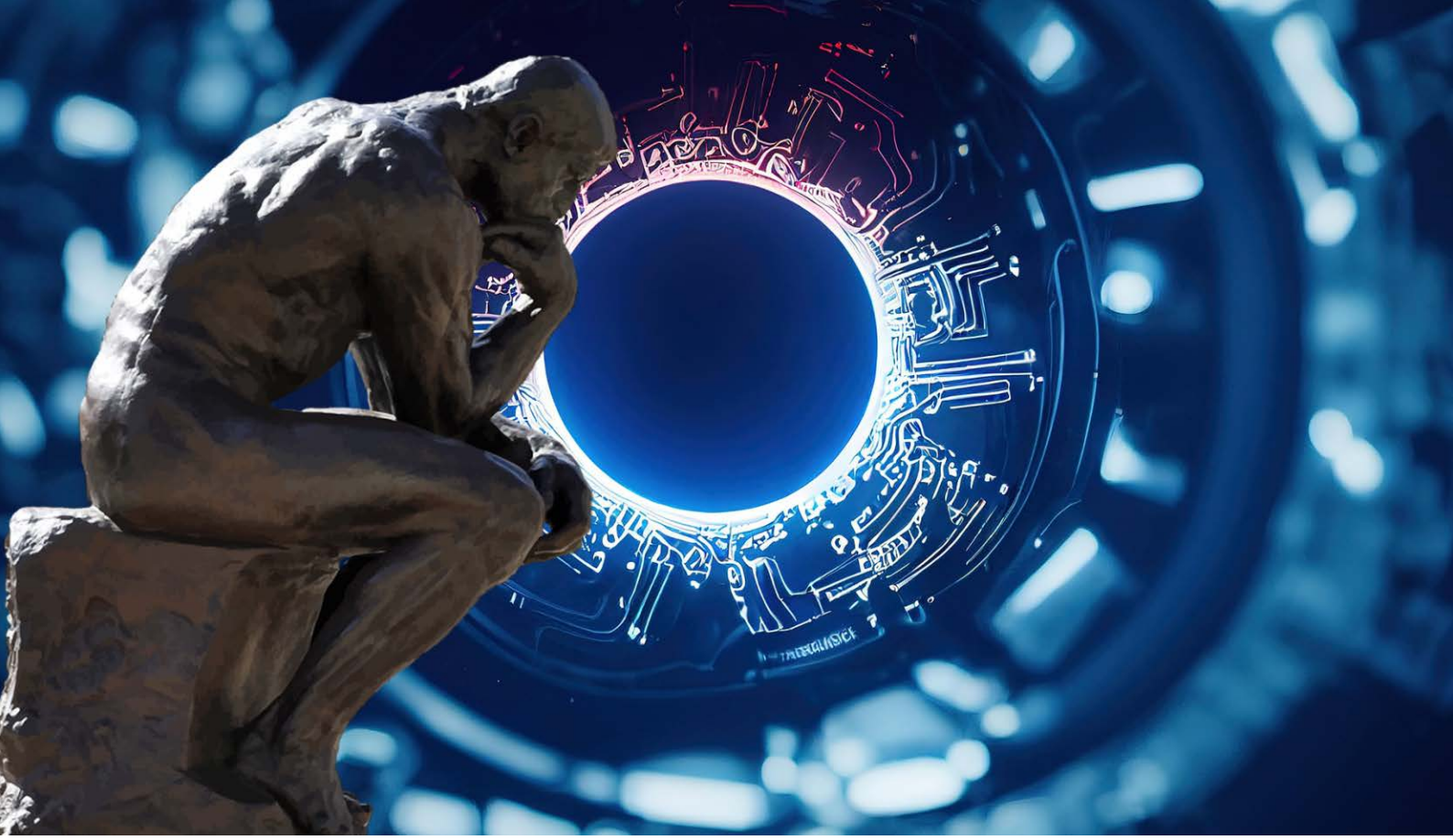


La **inteligencia artificial** y la **enseñanza** de la **filosofía***

*En este texto queremos plantear una serie de **cuestiones filosóficas previas** o preliminares (no técnicas) que permitan **entender mejor** de qué se está hablando **cuando se habla de inteligencia artificial (IA)**, así como afrontar algunas de sus **implicaciones epistemológicas**. De este modo, consideramos que la **filosofía** puede servir a **formar** al alumnado y **contribuir a la reflexión de la sociedad** en su conjunto con el objetivo de que se puedan encarar con **responsabilidad y conciencia crítica** los **enormes retos** que plantean actualmente los distintos **desarrollos de la IA**, y los que seguirán planteando **en el futuro**.*

