¿Puede la inteligencia artificial dar una “mente” a las máquinas? Reflexiones y preguntas desde la filosofía

Francisco José Soler Gil

Universidad de Sevilla

1. Introducción: ¿La era de la inteligencia artificial?

De un tiempo a esta parte se multiplican las voces que anuncian que nos encontramos a las puertas de una gran revolución. Quizás la mayor, o incluso la última, de las revoluciones de nuestra historia: la eclosión de la inteligencia artificial, o, como algunos denominan, la «singularidad».

¿Por qué la mayor de las revoluciones, o por qué incluso la última? Porque el poder está íntimamente ligado a la inteligencia. De modo que, así como el hombre ha podido dominar sobre el resto de las especies biológicas debido a su superior inteligencia, el desarrollo de máquinas de inteligencia superior a la humana arrebataría eventualmente a nuestra especie la posición central en la estructura de control del mundo.

¿Y por qué «singularidad»? Porque la creación de una máquina dotada de mayor inteligencia que el hombre implicaría que, entre otras capacidades, esa máquina podría desarrollar mejores sistemas de inteligencia artificial que los diseñados por sus propios creadores, con lo que se entraría en una espiral de crecimiento acelerado de inteligencia, cuyos únicos límites podrían ser los límites en la capacidad de almacenamiento y procesamiento de datos impuestos por las leyes de la naturaleza.

Un lector profano, como sin duda lo es el autor de estas líneas, atiende a estas voces con cierto escepticismo, pero no con desdén. Con cierto escepticismo, porque los escenarios que nos describen los profetas de la inteligencia artificial poseen en parte rasgos que parecen extraídos de los relatos de ciencia ficción. Y también porque los que ya vamos acumulando cierta edad guardamos en la memoria un rinconcito de anuncios no cumplidos, entre los que se incluyen los coches voladores, la colonización del espacio, y otras maravillas que en nuestra adolescencia se decían inminentes.

Con cierto escepticismo pero no con desdén, porque también se nos decía, por ejemplo, que ninguna máquina podría jugar al ajedrez tan bien como un maestro, y ya vemos lo que ha ocurrido finalmente (... y no sólo con el ajedrez, sino incluso con el juego del go, que es un juego incomparable más complicado, y que no puede dominarse por mero cálculo exhaustivo de posibilidades). Y también recordamos las bromas que gastábamos a propósito de las primeras versiones del traductor de google, y de otros intentos de reproducir artificialmente habilidades que nos parecían que sólo estaban al alcance de la inteligencia humana.

Por eso, cuando consideramos los logros obtenidos por las diversas líneas de investigación en el campo de la inteligencia artificial en los últimos años, es normal que surjan preguntas: ¿Hasta dónde puede llegar este desarrollo? ¿Existe alguna habilidad tan propia de la inteligencia humana que no pueda en modo alguno ser implementada por la inteligencia artificial? ¿O más bien debemos ir acostumbrándonos a la idea de que poco a poco seremos superados por las máquinas en todos los ámbitos asociados con el uso de la inteligencia? ¿Se mantendrá la actual limitación de los programas de inteligencia artificial a la realización de tareas específicas, o se llegará a desarrollar una máquina que pueda integrar todas esas habilidades particulares, y que actúe por tanto cuasi-humanamente de modo global?

En líneas generales hay tres tipos de respuesta a estas preguntas:

(1) La respuesta **abiertamente optimista** con respecto a las posibilidades de la inteligencia artificial, a la que se adscriben aquellos autores que consideran que es factible diseñar una máquina de inteligencia global no inferior a la humana, y que además estamos ya relativamente cerca de conseguir ese objetivo.

(2) La respuesta **moderadamente optimista** con respecto a las posibilidades de la inteligencia artificial, a la que se adscriben aquellos autores que consideran que es factible en principio diseñar una máquina de inteligencia global no inferior a la humana, pero que se trata de un objetivo difícil. Tan difícil que puede que aún se tarde siglos en alcanzar, o que de hecho no llegue a alcanzarse nunca.

