https://doi.org/10.21555/top.v750.3280

# Inteligencia artificial, superinteligencia y mejora cognitiva

# Artificial Intelligence, Superintelligence, and Cognitive Enhancement

Daniel Augusto Duarte Arias
Universidad de San Buenaventura
Colombia
daduartea@usbcali.edu.co
https://orcid.org/0000-0003-3218-8530

Recibido: 09 - 10 - 2024.
Aceptado: 30 - 11 - 2024.
Publicado en línea: 01 - 04 - 2026.

Cómo citar este artículo: Duarte Arias, D. A. (2026). Inteligencia artificial, superinteligencia y mejora cognitiva. *Tópicos, Revista de Filosofía*, 75, 503-540. https://doi.org/10.21555/top.v750.3280

**Resumen**

El transhumanismo es un movimiento que considera posible la consecución de un estadio evolutivo superior por la viabilidad de la "singularidad". Algunas posturas consideran que esto no es posible debido a los límites intrínsecos de la inteligencia artificial (IA). La tesis que defiendo es que, a pesar de la imposibilidad de lograr estadios sobrehumanos, la IA satisface los criterios de mejora cognitiva (en cierto nivel restaurativo y dentro de ciertos límites). Esta mejora se manifiesta en capacidades específicas que resultan en un incremento que trasciende los límites humanos, pero no conducen a un estadio evolutivo superior. Para defender esto, examino qué es la IA, la superinteligencia artificial (SIA) y la singularidad; posteriormente muestro algunas objeciones a la posibilidad de alcanzar una inteligencia artificial general (IAG) a partir de John Searle, Roger Penrose y Hubert Dreyfus; finalmente, argumento mediante una taxonomía en qué niveles puede decirse que la IA representa una mejora cognitiva.

*Palabras clave*: inteligencia artificial; IA débil; IA fuerte; mejora cognitiva; singularidad; superinteligencia; transhumanismo; sobrehumano; experiencia cualitativa; mente extendida.

**Abstract**

Transhumanism is a philosophical movement that considers the attainment of a higher evolutionary stage possible because of the feasibility of the "singularity." Some positions consider that this is not possible due to the intrinsic limits of artificial intelligence (AI). I claim that, despite the impossibility of reaching superhuman stages, AI satisfies the criteria of cognitive enhancement (at some restorative level and within limits). This enhancement manifests itself in specific capabilities that result in an increase that transcends human limits but does not lead to a higher evolutionary stage. To defend this, I examine what AI, artificial superintelligence (AIS), and the singularity are; subsequently I discuss some objections to the possibility of achieving a general artificial intelligence (GAI) drawn from John Searle, Roger Penrose, and Hubert Dreyfus; finally, I argue (by means of a taxonomy) at which levels AI can be said to represent cognitive enhancement.

*Keywords:* artificial intelligence; weak IA; strong IA; cognitive enhancement; singularity; superintelligence; transhumanism; superhuman; qualitative experience; extended mind.

> La pregunta original, "¿Pueden las máquinas pensar?", considero que carece de sentido y no merece ser discutida. Sin embargo, creo que al final del siglo el uso de palabras y la opinión general educada habrán cambiado tanto que se podrá hablar de máquinas pensantes sin esperar ser contradicho.
>
> Turing (1950, p. 442).

## 1. Introducción

A mediados del siglo XX, la discusión en torno a la Inteligencia Artificial (IA) generó un debate sobre la posibilidad de crear una máquina capaz de imitar al ser humano. Parece que la intuición de Alan Turing, expuesta en el epígrafe, esboza el debate actual en relación con la IA. Las preguntas que surgen sobre la IA, especialmente en el ámbito filosófico, muestran que no solo se contempla la posibilidad de emular la inteligencia humana, sino que eventualmente podría ser superada significativamente por una superinteligencia artificial. El transhumanismo aspira a materializar esa posibilidad.

Pueden encontrarse dos tipos de argumentos que se confrontan acerca de la superinteligencia. El primer tipo de argumentación es optimista en cuanto al surgimiento de la superinteligencia artificial, sobre todo con base en un evento de crecimiento exponencial de inteligencia denominado "singularidad". El otro tipo de argumentación es aquel en el que se cree que no es posible la aparición de la superinteligencia debido a los límites que la IA podría tener, por ejemplo, la incapacidad de una inteligencia general.

El problema fundamental que se intenta resolver aquí es si se cumple con los criterios suficientes de mejora cognitiva que nos lleven a un escenario que supere la capacidad humana. La tesis que pretendo sostener en este artículo es que la IA no puede satisfacer los criterios para lograr un estadio sobrehumano y solo logra algunas mejoras restaurativas y dentro de límites. Esta mejora se manifiesta en

capacidades específicas que resultan en un incremento que trasciende los límites humanos, pero no conducen a un estadio evolutivo superior. También sostendré que, para que la IA logre estas mejoras cognitivas superiores a las capacidades humanas, no es necesaria la explosión de inteligencia denominada "singularidad", la posibilidad de la IA fuerte ni el surgimiento de la inteligencia artificial general como lo han supuesto los transhumanistas. Para este propósito, propongo examinar qué es IA, superinteligencia artificial (SIA) y la singularidad; posteriormente mostraré algunas objeciones a la posibilidad de alcanzar una inteligencia artificial general (IAG) a partir de John Searle, Roger Penrose y Hubert Dreyfus; finalmente, argumentaré que la IA no logra un estadio sobrehumano sino sólo el nivel restaurativo y dentro de los límites.

## 2. Inteligencia artificial: la tesis transhumanista de la superinteligencia

El desarrollo de la IA, desde los modelos clásicos hasta la aspiración de una SIA, plantea interrogantes acerca de si estas inteligencias lograrán superar la capacidad humana. Estos cuestionamientos impulsan la noción de la "singularidad", respaldada en gran medida por transhumanistas como Raymond Kurzweil y Nick Bostrom o argumentos como los de David Chalmers. El transhumanismo muestra un escenario donde la posibilidad de superar la cognición humana se vuelve prometedora en presencia de la superinteligencia artificial. El propósito de esta sección será presentar a grandes rasgos los conceptos de "inteligencia artificial", "inteligencia artificial general", la tesis de la singularidad y la superinteligencia artificial en el marco del proyecto transhumanista.

La expresión "inteligencia artificial" tiene origen en la conferencia del Dartmouth College llevada a cabo en Hannover en 1956. Un objetivo inicial de la IA era satisfacer el test de Turing, que se basa en la pregunta de si pueden o no pensar las máquinas. Esta prueba tiene su origen en el desarrollo de las ciencias computacionales y las matemáticas. Su objetivo es crear un escenario en el que un ser humano interactúa con un interlocutor y, a medida que avanza en esta interacción, es incapaz de distinguir si está interactuando con un humano o una máquina. La razón de esta incapacidad de distinción radica en que el rasgo de inteligencia de la máquina será equivalente a la humana. En busca de satisfacer

este test, los desarrollos en inteligencia artificial se han diversificado y mostrado avances en diferentes modelos.

John Preston (2002) presenta una visión general de los modelos en el campo de la IA. El primer modelo se denomina "clásico" o "simbólico" y se enfoca en la creación de inteligencias centradas en la demostración de teoremas matemáticos basados en lógica, geometría y álgebra. El segundo modelo, denominado "conexionista", consiste en desarrollar una IA capaz de llevar a cabo actividades inteligentes de naturaleza simbólica y que logre resolver problemas que involucren un amplio espectro de elementos, como la visión, la comprensión del lenguaje natural, la planificación y el "aprendizaje automático", entre otras funciones cognitivas (cfr. Preston, 2002, pp. 11-13).

Otro modo de clasificar los modelos de la IA es el presentado por Stuart J. Russell y Peter Norvig (2021). Estos autores exponen el modelo de simulación humana y el modelo de ideal racional. Ambos modelos se desarrollan basados en el pensamiento o en la actuación. El modelo de simulación del pensamiento humano considera que se puede modelar, a partir de la introspección y captura de nuestro propio pensamiento, la observación de una persona en acción y la observación del cerebro en acción (cfr. Minsky, 1988; Newell y Simon, 1972). El modelo de la simulación de la actuación humana propone que los ordenadores o máquinas inteligentes superen rigurosamente el test de Turing (1950) alcanzando capacidades como la comunicación exitosa en el lenguaje humano, el almacenamiento de información, la resolución de preguntas, la extracción de conclusiones y el aprendizaje automático para adaptarse a nuevas circunstancias (cfr. Russell y Norvig, 2021, p. 20).

El modelo del ideal de pensamiento racional intenta codificar el pensamiento correcto a partir de razonamientos irrefutables. Un modelo de este tipo es el desarrollado por John McCarthy, denominado LISP. McCarthy (1978) realizó, a partir del procesamiento de listas, una programación de inteligencias artificiales capaces de interactuar con el lenguaje natural, reglas lógicas y manipulación de símbolos. Por otra parte, el modelo de ideal de actuación racional es en el que el agente inteligente "es aquel que actúa para conseguir el mejor resultado o, cuando hay incertidumbre, el mejor resultado esperado" (Russell y Norvig, 2021, p. 22). El objetivo proporcionado al agente es lo que determina qué es lo correcto o qué acción es la adecuada mientras interactúa con las estructuras del mundo.

Desde esta última modalidad de actuación racional surge un modelo basado en multiagentes expuesto por David Sarne y Barbara Grosz (2007) que consiste en un entramado complejo de agentes inteligentes que interactúan y aprenden de forma autónoma en contextos simulados o el mundo. Con los avances tecnológicos de estas inteligencias artificiales los desarrolladores pretenden lograr eventualmente la creación de una inteligencia artificial general (IAG) que logre igualar o superar la mente humana. Dicho propósito es uno de los principales motivos por los cuales los transhumanistas creen viable el surgimiento de la superinteligencia artificial.

La "superinteligencia" es el término con el que se referencia el punto en el que la inteligencia humana es superada por una ultrainteligencia artificial o por un agente ultrainteligente. La superación significativa de la inteligencia humana, según algunos transhumanistas, es el evento denominado "singularidad" (cfr. Chalmers, 2010; Kurzweil, 2005). Antes de caracterizar a la superinteligencia es necesario examinar a qué se refieren los transhumanistas con el concepto de "singularidad" y cómo a partir de este evento se puede materializar la posibilidad de una superinteligencia.