ERNESTO BALTAR
Universidad Rey Juan Carlos

Nuestra tesis es que la filosofía puede ayudar a coordinar y organizar ámbitos de estudio tan amplios y diferenciados como los que se concitan en el campo de la IA, merced a su visión global, externa y de conjunto, y a su capacidad de aplicar conceptos, clasificaciones y categorías racionales para tratar de comprender toda la complejidad inherente a este tema.

* **Nota:** Este artículo es una versión modificada (reducida en algunos apartados y ampliada en otros) del capítulo de Ernesto Baltar "La necesidad social y educativa de la Filosofía ante los retos de la Inteligencia Artificial", publicado por Álvaro Ledesma y Herminio Pagola (eds.), *Para todo y para nada. Miradas contemporáneas de la filosofía*, Servicio de Publicaciones de la Universidad de La Rioja, Logroño, 2023, pp. 151-168.

INTRODUCCIÓN: UN ENFOQUE FILOSÓFICO DE LA IA

Cuando defendemos la utilidad de un enfoque filosófico de la inteligencia artificial (IA) nos referimos a la necesidad de plantear una serie de cuestiones preliminares o previas, no estrictamente técnicas o científicas, sobre este tema. La IA plantea cuestiones importantes a muchos niveles: a nivel epistemológico, metodológico, de teoría del conocimiento, de lógica, de teoría de la ciencia, de teoría de la mente, así como a nivel de la concepción antropológica, de la teoría de la acción y de la decisión, de las consecuencias éticas, e incluso hay un trasfondo ontológico de la cuestión. Además, al ser un objeto de estudio interdisciplinar, contiene elementos de las matemáticas, de la computación, de la psicología, de la lingüística, de la neurociencia, la cibernética, etc. Por ejemplo, en el campo de la "ciencia cognitiva" (que puede ser un buen punto de referencia interdisciplinar para abordar las cuestiones de la IA que tienen que ver con la cognición) se entrecruzan y solapan muchos de esos campos de estudio.

En primera instancia, parece imposible manejar y conciliar ámbitos tan distintos, pero nosotros consideramos que si se puede intentar es desde la filosofía, porque ve las cosas desde fuera y puede aplicar conceptos, clasificaciones y categorías racionales para organizar todo ese panorama tan complejo y así tratar de comprenderlo. Dentro de las numerosas cuestiones filosóficas que plantea la inteligencia artificial, en este texto queremos centrarnos en tres:

> La filosofía puede manejar y conciliar los distintos ámbitos que se concitan en la IA, porque ve las cosas desde fuera y aplica conceptos, clasificaciones y categorías racionales para organizar y comprender ese panorama tan complejo

1) En primer lugar, queremos aclarar el concepto mismo de "inteligencia artificial". Como es una expresión ya establecida y generalizada la damos por supuesta y por conocida y no nos la planteamos, pero en el momento en que uno se pone a analizarla se da cuenta de que es realmente muy problemática. Creemos que aquí se encuentra el origen de muchas confusiones, pues se utiliza de forma equívoca un término que es claramente analógico, pero no se fundamenta ni explica la analogía. En consecuencia, se utiliza la misma palabra para designar cosas cualitativamente distintas, de modo que se cae en constantes malentendidos y se hace imposible el diálogo o el contraste de argumentos.

2) En segundo lugar, y directamente relacionado con lo anterior, queremos abordar muy brevemente las distintas concepciones o planteamientos que hay sobre la inteligencia artificial, que se pueden clasificar fundamentalmente en cuatro tipos. Nos centraremos sobre todo en uno de ellos, que es el más amplio y extendido y el que resulta más interesante: el llamado "modelo del agente racional" o "del agente inteligente". Y enumeraremos algunos problemas filosóficos de primera magnitud que plantea, como la racionalidad, el aprendizaje, la toma de decisiones, la intencionalidad, la conciencia, la introspección, la autoconciencia, el lenguaje, la memoria, la percepción, etc.