(3) La **respuesta pesimista** con respecto a las posibilidades de la inteligencia artificial, a la que se adscriben aquellos autores que consideran que es imposible diseñar una máquina de inteligencia global no inferior a la humana.

De las tres respuestas, la que he denominado «pesimista» puede resultarnos de especial utilidad, de cara a explorar las posibilidades y los límites de la inteligencia artificial, puesto que afirma que tales límites existen, y que se encuentran por debajo del nivel de la inteligencia humana. Y esta es una posición que invita a reflexionar más despacio sobre las similitudes y diferencias entre la mente humana y los procesos que tienen lugar en las máquinas dotadas de inteligencia artificial.

Dirijamos, pues, nuestra atención hacia esa línea argumentativa, planteando la cuestión de cuál es la raíz del escepticismo con respecto a las posibilidades del campo de la inteligencia artificial de cara a producir una réplica (quizás incluso mejorada) de la inteligencia humana.

Si uno atiende a las voces que defienden esta postura pesimista, o escéptica, encuentra sobre todo un planteamiento que se repite con frecuencia: la insistencia en que la composición y arquitectura del cuerpo humano en general, y del cerebro en particular, son muy diferentes a la composición y arquitectura de cualquier máquina, incluyendo, por supuesto, los más avanzados ordenadores. De manera que los procesos que tienen lugar en unas y otras entidades tienen que ser por fuerza diferentes, y no equiparables. En consecuencia, cuando hablamos de «inteligencia» en referencia a una máquina, no estamos usando el término «inteligencia» unívocamente, sino analógicamente. Todo el campo de investigación de la inteligencia artificial estaría edificado sobre una serie de decisiones terminológicas como esta, que introducen analogías entre lo humano y ciertos artificios mecánicos, y con ello introducen también ambigüedades, y dan lugar a malentendidos.

Desde luego, la referencia al uso analógico de términos mentales en el marco de las investigaciones de la inteligencia artificial apunta a un tema importante, y conviene seguir esa pista, para ver hasta dónde puede llevarnos. Pero antes de seguirla, conviene que subrayemos que el hecho de que el discurso de la inteligencia artificial se construya sobre analogías no es en sí algo problemático. Más bien se trata de un procedimiento legítimo y frecuente en la investigación científica y filosófica. De hecho, la mayor parte de los conceptos que se emplean en las descripciones filosóficas y físicas de la naturaleza nacieron en su día como analogías. No obstante, sí es cierto que cualquier analogía, por útil que resulte, tiene sus limitaciones. Y que conviene no pasarlas por alto.

2. Analogías en el campo de la inteligencia artificial y sus limitaciones: el problema de la subjetividad

Como no cabía esperar de otro modo, también en el ámbito de la inteligencia artificial se emplean conceptos y modelos analógicos.

La primera de las analogías, seguramente la principal de ellas, y también la menos problemática, es la que se establece entre cerebro y ordenador, afirmando que, en esencia, cerebro y ordenador operan de un modo tan similar que podemos decir que el cerebro es como un ordenador biológico, o que el ordenador es como un cerebro sintético.

Evidentemente, las enormes diferencias materiales y estructurales entre un cerebro y un ordenador saltan a la vista desde el primer momento. Empezando por la composición química (orgánica en un caso, inorgánica en el otro), y siguiendo por los aspectos estructurales. No obstante, si hacemos abstracción de todas estas diferencias, nos sigue quedando un núcleo común suficientemente significativo como para dar lugar a una analogía robusta: En uno y otro caso nos estamos refiriendo a entidades que transforman estímulos de entrada en señales eléctricas que son incorporadas al flujo de impulsos que se intercambian los componentes de una compleja red de interacciones, dando lugar a ciertos cambios en la estructura de dicha red, y finalmente a una acción hacia el exterior, por parte de la entidad en cuestión.

Desde esta perspectiva, podemos decir que cerebro y ordenador «trabajan» de un modo análogo. Y esta analogía está siendo fructífera en muchos sentidos, como por ejemplo el de orientar en parte el diseño de los ordenadores de forma que su funcionamiento reproduzca el funcionamiento de las redes neuronales en el cerebro, con resultados interesantes en muchos ámbitos (desde los recientes progresos en el reconocimiento de imágenes, voz y texto, al análisis en general de situaciones en las que no es factible un cálculo exhaustivo de todas las posibilidades).