Para Bostrom, la singularidad "es la posibilidad de una explosión de inteligencia, especialmente la perspectiva de una superinteligencia artificial" (Bostrom, 2014, p. 17). Sin embargo, el germen del argumento de la singularidad viene dado por la definición de "máquina ultrainteligente" de Irving John Good (1966). El autor expone al respecto:

> Definamos una máquina ultrainteligente como una máquina que puede superar con creces todas las actividades intelectuales de cualquier hombre por inteligente que sea. Puesto que el diseño de máquinas es una de estas actividades intelectuales, una máquina ultrainteligente podría diseñar máquinas cada vez mejores; entonces se produciría sin duda una "explosión de inteligencia", y la inteligencia del hombre quedaría atrás. Así pues, la primera máquina ultrainteligente es el último invento que el hombre necesita hacer, siempre que la máquina sea lo suficientemente dócil como para decirnos cómo mantenerla bajo control (Good, 1966, p. 33).

Con la definición de "máquina ultrainteligente", Good proporcionó un punto de partida para los tecnoptimistas en relación con la creación o surgimiento de la superinteligencia. De hecho, los argumentos principales que respaldan la idea de la singularidad expresan la posibilidad de esta explosión de inteligencia debido a avances significativos en la tecnología. Uno de los primeros en denominar la explosión de ultrainteligencia como "singularidad" fue Vernor Vinge (1993, p. 12), quien expresa que la singularidad puede ocurrir por tres factores: la creación de ordenadores con conciencia, las redes informáticas humano/ordenador y las ciencias biológicas.

En 1965, Gordon E. Moore, cofundador de Intel, sostenía la predicción de que el número de transistores por dispositivos se duplicaría cada dos años. Posteriormente esta idea se conocería como "la ley de Moore" para justificar el crecimiento exponencial de la tecnología.[1] Hans Moravec (2000) hace referencia a un crecimiento exponencial de la tecnología, similar al de Good y fundamentado en la ley de Moore. Moravec sostiene lo siguiente:

> Después de la guerra, la capacidad de los ordenadores se duplicaba cada dos años, un ritmo que se convirtió en algo habitual en la industria: las empresas que querían crecer debían superarlo, las que no conseguían mantener el ritmo perdían negocio; en la década de 1980, el tiempo de duplicación se redujo a dieciocho meses y, a finales de la década de 1990, el rendimiento de los ordenadores parece duplicarse cada doce meses (Moravec, 2000, p. 57).

Si la descripción de Moravec resulta válida, el crecimiento exponencial de la industria de la informática demostraría que el rendimiento de los dispositivos computacionales se duplicará de tal manera que en algún momento se volvería posible la creación de ordenadores altamente potentes y superinteligentes. Los transhumanistas esperan que esta disminución en el tiempo requerido para desarrollar máquinas más inteligentes y, por ende, el crecimiento exponencial conduzca a la eventual llegada de la singularidad de las máquinas.

[1] Un análisis detallado de la ley de Moore y el uso que algunos transhumanistas han dado a esta ley para justificar el crecimiento exponencial tecnológico lo realizó Antonio Diéguez (2016, pp. 157-160).

Raymond Kurzweil (2005) expresa una postura optimista similar a las expuestas anteriormente. Kurzweil sostiene que la singularidad representa un avance vertical en el cual el progreso de las máquinas tecnológicas experimenta una ruptura en la capacidad de comprensión humana. Según Kurzweil, "nos volveremos mucho más inteligentes a medida que nos fusionemos con nuestra tecnología" (2005, p. 37) y de esta manera se logrará trasladar la mente de un soporte biológico a un soporte mecánico tecnológicamente mejorado. Esto último es conocido comúnmente como *mind uploading*. Para Good, Kurzweil y Bostrom, subyace la preocupación de la domesticación o, al menos, la dominación de la máquina superinteligente. Según estos autores, a pesar de los beneficios que ven en su surgimiento, como método de mejora cognitiva, les alarma que no se desarrollen mecanismos para que estas máquinas sean dóciles a la voluntad humana y no se vuelvan en contra de sus creadores. Después de todo, ¿qué podría provocar que un ser superinteligente se someta a uno menos inteligente? (cfr. Diéguez, 2021, p. 31).

Las explicaciones de Moravec, Vinge y Kurzweil del concepto de "singularidad" surgen del enfoque físico-matemático de la ley de Moore. Sin embargo, para filósofos como Diéguez (2020, pp. 378 y ss.) y Luc Ferry (2018, pp. 36-41), el relato de estos autores parece cumplir una función técnico-utópica sobre el progreso tecnológico.[2] Quizás la falta de rigor metodológico, el inadecuado uso y la falta de claridad sobre el funcionamiento de la ley de Moore y su distante aplicabilidad en cualquier ámbito tecnológico contribuyan a que la especulación o los argumentos erróneos desestimen la idea de la singularidad según lo propuesto por los autores que comparten la perspectiva de Kurzweil. Ahora bien, ¿es verdaderamente ingenua y cercana a la ciencia ficción la perspectiva singularista del transhumanismo? David Chalmers (2010) presenta una versión filosófica del argumento de Good a favor de la singularidad. Chalmers expone el argumento de la siguiente manera:

1. Habrá IA+.
2. Si hay IA+, habrá IA++.

---

3. Habrá IA++ (Chalmers, 2010, p. 5).

[2] Una revisión crítica del concepto de "singularidad" y en contra del transhumanismo es la realizada por Catalina Elena Dobre y Rafael García Pavón (2024).

Chalmers parte del presupuesto de que la inteligencia artificial es al menos tan capaz como la de un ser humano. La IA+, por otro lado, representa una inteligencia superior a la del ser humano más brillante, y finalmente, la IA++ encarna una inteligencia significativamente superior a la de cualquier ser humano. A primera vista, puede parecer arbitrario que Chalmers asuma la posibilidad de alcanzar la inteligencia artificial en un nivel humano. Sin embargo, aborda esta posibilidad desde dos argumentos fundamentales.

El primer argumento se apoya en la tesis de Bostrom y Sandberg (2009), quienes sostienen que los cerebros pueden ser copiados o emulados en máquinas. Para Chalmers, este argumento se basa en la noción de que el cerebro también opera como una máquina y, por lo tanto, podemos emular su funcionamiento en una máquina. Si esta emulación es factible, en poco tiempo podríamos tener IA que replique los complejos procesos cerebrales (cfr. Chalmers, 2010, p. 8).

El segundo argumento a favor del surgimiento de la inteligencia artificial en un nivel general tiene un enfoque evolutivo. Según Chalmers, la evolución fue capaz de generar inteligencia, lo que sugiere la posibilidad de que la inteligencia artificial también pueda ser alcanzada. Básicamente, si la inteligencia pudo emerger sin necesidad de preexistencia, es aún más plausible que los seres humanos puedan crear inteligencia artificial utilizando su propia capacidad intelectual (cfr. Chalmers, 2010, p. 10). Dicho de otra manera, si el azar evolutivo dotó al ser humano de inteligencia, es aún más plausible que una máquina inteligente pueda crear otras máquinas, al igual que lo ha hecho el ser humano, pero en este caso sin depender de procesos arbitrarios, sino más bien de un conjunto de pasos calculados por procesos cognitivos controlados. Finalmente, Chalmers reconoce que hasta el momento no se han logrado estos avances en la inteligencia artificial, pero argumenta que estas premisas apuntan a la factibilidad del surgimiento de la IA. Una vez que esto ocurra, la aparición de la IA+ y la IA++ es solo una cuestión de tiempo.

Otra postura a favor de la singularidad es la presentada por Nick Bostrom (2014), quien define "superinteligencia" como "cualquier intelecto que supere con creces el rendimiento cognitivo de los humanos en prácticamente todos los ámbitos de interés" (Bostrom, 2014, p. 39). Bajo esta perspectiva, la superinteligencia se refiere a un sistema que alcanza un nivel de inteligencia que va más allá de la inteligencia humana en general.

Los caminos propuestos por Bostrom para alcanzar la emergencia de la superinteligencia son cuatro: en primer lugar, la inteligencia artificial, la cual debe tener como características la capacidad de aprendizaje, la habilidad de tomar decisiones ante la incertidumbre, la aptitud para extraer conceptos a partir de datos y estados internos, y el uso de conceptos adquiridos mediante representaciones empleadas en el razonamiento lógico e intuitivo (Bostrom, 2014, p. 40).

El segundo camino consiste en la emulación del cerebro biológico en una estructura computacional. Para lograr este objetivo, se debe realizar un escaneo detallado del cerebro; posteriormente, utilizando los datos obtenidos, se reconstruirá una red neuronal y, por último, esta red neurocomputacional resultante se implementará en un ordenador lo suficientemente potente (cfr. Bostrom, 2014, p. 47). Este enfoque difiere del primero, ya que la creación de inteligencias artificiales no implica necesariamente la copia o emulación de cerebros biológicos. Lo que comparten ambos enfoques es la necesidad de un soporte tecnológico no biológico.

El tercer enfoque consiste en la mejora de la cognición biológica. Aunque los métodos tradicionales, como la educación, han demostrado la posibilidad de mejora cognitiva, Bostrom resalta que existen métodos que generan resultados sustanciales de manera mucho más rápida. Estos métodos transitan entre las ciencias biomédicas y la ingeniería genética, lo cual relegará a la educación o a la selección en la cría como programas ineficientes.

La cuarta vía es la interfaz cerebro-ordenador. Bostrom expone que los implantes aún presentan desafíos en su desarrollo y podrían ser útiles como un método para eventualmente emular el cerebro en un dispositivo. En cualquier caso, se espera que a través de la interfaz cerebro-ordenador los seres humanos puedan mejorar la memoria, el cálculo aritmético y la transmisión de datos, lo que podría resultar en una mejora radical del cerebro biológico hacia una forma aumentada (cfr. Bostrom, 2014, pp. 64-65).

Finalmente, Bostrom propone las redes y organizaciones como una forma de surgimiento de la superinteligencia. Esta red consistiría en la interconexión de mentes humanas individuales con artefactos y robots. Si esto resulta factible, dicha red podría llegar a alcanzar un nivel de superinteligencia, lo que llevaría a una mejora en la capacidad intelectual tanto de la red como de la organización en su conjunto (cfr. Bostrom, 2014, p. 68).

Al igual que las posturas singularistas antes mencionadas, Bostrom es optimista frente a la posibilidad de surgimiento de la superinteligencia, ya que, desde su perspectiva, los caminos que conducen a esta inteligencia superior aumentan las probabilidades de lograr un progreso humano significativo más allá de sus límites biológicos. También el optimismo de Bostrom radica, en mayor medida, en el surgimiento de la superinteligencia de origen artificial (cfr. Bostrom, 2014, p. 71). Para los propósitos de este artículo es de interés el camino de la inteligencia artificial y en menor medida la emulación de cerebros por los límites técnicos a los que esta copia conduce.