3) En tercer lugar, hay consecuencias o derivadas éticas de la IA (en las que no en-

traremos aquí por cuestiones de espacio, aunque no son menos relevantes). Las nuevas tecnologías crean situaciones nuevas, que antes no existían, y muchas veces surgen nuevos problemas, y hay que analizarlos y tratar de darles una solución. Por supuesto, no son sólo problemas éticos, sino que también crean problemas nuevos a nivel legal, social, político, etc. Generalmente se suele decir que las leyes van por detrás de la realidad, pero hay cuestiones en las que no se puede esperar a ver qué pasa porque son irreversibles y no tienen vuelta atrás, o causan un daño irreparable y que ya no se puede subsanar.

Evidentemente, quienes piensen que se pueden dejar todos estos temas en manos de los expertos, de los técnicos o de los ingenieros informáticos, seguramente ni se planteen las derivadas éticas. En el ámbito de la técnica parece imperar una lógica, una *tecno-lógica* podríamos decir, que asume que todo lo que sea técnicamente posible se acabará haciendo más tarde o más temprano[1]. En la asunción de esa actitud fatalista se concitan los intereses de muchos grupos sociales, empresas y personas.

EL CONCEPTO DE INTELIGENCIA ARTIFICIAL

Una primera causa de confusión cuando hablamos de "inteligencia artificial" es la falta de clarificación del término. En el lenguaje corriente y en los debates públicos que se difunden en los medios de comunicación, cuando se habla de inteligencia artificial se está dando por sentada de manera acrítica la *hipótesis de la IA fuerte* y el llamado "*gran sueño* de la IA" (Wooldridge 2020, 2-3), es decir, la posibilidad de construir máquinas que tengan las mismas capacidades de acción inteligente que poseemos los seres humanos; entre otras cosas, que tengan conciencia, que sean autoconscientes y que operen de manera autónoma como lo hacemos nosotros. Es normal que esta versión sea la más extendida en la sociedad en general, pues es la que predomina en las novelas, películas, videojuegos y series de televisión, pero hay que ser conscientes de que no deja de ser eso: una figuración que se sostiene en el ámbito de la ciencia-ficción.

Aunque desde un punto de vista filosófico la *hipótesis de la IA fuerte* resulte más sugestiva y estimulante, por los profundos desafíos que plantea, hasta ahora buena parte de sus contenidos están confinados en la pura especulación

Lo cierto es que la inmensa mayoría de los investigadores de la IA no se dedican a trabajar para lograr ese "gran sueño", sino que se ocupan en otras tareas –quizá menos excitantes o glamurosas, pero seguramente más útiles para todos– y se centran en desarrollar en las máquinas determinadas tareas muy específicas que hasta ahora han requerido cerebros humanos (y también en ocasiones cuerpos humanos) y que las técnicas de computación convencionales no han logrado desarrollar; por ejemplo, las herramientas de traducción automática, los instrumentos de realidad aumentada, los coches autónomos, distintas aplicaciones en el ámbito de la salud (sobre todo para el diagnóstico de enfermedades mediante el reconocimiento de imágenes), etc. Estas son las principales áreas de investigación de la IA en las que se trabaja actualmente.

Aunque desde un punto de vista filosófico la *hipótesis de la IA fuerte* resulte seguramente más sugestiva y estimulante, por los profundos

desafíos que plantea, hasta ahora buena parte de sus contenidos están confinados en el terreno de la pura especulación. A día de hoy, no sólo no se ha conseguido desarrollar en las máquinas esas capacidades "fuertes" sino que no se sabe si será posible alcanzarlas en algún momento, pese a que los publicistas de la IA estén poniendo constantemente *deadlines* o fechas de cumplimiento que después siempre tienen que ir aplazando o retrasando ante la evidencia de los hechos (no hay que minusvalorar la fuerza recaudatoria de esos anuncios, pues estas investigaciones requieren de enormes cantidades de financiación). Además, no hay consenso en que ese deba ser el fin último de la IA, ni tampoco está claro, en caso de que fuera posible su cumplimiento, que se trate de un objetivo bueno o deseable para la sociedad. No en vano, se suele hacer referencia a posibles escenarios distópicos e incluso apocalípticos que podría plantear la IA fuerte, como la desaparición masiva de puestos laborales a manos de la robótica, el desarrollo de superinteligencias que pudieran tomar el control sobre los seres humanos y que incluso pudieran provocar nuestra desaparición, etc.