Ahora bien, la analogía entre cerebro y ordenador, nos deja a las puertas de todo un paquete de analogías que enlazan de modo natural con ella, y que, aun siendo también muy útiles para orientar heurísticamente las investigaciones en inteligencia artificial, conllevan un alto precio. Este paquete tiene como núcleo la denominada «analogía de la información», que

establece un puente entre los procesos mentales conscientes en el hombre y el procesamiento de la información en el ordenador.

«Información» es el concepto clave, y se trata de un concepto muy importante, porque apunta a un contenido inmaterial, y de este modo parece ofrecernos una vía para afrontar el enigma de la relación que se da en el hombre entre mente y materia: ¿Cómo puede algo inmaterial, como parece ser la mente (invisible, intangible, inapresable por medio de magnitudes físicas), estar ligado tan íntimamente con algo tan material como el cerebro y su red de neuronas conectadas por impulsos eléctricos y químicos?

Para responder a esta pregunta, se sugiere que los procesos mentales son como el procesamiento de información en un ordenador. Y que así como la información también parece responder claramente a lo que entendemos por una realidad inmaterial, intangible, invisible, etc., y sin embargo es justo aquello que almacenan y elaboran y transmiten los ordenadores, del mismo modo podríamos entender que las operaciones de la mente humana son operaciones de almacenamiento, elaboración y transmisión de información. De manera que «información» sería por tanto una especie de concepto puente, que nos permitiría anclar lo mental a lo material.

Desde esta perspectiva, la pregunta que estamos tratando de abordar en este texto —es decir, la pregunta de si podrá el hombre dotar de mente a las máquinas— tendría una respuesta claramente afirmativa como cuestión de principio. Es decir, que no habría ningún misterio en ese dotar de mente a las máquinas; no habría ningún misterio, sino un simple problema técnico: el problema de diseñar los programas informáticos adecuados capaces de conseguir que los ordenadores recojan y procesen la información del entorno tan bien como los seres humanos, no sólo en lo relativo a determinadas tareas (como jugar al ajedrez, o al go, o como reconocer voces o rostros etc.) sino con relación a todas las tareas de las que se ocupa la mente humana. De manera que, puesto que el pensamiento sería procesamiento de información, una máquina que procesara información de manera general al nivel del hombre (o por encima incluso de este nivel) sería automáticamente una máquina pensante.

Ahora bien, aunque no se puede negar que esta analogía entre el pensamiento y el tratamiento de información ha sido muy estimulante, y está sirviendo de guía en el desarrollo del campo de la inteligencia artificial, lo cierto es que en este caso el puente conceptual

establecido se paga a un alto precio, y concretamente al precio de dejar fuera del análisis el fenómeno de la conciencia subjetiva (con su autoconciencia incluida), o, como suele denominarse en la jerga filosófica, la «perspectiva de la primera persona».

Para entender lo anterior, fijémonos en la diferencia existente entre una persona que se informa y un ordenador que procesa información. Consideremos una vieja rutina, que va camino poco a poco de la desaparición: la lectura de un periódico.

Este proceso puede analizarse desde un punto de vista objetivo, en la perspectiva de la tercera persona, con ayuda de la teoría de la información. Desde esta perspectiva, lo que proporciona el periódico es un conjunto de datos visuales (letras dispuestas de cierto modo), que son codificados como impulsos eléctricos en el ojo, y pasan al cerebro, donde esos datos son sometidos a diversos procesos de análisis, que permiten reajustar la representación de la realidad que sirve de base para producir impulsos cerebrales que determinarán acciones de respuesta. Considerado el proceso así, lo esencial de una información es que se trata de un dato que permite reducir en alguna medida la cantidad de incertidumbre que incluye el mapa cerebral de la realidad en un momento dado. Y, ciertamente, así considerado, lo mismo podemos estar hablando de un hombre que de un ordenador.