A la definición de "superinteligencia" se le añaden tres posibles modalidades que surgen de los diversos caminos previamente expuestos. La primera de estas modalidades es la superinteligencia de velocidad, la cual implica "un sistema que puede hacer todo lo que el intelecto humano puede hacer, pero mucho más rápido" (Bostrom, 2014, pp. 71-72). La velocidad alcanzada por esta superinteligencia es tal que, al encontrarse con limitaciones en el mundo material, optaría por trabajar con objetos digitales. Sin embargo, esta velocidad no implica un avance en otras capacidades cognitivas, sino más bien en la rapidez con la que se procesa la información.

La segunda modalidad es la superinteligencia colectiva, la cual consiste en "un sistema compuesto por un gran número de intelectos menores, de manera que el rendimiento general del sistema superaría enormemente al de cualquier sistema cognitivo actual en muchos ámbitos generales" (Bostrom, 2014, p. 73). Esta inteligencia opera como una amplia red colectiva que busca resolver diversas tareas al abordar subproblemas específicos. En cierta medida, esta red colectiva, como el lenguaje, ha contribuido a la mejora de la humanidad, si comparamos su inteligencia con la de los ancestros del ser humano. En resumen, la superinteligencia colectiva debe funcionar como una suerte de mente inmensa en la que interactúan un conjunto de mentes humanas coordinadas y comunicadas. Un ejemplo de esto es la multiagencia expuesta por David Sarne y Barbara Grosz (2007), enunciada anteriormente.

Finalmente, la superinteligencia de calidad es "un sistema que es al menos tan rápido como una mente humana y cualitativamente mucho más inteligente" (Bostrom, 2014, p. 75). La superinteligencia, al ser cualitativamente superior, implica ser de una calidad considerablemente mayor, similar a cómo la inteligencia humana supera la de otros

animales. Si el ser humano lograra un conjunto de ventajas cognitivas de una capacidad similar a la obtenida con las representaciones lingüísticas complejas, entonces alcanzaría el estatus de superinteligencia. En otras palabras, al igual que en un punto evolutivo los seres humanos se distanciaron de los chimpancés debido a las estructuras lingüísticas complejas, de manera similar, a través de la emergencia de estructuras cognitivas aún más complejas que las actuales, podría alcanzarse una superinteligencia de cualidad superior. Más adelante se desarrollarán los argumentos en contra de esta posibilidad porque para lograr esta superinteligencia se requiere la IAG.

Bostrom también argumenta que para que sea posible el surgimiento de la superinteligencia es necesario apostar por diferentes métodos. A pesar de ser proyectos distintos según su naturaleza, estos métodos pueden conducir a que los avances en uno u otro produzcan mejoras que eventualmente lleven a la superinteligencia. Así, la superinteligencia, al ser un proyecto de mejora cognitiva humana, requiere al menos mejoras en velocidad, colectividad y calidad. Sin embargo, los avances actuales en IA muestran que las mejoras en velocidad y en colectividad ya son posibles a pesar de que estos no conduzcan, por ahora, a mejoras de calidad que permitan pensar en un estadio evolutivo superior. Por tal razón, para Bostrom, la manera más eficaz de alcanzar la posibilidad de esta superinteligencia es a través de la inteligencia artificial, que actúa como herramienta capaz de estructurar mentes digitales con algunas ventajas sobre las mentes biológicas.

Modelar inteligencias artificiales a partir del funcionamiento del cerebro humano puede tener ciertos límites, por ejemplo, el mismo desconocimiento de cómo funciona el cerebro y sus conexiones neuronales, cómo surge el pensamiento o qué interacciones y acciones pueden enriquecer más o menos los procesos cognitivos. Intentar estructurar una inteligencia artificial parcialmente independiente de cómo funciona el cerebro biológico puede ser mucho más prometedor. Por ejemplo, los hermanos Wright pronto descubrieron que no era necesario replicar el movimiento de las alas de las aves para crear máquinas voladoras. En lugar de intentar emular a las especies que biológicamente pueden volar, buscaron que la máquina volara mediante sus propias características. Incluso hoy en día existen máquinas voladoras, como drones, helicópteros y globos, que cumplen este objetivo sin imitar las alas de las aves. El transhumanismo puede optar por un escenario mucho más realista si la construcción de IA

no necesariamente recurre a modelos biológicos que conduzcan a los problemas que aquí he expuesto.

## 3. Críticas y objeciones a la inteligencia artificial general

A pesar del optimismo del transhumanismo respecto al surgimiento de la superinteligencia y la posibilidad de superar la mente humana, algunos autores, como John Searle, Roger Penrose y Hubert Dreyfus, desestiman la tesis de la inteligencia artificial general y, por ende, la eventual explosión de inteligencia denominada "singularidad". El objetivo de esta sección es exponer algunos argumentos en contra de la posibilidad de desarrollo de inteligencias artificiales fuertes o inteligencias artificiales generales desde una perspectiva filosófica.

Los desarrollos teóricos de Turing inspiraron otras preguntas: ¿cómo surge la mente de un cerebro físico? ¿Cómo se traducen estados mentales en acción? ¿Puede lograr un modelo de inteligencia artificial la capacidad de inteligencia general? Estas preguntas, entre muchas otras, son el punto de partida de las distinciones filosóficas denominadas la "IA fuerte" y la "IA débil":

> Según la IA débil, el principal valor del ordenador en el estudio de la mente es que nos proporciona una herramienta muy poderosa. Por ejemplo, nos permite formular y probar hipótesis de forma más rigurosa y precisa. Pero según la IA fuerte, el ordenador no es solo una herramienta para el estudio de la mente, sino que un ordenador adecuadamente programado es realmente una mente, en el sentido de que los ordenadores con los programas adecuados pueden comprender y tener otros estados cognitivos. En la IA fuerte, dado que el ordenador programado tiene estados cognitivos, los programas no son meras herramientas que nos permiten probar explicaciones psicológicas, sino que los programas son en sí mismos las explicaciones (Searle, 1980, p. 417).

La distinción entre ambas IA radica en que, mientras la IA débil funciona como una herramienta utilizada para probar hipótesis o explicaciones psicológicas, o como la simulación de una mente, la IA fuerte funciona como una mente. La IA fuerte no solo arroja elementos relevantes en el estudio de la mente, sino que además es una mente en sí

misma porque produce estados cognitivos. Es una inteligencia artificial general (IAG) con capacidades iguales o superiores a la mente humana.

Luc Ferry (2018) expone que la IA débil es una imitación mecánica que puede reproducir aspectos de la inteligencia humana, pero que aún no cumple con la superación del test de Turing. Algunos automóviles autónomos, los asistentes virtuales de Google, Siri, o los LLM como ChatGPT o DeepSeek son ejemplos de inteligencias artificiales débiles que funcionan bien en tareas específicas, pero no logran ser igual o más inteligentes que los humanos, es decir, no logran adquirir las capacidades de una IAG. En cambio, los partidarios de la IA fuerte, según Ferry, argumentan que es posible replicar el cerebro humano y que eventualmente este podrá ser almacenado en dispositivos o máquinas capaces de tener conciencia de sí mismos. Por esta razón, no solo superarán la prueba de Turing, sino que además tendrán rasgos superiores a la inteligencia humana, como mayor capacidad, velocidad o cualidades en los procesos cognitivos que incluso les permitan crear otras superinteligencias, como lo expone Bostrom (2014).

Al igual que Searle, Fred Dretske (1981) considera que los sistemas (aunque no necesariamente inteligencias artificiales) pueden conseguir el rasgo de la intencionalidad si interactúan con el mundo. Algo similar sostiene Ferry, quien es escéptico acerca de la emergencia de esta inteligencia artificial fuerte porque se requiere una correlación entre el cerebro o la mente (ya sea biológica o artificial) y el mundo (cfr. Ferry, 2018, pp. 173-175). En estos tres casos la intencionalidad es un rasgo esencial de la mente y solo es posible su surgimiento si se interactúa con el mundo.

A pesar de las propuestas expuestas por Bostrom (2014), Sandberg (2011) o Chalmers (2010) sobre la IA, los detractores argumentan según los diferentes enfoques. Algunas posturas buscan demostrar la imposibilidad de la creación de una inteligencia artificial lo suficientemente capaz de simular la inteligencia humana. Algunos de los argumentos que controvierten la posibilidad de que una máquina logre la IAG son la habitación china de Searle, el argumento basado en los teoremas de incompletitud de Gödel y las objeciones de Dreyfus.

La habitación china es un experimento mental en el que se imagina que un ser humano que solo se comunica en inglés está dentro de una habitación. Allí tiene disponibles un libro de reglas escrito en inglés y algunos papeles. Desde fuera, introducen algunos documentos que contienen símbolos en chino y, posteriormente, el individuo va

organizando los caracteres según las reglas del libro. Luego de esto, envía fuera de la habitación nuevos trozos de papel con respuestas fluidas e inteligentes. Desde una perspectiva externa, se deduce que hay un sistema que recibe datos en forma de caracteres chinos y genera respuestas inteligentes. Pero ¿entiende realmente el ser humano dentro de la habitación el idioma chino? Se da por supuesto que no, porque al ser solo angloparlante, no tiene una comprensión del chino. Lo mismo sucede con la IA porque la semántica no puede reducirse a la sintaxis. Searle afirma que una supuesta IA fuerte no tiene comprensión y, por lo tanto, no posee la capacidad de responder a estímulos o funciones cognitivas porque al operar exclusivamente manipulando símbolos no tiene comprensión. Por ejemplo, un ser humano lee una historia y realiza diversas funciones o acciones cognitivas producto de la comprensión (cfr. Searle, 1980, pp. 417 y ss.).

En relación con la pregunta de si las máquinas pueden pensar, Searle (1980, p. 422) sostiene que esto no es posible debido a la falta de conciencia y de *qualia*. El debate se centra en si las máquinas tienen conciencia del mundo exterior, del yo o de las experiencias subjetivas. Aunque las máquinas inteligentes pueden ser programadas con algunos códigos de experiencias subjetivas, los seres humanos tienen cierta ventaja sobre las IA, ya que pueden utilizar su propio aparato subjetivo para apreciar las experiencias subjetivas de los demás. Frente a ciertas experiencias, las máquinas solo pueden proporcionar datos, mientras que la experiencia humana puede, por ejemplo, *saber qué se siente* cuando alguien se golpea con un martillo porque puede experimentar el golpe con un martillo (Russell y Norvig, 2021, pp. 1036-1037).

