivo de comunicación entre las conciencias, lo cual no es una afirmación realmente sorprendente. Aquí es importante precisar que en el lenguaje la comunicación es transmisión de significados, pero también intersubjetividad, en un sentido no solo conductista sino sobre todo intencional, como lo plantea Husserl con su concepto de "función comunicativa" (*kundgebende Funktion*). Con esta función, el hablante tiene la intención de expresarse sobre algo y el receptor entiende esa intención, más allá del simple significado o contenido del mensaje. Como Husserl señala: "lo que hace posible el intercambio mental (*geistiger Verkehr*) y que el habla sea precisamente habla, se encuentra en la correlación transmitida físicamente, entre las vivencias físicas y psíquicas de las personas que se comunican entre sí" (Husserl, 1913, p.33).

Al ser transmitida físicamente, esta correlación puede ser percibida como una conducta, pero además implica una experiencia subjetiva compartida, con una entidad propia, más allá de la de los "qualia" a nivel individual. Estas vivencias comunes generan un mundo experimentado de manera colectiva, o como le llama el teórico social Habermas, con un término tomado también de Husserl, un "mundo vital" (*Lebenswelt*) (Habermas, 2014, p. 123). Los significados no solo se transmiten, sino que, sobre todo, se comparten. No hay porqué pensar que esta dimensión sea menos empírica que la de la biología o la de las neurociencias.

A continuación, cabría otra pregunta: ¿cómo puede haber una certeza de que el carácter subjetivo de las experiencias, lo que hemos denominado como *qualia*, es realmente compartido y comunicado? ¿cómo puedo saber yo que lo que llamo "café" para designar el color del chocolate o "amarillo" (o "blanco", si es natural) para designar el color de la vainilla, es igualmente experimentado por otra persona? En este caso no tenemos manera de discernir lo que es netamente subjetivo de lo que es objetivo, aunque queramos argumentar que a los colores corresponden ciertos espectros luminosos comprobables, lo cual sería solo desplazar el problema a otro nivel empírico. Para esclarecer esta cuestión debemos explorar otros aspectos de la conciencia, que tampoco son ni pueden ser compartidos por la IA y que algunos psicólogos y filósofos llaman *inteligencia emocional* y *reconocimiento*.

La inteligencia emocional, como la intencionalidad, no puede ser explicada de manera exhaustiva por criterios conductistas y por lo mismo tampoco atribuida a computadoras o robots que aparentemente muestren empatía, como llorar, cuando alguien

llora, o reír cuando alguien ríe. Algunos psicólogos experimentales han intentado explicar esta empatía a nivel biológico mediante lo que han denominado "neuronas espejo". Para ello han realizado experimentos a nivel neuronal y con sujetos experimentales, según los cuales, una persona, cuando observa a otra persona que muestra enojo o alegría, por ejemplo, mediante expresiones del rostro, experimentará de manera correspondiente enojo o alegría (Bauer 2016, pp. 25-60). Estos resultados mostrarían que hay una reacción subjetiva instintiva de comunicación o empatía entre los sujetos, o sea entre las conciencias.

No obstante, dicho planteamiento, si bien contribuye a un entendimiento de los "poderes causales" de la inteligencia emocional, no explica de manera adecuada su conexión con la intencionalidad. Si existen dichas neuronas "espejo", lo cual hasta ahora no ha sido plenamente comprendido, su existencia no podría explicar que dos sujetos entendieran lo que son los conceptos o significados de enojo o alegría a nivel abstracto. Esta objetividad de la comunicación solo es posible mediante el empleo de los lenguajes, la transmisión de significados y las vivencias compartidas. La empatía instintiva, que consiste en experimentar un sentimiento cuando alguien más lo experimenta no puede sustituir la capacidad semiótica e intencional de la conciencia, pues ésta es mucho más amplia, por ejemplo, cuando hablamos de conceptos abstractos o carentes de emocionalidad.

Por otra parte, el lenguaje, que es instrumento básico de la intersubjetividad, tampoco la explica completamente. Para ello requerimos el concepto de *reconocimiento*. Este puede ser explicado fenomenológica o conceptualmente, como una consecuencia natural de la autoconciencia a la manera en que lo hicieron filósofos idealistas como Hegel, cuyos planteamientos básicos se presentaron anteriormente y han sido retomados por teóricos sociales como Marx o Habermas. De acuerdo con este último, el reconocimiento es una consecuencia de la subjetividad, en tanto reflexiona sobre sí misma e identifica otras autoconciencias en esta reflexión. En alusión al concepto hegeliano de *espíritu*, que tiene una connotación claramente histórica y social, además de epistemológica, Habermas (1969) afirma que "la conciencia existe como el centro en el que los sujetos se encuentran, de tal manera que sin ello no pueden existir como sujetos" (p. 13). Gracias a que un sujeto identifica su propia subjetividad es que puede identificar o reconocer otros sujetos como él, para vincularse con ellos en una suerte de solidaridad que rebasa la mera comunicación y el entendimiento que proporciona el lenguaje.