Pero cuando una persona repasa el periódico, junto al proceso anterior (y paralelamente a él) están ocurriendo otras cosas: el proceso de información va acompañado de experiencias subjetivas: sentimientos, representaciones mentales de personajes y escenas, divagaciones mentales, evocaciones de colores, de sonidos, etc. Informarse, desde este punto de vista, es un «darse cuenta» de las cosas. Un darse cuenta a sí mismo.

Todo este mundo de la perspectiva subjetiva, de la perspectiva en primera persona, resulta informulable desde la perspectiva objetiva que sirve de base a la analogía de la información. Y por eso, podemos decir que esa analogía se paga al precio de reducir la realidad de la experiencia humana del informarse a la dimensión objetivable de dicha experiencia. Y esa reducción al plano de lo objetivable lastra toda una cadena de analogías asociadas con la anterior. Pues en el contexto de los ordenadores se hablará entonces de «memoria», de «inteligencia», de «aprendizaje» etc., pero si uno presta atención a las definiciones que se

ofrecen usualmente de dichos conceptos en este ámbito se dará cuenta enseguida de que tales definiciones excluyen la dimensión subjetiva asociada a los mismos en la experiencia humana.

Y así, por ejemplo, definiremos en este contexto «memoria» como capacidad física de almacenar datos haciendo que un sistema físico se encuentre situado en un estado de entre varios en los que puede hallarse. Y hacer uso de la memoria simplemente consistirá en comprobar en qué estado se encuentra ese sistema, para recuperar así el dato que codificamos al situar en dicho estado al sistema. No se trata de negar que algo de esto hace también nuestro cerebro. Pero claro está que semejante perspectiva no puede sino dejar fuera todo el aspecto subjetivo del proceso humano de recordar, de evocar acontecimientos o pensamientos que uno tuvo en un momento dado.

Pues bien, si tenemos en cuenta que la analogía de la información y las analogías asociadas a ella (de la memoria, del aprendizaje, de la inteligencia...) que se emplean en el campo de la inteligencia artificial suponen una reducción de la realidad humana al plano de lo objetivable, la cuestión de la que tenemos ocuparnos aquí, la de si podremos dotar de mente a las máquinas, se presenta como mucho más complicada. Y el objetivo de dotar de mente a las máquinas puede comenzar a ser visto como algo tal vez inalcanzable, al menos en el marco de las investigaciones actuales en el campo de la inteligencia artificial.

El problema, en definitiva, es la subjetividad. Y es la conciencia que abre esa perspectiva subjetiva.

¿Podríamos dotar a las máquinas de una conciencia que las convirtiera en sujetos, es decir, que les dotara de subjetividad? ¿Cómo?

Y si no podemos hacer esto, ¿podríamos al menos construir máquinas que se comporten en todos los sentidos y contextos como un ser humano, pero sin tener conciencia, y por tanto subjetividad? ¿O la diferencia entre tener perspectiva subjetiva y no tenerla se manifestará necesariamente en diferencias objetivas (determinadas habilidades, determinadas conductas) entre la máquina y el hombre? Y si se tienen que dar ciertas diferencias insalvables entre seres dotados de subjetividad y seres que no la tienen, ¿en qué habilidades o conductas se percibirán estas diferencias?

Ninguna de estas preguntas es retórica. Se trata de interrogantes reales que se abren ante el que trata de acercarse al problema de la mente y la inteligencia artificial. Y son muchos interrogantes. Y arduos. No puedo ofrecer una respuesta a los mismos. Pero sí que puedo al menos repasar en qué consisten las alternativas principales.

3. De apariciones inesperadas, zombis y habilidades imposibles

Teniendo en cuenta las preguntas que acabo de formular, se abren ante nosotros tres escenarios principales:

(1) El primero de ellos es que una máquina dotada de inteligencia a nivel humano en el sentido analógico mencionado en el apartado anterior, resulte que tiene conciencia y subjetividad. A este escenario podríamos denominarlo «aparición inesperada de la conciencia».

(2) El segundo escenario consiste en que sea posible construir una máquina indistinguible de un ser humano, pero sin conciencia, y por tanto sin perspectiva subjetiva. A este escenario podemos denominarlo «creación de un zombi».