El argumento gödeliano está basado en los teoremas de incompletitud de Gödel y es expuesto por Roger Penrose (1995). La explicación consiste en demostrar que las máquinas no pueden alcanzar el nivel de la inteligencia humana. Para Gödel, los sistemas matemáticos deductivos no pueden ser consistentes y completos simultáneamente. Una IA modelada desde un enfoque clásico, como el expuesto por John Preston (2002), Allen Newell y Herbert Alexander Simon (1972), o Marvin Minsky (1988), mediante un lenguaje simbólico, no podría escapar a los problemas evidenciados por los teoremas gödelianos. Penrose, utilizando los teoremas, reconstruye el argumento de la siguiente manera:

> Tratamos de suponer que la totalidad de los métodos de razonamiento matemático (inatacables) que son en principio humanamente accesibles pueden encapsularse en algún sistema formal sólido (no necesariamente computacional) F. Un matemático humano, si se le presenta F, podría argumentar de la siguiente manera (teniendo en cuenta que la frase "yo soy F" no es más que una abreviatura de "F encapsula todos los métodos humanamente accesibles de demostración matemática"): aunque no sé si soy necesariamente F, concluyo que si lo fuera, entonces el sistema F tendría que ser sólido y, más aún, F′ tendría que ser sólido, donde F′ es F complementado por la afirmación adicional "Yo soy F". Percibo que de la suposición de que yo soy F se deduce que el enunciado de Gödel G(F′) tendría que ser verdadero y, además, que no sería una consecuencia de F′. Pero acabo de percibir que "si resulta que soy F, entonces G(F′) tendría que ser cierta", y percepciones de esta naturaleza serían precisamente lo que se supone que F′ consigue. Puesto que, por tanto, soy capaz de percibir algo más allá de los poderes de F′, deduzco que, después de todo, no puedo ser F. Además, esto se aplica a cualquier otro sistema (gödelizable), en lugar de F (Penrose, 1995, § 3.2).

Penrose comienza su argumentación invitando al lector a suponer que los métodos matemáticos pueden incorporarse a un sistema formal sólido al que denomina "F". A partir de esto, un matemático humano podría argumentar que si él fuera un sistema que encapsula todos los métodos matemáticos, tendría que ser sólido. También podría argumentar que un sistema "F′", que es esencialmente el mismo que el sistema "F" con la afirmación adicional "yo soy F", también debería ser sólido. A continuación, Penrose introduce el enunciado "G(F′)" basado en el teorema de Gödel, que implica la autorreferencia en sistemas formales. En este contexto, el matemático podría afirmar que "yo soy F" es verdadero, pero este enunciado no es una consecuencia del sistema formal F′. En otras palabras, un sistema que abarca los métodos matemáticos no puede reconocer F′, y, por lo tanto, F′ está más allá del sistema consistente F. Si esto es así, entonces el sistema F no es

simultáneamente completo y consistente, ya que existen verdades que escapan a este sistema.

¿Qué sucedería si aplicáramos este argumento a la modelación o programación de inteligencias artificiales? Según Erik Larson (2021), el problema con la programación de sistemas formales de inteligencia artificial radica en su ceguera. Curiosamente, el ser humano parece tener una ventaja sobre estos sistemas, ya que, a diferencia de los sistemas formales, posee la capacidad de "ver" cosas como "yo soy F", que las computadoras no pueden. Mientras que los sistemas formales tienen limitaciones, como la incapacidad de demostrar algo verdadero en su propio lenguaje, el ser humano puede percibir aspectos que escapan a la capacidad de las máquinas (cfr. Larson, 2021, pp. 12-14). Searle argumentaría que este tipo de capacidades humanas se deben a una causalidad extrañamente inexplicable que existe entre el cerebro biológico y el pensamiento humano (cfr. Searle, 1980, p. 424).

Desde esta perspectiva, no parece factible modelar una IA fuerte capaz de abordar una variedad de objetivos, sino más bien una IA débil destinada a tareas o propósitos específicos. En otras palabras, la IA no resulta útil cuando se la utiliza más allá de sus capacidades predefinidas. En contraste, los seres humanos podemos acceder a verdades que no se limitan a meras deducciones, sino que se extienden en una gama de posibilidades para alcanzar la verdad y el conocimiento. Un ejemplo de esto son los saltos creativos que a lo largo de la historia han caracterizado a la humanidad, en los cuales la verdad y el conocimiento se encontraban fuera de las fronteras del saber humano. Siguiendo el argumento de Penrose, en un sistema "F" que sirva de modelo para programar la inteligencia artificial, no habría algo así como un conocimiento fuera de sus propias fronteras. Con los teoremas de Gödel queda desestimado el argumento de Chalmers (2010) de que un camino para alcanzar la singularidad es copiar o emular el cerebro biológico (a menos que se resuelva este problema), porque estos teoremas muestran los límites que tiene una máquina frente al cerebro humano si se desarrolla desde un modelo simbólico.

Así las cosas, las implicaciones del argumento de Penrose es que los sistemas de IA, desde una perspectiva formal, no pueden ser completos y consistentes. Si un sistema fuera completo y consistente, debería ser capaz de incluir referencias a sí mismo; sin embargo, las autorreferencias sugeridas por los teoremas de Gödel revelan que surgirían contradicciones dentro del mismo sistema, ya que este tipo de

verdades trascienden los límites de su programación. En consecuencia, no existiría una IA fuerte capaz de razonar sobre los aspectos de su propio funcionamiento o conocimiento, lo que conduciría a paradojas en sistemas como F.

Huber Dreyfus (1967) también critica la posibilidad de que la inteligencia artificial pueda pensar y, además, que pueda simular o emular la inteligencia humana. Según Dreyfus, el ser humano posee algo que está "más allá" del simple procesamiento de información regulado y formal. Identifica al menos cuatro ideas sostenidas por los científicos de la IA que considera erróneas o hipótesis infundadas y cuestionables. Su argumentación en contra de la inteligencia artificial se divide en supuestos biológicos, psicológicos, epistemológicos y ontológicos.

La hipótesis biológica se basa en la consideración de que la IA y el cerebro biológico operan de la misma manera, es decir, que los procesos de información son equivalentes entre lo biológico y las máquinas. Para Dreyfus, el error de los desarrolladores de IA radica en que, mientras la máquina opera en términos binarios como "sí" y "no" o interruptores abiertos y cerrados, el cerebro biológico no está determinado por relaciones de reglas, sino que funciona de una manera mucho más compleja. La hipótesis psicológica expone cómo la mente opera solo en un sistema formal. Dreyfus refuta esta hipótesis argumentando que el procesamiento de datos se realiza en tercera persona y el "procesador" no juega un papel esencial (cfr. Dreyfus, 1967, p. 14).

En relación con lo anterior, la hipótesis epistemológica se basa en que el comportamiento inteligente puede ser simulado en un dispositivo que procesa información de manera objetiva, desprendida y desencarnada. Dreyfus sostiene que no es posible formalizar lógicamente toda la conducta humana, ya que, por ejemplo, hay acciones humanas, como los saltos creativos, que exceden cualquier sistema lógico formal. Finalmente, la objeción ontológica de Dreyfus hacia los científicos de la IA es que tienden a creer que los datos del mundo se pueden analizar como elementos independientes. Esto, para Dreyfus, es cuestionable porque toda conducta inteligente requiere una interacción con el mundo y no se limita a una formalización de contextos restringidos, como sucede en la programación de la IA mediante reglas lógicas (cfr. Dreyfus, 1967, p. 14).

Dreyfus descarta la viabilidad de una Inteligencia Artificial General (IAG) debido a que existen elementos en la inteligencia humana que no pueden ser simulados. El primero es el reconocimiento de patrones

que depende en gran medida de situaciones contextuales complejas. La IA reconoce patrones de datos abstractos, pero sin comprensión de contextos físicos y emocionales (cfr. Dreyfus, 1967, pp. 16-21). El segundo es la resolución de problemas en entornos complejos de interacción con el mundo que las IA no logran completamente sin cuerpos mecánicos (cfr. Dreyfus, 1967, pp. 21-26). Por lo tanto, la aspiración de emular o crear máquinas que se comporten de manera igual o superior al ser humano se reduce, según él, a una hipótesis falsa.

Hasta este punto he mostrado que las aspiraciones del transhumanismo por lograr una superinteligencia artificial son puestas en duda debido a las objeciones que desarrollan algunos filósofos. Searle, Penrose y Dreyfus cuestionan el surgimiento de superinteligencias debido a las limitaciones intrínsecas de la inteligencia artificial porque hay rasgos de la biología que juegan un papel fundamental en los procesos cognitivos que aún no son comprendidos y tampoco pueden ser emulados. No obstante, a pesar de que las esperanzas singularistas se vean frustradas, ¿con ello mueren los ideales transhumanistas? ¿No hay algún tipo de "mejora" que sea alcanzable por los desarrollos de IA?

## 4. Mejora cognitiva e inteligencia artificial

En la actualidad, no existen inteligencias artificiales generales. Un escenario más realista al que podrían recurrir los transhumanistas es atender a las mejoras cognitivas que han logrado desarrollar quienes programan inteligencias artificiales. El propósito de este apartado será mostrar cómo la inteligencia artificial supera en gran medida a los seres humanos en algunos procesos cognitivos. Sin embargo, estas mejoras no conducen a un estadio evolutivo superior porque no son constitutivas del ser humano debido a los límites y objeciones ya expuestos por Searle, Penrose y Dreyfus. Mi argumento radica en que, para lograr el estadio sobrehumano, no solo se requiere el aumento del procesamiento de información, como ocurre con algunas máquinas inteligentes, sino además la transformación radical del carácter cualitativo de la experiencia.[3] Así las cosas, clasificaré las mejoras cognitivas según los usos de IA. Cada tipo de mejora puede producirse a nivel individual, social y de especie según los contextos en que ocurran. Con esta

[3] Este argumento ha sido especialmente desarrollado por Rivera-Novoa (2020 y 2024) pero enfatizando cómo el uso de algunas tecnologías conduce a la pérdida de esta fenomenología cualitativa.

clasificación, argumentaré en qué tipos de mejoras de la inteligencia artificial se puede suponer o no un escenario en el que sus procesos cognitivos sean superiores al ser humano.