Por otra parte, si analizamos el concepto de Inteligencia Artificial en sí mismo, nos damos cuenta de que es un concepto analógico que se establece por comparación con la Inteligencia Natural, en concreto con la Inteligencia Humana, que es la única que conocemos. Por debajo de esta dicotomía se encuentra la distinción clásica entre lo natural y lo artificial, entre lo que es por naturaleza y lo que el hombre fabrica o produce (los artefactos)[2]. Lo que queremos subrayar aquí, en primer lugar, es que si el fundamento de la analogía simplemente se da por supuesto y no se explicita, como suele suceder, entonces no logramos entender la comparación y caemos en la confusión. La analogía es una relación entre relaciones e implica proporcionalidad, orden, armonía, semejanza, límite, equilibrio... Consiste en la semejanza a pesar de las diferencias: como decían los clásicos, es "simplemente desemejante" (*simpliciter diversa*) y "según algún aspecto semejante" (*secundum quid eadem*).

Si analizamos el concepto de Inteligencia Artificial, nos damos cuenta de que es un concepto analógico que se establece por comparación con la Inteligencia Natural, en concreto con la Inteligencia Humana, única que conocemos

Si nos atenemos a la clasificación escolástica tradicional de los nombres comunes (recordemos que el nombre común es aquel que se predica de muchos seres distintos), podemos distinguir tres tipos fundamentales de términos:

1) *Equívoco*: es aquel que tiene significados totalmente distintos.

2) *Unívoco*: es aquel que tiene un sentido perfectamente idéntico, sin orden de prioridad y posterioridad.

3) *Análogo*: es aquel que tiene en parte un sentido idéntico y en parte un sentido diverso o desigual, con orden de prioridad y de posterioridad.

En cuanto al concepto de IA, aunque se utilice muchas veces de manera equívoca (sin ni siquiera plantearse el significado ni ser consciente de las dificultades que conlleva), parece evidente en principio, como hemos dicho, que se trata de un concepto análogo. El problema viene cuando se quiere explicar o determinar qué tipo de analogía es la que se está estableciendo en cada una de las concepciones de la IA: ahí es donde reside la clave de comprensión del concepto y la dificultad más importante.

Una tarea pendiente que convendría hacer sobre cada una de las distintas concepciones existentes de la IA y que la filosofía puede ayudar a clarificar es determinar en cada caso cuál es el tipo de analogía establecida, si realmente hay una analogía y no se ha caído en equivocidad. En la mayoría de los casos habría que encontrar el término medio o los términos medios que sirven de mediación entre ambas concepciones de la inteligencia; la clave sería captarlos y analizarlos. También habría que distinguir si es una analogía, a) según el conocimiento (*secundum intentionem*) y no según el ser (*secundum esse*), b) según el ser y no según el conocimiento, o c) si es según el ser y según el conocimiento. Todas estas indicaciones previas servirían para clarificar enormemente los debates sobre la IA.

Como exponen Stuart J. Russell y Peter Norvig en *Inteligencia Artificial: Un Enfoque Moderno*, existen cuatro modelos principales o paradigmas de comprensión de la IA: el cognitivo, el lógico, el conductista y el del agente racional

Veamos a continuación cuáles son los modelos más importantes de comprensión de la IA.