(3) Finalmente, cabe la posibilidad de que el estar dotado o no de subjetividad implique necesariamente diferencias entre lo que hombres y máquinas puede llegar a hacer. En este caso, hablaríamos de «habilidades imposibles» para una máquina sin conciencia.

Concluyamos, pues, el texto, esbozando, por encima, estos tres escenarios.

3.1 Aparición inesperada de la conciencia

En el apartado anterior he insistido en que los términos «inteligencia», «memoria», «aprendizaje», etc., tal y como se emplean en el ámbito de la inteligencia artificial, constituyen en realidad proyecciones al plano de la objetividad de realidades más ricas del mundo de lo humano. Pero ¿no podría ocurrir que una máquina construida de tal modo que poseyera una inteligencia objetivo-funcional semejante a la humana, inmediatamente poseyera también una inteligencia semejante a la humana en todo lo demás? Es decir, ¿no cabría conjeturar, por ejemplo, que la conciencia, y con ella la perspectiva de la primera persona aparece

espontáneamente como una propiedad emergente de los sistemas físicos que son capaces de manejar información a la manera en que lo hace el cerebro?

Como no sabemos realmente el origen de la conciencia, no podemos descartar de entrada esta hipótesis, y en realidad, poco más podemos decir de ella. Si acaso advertir que la conjetura de la «emergencia» de la conciencia puede ser formulada en diversas variantes, y no todas son prometedoras de cara al objetivo de dotar de mente a una máquina. Pues, por ejemplo, hay también quien sostiene que la emergencia de la conciencia sólo tiene lugar en sistemas físicos que posean propiedades muy determinadas, entre ellas una determinada composición química, orgánica, y una determinada arquitectura (como la red neuronal del cerebro). Y si esto fuera así, entonces ninguna de las máquinas que se están construyendo actualmente, o que se construirán en el futuro previsible, podría llegar nunca a estar dotada de mente.

En realidad, el problema de esta hipótesis es que no parece haber forma de justificar que la «emergencia» de la mente tenga que ocurrir sobre una base material determinada en lugar de sobre otra. Más concretamente: no sabemos si la emergencia se produciría allá donde se den determinadas propiedades funcionales de la materia a cierto nivel (por ejemplo, capacidad de manejar información de tales o cuales modos), o si más bien requeriría un soporte material determinado, más allá de esas propiedades funcionales. Por eso, cualquier especulación en esta línea parece un tanto arbitraria.

3.2 La creación de zombis

En vez de intentar justificar que estas o aquellas configuraciones materiales tienen de suyo conciencia, supongamos entonces que es probable que nuestras máquinas no la posean, puesto que no sabemos cómo surge, y sería más bien una rara casualidad el que diéramos con ella por mero tanteo.

¿Podríamos, aun así, construir una máquina tal que fuera capaz de actuar en todas las situaciones como lo haría un ser humano, es decir, cabría construir una máquina con inteligencia artificial general de nivel humano?

Si esto fuera posible, lo que habríamos creado es un «zombi»: un ser que actúa, que interacciona con los hombres y con su entorno en general, que analiza problemas, se propone objetivos, traza planes, etc., pero sin que posea subjetividad. No tiene perspectiva de la primera persona, pero visto desde el plano objetivo de la tercera persona, no se distingue de un ser con subjetividad. ¿Es esto posible?

Hay quienes afirman que no es posible que se den diferencias ontológicas donde no se dan diferencias empíricas, y en ocasiones remiten como apoyo al llamado «principio de identidad de los indiscernibles», formulado por Leibniz, que vendría a decir que dos entidades que no presentan ninguna diferencia empírica son en realidad la misma.

Pero es dudoso que esto tenga que ser necesariamente así. En física, sin ir más lejos, abundan los ejemplos de escenarios ontológicamente distintos pero empíricamente equivalentes. Por mencionar tan solo dos de ellos: (1) la mecánica cuántica y la mecánica bohmiana describen dos escenarios ontológicos muy diferentes, y sin embargo no se puede distinguir empíricamente entre ellas; y (2) el modelo cosmológico estándar es empíricamente equivalente a una infinidad de modelos cuya métrica difiere de la del modelo estándar más allá de los límites del universo observable.