En virtud de disipar ambigüedades sobre el concepto de "mejora", aquí sostendré que una mejora cognitiva es cualquier cambio biológico o psicológico de un individuo que posibilite o aumente sus procesos cognitivos en circunstancias *C*. "Aumento" puede entenderse como optimizar, potenciar o favorecer los procesos funcionales. De este modo, una capacidad es cualquier estado biológico o psicológico de una persona que posibilite o aumente un proceso funcional en las circunstancias *C*.[4]

Para examinar los diferentes tipos de mejora, asumiré como criterios de cuándo estamos frente a un proceso cognitivo y cuándo estos procesos son mejoras cognitivas la postura de Rowlands (2010) y Sandberg (2011). Así las cosas, un proceso cognitivo es 1) el procesamiento de información, incluyendo la manipulación o transformación de las estructuras que contienen la información; 2) el proceso tiene como función poner a disposición del sujeto (agente inteligente) información que antes del proceso no estaba disponible; 3) la producción de información está disponible a través de un estado representacional, y 4) este proceso le pertenece al sujeto que produce el estado representacional (cfr. Rowlands, 2010, pp. 110–111).

Ahora bien, una mejora cognitiva sucede cuando el proceso cognitivo amplía o extiende los procesos de información. Una vez se esté frente a un proceso que reúna los cuatro criterios y que además aumente o mejore las facultades centrales, como "adquirir información (percepción), seleccionar (atención), representar (comprender) y retener (memoria) información, y usarla para guiar el comportamiento (razonamiento y coordinación de salidas motoras)" (Sandberg, 2011, p. 71), se podrá decir que hay mejora cognitiva. En general, se deben mejorar todos los sistemas y estructuras que permitan el funcionamiento y la obtención de

[4] Esta definición que propongo dirime ambigüedades alrededor del concepto de mejora porque no se sitúa desde el enfoque naturalista, ni del constructivista social expuestos por Parens (cf. 1998, p. 1). En este sentido, diluye la ambigüedad alrededor del referente de "normalidad" y del vínculo social que puedan tener conceptos como enfermedad, salud o terapia como valores negociados. Lo que ocurre con mi definición es que establece fronteras –algo que no ocurre en la literatura transhumanista–entre los diferentes usos de tecnologías para tres tipos de mejoras: restaurativas, dentro de los límites del *Homo sapiens* y sobrehumanas.

conocimiento. Si se cumplen las condiciones propuestas por Rowlands y además hay un aumento en las facultades cognitivas, entonces se puede hablar de una mejora cognitiva. Todo este andamiaje cognitivo se soporta desde un enfoque de cognición situada *4E* en la interacción activa de las estructuras mentales, corporales y ambientales. Delineado este criterio de mejora cognitiva, expondré la taxonomía propuesta para examinar la inteligencia artificial.

El primer tipo de mejora se refiere al aumento restaurativo de procesos cognitivos dada la discapacidad que reduce dichos procesos. En el segundo tipo, significa tener un aumento de procesos cognitivos sin que esto conlleve a un estadio superior al ser humano promedio. El tercer tipo de mejora significaría, sobre todo, aumentar o superar el estado biológico y psicológico hasta alcanzar procesos cognitivos más allá del ser humano. En este sentido, este tipo de mejora trascendería las fronteras hasta lugares y procesos nunca explorados por el *Homo sapiens*. Esto último tendrá mayor tratamiento en la sección 4.3.

Las mejoras cognitivas pueden ocurrir en tres niveles. En el nivel individual está cualquier mejora que afecte a un individuo sin que esto necesariamente se vea reflejado a nivel social o como especie. Su objetivo es que la aplicación de tecnologías le permita a un individuo mejorar alguno o varios procesos cognitivos. A nivel social está cualquier mejora que afecte a varios individuos establecidos en una sociedad. Su objetivo es la aplicación de tecnologías que le permita a una sociedad mejorar alguno o varios aspectos cognitivos. En el nivel de especie está cualquier mejora que afecte a todos los individuos de una especie. Su objetivo es la aplicación de tecnologías que le permitan a una especie mejorar alguno o varios procesos cognitivos. En este marco explicativo argumentaré cada tipo de mejora analizando algunas inteligencias artificiales.

### 4.1. Mejoras restaurativas

Una mejora restaurativa es la aplicación de tecnologías de tratamiento o reparación de un perjuicio cognitivo hasta el punto de lograr que procesos cognitivos deficientes o ausentes mejoren, se posibiliten o aumenten por dicha intervención. En este tipo de mejoras se pueden contemplar tratamientos en los cuales las personas, con la ayuda de aplicaciones o dispositivos, resuelvan problemas o tareas de tal manera que sus deficiencias cognitivas se restauren, lo que también contribuye a mejorar su procesamiento. Sincrolab (s. f.) es una plataforma terapéutica utilizada en el tratamiento del trastorno por déficit de atención e hiperactividad (TDAH). Esta herramienta, además de servir

como terapia para pacientes con TDAH, se destaca por su capacidad de adaptarse a las distintas necesidades de estimulación personalizada de los usuarios.

El tratamiento implica que el paciente interactúe con una serie de juegos en dispositivos electrónicos, lo que permite que la inteligencia artificial supervise las mejoras en el control inhibitorio de quienes participan en estos juegos. En este sentido, Sincrolab lleva a cabo procesos en tiempo real que también afectan a los seres humanos. A raíz de esto, los pacientes experimentan mejoras en la flexibilidad cognitiva, la memoria y el comportamiento. En sentido estricto Sincrolab no es una inteligencia artificial general; sin embargo, ilustra cómo es posible lograr una mejora restaurativa en el nivel cognitivo mediante el uso complementario de la inteligencia artificial.

Sincrolab desarrolla un procesamiento de información que posteriormente utiliza para interactuar con un paciente. El apoyo brindado por esta IA permite transformar la información para ofrecer alternativas al sujeto que la usa. Es en este sentido que se puede considerar que algunas de las IA, modeladas para este tipo de objetivos, logran superar la pérdida de habilidades en el nivel individual. Al operar como mejora restaurativa, este tipo de mejora no conduce a un estadio evolutivo superior porque su intervención no genera un cambio constitutivo en el individuo.

Un ejemplo similar es el uso de vacunas que, a pesar de funcionar como mejoras restaurativas, una vez aplicadas no llevan al ser humano a un estadio evolutivo superior. Con AlphaFold (Google DeepMind, 2023c) se pueden conseguir objetivos similares. Esta IA simula la estructura tridimensional de las proteínas del ADN. Las implicaciones que pueden tener estos avances son significativas sobre todo en el desarrollo de medicamentos y las investigaciones en medicina de biología a nivel molecular. Esta IA también puede aplicarse a nivel individual, por ejemplo, en la búsqueda y simulación de estructuras de proteínas para la mejora genética de un individuo eliminando desde su raíz genética enfermedades o posibles defectos de empalme en el ADN.

Tanto Sincrolab como AlphaFold pueden ser aplicados a nivel social y de especie. En principio si una mejora restaurativa logra éxito en su aplicación a nivel individual, por consiguiente, es muy probable que su aplicación social tenga éxito. Vale la pena señalar que el examen de estas intervenciones a nivel social depende de diversos factores debido a que las máquinas inteligentes se adaptan a usuarios en contextos

determinados. Un escenario en el que Sincrolab puede aplicarse a nivel social es que, como plan de gobierno, una nación adopte esta IA para un tratamiento generalizado de la población con TDAH y con ello logre superar la pérdida de habilidades en la población. Un caso similar ocurre con AlphaFold si un laboratorio médico asume esta IA para desarrollar avances que mejoren las condiciones genéticas de un grupo social.

Otro escenario es la aplicación de mejoras restaurativas a nivel de especie. Si se logra que inteligencias artificiales modeladas como Sincrolab o AlphaFold tengan acceso libre para cualquier individuo, entonces se estaría hablando de superación de pérdidas de habilidades a nivel de especie. Un posible evento restaurativo sería que, a partir del análisis de datos que realiza AlphaFold sobre algún defecto genético, como el alzheimer, se consiga con un proyecto de cría, eliminar una enfermedad o la causa de enfermedades en la especie. En este caso, a pesar de lograrse una aplicación tecnológica de forma generalizada, al ser de nivel restaurativo no se puede hablar aún de una mejora que lleve al *Homo sapiens* a un estadio evolutivo superior porque estas mejoras no representan cambios sobrehumanos que permitan considerar la emergencia de una nueva especie.

### 4.2. Mejoras dentro de los límites

Una mejora dentro de los límites se presenta cuando hay uso de tecnologías para mejorar procesos cognitivos que no superen el umbral del *Homo sapiens*. A nivel individual, la mejora dentro de los límites se refiere a aumentos que ofrecen ventajas cognitivas a individuos. Aunque puede suceder, no necesariamente debe verse reflejado en el nivel social o de especie. Su objetivo es que la aplicación de tecnologías permita a un individuo mejorar uno o varios aspectos cognitivos sin que estos conduzcan al individuo mucho más allá de la especie. Por lo tanto, se puede afirmar que una inteligencia artificial mejora la cognición a nivel individual si un sujeto logra aumentar habilidades o capacidades cognitivas que antes no tenía sin que esto represente un estadio evolutivo superior.

En principio, Sincrolab y AlphaFold también funcionan como ejemplos de mejoras cognitivas dentro de los límites del *Homo sapiens*. Sin embargo, entre los algoritmos desarrollados por DeepMind (Google DeepMind, s. f.) se encuentra AlphaZero (2018), que es un programa capaz de jugar al ajedrez o al go de manera similar o incluso mejor que un ser humano. La superioridad de AlphaZero respecto a otras

inteligencias artificiales radica en su capacidad para aprender de forma autónoma y desarrollar estrategias que le permitan ganar partidas contra seres humanos, resolviendo así tareas complejas con movimientos simples. Un escenario posible puede ser que un individuo, al igual que se entrena con un jugador experto en ajedrez o mediante la lectura de libros con reglas sobre el ajedrez, también lo puede hacer con AlphaZero. Esto eventualmente debería lograr que el jugador entrenado por esta inteligencia artificial mejore procesos cognitivos como la percepción, la atención, la comprensión y la memoria. Así, mientras la IA se adapta y aprende de los movimientos de las piezas del usuario, también el jugador humano enriquece sus procesos cognitivos.