El reconocimiento también puede ser explicado mediante teorías genéticas culturales. En este sentido sería una facultad subjetiva aprendida o histórica, a diferencia de la intencionalidad. Con ello llegaríamos a un criterio de objetividad o de comunicabilidad de la conciencia en el que sus aspectos natural y cultural son indisociables. El reconocimiento permitiría alcanzar un nivel más profundo de intersubjetividad que el lenguaje y en este sentido se vincularía como empatía *aprendida* con la inteligencia emocional y su carácter moral o ético.

Como cultura, la subjetividad y la intersubjetividad se objetivan en las obras y acciones humanas (incluidas las computadoras), sistemas de conocimientos y creencias,

valores, normas y en los lenguajes orales o escritos, que podemos llamar memoria cultural. Como conciencia colectiva, la intersubjetividad posibilita la transición de un "yo" a un "nosotros", que aglutina una base común de experiencias, reconocimiento e identidad colectiva. En ese sentido, el sociólogo francés Maurice Halbwachs considera que el término "memoria colectiva" debe entenderse en sentido literal y no metafórico, pues "no hay memoria posible fuera de aquellos marcos de referencia de los que los seres humanos que viven en una sociedad se sirven, para fijar y reencontrar sus recuerdos" (Assmann 2018, p. 35). Al respecto podríamos añadir que a nivel colectivo (e incluso individual) tampoco hay conciencia, si no hay memoria que la sustente.

Por otra parte, puede argumentarse que la cultura es un sistema de significados y que el significado se construye mediante la interacción entre los humanos y las tecnologías, incluida la IA (Coeckelbergh, 2024, p. 2228). No obstante, ello ocurre precisamente en contextos culturales, es decir intersubjetivos y entendidos como colectividades de conciencias, en las cuales no habría manera de distinguir que interpretaciones o generaciones semánticas corresponderían exclusivamente a las computadoras o aplicaciones de IA "per se".

Una computadora o un LLM pueden ser vistos como obras culturales, en el sentido manejado anteriormente, es decir como la objetivación de una subjetividad. Pero no habría un criterio claro para identificar las contribuciones independientes de dichos artefactos a la dimensión cultural del fenómeno, como lo hay al pensar que la intersubjetividad existía antes que las computadoras. La especificidad de las aplicaciones de IA estaría más vinculada a su carácter de herramientas o medios y como constelaciones semánticas "autónomas", sería difícil, por no decir imposible, concebirlas sin la intervención de las conciencias humanas, en tanto desarrolladoras o usuarias de dichas herramientas.

Aquí es importante distinguir entre obras culturales como tales y obras artificiales: Una computadora, como otras herramientas, es una obra artificial en tanto es construida deliberadamente para realizar una cierta función o funciones (en tanto cosa, con una causa humana), pero solo es cultural en la medida en la que se la interpreta o se le atribuye un significado (en tanto signo), que puede estar o no vinculado con dicha función. Pero ese significado es encontrado o generado por un sujeto, individual o colectivo. La computadora o la aplicación de IA no interactúa de manera intersubjetiva y no genera por sí misma un significado, pues carece precisamente de subjetividad.

Llegamos aquí al punto en que nos podemos preguntar si el dualismo básico de la existencia del *hardware* y el *software* de la IA tiene un equivalente en un dualismo cuerpo-mente. El hecho de que el cerebro sea causa de la intencionalidad de la conciencia y las obras objetivas de la cultura el medio y soporte de la conciencia colectiva, no nos permite concluir que la conciencia se reduce al cerebro y la intersubjetividad a una cultura determinada. Tampoco sería procedente una reductibilidad en sentido inverso. De hecho, la cultura, incluido el concepto de "cerebro" y la metodología para estudiarlo, existe más allá de la existencia individual de las conciencias pensantes, consideradas individualmente, pero son las conciencias individuales las que la piensan, recrean y desarrollan. Vemos en todo caso que existe una interrelación o una

influencia mutua entre ambas instancias, por lo que un dualismo de sustancias o principios separados y coincidentes solo por una especie de armonía preestablecida para la conciencia y el cuerpo o para la intersubjetividad y la cultura tampoco es justificable. La cultura y el conocimiento no son un tipo de software que pueda ser instalado o reemplazado en distintos cerebros, sino que son un proceso de apropiación y desarrollo llevado a cabo por la intencionalidad de cada sujeto en su biografía y por la colectividad de sujetos en su historia.

