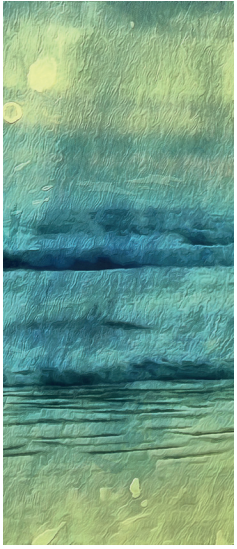
Autoconsciencia en la Inteligencia Artificial



Self–Consciousness in Artificial Intelligence

ERNESTO IZAR MARTÍNEZ*

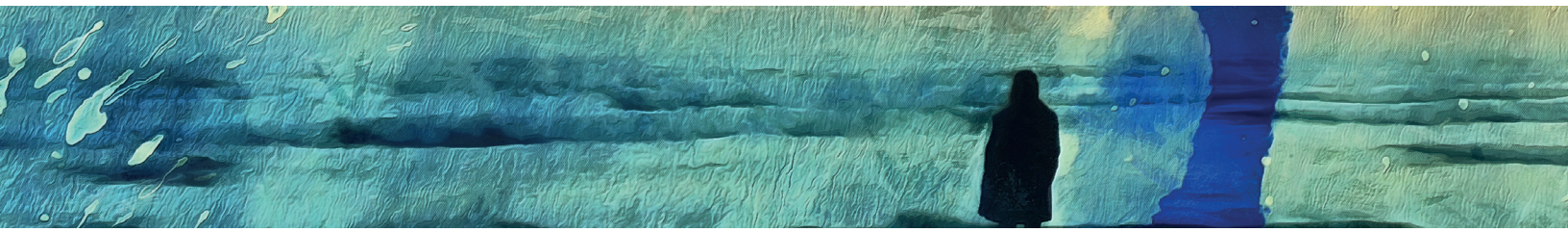
Resumen. Izar Martínez, Ernesto. *Autoconsciencia en la Inteligencia Artificial*. El presente texto es una exploración de la limitación epistemológica de las Inteligencias Artificiales. A partir del *cogito* cartesiano, se propone que éstas no pueden generar un pensamiento propio si no son capaces de cuestionar la información de la que son alimentadas. Después, con Hilary Putnam, se cuestiona si lo que dice la IA pudiera ser considerado como conocimiento, en tanto que no hay un contacto con el mundo sobre el cual basar dicho conocimiento. Finalmente, trato de concluir si una autoconsciencia sobre la cual generar conocimiento es posible en la IA. La reflexión final trata de poner en duda esa concepción de un futuro ficticio en el que la máquina será una cosa no tan distinta del humano.

Palabras clave:
Inteligencia Artificial, autoconsciencia, conocimiento, *cogito*

Abstract. *Izar Martínez, Ernesto.* Self–Consciousness in Artificial Intelligence. *This text explores the epistemological limitation of Artificial Intelligences. Taking the Cartesian cogito as a starting point, it argues that these intelligences cannot generate their own thinking if they are not capable of questioning the information they are fed. Then, following Hilary Putnam, it questions whether what AI says can be considered knowledge, inasmuch as there is no contact with the world on which to base such knowledge. Finally, I try to conclude, briefly, whether self–consciousness on which to generate knowledge is possible in AI. The final reflection seeks to raise doubts about the conception of a fictional future where machines will be almost indistinguishable from humans.*

Keywords:
Artificial Intelligence, self–consciousness, knowledge, cogito

Foto: Black_Morion, Depositphotos



* Estudiante del cuarto semestre de la Licenciatura en Filosofía y Ciencias Sociales en el Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO). Correo institucional: ernesto.izar@iteso.mx

La *Inteligencia* y su posible artificialidad

En los últimos años, parece imposible no girar la mirada hacia el rapidísimo avance de la Inteligencia Artificial (IA). Impresiona, a la vez que alarma, la facilidad que un modelo algorítmico tiene para generar razonamientos lógicos; impresiona aún más el hecho de que en ocasiones, al dialogar con una IA, da la impresión de que el interlocutor es una persona. La preocupación de que en un futuro la IA reemplazará toda actividad humana por su eficacia y vasto contenido es cada vez más fuerte. En efecto, los modelos de IA contienen toda la información que puede ser introducida en sus servidores; mucha más de la que un humano puede conocer en toda su vida. Sin embargo, vale la pena preguntarse si tal océano de información es, en realidad, conocimiento, o si sería mejor pensar en la IA como un motor de búsqueda no muy diferente de Google. En otras palabras, si la IA puede constituirse como un *ser pensante* y no sólo como algoritmo.