CUATRO MODELOS DE COMPRENSIÓN DE LA IA

Como exponen Stuart J. Russell y Peter Norvig en *Inteligencia Artificial: Un Enfoque Moderno*, existen cuatro modelos principales o paradigmas de comprensión de la IA: el cognitivo, el lógico, el conductista y el del agente racional. Veamos muy brevemente en qué consiste cada uno de ellos y qué problemas filosóficos plantean:

1) Modelo cognitivo: sistemas que piensen como los seres humanos

Este modelo se ha formulado de varias maneras: unos autores plantean automatizar actividades como la toma de decisiones, la resolución de problemas o el aprendizaje (actividades que todos asociamos al pensamiento humano); otros hablan directamente de hacer que los ordenadores piensen, construir máquinas que tengan mentes. Para todo esto es necesario entender primero el funcionamiento de la mente humana, ya sea por introspección o por experimentos psicológicos. Una vez que se tenga una teoría suficientemente precisa de cómo trabaja la mente humana, entonces se podrá trasladar a un programa de ordenador, con datos de entrada y salida del programa (*inputs/outputs*) y tiempos de reacción similares a los de un ser humano (cfr. Russell y Norvig 2011, 4)[3].

La *hipótesis de la IA fuerte* defiende que las máquinas sí piensan realmente (se oponen al pensamiento simulado), mientras que la *hipótesis de la IA débil* defiende que es posible que las máquinas actúen con inteligencia o "como si" fueran inteligentes

El problema de este enfoque es que por un lado el funcionamiento de la mente humana y del cerebro sigue siendo un misterio (a día de hoy no hay una teoría unitaria ni suficientemente definida al respecto) y, por otro lado, ya el hecho de limitar el pensamiento a resolución de problemas es un enfoque tremendamente reduccionista (además, se fijan sólo en determinado tipo de problemas; otros ni siquiera los consideran).

2) Modelo lógico o de las "leyes de pensamiento": sistemas que piensen racionalmente

Corresponde a la tradición logicista de la IA, que trata de construir sistemas inteligentes a partir de programas informáticos que resuelvan cualquier problema que pueda ser descrito en notación lógica (el problema tiene que ser resoluble en términos lógicos). Ponen el énfasis en hacer inferencias correctas y se centran en el estudio de los cálculos que hacen posible percibir, razonar y actuar, o en el estudio de las facultades mentales mediante el uso de modelos computacionales (cfr. Russell y Norvig 2011, 5).

Un problema evidente de este enfoque es que hay formas de actuar racionalmente que no implican necesariamente realizar inferencias lógicas (cuando retiramos la mano del fuego para no quemarnos es un acto reflejo, y resulta mucho más eficiente que ponerse a pensar si la quitamos o no). Además, es muy difícil expresar en notación lógica cierto tipo de conocimiento, como los conocimientos informales. Ahí está todo el campo de la lógica difusa (*fuzzy logic*), que tanto desarrollo está teniendo estos últimos años. En cualquier caso, una cosa es resolver un problema en la teoría y otra muy distinta resolverlo en la práctica. Y

un sistema de IA tendrá que saber manejarse en la realidad, en el mundo operativo.

3) Modelo conductista o de simulación: sistemas que actúen como seres humanos

Se trata de desarrollar máquinas capaces de realizar funciones que normalmente hacen las personas aplicando su inteligencia. Otros lo formulan diciendo que se trata de lograr que los ordenadores hagan tareas que parecen específicamente humanas, o que por ahora hacen mejor los seres humanos.

El problema es que, en cualquier caso, lo máximo que podríamos decir es que una máquina llega a simular determinados comportamientos humanos. En general, la *hipótesis de la IA fuerte* defiende que las máquinas sí piensan realmente (se oponen al pensamiento simulado), mientras que la *hipótesis de la IA débil* defiende que es posible que las máquinas actúen con inteligencia o "como si"[4] fueran inteligentes.

4) Modelo del "agente racional": sistemas que actúen racionalmente

Consiste en el "estudio de los agentes que reciben percepciones del entorno y llevan a cabo las acciones, según determinadas funciones" (Russell y Norvig 2011, 33), poniendo por tanto el énfasis en la agencia inteligente –o inteligencia agente– de un sistema que es capaz de decidir qué hacer y a continuación emprender la acción.