En el caso de la IA dicho dualismo es un requerimiento de su existencia, aunque el orden de causalidad pueda ser alternado, por ejemplo, cuando un programa es diseñado para un tipo de computadora o cuando una computadora es fabricada para cumplir con ciertos requisitos de software.

8. Conclusión y perspectivas

Es frecuente encontrar en las ciencias cognitivas actuales y la *Philosophy of Mind* la afirmación de que la mente se define e investiga a nivel psicológico como la causa de comportamientos y acciones, entre los que se encuentran manifestaciones lingüísticas o actos del lenguaje. Generalmente dichas afirmaciones van acompañadas de argumentos en el sentido de que lo distintivo y exclusivo de la conciencia humana es el carácter subjetivo de las experiencias (Chalmers 1996, pp. 104-106) y que la simple posesión de significados o conceptos no implica la experiencia subjetiva de los mismos, por lo que una computadora, en su versión de IA "fuerte", puede llegar a poseer en su memoria dichos significados y por ello disponer de una intencionalidad y de una dimensión semántica. Con ello se pretende demostrar la objetividad y carácter empírico de la mente y por tanto la posibilidad de replicarla mediante la IA.

No obstante, este argumento acude a una distinción arbitraria y *ad hoc* entre mente y conciencia, entre el carácter subjetivo de las experiencias y el objetivo de los significados y la intencionalidad. Pero ¿hay manera de distinguir realmente la posesión de un significado y su experiencia subjetiva? Parecería que no, a la luz de lo que se ha procurado mostrar en este trabajo. El mero criterio funcionalista o conductista de la exterioridad del comportamiento o de los signos del lenguaje, en principio replicables mediante la IA, no resuelve de una manera fundamental el problema de la causalidad y características esenciales de la conciencia.

Fundamentar esta última afirmación ha sido el propósito de este trabajo. Es en este contexto que hemos iniciado con una descripción y explicación de los principales rasgos del diseño y ejecución de computadoras y procesadores de datos, utilizando la lógica matemática, la probabilidad condicionada y las tecnologías de máquinas de estados finitos basadas en transistores y otros medios físicos. Dichas tecnologías cumplen fines específicos en las aplicaciones de la IA, pero el término ha adquirido un uso inflacionario, que se ha extendido a la expectativa de crear modelos capaces de simular o incluso replicar a la conciencia humana en un sentido amplio, por ejemplo, con la IA "fuerte". Esta expectativa supone en buena medida que la conciencia es un producto del cerebro y que, si se puede establecer una analogía sólida entre el

funcionamiento de éste y el de las computadoras, eventualmente éstas podrán replicar fenómenos o capacidades esenciales de la conciencia.

Sin embargo, estos enfoques evitan sistemáticamente el empleo del concepto de subjetividad. En este trabajo se ha buscado mostrar que no es posible abordar el concepto de conciencia o mente sin el de subjetividad, entendida y discutida en una larga tradición filosófica que llega hasta el presente, con distintas variaciones y enfoques, también alternativos a los del conductismo y el funcionalismo. Así, se ha procurado explicar características esenciales de la subjetividad.

Para ello, hemos establecido que la conciencia en sentido amplio no solo es intencionalidad o "qualia", sino también certeza subjetiva, que permite unificar ambas capacidades y experiencias en un "yo", que se puede reconocer a sí mismo y a otros sujetos. Pero aquí es importante destacar que la conciencia no es solo un yo sino una relación consigo misma y con el mundo. Esto puede plantearse tanto de manera fenomenológica como dialéctica, aunque ambos enfoques puedan ser considerados como excluyentes por algunos. La conciencia se relaciona con el mundo de una manera inmediata y no solo como un vínculo ontológico general o trascendental: su naturaleza misma es esta relación; es un sujeto objetivado y un objeto subjetivado, pero no simplemente como una relación simétrica. Se trata de un proceso donde existe un momento de correlación, otro de tensión, otro de irreductibilidad y otro de superación (Aufhebung). Es en este sentido inmanente que se plantea cómo se pueden vincular la posesión de significados, la intencionalidad y el carácter subjetivo de las experiencias (qualia).