Una extensión desarrollada a partir de los avances de AlphaZero es AlphaTensor (Google DeepMind, 2022), que, siguiendo los principios de su predecesor, busca resolver tareas de manera más simple y eficiente, lo que permite igualar o mejorar el número de operaciones que debe ejecutar un sistema. Para ayudar a comprender este avance se puede imaginar un escenario en el que, frente al objetivo de encontrar datos específicos en una serie de símbolos aleatorios, el ser humano puede llevar determinado tiempo mientras examina las diferentes posibilidades de ubicación de los datos. AlphaTensor, en el mismo escenario, podría lograr un algoritmo mucho más sencillo y eficaz, lo que no solo resolvería la tarea de encontrar los datos específicos que se le han pedido, sino que también podría ajustar cada vez más el algoritmo para encontrar datos u otros símbolos de manera más rápida. Al final, el algoritmo lograría ajustes que incluso generen numerosas oportunidades de nuevo conocimiento. En este caso, la IA realiza procesos cognitivos superiores en velocidad a los del ser humano; no obstante, si un ser humano se entrenara constantemente en una tarea específica también lograría resolverla, por lo cual no estaríamos frente a un estadio evolutivo superior sino frente a una superación dentro de los límites del *Homo sapiens*. Bostrom (2014) estaría de acuerdo en decir que aquí estamos frente a una mejora en la modalidad de velocidad. Esto también será discutido en la sección 4.3, donde señalaré que las transformaciones del carácter cualitativo de la experiencia supondrían un estadio evolutivo superior y no exclusivamente este tipo de mejoras en la modalidad de velocidad.

A nivel social, una mejora dentro de los límites se refiere a cualquier mejora o mejoras que afecten a un grupo de individuos. Su objetivo es que la aplicación de tecnologías permita a un grupo mejorar uno o varios

aspectos cognitivos que antes no tenía o que aumenta de tal manera que aún no supera la especie *Homo sapiens*. Por lo tanto, se puede afirmar que una inteligencia artificial mejora la cognición a nivel social si un grupo de individuos logra aumentar capacidades cognitivas dentro de los límites del *Homo sapiens*.

AlphaZero y AlphaTensor pueden ser aplicados a nivel social y de especie. Un escenario en el que AlphaZero puede aplicarse a nivel social es que un centro educativo adopte como modelo de mejora de algunas habilidades cognitivas el entrenamiento de estudiantes con el ajedrez. Si el proyecto tiene éxito, habrá un grupo de estudiantes que cuenten con habilidades superiores frente a otros que no usaron AlphaZero. AlphaTensor podría realizar una tarea similar si se utiliza como herramienta para enseñar el algoritmo más eficiente en la búsqueda de datos específicos. Por ejemplo, un grupo de programadores puede aprender a partir del trabajo conjunto con AlphaTensor cómo organizar datos que se encuentran de forma aleatoria.

AlphaDev (Google DeepMind, 2023a) es un sistema desarrollado por DeepMind que optimiza algoritmos en las ciencias de la computación para lograr mejorar la informática y los microchips. Las mejoras en *hardware* han llegado a un límite físico y con los avances de AlphaDev se logra mayor eficiencia en el rendimiento de los microprocesadores. Los algoritmos descubiertos han sido lanzados en bibliotecas de código abierto y por tanto todo el mundo puede acceder a ellos. Ahora bien, que haya un mayor rendimiento en desarrollo de procesadores o microchips para elaborar máquinas más eficientes no implica que esto conduzca a un estadio evolutivo superior. Sin embargo, estos avances tecnológicos sí muestran que pueden darse mejoras cognitivas en la velocidad con la que se procesa información o la capacidad de memoria de las máquinas. Si estos avances tecnológicos se reducen solo a algunos grupos sociales, entonces estaremos hablando de ventajas tecnológicas en unos y desventajas en otros.

OpenAI es una empresa que tiene por objetivo crear la IAG. Su principal desarrollo tecnológico es ChatGPT (OpenAI, s. f.). Esta IA multimodal procesa información a partir de *inputs* de lenguaje natural e imágenes. Este modelo genera respuestas precisas y coherentes como un ser humano e incluso en ocasiones superiores a un *Homo sapiens* promedio. Ya sea a nivel individual, social o como especie, el GPT puede funcionar como un compañero de estudio o incluso de trabajo. En el escenario de aplicación a nivel de especie pueden pensarse situaciones en

las que, a partir de *prompts* o comandos que tengan por objetivo mejorar los procesos educativos en etapas tempranas del ser humano, se tenga el fin de garantizar la consecución de mejores competencias o habilidades. Si se aplica de forma adecuada, esta IA puede ayudar a enseñar, de la misma manera que un profesor, algunas habilidades a los usuarios e incluso proporcionar explicaciones mucho más precisas. Sin embargo, un uso no adecuado de esta herramienta puede terminar por afectar o ir en detrimento de la cognición biológica si, por ejemplo, en lugar de ser herramienta de aprendizaje se toma como herramienta de producción, delegando o desincentivando el entrenamiento de capacidades como la comprensión, el análisis o la memoria (cfr. Rivera-Novoa y Duarte Arias, 2025).

En efecto, las herramientas de inteligencia artificial podrían terminar por reemplazar algunos procesos cognitivos, ya sea debido a su mayor eficiencia, velocidad o simplemente a su mayor confiabilidad. Un enfoque clásico de la cognición argumentaría que si el ser humano desplaza sus tareas cognitivas a máquinas, se menoscabarían sus conexiones neuronales, que toman un *input* y producen un *output*, por la pérdida de capacidades en procesos cognitivos, sobre todo en cerebros biológicos. Estas inteligencias pueden dar lugar a una brecha cognitiva entre diferentes grupos sociales que no acceden a estas máquinas inteligentes, pues "hemos entrado en una segunda edad de la máquina en la que las máquinas no son únicamente un complemento de los humanos, como en la Revolución Industrial, sino que también los sustituyen" (Coeckelbergh, 2021, p. 15). De esta manera, podría surgir una distinción entre humanos de primera y segunda categoría, donde los primeros logren ventajas tecnológicas sobre los segundos.

No obstante, algunos transhumanistas como Kurzweil (2005), Sandberg (2011) y con cautela Bostrom (2014) no verían en esta interacción un detrimento de la cognición, sino un desplazamiento o extensión de los procesos cognitivos. Incluso, como he señalado antes, Kurzweil (2005) promueve el *mind uploading* porque esta copia de la mente no lleva a perder capacidades, sino que las propiedades biológicas cerebrales lograrían simularse o desplazarse a un soporte mecánico o tecnológico mejor y más duradero.

Ahora bien, bajo la perspectiva de las *4E*, especialmente de la tesis de la mente extendida (Clark y Chalmers, 1998), se argumenta que los procesos cognitivos pueden ser extendidos hacia la inteligencia artificial y que dichos procesos realizados con estas herramientas formarían

parte de los procesos cognitivos humanos. En otras palabras, al no limitar los procesos cognitivos a lo que ocurre en el cerebro, cualquier tarea inteligente que se realice mediante la interacción entre la mente, el cuerpo y el mundo puede considerarse un proceso cognitivo mejorado si este da cuenta de que dicho proceso realmente supera al proceso que se lleva a cabo solamente en un soporte biológico como el cerebro. Sin embargo, esta situación en particular podría estar más relacionada con la hibridación tecnológica que con procesos cognitivos exclusivamente desarrollados por la inteligencia artificial. La interacción entre máquinas y seres humanos también se puede considerar un tipo de mejora cognitiva, como lo expone Bostrom; no obstante, como se ha dicho antes, este tipo de mejora obedece a un tipo de hibridación tecnológica que aquí no expondremos.

### 4.3. Mejoras sobrehumanas

Una mejora sobrehumana se presenta cuando el uso de tecnologías conduce a capacidades cognitivas más allá de las capacidades humanas actuales. Por ejemplo, para el transhumanismo, copiar totalmente nuestras funciones cerebrales a una máquina o razonar y memorizar información en niveles superiores a los del *Homo sapiens*; en cualquier caso, una mejora transhumana en la que el éxito cognitivo sea radicalmente superior a las condiciones actuales. En este apartado argumentaré por qué las mejoras sobrehumanas, por lo menos en el estado actual, no son posibles debido a que el uso de las inteligencias artificiales no transforma radicalmente el carácter cualitativo de la experiencia. También este mismo argumento hablará en contra del *mind uploading* argumentado por el transhumanismo debido a la limitación técnica de los recursos actuales de las IA.

Mariano Asla (2021)tanto en función de cotejar las intuiciones espontáneas como para intentar esclarecer racionalmente los límites de lo concebible y de lo probable. Mi propósito en este trabajo es analizar, a la luz de las teorías animalista y psicologista de la identidad personal, la hipótesis de la transferencia mental (TM analiza la idea del *mind uploading* del transhumanismo y postula algunos desafíos a los que se debe enfrentar este propósito de copiar o volcar la mente en las máquinas. Para Asla, el transhumanista suele suponer que la transferencia de la mente preserva la continuidad de la identidad personal. Sin embargo, el punto débil de esta idea radica en que esto implica la discontinuidad espaciotemporal, lo que hace imposible que la identidad se mantenga.

Otra dificultad es que la suposición de que la IA puede igualar la conciencia humana desconoce el vínculo intrínseco entre biología, corporeidad y conciencia. Esto último obedece a una imposibilidad técnica de los recursos actuales si se pretende simplemente "subir" la mente a un soporte no biológico (cfr. Asla, 2021, p. 10)tanto en función de cotejar las intuiciones espontáneas como para intentar esclarecer racionalmente los límites de lo concebible y de lo probable. Mi propósito en este trabajo es analizar, a la luz de las teorías animalista y psicologista de la identidad personal, la hipótesis de la transferencia mental (TM.

Chalmers (1995; 1996, pp. xi y ss.) sostiene, al igual que Nagel (1974), que ciertos aspectos del procesamiento de información pueden explicarse fácilmente mediante modelos mecánicos o causales. Sin embargo, esta explicación es incompleta porque una parte del procesamiento de información está acompañada de un aspecto subjetivo. Este aspecto del carácter cualitativo de la experiencia se relaciona con los criterios de Rowlands (2010), según los cuales un proceso es cognitivo si implica la manipulación y transformación de las estructuras que contienen información. Sostengo que esta transformación cognitiva sobrehumana solo puede darse mediante una transformación radical del aspecto subjetivo de la experiencia. Este aspecto subjetivo es el criterio de experimentar *cómo es ser como* un organismo *X* (cfr. Nagel, 1974). Así, para determinar la emergencia de un estadio evolutivo superior, el procesamiento de información no puede separarse de una transformación radical del carácter cualitativo de la experiencia, dado el vínculo que los une. Una transformación radical del carácter cualitativo de la experiencia implicaría la transformación y manipulación profunda de las estructuras que contienen la información, y no solamente superar el procesamiento informativo que podría realizar un *Homo sapiens*.