Y dejando aparte estos ejemplos, quizás un tanto alejados del tema humano, lo cierto es que desde un punto de vista meramente físico no hay diferencia entre un escenario en el que ciertas decisiones humanas son azarosas y otro en el que dichas decisiones son libres. Consideradas las dos situaciones desde el plano físico, no hay diferencia. Pero realmente un mundo en el que las acciones humanas son libres y otro en el que las acciones son azarosas no son en modo alguno idénticos.

Por tanto, ningún principio parece impedir de entrada la posibilidad de que también en el caso de las máquinas dotadas de inteligencia artificial general y los seres humanos ocurriera lo mismo. Es decir, que unas y otros fueran muy distintos (por carecer las máquinas de conciencia) y sin embargo resultaran empíricamente indistinguibles.

Sin embargo, aunque esta alternativa no parece de entrada imposible, sí que hay que reconocer que intuitivamente no resulta atractiva. Al examinarla, es difícil dejar de lado la sospecha de que la naturaleza no hace nada en vano, y que, por tanto, si estamos dotados de

conciencia, es porque con ellas se pueden hacer cosas que no pueden hacerse sin ella. Si tomamos en serio esta consideración, entonces tenemos que pasar al tercer escenario.

3.3 ¿Hay habilidades humanas imposibles de alcanzar por las máquinas?

Si las máquinas que fabricamos, por muy hábiles programas que ejecuten, no pueden tener conciencia, y si tiene que existir alguna diferencia empírica entre las cosas que puede hacer un ser sin conciencia y otro con ella, entonces debería de haber algún tipo de habilidad humana inalcanzable para las máquinas dotadas de inteligencia artificial. ¿Las hay? O al menos, ¿podemos sospechar que las hay? ¿Y cuáles podrían ser?

Desde luego, la experiencia pasada ha sido desalentadora, por lo que se refiere a la determinación de límites a la inteligencia artificial. Se dijo que ninguna máquina podría vencer a un campeón de ajedrez, y lo «imposible» ocurrió. Como acaba de ocurrir con el juego de go, y como quizá ocurra en breve con la conducción de automóviles, y tantas otras actividades...

Pero aun así, puede resultar interesante el ejercicio de cuestionarse qué podría resultar especialmente difícil de lograr por una máquina sin subjetividad.

Evidentemente, si queremos buscar habilidades exclusivamente humanas, tendríamos que hacerlo entre aquellas en las que la comprensión subjetiva parezca más insustituible.

En su famoso experimento mental de la «caja china», Searle comparó el ordenador con una máquina en la que hubiera una persona dentro que, sin saber chino, tuviera un manual de instrucciones en el que se indicara «cuando se introduzcan en la máquina tales signos chinos, responda proyectando en la pantalla tales otros». Con un manual de instrucciones suficientemente bueno (o sea: con un programa suficientemente bueno) la impresión que se tendría es que el operario (o el ordenador) sabe conversar en chino. Pero lo cierto, dice Searle, es que no sabe, porque se limita a aplicar unas reglas sin entender el significado de los términos que emplea.

Reflexionando sobre este experimento mental podríamos plantearnos si acaso es posible entender en profundidad el significado de las palabras sin poder contemplar desde una perspectiva subjetiva los objetos a los que se refieren. Lo que al operario en la máquina le falta

en el fondo es la conexión entre los signos chinos que recibe y el contenido subjetivo que asociamos con las palabras. Sin este conocimiento, que requiere la actividad de la conciencia, el operario podía mantener un simulacro de conversación en chino, con ayuda del manual de instrucciones. Pero ¿hubiera podido también usar el lenguaje de manera creativa? Quiero decir, ¿podría por ejemplo descubrir metáforas y analogías nuevas con sentido? ¿Podría hacer esto una máquina inconsciente?

Dicho de otro modo: ¿Dota la conciencia (y con ella la subjetividad) de una capacidad de generar ideas nuevas que no puede darse sin ella?

Lo cierto es que no lo sé. Y que por eso esta pregunta, como todas las que han ido apareciendo en las páginas anteriores, tengo que dejarla abierta.