Se ha buscado mostrar este proceso mediante una estructuración de enfoques: semiótico, para la dimensión semántica de la conciencia; fenomenológico, para la intencionalidad; y dialéctico para la autoconciencia. Dichos enfoques pueden ser vistos como complementarios para ilustrar aspectos de la conciencia que no pueden ser replicados por la IA. Se ha procurado también argumentar que estos enfoques muestran las limitaciones de los enfoques conductista y funcionalista, al intentar explicar a la conciencia exclusivamente como un fenómeno emergente del cerebro o de las redes neuronales naturales, que se buscan reproducir mediante diversos enfoques tecnológicos de la IA.

La conciencia no puede ser reducida a un producto del cerebro y no puede ser abstraída de la intersubjetividad. Ésta se manifiesta y reproduce como lenguaje y cultura. Si bien los poderes causales de la conciencia a nivel individual pueden ser referidos al cerebro, hemos buscado demostrar en este trabajo que la conciencia no se agota en el individuo, sino que, adicionalmente posee una dimensión colectiva, social e histórica, que, como cultura, le sirve de causa, soporte y medio. Ello no supone la existencia de una sustancia o algún principio metafísico ajeno o trascendente a la naturaleza o a la experiencia histórica. De hecho, en la conciencia misma comprobamos la imposibilidad de separar radicalmente naturaleza y cultura, o bien de querer reducir una a la otra. Estas características distintivas de la conciencia y esta irreductibilidad respecto al cerebro o procesos neuronales y electromagnéticos, muestran también las dificultades de su reducción a procesos computacionales como los de la IA, que ha empleado

como referencia desde sus inicios el análisis del funcionamiento del cerebro, como criterio básico de determinación de la conciencia humana y la mente. En un contexto más general y para concluir, podemos señalar que la IA no requiere ser equiparada con la conciencia para que sus distintas aplicaciones formen cada vez más parte de nuestra vida e interacción en sociedad. La IA es un instrumento y proporciona una diversidad de herramientas que pueden favorecernos o perjudicarnos, según el uso que les demos. Como en el caso de otras herramientas, lo determinante son las intenciones y fines que sus usuarios les adscriban. Son estos usuarios, tanto a nivel de propietarios, como de desarrolladores o consumidores de la IA, los que poseen una conciencia y un sentido moral (o ausencia del mismo) respecto a su uso. Está en nosotros y no en los sistemas de IA la opción. Más allá de visiones distópicas o de la ciencia ficción, la IA no es una fatalidad: nos ofrece ventajas y riesgos reales que podemos y debemos confrontar y moldear. La IA, por ejemplo, permite que los seres humanos dejen de llevar a cabo muchos trabajos peligrosos o tediosos y que puedan dedicarse a actividades más creativas o agradables. Esto representa una ventaja en sí, pero también el riesgo de desempleo para muchos. No obstante, ello puede ser contrarrestado mediante una distribución adecuada de los beneficios económicos y monetarios de este tipo de optimizaciones y está en la capacidad de los gobiernos y las sociedades regular una repartición más equitativa de dichos beneficios, especialmente en los países en desarrollo.

Igualmente, respecto a la garantización de las libertades de pensamiento y acción de las personas. La IA facilita la transportación y el acceso a información sobre servicios, productos y mercados a consumidores y empresas, pero ello no debe ser objeto de abuso ni explotación de datos e información privada de los usuarios, ni de violación de sus derechos de privacidad. Un riesgo asociado a esto mismo, es la utilización desproporcionada de datos de particulares por parte de empresas monopólicas en el ámbito de aplicaciones de IA y gobiernos para el espionaje o la manipulación ideológica o del comportamiento con fines políticos y comerciales. El riesgo de totalitarismos políticos y económicos derivados de la instrumentalización masiva de la IA debe ser también contrarrestado por sociedades y gobiernos capaces de proteger las garantías individuales y derechos de sus ciudadanos, con normas que, si bien no pueden anticiparse al ritmo de las innovaciones, sean por lo menos capaces de reaccionar oportunamente a las mismas.

Finalmente, si la utilización de la IA se ha extendido ya de una manera importante a los sistemas de guerra y defensa de varios países, no se debe cejar en el esfuerzo por limitar o incluso prohibir el desarrollo de aplicaciones para fines letales y sobre todo para permitir la autonomía de máquinas en la toma de decisiones sobre la vida o muerte de seres humanos.