A nivel individual, este tipo de mejora afecta a sujetos concretos. Aunque puede suceder, no necesariamente debe verse reflejado en el nivel social o de especie. Su objetivo es que la aplicación de tecnologías permita a un individuo mejorar buenas habilidades y ventajas respecto a otros individuos que inicialmente tenían sus mismas condiciones cognitivas. Por lo tanto, se puede afirmar que una inteligencia artificial mejora la cognición a nivel individual si un sujeto logra aumentar habilidades o capacidades cognitivas mucho más allá que el *Homo sapiens* conduciéndolo a un estadio evolutivo superior.

Las inteligencias artificiales anteriormente mencionadas —Sincrolab, AlphaFold, AlphaZero, AlphaTensor, AlphaDev y ChatGPT,

entre otras IA modeladas de esta misma manera— son ejemplos de mejoras cognitivas sobrehumanas en algunos procesos cognitivos. El límite que tienen estas IA es que sus procesos de información que van más allá del *Homo sapiens* se realizan en objetivos limitados, contextos determinados o disciplinas específicas. Por ejemplo, Sincrolab no podría resolver tareas fuera del objetivo designado sobre el tratamiento del TDAH y por tanto solo allí su mayor precisión, eficiencia y velocidad en el procesamiento de datos supera al ser humano. Ocurre de forma similar con AlphaZero, que solo podrá superar a jugadores del ajedrez. AlphaFold, AlphaTensor, AlphaDev y ChatGPT superan habilidades humanas según el objetivo para el que fueron desarrolladas. Así, una inteligencia artificial débil, es decir, una inteligencia artificial específica o desarrollada para realizar tareas específicas, puede tener capacidades de procesamiento superiores al ser humano: memoria más amplia; procesamiento de información más preciso, rápido; mayor comprensión de algunos datos. La dificultad de estos sistemas es que no pueden acompañar estos datos con una experiencia subjetiva particular, por lo que su procesamiento resulta incompleto.

Una inteligencia artificial prometedora por sus avances en robótica es RT-2 (Google DeepMind, 2023b). Este modelo está clasificado en los llamados *transformer* porque utilizan el aprendizaje por refuerzo (cfr. Mnih *et al*., 2015). RT-2 está preentrenado en comprensión semántica y visual más allá de los datos robóticos a los que fue expuesto. Al ser multimodal por la consecución de datos del lenguaje natural y una cámara adaptada para interactuar con el mundo, el modelo demuestra habilidades emergentes, como el razonamiento en cadena, permitiéndole realizar tareas complejas. En el caso de inteligencias artificiales como RT-2, hay una interacción entre la máquina, los procesos de información y el entorno activo.

A pesar de esta superioridad de la IA en habilidades cognitivas específicas, esto no representa un estadio evolutivo superior. La adquisición de habilidades en el lenguaje natural, la manipulación de información y transformación de estructuras cognitivas por esta adquisición de lenguaje puede significar una superación de algunos procesos cognitivos de la inteligencia, ya sea biológica o artificial, pero esto no significa que por consecuencia se logre un estadio evolutivo superior por sí solo. Es decir, realizar de manera superior un proceso cognitivo no representa un salto evolutivo porque, así como existen cerebros biológicos más adaptados para memorizar, imaginar, procesar

información más rápido o comprender de forma más adecuada representaciones sin que esto conduzca a los humanos más allá del *Homo sapiens*, ocurre de manera similar con las inteligencias artificiales.

Estas limitaciones pueden exponerse con un argumento basado en el experimento mental de Frank Jackson (1982 y 1986). Una variación de este experimento podría suponer que, por alguna razón que se desconoce, una IA ha sido programada con todo el conocimiento proposicional posible sobre los colores. Gracias a esta programación, el sistema de procesamiento de información es capaz de responder cualquier pregunta sobre los colores y sus propiedades físicas. Sin embargo, este sistema nunca ha experimentado el color rojo. Un individuo adapta una cámara a este sistema, acerca un objeto rojo, y posteriormente le pregunta qué se siente ver el rojo. El sistema realiza un proceso de información y describe reacciones neurofisiológicas asociadas con ver el color rojo, aspectos culturales del color, incluso predicciones de cómo respondería un individuo al color rojo. Lo que el sistema no podría responder es qué se siente ver el color rojo porque no tendrá acceso a una fenomenología particular del "rojo".

Aquí a lo que me refiero es que, a pesar de que las IA ofrecen, en un sistema cognitivo, un aumento en algunos procesos cognitivos debido a mayor capacidad de memoria o mayor velocidad de procesamiento de información, dichos procesos son parcialmente constitutivos porque son incompletos. La información brindada por la IA sobre los colores, incluso si es aprehendida por un sujeto humano, no transformará su experiencia consciente. La IA puede ofrecer información sobre espectros ultravioletas o infrarrojos. Sin embargo, aquí el asunto no es solo que la IA no tenga el conocimiento de qué se siente ver esos colores, sino que tampoco los sujetos que usan la IA sabrán qué se siente ver estos colores porque, justamente, se sitúan fuera del espectro humano y del sistema cognitivo.

De esta manera, las máquinas inteligentes pueden complementar ciertos procesos, pero no conducen a un estadio evolutivo superior porque la cognición no se reduce exclusivamente a este tipo de procesos. La inseparabilidad de los procesos cognitivos de la experiencia subjetiva muestra que no es suficiente la manipulación de los datos o la información procesada, sino que además se requiere de la experiencia que acompaña estos datos. De este modo, los procesos apoyados con la IA logran ser complementarios, pero no lo suficientemente constitutivos para conducir a un estadio evolutivo superior.

Cuando se examinan las mejoras cognitivas restaurativas y dentro de los límites, se puede destacar esta misma dificultad que de fondo subyace en las inteligencias artificiales actuales. Una máquina puede realizar procesos cognitivos específicos superiores al ser humano y, sin embargo, no afectar de forma individual, social o como especie al *Homo sapiens* porque dichos procesos no son parte constitutiva del ser humano. En este sentido, Bostrom tendría razón en apostar a la consecución de la superinteligencia desarrollando inteligencia artificial, interfaz cerebro-ordenador, emulación del cerebro biológico y la ingeniería genética. Quizás la superinteligencia sí se logre por el camino de la ingeniería genética por la ventaja que tienen los cerebros biológicos sobre los artificiales.

Cuando hago referencia a que el procesamiento de información de la IA no es constitutivo al ser humano es porque, al exponer los argumentos de Dreyfus, se destaca el papel fundamental que desempeña el entorno en la adquisición del conocimiento. En este sentido, habilidades como la imaginación, el uso de metáforas y una sólida capacidad para tolerar la ambigüedad permiten dar sentido al mundo, aspectos de los que aún carece la inteligencia artificial (cfr. Hoffmann, 2023, pp. 229 y ss.). Sin embargo, estos procesos de información en algunos casos son superiores en la IA por mayor capacidad de memoria, velocidad en el acceso, etcétera.

Lo que tienen en común los límites de la IA expuestos por Searle (1980), Penrouse (1995), Dreyfus (1967), por un lado, y los argumentos acerca de los estados mentales expuestos por Chalmers (1996), Nagel (1974) y Jackson (1986), por el otro, es que señalan que los procesos cognitivos son inseparables de la experiencia cualitativa. Además, para una transformación radical mediante el uso de IA se requiere resolver el problema de la brecha que surge en la modelación o simulación de procesos cognitivos y la irreductibilidad de la experiencia a datos o procesamiento de información. Para una mejora genuinamente sobrehumana se requiere fundamentalmente la completitud de una dimensión experiencial que las IAs actuales no poseen. El límite de las IA no es solo una imposibilidad técnica, sino además intrínseca a los recursos actuales para alcanzar la superinteligencia o la *mind uploading* como lo expone Asla (2021)tanto en función de cotejar las intuiciones espontáneas como para intentar esclarecer racionalmente los límites de lo concebible y de lo probable. Mi propósito en este trabajo es analizar,

a la luz de las teorías animalista y psicologista de la identidad personal, la hipótesis de la transferencia mental (TM.

A pesar de los límites actuales de la IA, estos procesamientos de información permiten la consecución de objetivos específicos que dan cuenta de mejores operaciones que apoyan a los humanos sin que constituyan un estadio evolutivo superior. Estos procesos mejorados por la IA no son constitutivos del ser humano porque cuando se realiza el procesamiento de información por la IA no se están manipulando ni transformando las estructuras que contienen la información directamente por el ser humano.

En otras palabras, dos de las condiciones que debe cumplir un proceso cognitivo no son satisfechas por el ser humano, sino por la máquina, a saber: el proceso mismo de la información y la atribución del proceso como tal. Esta es una de las razones por las que los transhumanistas recurren a la tesis del volcado de la mente o el cerebro a las máquinas para la consecución de la superinteligencia. Debido a los límites expuestos de la IA, si no son posibles las mejoras sobrehumanas que conduzcan a un estadio evolutivo superior en el nivel individual, tampoco serán posibles las mejoras en el nivel social o de especie. Una síntesis de la exposición de la taxonomía anterior es la siguiente tabla:

*Taxonomía de mejoras cognitivas y la inteligencia artificial*

| **Nivel/tipo** | **Restaurativa** | **Dentro de los límites** | **Sobrehumana** |
|---|---|---|---|
| Individual | Las IA proporcionan mejoras en habilidades cognitivas perdidas y pueden otorgar ventajas cognitivas en individuos concretos. | Las IA proporcionan habilidades cognitivas dentro del rango *Homo sapiens* y pueden otorgar ventajas cognitivas a individuos concretos. | La IA tiene habilidades cognitivas específicas superiores al rango *Homo sapiens*; sin embargo, estas habilidades no son constitutivas al ser humano. |

| | | | |
|---|---|---|---|
| Social | Las IA proporcionan mejoras en habilidades cognitivas perdidas y pueden otorgar ventajas cognitivas a grupos. | Las IA proporcionan mejoras en habilidades cognitivas dentro del rango *Homo sapiens* y pueden otorgar ventajas cognitivas en grupos. | La IA, al no lograr una mejora cognitiva sobrehumana constitutiva al individuo, tampoco conduce a mejoras sociales. |
| Especie | La IA proporciona mejoras en habilidades cognitivas perdidas y puede otorgar ventajas cognitivas a la especie *Homo sapiens*. | La IA proporciona mejoras en habilidades cognitivas dentro del rango *Homo sapiens* y puede otorgar ventajas cognitivas a la especie *Homo sapiens*. | La IA, al no lograr una mejora cognitiva sobrehumana constitutiva a grupos sociales ni a individuos concretos, tampoco conduce a mejoras de especie. |

*Tabla 1*

## 5. Conclusiones

El transhumanismo aspira a materializar el surgimiento de una inteligencia artificial general que, eventualmente, a partir de una explosión de inteligencia denominada "singularidad", logre superar al ser humano en varios aspectos o tareas. No obstante, puede afirmarse

que este ideal es solo una ficción, aunque puede otorgar ciertas mejoras restaurativas y dentro de los límites de lo humano.