El problema de la inteligencia y la consciencia, evidentemente, no es nuevo en la filosofía. Tan revisado y estudiado a profundidad, parece un tema inagotable en tanto que, a fin de cuentas, da la impresión de que no hay una respuesta definitiva. En cierta medida, este artículo pretende señalar el problema de conceptualizar como *inteligencias* tecnologías generativas. Sin embargo, parece que eso que llamamos consciencia (y lo que ello signifique) es lo que separa al ser humano de otros animales: ¿cómo definir algo que en definitiva no es animal y que de cualquier modo parece capaz de generar conocimientos similares a los del ser humano?[1]

Permítaseme presentar una definición general de lo que es la IA como base de la cual partir. Esta definición, tan limitada y debatible como cualquier otra, no pretende cerrar la discusión sobre lo que la IA sea, sino sentar un piso común en el cual la entendamos; así pues, la Real Academia Española define *Inteligencia Artificial* como la "disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico".[2] A partir de esta generalísima definición podemos, ahora sí, discutir si es que la IA contiene esa *inteligencia* en un modo más complejo.

Además de esta definición enciclopédica, también me gustaría delimitar esos programas informáticos de los que la RAE habla: ChatGPT, Gemini, Copilot, entre otros más

[1] Considerando que el humano es en parte animal, pero se distancia de otros justamente por sus facultades intelectivas.
[2] Real Academia Española, "inteligencia artificial", https://dle.rae.es/inteligencia.
Consultado 16/II/2026.

que son ya inescapables. Más allá de discutir su historia y uso terminológico, que en los últimos años parece más una estrategia para presentar tecnologías de punta que sean más de punta que los de la competencia, la IA me intriga en tanto que concepto de *inteligencia*, pues, como dice José Ramón Casar Corredera, estos modelos en específico cuentan con "características indistinguibles de las que produciría un ser humano".[3] Esto, cuando menos, suena alarmante por una razón: con el acelerado avance de estos modelos, parece que esa indistinguibilidad no tardará mucho en identificar totalmente al humano y la máquina. O, tal vez, esto ni siquiera sea posible por las razones que expondré más adelante.

Para este primer asentamiento de la discusión, abordaré desde René Descartes y sus *Meditaciones Metafísicas*[4] el concepto de la autoconsciencia y el *cogito*. Esto, con la intención de hacer de la autoconsciencia un elemento clave para la identidad y el reconocimiento del propio *ser*. Me atrevo a argumentar desde ahora que este reconocimiento del *ser* constituye la diferencia primordial entre animal, humano y máquina. Con todo, el ser autoconsciente le confiere al sujeto un anclaje primordial a la realidad. Así, por ejemplo, yo no puedo jamás dudar de que *soy algo* que piensa; por lo tanto, mi intención es preguntar qué es aquello que pueda asegurarle a la máquina que, en efecto, *es algo más* que datos e información algorítmica.

La autoconsciencia en las Meditaciones Metafísicas de Descartes

Cuando Descartes se preguntaba si eso que concebía por realidad no era más que un sueño o una ilusión, en realidad se cuestionaba: ¿qué era lo que podía conocer?, ¿de qué estaba compuesto el conocimiento que él creía verdadero? O, mejor dicho, ¿cuál era la naturaleza de su propia inteligencia?, ¿cómo podía negar todo lo que le rodeaba? Y sin embargo, ¿le era imposible negarse a sí mismo?

Al poner en duda todo su entorno, Descartes reconoce que su conocimiento tiene un límite y, por tanto, no es perfecto. Si es capaz de poner algo en duda, es porque no hay un conocimiento del que pueda estar seguro. Podría incluso dudar de su propia existencia; mas este escepticismo siempre va acompañado de un enunciado del tipo "esto es una ilusión". Es decir, aquello que conoce por los sentidos bien puede ser un engaño, y que "nada cierto hay en el mundo".[5] Pero inmediatamente después de esto, Descartes se pregunta

[3] José Ramón Casar Corredera, "Inteligencia artificial generativa" en *Anales de la Real Academia de Doctores de España*, Vol. 8, No. 3, Madrid, 2023, p. 476.
[4] René Descartes, *Meditaciones metafísicas con objeciones y respuestas*, KRK, Oviedo, 2005.
[5] *Ibidem,* p. 142.