Referencias

Assman, J., (2018), *Das kulturelle Gedächtnis. Schrift, Erinnerung und politische Identität in frühen Hochkulturen*, Verlag C.H. Beck, München.

Bauer, J., (2016), *Warum ich fühle, was du fühlst. Intuitive Kommunikation und das Geheimnis der Spiegelneurone*, Wilhelm Heyne Verlag, München.

Bishop, J. M., Nasuto, S. J., Coecke, B., (2013), "Quantum Linguistics' and Searle's Chinese Room Argument", en Müller, Vincent C. (Ed.), *Philosophy and Theory of Artificial Intelligence*, pp. 17-28, Springer, Berlin

Calvo, P., (2024), "Hiperética artificial: crítica a la colonización algorítmica de lo moral", *Revista de Filosofía*, Ediciones Complutense, Madrid, pp. 71-91. https://dx.doi.org/10.5209/resf.81655.

Coeckelbergh, M.,·David J. Gunkel, D. J., (2024), "ChatGPT: deconstructing the debate and moving it forward", *AI & SOCIETY* 39, pp. 2221–2231. https://doi.org/10.1007/s00146-023-01710-4

Chalmers, D. J., (1996), *The Conscious Mind, In Search of a Fundamental Theory*, Oxford University Press.

Dennett, D. C., (1991), *Consciousness Explained*, Back Bay Books, New York.

Eco, U., (1994), *Einführung in die Semiotik*, Wilhelm Fink Verlag, München.

Ertel, W., (2017), *Introduction to Artificial Intelligence*, Springer International Publishing, Switzerland.

Gödel, K., (1931), "Über formal unentscheidbare Sätze der 'Principia Mathematica' und verwandter Systeme I" , *Monatshefte für Mathematik und Physik* 38, pp. 173-198.

Habermas, J., (1969), *Technik und Wissenschaft als Ideologie*, Suhrkamp Verlag, Frankfurt am Main.

——, (2014), *Theorie des kommunikativen Handelns*, Band 1, Suhrkamp Verlag, Frankfurt am Main.

Heidegger, M., (1989), *Beiträge zur Philosophie (Vom Ereignis)*, Vittorio Klostermann, Frankfurt am Main.

Hegel, G. W. F., (1986), *Wissenschaft der Logik* I, Suhrkamp Verlag, Frankfurt am Main.

——, (1988), *Phänomenologie des Geistes*, Felix Meiner Verlag, Hamburg.

Husserl, E., (1913), *Logische Untersuchungen*, Zweiter Band, Max Niemayer, Halle a. d. Salle.

——, (1998), *Die phänomenologische Methode, Ausgewählte Texte I*, Reclam, Stuttgart.

Hwang, E. O., (2005), *Digital Logic and Microprocessor Design*, CL Engineering.

Kant, I., (1998), *Kritik der reinen Vernunft*, Felix Meiner Verlag, Hamburg.

Malle, B. F., (2022), Attribution Theories: How People Make Sense of Behavior, en Chadee, D. (Ed.), *Theories in Social Psychology*, (2nd edition, pp. 93-119), Wiley-Blackwell.

Nagel, T., (1974). "What Is It Like to Be a Bat?", *The Philosophical Review*, Vol. 83, No. 4, pp. 435-450, Duke University Press. (Disponible en http://www.jstor.org/stable/2183914)

Pérez, R., (2022), "La conquista de la aritmética binaria y la conquista espiritual de América según Juan Caramuel y Lobkowitz", *Revista Interpretatio*, 7.1, marzo-agosto 2022, UNAM, México, pp. 109-126. doi.org/10.19130/irh.2022.1.2701X46.

Roth, G., Tuggener, L. Tuukas, Roth, F. C., (2024). *Natürliche und künstliche Intelligenz, Ein kritischer Vergleich*, Springer, Berlin.

Ryle, G., (2009), *The Concept of Mind*, Routledge, New York.

Searle, J. R., (1980), "Minds, Brains, and Programs", *Behavioral and Brain Sciences* 3, pp. 417-457, Cambridge University Press, Cambridge Massachusetts.

Shannon, Claude, (1938), "A Symbolic Analysis of Relay and Switching Circuits", *Transactions American Institute of Electrical Engineers*, Vol. 57. Washington D. C.

Turing, A. M., (1950), "Computing Machinery and Intelligence", *Mind* 49, pp. 433-460

Wiener, N., (1985), *Cybernetics or Control and Communication in the Animal and the Machine*, The M. I. T. Press, Cambridge Massachusetts.