Este enfoque del "agente racional" es el más amplio y omnicomprensivo de los existentes, pues reúne todas las habilidades necesarias recogidas en el test de Turing[5] y considera que un agente es racional cuando actúa con la intención de alcanzar el mejor resultado (o, en casos de incertidumbre, cuando logra el mejor resultado esperado). Para ser considerados como tales, a diferencia de los *software* convencionales, los agentes racionales tienen que estar dotados de controles autónomos, percibir su entorno, persistir durante un período de tiempo prolongado, adaptarse a los cambios y ser capaces de alcanzar objetivos diferentes.

El enfoque del "agente racional" es el más amplio y omnicomprensivo de los existentes, pues reúne las habilidades recogidas en el test de Turing y considera que un agente es racional cuando actúa con la intención de alcanzar el mejor resultado

En el siguiente apartado vamos a recoger algunos de los problemas filosóficos fundamentales que plantea este modelo del agente racional, que es el más extendido.

PROBLEMAS FILOSÓFICOS DE LA IA

Una primera cuestión filosófica que se está jugando aquí es qué se entiende por *racionalidad*.

De entrada, se parte de un concepto de la racionalidad como "la capacidad de elegir la mejor acción posible para alcanzar un objetivo concreto, dados unos determinados criterios que es necesario optimizar y además teniendo en cuenta los recursos disponibles" (Russell y Norvig 2011, 41). Y uno de los elementos centrales es la "percepción" del entorno en el que se encuentra inmerso el sistema a través de sensores que recopilan e interpretan datos, procesan la información derivada de esos datos, seleccionan la acción más óptima que se puede realizar y actúan en consecuencia mediante los oportunos accionadores, logrando así modificar el entorno.

Por tanto, las capacidades fundamentales de este tipo de sistemas son la percepción, el "razonamiento" (dicho con más propiedad, el tratamiento de información)[6] y la acción. Entendemos que tanto "racionalidad" como "percepción" hay que ponerlos aquí entre comillas, porque al menos a día de hoy no son comparables con la racionalidad o la percepción humanas, que son el punto de referencia. En todas estas cuestiones se reproduce la eterna confusión entre información y conocimiento, omnipresente en los contextos tecnológicos.

Lo mismo sucede con el *aprendizaje*. Se habla de *aprendizaje supervisado* (que introduce suficientes ejemplos de comportamiento de entrada y salida) y de *aprendizaje por refuerzo* (que cada vez que adopta una decisión proporciona una señal de recompensa que informa al sistema si la decisión fue acertada o no). Algunos métodos de aprendizaje automático adoptan algoritmos basados en el concepto de *redes neuronales*, que tratan de replicar el funcionamiento del cerebro humano; en concreto, en el denominado *aprendizaje profundo* la red neuronal cuenta con varias capas entre la entrada y la salida que le permiten aprender la relación general entre ellas en pasos sucesivos. Este último enfoque resulta más preciso y requiere un menor grado de orientación humana. Por último, también está *la robótica (o inteligencia artificial integrada)*, que se ocupa de las máquinas físicas que tienen que enfrentarse a la dinámica, la incertidumbre y la complejidad del mundo físico. Además de la IA, en el diseño y funcionamiento de robots desempeñan algún tipo de función la ingeniería mecánica o la teoría del control[7].

En definitiva, cuando uno se pone a examinar las ideas que se expresan desde el modelo del "agente inteligente" se encuentra con problemas filosóficos muy graves que no han sido analizados suficientemente, como la "racionalidad", el "aprendizaje", la "resolución de problemas" o la "toma de decisiones" (esta última nos parece la cuestión más grave y que más implicaciones tiene a nivel ético y social). En el fondo el problema es que los enfoques de la IA sobre esos conceptos suelen ser muy reduccionistas: son enfoques que toman la parte por el todo, o que se fijan sólo en un aspecto y obvian otras dimensiones que también están ahí y que hay que tener en cuenta. Además, hay una especie de "adanismo conceptual", sobre todo en los ámbitos más técnicos de la IA, que

Algunos métodos de aprendizaje automático adoptan algoritmos basados en el concepto de *redes neuronales*, que tratan de replicar el funcionamiento del cerebro humano; en el *aprendizaje profundo* la red neuronal cuenta con varias capas entre la entrada y la salida que le permiten aprender la relación general entre ellas en pasos sucesivos

es fruto básicamente de la ignorancia: se plantean estas cuestiones como si fuesen nuevas, empezando de cero, y las tratan de resolver en unos términos a veces bastante superficiales, saltándose siglos y siglos de debate filosófico sobre este tipo de conceptos.