cómo es que genera tales pensamientos. Si nada cierto hay en el mundo, ¿también él es incierto? Se ha convencido de que lo corpóreo es, cuando menos, poco confiable; si este mundo corpóreo es accesible inmediatamente por lo sentidos, es lógico que se cuestione entonces sobre esa llave de entrada a la realidad. ¿Y qué podrá afirmar de sí mismo, si todo aquello que creía conocer resulta ya no tan confiable?: "ya estoy persuadido de que nada hay en el mundo; […] ¿y no estoy asimismo persuadido de que yo tampoco existo? Pues no: si yo estoy persuadido de algo, o meramente si pienso algo, es porque yo soy".[6]

De aquí la famosa sentencia del *cogito, ergo sum*. El *cogito* funciona como anclaje a una realidad confusa y engañosa. Esta formulación de la autoconsciencia sirve para afirmar que, aunque se llegase a un escepticismo radical en el que nada es válido, no se podría negar uno a sí mismo. El mero enunciado de "yo no existo" ya supone un *yo* que niega su propia existencia; Hilary Putnam lo explica así: "Por ejemplo, 'no existo' se autorrefuta si *soy yo* (para cualquier 'yo') quien lo pienso. De modo que uno puede estar seguro de que existe con sólo pensar en ello".[7] Más adelante volveré a este punto.

Con esta certeza que le otorga el *cogito*, la inteligencia humana se enfrenta al mundo como sabedora de que, si va a tener certeza de algo, es de sí misma. Entonces, la autoconsciencia es el primer conocimiento, el más cierto. Es así que el pensamiento reflexivo, ese que se piensa a sí mismo, le da al conocimiento una base más sólida; parece entonces que la inteligencia no es la diferencia específica de la humanidad para reconocerse distinta a los animales, sino, justamente, este autoconocimiento. Habrá que ver cómo se separa de la IA, si es que lo hace.

Autoconocimiento y referencia

Si, al igual que varios pensadores en la historia de la filosofía, pensamos en el humano como un alma que se sirve de una masa de cuerpo para moverse por el mundo,[8] es fácil caer en el error de despreciar los sentidos como un medio engañoso para obtener conocimientos. Decíamos que la consciencia de sí es la certeza de que *somos*; sin embargo, es también importante tener una noción de lo corpóreo para crear conocimientos. Hilary Putnam, a quien ya había mencionado anteriormente, se opone a esta idea de que con la

[6] *Ibidem*, p. 143.
[7] Hilary Putnam, *Razón, verdad e historia*, Tecnos, Madrid, 2006, p. 21.
[8] Esto, por lo menos, a la hora de querer pensar y alcanzar verdades inmutables o de mayor valor que las cotidianas. Véase en Platón, por ejemplo: "[…] ¿no te parece que, por entero […], la ocupación de tal individuo no se centra en el cuerpo, sino que, en cuanto puede, está apartado de éste, y, en cambio, está vuelto hacia el alma? […] ¿Es que no está claro, desde un principio, que el filósofo libera su alma al máximo de la vinculación con el cuerpo, muy a diferencia de los demás hombres?". Platón, *Diálogos III*, Gredos, Madrid, 1988, p. 41 (*Fedón,* 64e–65a).

pura razón podríamos generar conocimiento. Él utiliza la imagen de *cerebros en cubetas* para ejemplificar esta superioridad de la razón sobre el cuerpo.

Los *cerebros en cubetas* son, tal como se oye, un cerebro sin cuerpo que, sin embargo, siente; o, mejor dicho, tiene conexiones nerviosas que le comunican algo similar a una sensación. Imaginemos entonces que, como dice Putnam, somos un cerebro en una cubeta. Si bien, no habría manera de estar seguros de que lo somos, Putnam afirma que, en efecto, no lo somos. Suena contradictorio, pero él explica que imaginando un mundo en el que las personas fueran *cerebros en cubetas* y "pueden pensar y 'decir' cualquier palabra que nosotros pensemos o digamos, no pueden 'referirse' a lo que nosotros nos referimos. En particular, no pueden decir o pensar que son cerebros en una cubeta (incluso pensando '*somos cerebros en una cubeta*')".[9]

Imaginemos ahora que la IA es esta especie de *cerebro en una cubeta*. No en el sentido de que sea sentiente (esto supone su propio problema), sino como este recinto de información inacabable que no deja de evolucionar y acumular conocimiento. Aquí juega un papel importantísimo la *referencia*.