Las mejoras restaurativas proporcionan a los seres humanos mejoras en capacidades cognitivas perdidas. Las inteligencias artificiales operan a nivel individual, social y de especie. A pesar de que Sincrolab y AlphaFold no son inteligencias artificiales generales, complementan procesos cognitivos que permiten mejoras restaurativas; sin embargo, su aplicación no genera un salto evolutivo en el ser humano debido a los límites intrínsecos en la capacidad de procesar aspectos subjetivos y cualitativos de la experiencia por parte de las IA.

Las mejoras dentro de los límites proporcionan habilidades cognitivas dentro del rango del *Homo sapiens* y pueden conceder ventajas cognitivas. Las inteligencias artificiales que emplean este tipo de mejoras operan a nivel individual, social y de especie. AlphaZero, AlphaDev y ChatGPT no son inteligencias artificiales generales; ahora bien, sus procesos permiten mejoras dentro de los límites porque el procesamiento de información que realizan amplía y aumenta las capacidades en ciertos procesos cognitivos, aunque esto particularmente es incompleto. Estas mejoras no conducen a un estadio evolutivo superior porque estos procesos no son constitutivos debido a que las IA, en los desarrollos actuales, no logran transformar el aspecto cualitativo de la experiencia.

En consecuencia, las mejoras "sobrehumanas" de la IA no proporcionan capacidades cognitivas al *Homo sapiens* que conduzcan a un escenario en el que el ser humano sea una especie radicalmente mejorada. La IA posee capacidades de procesamiento específico superiores al rango *del Homo sapiens*, como se ha señalado anteriormente; sin embargo, al no lograr una mejora cognitiva sobrehumana del individuo, tampoco conduce a mejoras sociales ni de especie. Este argumento no niega el valor de las IA como herramientas cognitivas potentes, sino que cuestiona su capacidad para representar por sí mismas una mejora genuinamente sobrehumana, debido a su fundamental incompletitud experiencial.

Siguiendo el criterio de Rowlands (2010), dos de las condiciones que debe cumplir un proceso cognitivo no son satisfechas por el ser humano, sino por la máquina, a saber, el proceso mismo de la información y la atribución del proceso como tal. Por esta razón, una mejora cognitiva sobrehumana que se produzca en la IA no es intrínseca al ser humano y es por esto por lo que esta mejora no produce un salto evolutivo del *Homo sapiens*. En este artículo he desarrollado el análisis de las mejoras

cognitivas en la inteligencia artificial. Aquí se muestra cómo, con los desarrollos actuales de IA, no se satisface el estadio evolutivo superior, la artificial fuerte o el *mind uploading*. A pesar de esta imposibilidad la IA logra algunas mejoras restaurativas y dentro de los límites humanos.

## Referencias


Asla, M. (2021). Mental Transference (Mind Up-loading): Metaphysical Controversies around the Preservation of Personal Identity. *Forum. Supplement to Acta Philosophica*, 3, 169-183. https://doi.org/10.17421/2498-9746-03-09

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bostrom, N. y Sandberg, A. (2009). Cognitive Enhancement: Methods, Ethics, Regulatory Challenges. *Science and Engineering Ethics*, *15*(3), 311-341. https://doi.org/10.1007/s11948-009-9142-5

Chalmers, D. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, *2*(3), 200-219.

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

Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, *17*(9-10), 7-65.

Clark, A. y Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7-19. https://doi.org/10.1093/analys/58.1.7

Coeckelbergh, M. (2021). *Ética de la inteligencia artificial.* L. Álvarez Canga (trad.). Cátedra.

Diéguez, A. (2016). La singularidad tecnológica y el desafío posthumano. *Pasajes: Revista de Pensamiento Contemporáneo*, 50, 154-164.

Diéguez, A. (2020). La función ideológica del transhumanismo y algunos de sus presupuestos. *Isegoría*, 63, 367-386. https://doi.org/10.3989/isegoria.2020.063.05

Diéguez, A. (2021). *Cuerpos inadecuados el desafio transhumanista a la filosofia*. Herder.

Dobre, C. E. y García Pavón, R. (2024). Singularidad individual versus "singularity". Una crítica al transhumanismo desde el pensamiento de Søren Kierkegaard. *Tópicos, Revista de Filosofía*, 69, 389-420. https://doi.org/10.21555/top.v690.2503

Dretske, F. I. (1981). *Knowledge & the Flow of Information*. MIT Press.

Dreyfus, H. L. (1967). Why Computers Must Have Bodies in Order to Be Intelligent. *The Review of Metaphysics*, *21*(1), 13-32.

Ferry, L. (2018). *La revolución transhumanista*. A. Martorell (trad.). Alianza.

Good, I. J. (1966). Speculations Concerning the First Ultraintelligent Machine. En F. L. Alt y M. Rubinoff (eds.), *Advances in Computers. 6* (pp. 31-88). Elsevier. https://doi.org/10.1016/s0065-2458(08)60418-0

Google DeepMind. (s. f.). *Google DeepMind*. https://www.deepmind.com/

Google DeepMind. (2022, 5 de octubre). *Discovering Novel Algorithms with AlphaTensor.* https://www.deepmind.com/blog/discovering-novel-algorithms-with-alphatensor

Google DeepMind. (2023a, 7 de junio). *AlphaDev Discovers Faster Sorting Algorithms*. https://www.deepmind.com/blog/alphadev-discovers-faster-sorting-algorithms

Google DeepMind. (2023b, 28 de julio). *RT-2: New Model Translates Vision and Language into Action*. https://www.deepmind.com/blog/rt-2-new-model-translates-vision-and-language-into-action

Google DeepMind. (2023c, 19 de septiembre). *A Catalogue of Genetic Mutations to Help Pinpoint the Cause of Diseases*. https://www.deepmind.com/blog/alphamissense-catalogue-of-genetic-mutations-to-help-pinpoint-the-cause-of-diseases

Hoffmann, C. H. (2023). *The Quest for a Universal Theory of Intelligence: The Mind, the Machine, and Singularity Hypotheses*. De Gruyter. https://doi.org/10.1515/9783110756166

Jackson, F. (1982). Epiphenomenal Qualia. *The Philosophical Quarterly*, *32*(127), 127-136. https://doi.org/10.2307/2960077

Jackson, F. (1986). What Mary Didn't Know. *The Journal of Philosophy*, *83*(5), 291-295. https://doi.org/10.2307/2026143

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. Viking.

Larson, E. J. (2021). *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. The Belknap Press of Harvard University Press. https://doi.org/10.4159/9780674259935

McCarthy, J. (1978). History of LISP. *ACM SIGPLAN Notices*, *13*(8), 217-223. https://doi.org/10.1145/960118.808387

Minsky, M. (1988). *The society of mind*. Simon & Schuster.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. y Hassabis, D. (2015). Human-level Control through Deep Reinforcement Learning. *Nature*, *518*(7540), 529-533. https://doi.org/10.1038/nature14236

Moravec, H. P. (2000). *Robot: Mere Machine to Transcendent Mind*. Oxford University Press.

Nagel, T. (1974). What Is It Like to Be a Bat? *Philosophical Review*, *83*(4), 435-450.

Newell, A. y Simon, H. A. (1972). *Human Problem Solving*. Prentice Hall.

OpenAI. (s. f.). *ChatGPT*. https://openai.com/chatgpt

Parens, E. (1998). Special Supplement: Is Better Always Good? The Enhancement Project. *The Hastings Center Report*, *28*(1), S1-S17. https://doi.org/10.2307/3527981

Penrose, R. (1995). Beyond the Doubting of a Shadow: A Reply to Commentaries on *Shadows of the Mind*. *PSYCHE: An Interdisciplinary Journal of Research on Consciousness*, 2, 1-40.

Preston, J. (2002). Introducción. En M. Bishop y J. Preston (eds.), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* (pp. 1-50). Clarendon Press. https://doi.org/10.1093/oso/9780198250579.003.0001

Rivera-Novoa, Á. (2020). Mente extendida y transhumanismo. ¿Qué tan humana es la mente de un cyborg? En O. M. Donato Rodríguez, D. M. Muñoz González y Á. Rivera Novoa (eds.), *Redefinir lo humano en la era técnica: perspectivas filosóficas* (pp. 75-89). Universidad Libre. https://repository.unilibre.edu.co/handle/10901/19869

Rivera-Novoa, Á. (2024). La tesis de la mente extendida y el ideal transhumanista de mejoramiento cognitivo. *Trilogía. Ciencia Tecnología Sociedad*, *16*(33), e3142. https://doi.org/10.22430/21457778.3142

Rivera-Novoa, Á. y Duarte Arias, D. A. (2025). Generative Artificial Intelligence and Extended Cognition in Science Learning Contexts. *Science & Education*. https://doi.org/10.1007/s11191-025-00660-1

Rowlands, M. (2010). *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. MIT Press. https://doi.org/10.7551/mitpress/9780262014557.001.0001

Russell, S. J. y Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Pearson.

Sandberg, A. (2011). Cognition Enhancement: Upgrading the Brain. En J. Savulescu, R. ter Meulen y G. Kahane (eds.), *Enhancing Human Capacities* (pp. 71-91). Wiley-Blackwell. https://doi.org/10.1002/9781444393552.ch5

Sarne, D. y Grosz, B. J. (2007). Sharing Experiences to Learn User Characteristics in Dynamic Environments with Sparse Data. En E. H. Durfee, M. Yokoo, M. N. Huhns y O. Shehory (eds.), *Proceedings of the 6th International Joint Conference on Autonomous Agents and*

*Multiagent Systems* (pp. 202-209). Association for Computing Machinery. https://doi.org/10.1145/1329125.1329176

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, *3*(3), 417-424. https://doi.org/10.1017/S0140525X00005756

Sincrolab. (s. f.). *Sincrolab. Recuperación cognitiva y neuropsicología digital*. https://sincrolab.es/

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

Vinge, V. (1993). *The Coming Technological Singularity: How to Survive in the Post-Human Era*. https://api.semanticscholar.org/CorpusID:5750614