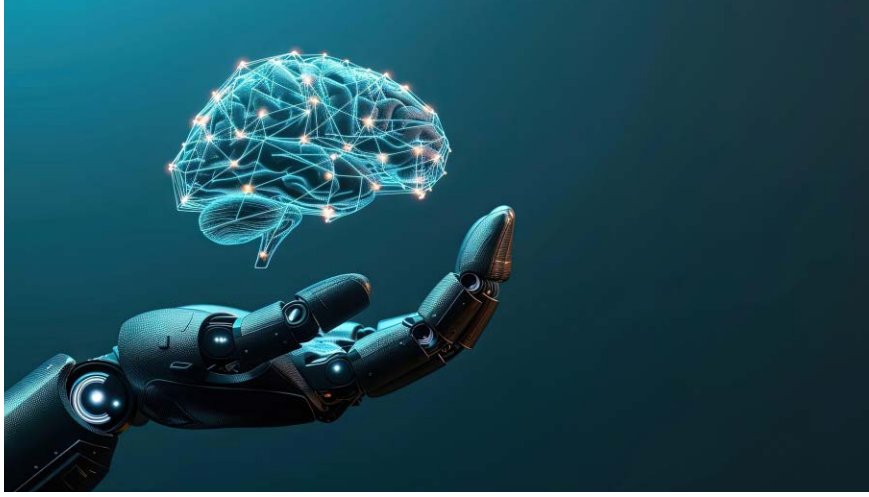
Además, la consecuencia de la tesis del procesamiento de información o de la metáfora computacional, que caracteriza a la mente como un procesador de información similar a un ordenador, es "una relación de circularidad improductiva, carente de tensión, entre la psicología cognitiva y la IA simbólica: la primera supone que la mente es como una computadora, y la segunda pretende utilizar computadoras para duplicar el funcionamiento de la mente" (Carabantes 2016, 14).

Aparte de los temas que acabamos de mencionar, nos encontramos con otros problemas filosóficos gravísimos como la intencionalidad, la conciencia, la introspección, la autoconciencia, el lenguaje, la memoria, la percepción, etcétera[8]. Asimismo, aparece la cuestión de los límites y el alcance del conocimiento humano, los procesos psicológicos de la cognición: ¿de dónde viene el conocimiento?, ¿cómo se genera la inteligencia o la conciencia a partir de un cerebro físico?[9], ¿qué reglas formales son las adecuadas para obtener conclusiones válidas? (es decir, todo el campo de la lógica y la matemática). Otro campo que también estaría involucrado aquí es el de la teoría de la acción: las razones para actuar, la motivación, el deseo, la influencia de las pasiones, etcétera[10]. Por no mencionar problemas filosóficos tan importantes como la libertad y la voluntad, o la relación mente-cuerpo, que es uno de los asuntos centrales de la filosofía moderna[11].

Cuando se examinan las ideas expresadas desde el modelo del "agente inteligente" se encuentran problemas filosóficos muy graves no analizados suficientemente, como la "racionalidad", el "aprendizaje", la "resolución de problemas" o la "toma de decisiones"

¿Se puede pretender saltar por encima de todos estos debates filosóficos sin tenerlos en cuenta? No parece adecuado, pero es lo que ocurre con la mayoría de los expertos en IA. Frente a esas concepciones estrechas de la "racionalidad" y el "aprendizaje", se puede negar la capacidad de las máquinas y ordenadores para actuar inteligentemente, y se puede hacer utilizando distintos argumentos:

- El argumento de incapacidad, que parte de la evidencia de que hay cosas que las máquinas nunca podrán hacer. Aquí está envuelta también toda la cuestión emocional o afectiva.
- La objeción matemática por el teorema de la incompletitud de Gödel: las máquinas son sistemas formales que están limitados por este teorema, mientras que las personas no.
- El argumento de la informalidad del comportamiento y de las decisiones: no todo es reducible a notación lógica o a algoritmo.
- El problema de la cualificación, que para nosotros es clave: el comportamiento humano es demasiado complejo para poder captarlo mediante un conjunto de reglas, y no todo es cuantificable. Esto es fundamental para entender un fenómeno tan pujante en los últimos años como es el Big Data.