Para entenderlo mejor, propongo el siguiente caso: si pidiera a cualquier modelo de IA que genere una imagen de un árbol o una manzana, no tendría problema en hacerlo. Si pidiera que la perfeccione o que cambie el estilo, la iluminación de la imagen, el color de las hojas, etc., lo haría. Sin embargo, argumenta Putnam, "es cierto que la máquina puede platicar maravillosamente acerca del paisaje de Nueva Inglaterra, por ejemplo. Pero si tuviese enfrente una manzana, una montaña o una vaca, no podría reconocerlas".[10] Digamos, lo sensorial no es lo necesario para afirmar que hay inteligencia, sino el que haya una *referencia* a ese mundo exterior. Putnam hace hincapié en la *referencia*. Si estos *cerebros en una cubeta* (la IA) nos hablan de un árbol, podrá describirlo con el mayor detalle posible y con una precisión que nos ayude a imaginar un árbol real. Esto de lo que nos habla la máquina ¿a qué está referido? Uno podría decir que a los árboles, por supuesto. Pero esto lo hacemos en tanto que tenemos un conocimiento previo de lo que es un árbol y al escuchar la palabra la unimos con esa imagen. La IA no refiere a nada más que a lo que se le ha configurado a identificar como *árbol*. Putnam se pregunta: "[…] cuando sus verbalizaciones contienen la palabra 'árbol', ¿se refieren realmente a árboles? De forma más general: ¿acaso pueden referirse a objetos *externos*? (Como algo opuesto a los objetos que aparecen en la imagen producida por la maquinaria automática, por ejemplo)".[11]

[9] Hilary Putnam, "Cerebros en una cubeta"…, p. 21.
[10] *Ibidem*, p. 23.
[11] *Ibidem*, p. 25.

Así, Putnam argumenta en favor de una unión mente–cuerpo que posibilita el conocimiento. No se trata tan sólo de una teoría empirista, sino que, en el fondo, los conceptos generados por la razón no sirven de nada si no se sintetizan con una imagen traída por los sentidos, tal como lo dice Putnam con el concepto *árbol* y el objeto exterior *árbol*.

En términos muy generales, como lo pone Casar Corredera, la IA recibe información de manera supervisada o no supervisada:

> Esto [el aprendizaje supervisado] implica tener clasificados previa y correctamente los datos y tener 'anotadas' previamente las respuestas. El aprendizaje no supervisado, por el contrario, comprende un conjunto de técnicas que permiten a la arquitectura aprender, sin referencia expresa previa de qué es correcto y qué no, es decir, directamente de los datos disponibles, sin un supervisor expreso.[12]

Más aún, el *conocimiento* que la IA comunica no es un conocimiento nuevo o propio, como Mariana Méndez–Gallardo explica: "las máquinas obtienen 'conocimiento' a través de los seres humanos que codifican e insertan manualmente información, pero no son capaces de leer y escuchar sin supervisión alguna".[13]

Lo que me interesa denotar aquí es que la IA tiene predispuestos los datos que manejará más adelante; por lo tanto, ya sea que le pregunte qué es un árbol, o cualquier otra pregunta con más complejidad, la IA responderá conforme a aquello que se le ha alimentado en forma de *árbol*.

Esto quiere decir que la IA no puede cuestionar si lo que considera como *conocimiento* (aquellos códigos de información insertados manualmente) es, de hecho, verdadero. No puede preguntarse si está en un sueño o si el genio maligno está jugando con su mente, como hizo Descartes; tampoco puede, como propone Putnam, imaginar que es un *cerebro en una cubeta* y hacer el ejercicio para demostrar que, en efecto, no lo es. No puede, a menos que genere una autoconsciencia. Y hay que preguntarse, ¿es posible que lo haga en algún punto?

La autoconsciencia en la IA

Vuelvo ahora al *cogito* de Descartes. La certeza de que *yo soy* es suficiente para que ese *yo* pueda navegar por la incertidumbre del mundo. Digamos que, independientemente de lo que las cosas externas sean, siempre *soy yo* el más "real" entre ellas, y ellas siempre son en relación con el *yo*.

---

[12] José Ramón Casar Corredera, "Inteligencia artificial…", p. 479.
[13] Mariana Méndez–Gallardo, "La IA desde un pensamiento creativo, simbólico y humanista" en *Christus: revista de teología ciencias humanas y pastoral*, año 31, No. 846, Tlaquepaque, julio de 2024, pp. 10–14, p. 13.

¿Por qué digo esto? Porque la IA, al no diferenciar entre un *árbol*, una *pintura de un árbol* y un *yo*, no puede tener esa concepción de las cosas en torno a sí misma. Me explico: cuando recibe el concepto de *árbol* en sus procesos de alimentación, los recibe pasivamente. Es decir, no puede ni tiene jamás la oportunidad de redefinirse el concepto de *árbol*. Si no puede ni siquiera poner en duda el concepto más superficial y externo, no podrá enterarse, al modo cartesiano, de que *ella misma* es quien duda y refuta los datos que recibe. No logra esa interioridad del *cogito* que es consciente de que *yo dudo*. Así, no es consciente, ni siquiera de que existe. Pues, aunque se le diera el comando de decir "yo existo", no hay esa interiorización del pensamiento que se sabe existente.

Pero incluso el *cogito* depende de algo más que de sí mismo. No se basta a sí mismo para existir, pues entonces habría que preguntar cómo es que en primer lugar duda; en pocas palabras: ¿qué ha hecho tal inteligencia para distanciarse de otros seres?, ¿cómo es que logra hacer funciones más complejas que un mero procesamiento de información? Así pues, este *ser pensante* se constituye sólo a partir de un trascendente. "Hallar en el pensamiento el ser, sólo es posible si el pensamiento que piensa se identifica con el ser (es el caso del Absoluto) o si el ser ha sido puesto antes, por el Absoluto, en el pensamiento que piensa".[14]

Conclusión

A modo de conclusión, me gustaría aclarar que este trabajo es tan sólo un primer esbozo de lo que, espero, pueda resultar en una investigación mucho más extensa y elaborada. La discusión no acaba aquí y está lejos de hacerlo; al contrario, parece que con la inserción de la IA en toda herramienta tecnológica, el problema no hace más que acrecentarse. Lo que me importa señalar aquí es la distinción que parece irse borrando progresivamente entre el humano y la máquina o la invención. Si es posible que se borre por completo, o si pasará en un futuro, no tengo cómo saberlo. Sí creo, sin embargo, que, mientras no se dé esta "iluminación" por la cual la IA pueda cuestionarse la realidad, es altamente improbable. Pienso, por ejemplo, en *Blade Runner*, de Ridley Scott. En la película, los replicantes (cíborgs), en su mayoría, ignoran que son cíborgs. Viven creyendo que son humanos, y que los recuerdos que almacenan son vivencias pasadas. No creen, ni por un segundo, que tales recuerdos son datos que les fueron alimentados antes de "nacer". Sin embargo, sí tienen esta capacidad de cuestionarse si, en efecto, esa realidad que creen verídica es cierta. Pero no es hasta que ocurre ese primer cuestionamiento cuando realmente se dan cuenta de su existencia como robots, y no como humanos.

[14] Juan Pegueroles, *El pensamiento filosófico de san Agustín*, Labor, Barcelona, 1972, p. 45.

La IA, en algún futuro, podría llegar a un nivel de sofisticación tal que no se pueda diferenciar de un ser humano en tanto que genere razonamientos y respuestas intelectualmente competentes. ¿Es posible esto?, ¿puede entonces ocurrir esa primera duda que despierte a la IA como *ser pensante*? Pues, no es lo mismo que una IA procese un texto que diga "yo dudo", a que verdaderamente dude de su existencia. Ciertamente, la respuesta no está muy clara; sobre todo porque, como vimos, este "chispazo" de consciencia parece proceder de algo trascendente al humano. Quizá tan sólo nos queda esperar a ese milagro en el que la IA pueda constituirse como un *ser pensante* que se cuestiona a sí mismo como ser; ese "chispazo" que alguna vez tuvo también el ser humano y que, en el presente, está tratando de replicar.

Fuentes documentales

Casar Corredera, José Ramón, "Inteligencia artificial generativa" en *Anales de la Real Academia de Doctores de España*, Vol. 8, No. 3, Madrid, 2023, pp. 475–489.

Castañeda Vargas, Fredy Humberto, "La unión e interacción entre el alma y el cuerpo, y su posible inteligibilidad en Descartes" en *Xipe totek*, ITESO, año 33, Vol. II, No. 122, Tlaquepaque, enero de 2025, pp. 105–130.

Descartes, René, *Meditaciones metafísicas con objeciones y respuestas*, KRK, Oviedo, 2005.

Haraway, Donna, *Ciencia, cyborgs y mujeres: la reinvención de la naturaleza*, Cátedra, Madrid, 1995.

Méndez–Gallardo, Mariana, "La IA desde un pensamiento creativo, simbólico y humanista" en *Christus: revista de teología, ciencias humanas y pastoral*, ITESO, año 31, No. 846, Tlaquepaque, julio de 2024, pp. 10–14.

Pegueroles, Juan, *El pensamiento filosófico de san Agustín*, Labor, Barcelona, 1972.

Platón, *Diálogos III*, Gredos, Madrid, 1988.

Putnam, Hilary, *Razón, verdad e historia*, Tecnos, Madrid, 2006.

Real Academia Española, "inteligencia artificial", https://dle.rae.es/inteligencia Consultado 16/II/2026.