En cualquier caso, al margen de en qué términos se plantee la cuestión, no hay que olvidar que en las máquinas siempre faltará la autoconciencia: la conciencia de sus propias acciones y estados mentales. También se pueden plantear aquí las cuestiones de la emocionalidad, la intencionalidad, las "inteligencias múltiples" (Gardner), la "inteligencia sentiente" (Zubiri), etc.

CONCLUSIÓN: LA URGENCIA DE UNA REFORMA UNIVERSITARIA

En cuanto a la indefinición y limitaciones que hemos visto del concepto de IA, quizá la cuestión de fondo que se está jugando ahí es que seguramente no tiene sentido la pregunta tal y como la formuló Alan Turing en el título de su famoso artículo: "¿Puede pensar una máquina?", pues esa pregunta implica utilizar en sentido analógico –y sin explicar– un término ("pensar") que están tratando de aplicar en sentido unívoco, identificándolo con el pensamiento humano, cuando en realidad lo utilizan de manera equívoca, y es ahí –insistimos– donde se producen las confusiones y los malentendidos. Hay quien considera que debería transformarse ese interrogante en este otro: ¿pueden las máquinas hacer como que piensan e igualar o mejorar los resultados del pensamiento humano? Por su parte, John Searle dijo: "¿Podría pensar una máquina? La respuesta es obviamente, sí. Nosotros somos precisamente esas máquinas".

En cuanto a las consecuencias éticas (en las que no hemos entrado en este artículo), es evidente que la IA va a afectarnos en el futuro de formas que aún desconocemos, por lo que la propia sociedad, los expertos, los investigadores y los dirigentes gubernamentales deben estar en alerta constante. Además de las posibles derivaciones totalitarias de las técnicas biotecnológicas o biopolíticas por su uso desde instancias de poder antiliberales, ha de señalarse el posible menoscabo de las libertades individuales, en términos de control, vigilancia, invasión de la privacidad, etc. Y la cuestión que hemos comentado de la responsabilidad y la toma de decisiones es desde nuestro punto de vista lo más importante, pues afecta a lo más profundo que somos como seres humanos.

La disponibilidad de datos digitales masivos, los avances biotecnológicos, la capacidad de almacenamiento y las innovaciones de la ciencia y la ingeniería en relación con los métodos y herramientas de la IA, dibujan un futuro muy interesante con un fuerte componente de incertidumbre

La disponibilidad de datos digitales masivos, los avances biotecnológicos, la capacidad de almacenamiento informático y las innovaciones en el campo de la ciencia y la ingeniería en relación con los métodos y herramientas de la IA, dibujan un futuro muy interesante pero que también incluye un fuerte componente de incertidumbre. Un futuro que, en cualquier caso, debe ser pensado desde la filosofía.

Por eso consideramos que es necesaria una reforma universitaria que no distinga tan marcadamente entre letras y ciencias, entre humanidades y tecnología, superando la di-

visión de las "dos culturas" (C. P. Snow). Por ejemplo, a nuestro parecer sería imperioso que las universidades españolas incluyeran entre su oferta académica el Doble Grado en Filosofía e Inteligencia Artificial, donde los estudiantes puedan formarse en la historia de la filosofía, la teoría del conocimiento, la lógica, la ética, la estética, la metafísica, la psicología, la lingüística, las matemáticas, la informática, la estadística, la tecnología, la neurociencia, etc. (además de inglés y otros idiomas, incluida alguna lengua muerta).

De hecho, creemos que de esta "nueva" universidad podrían salir al mercado laboral y profesional personas muy potentes –una nueva generación formada[12], inteligente y comprometida con su país– que puedan cambiar la situación de España, francamente mejorable en tantos aspectos. ■

BIBLIOGRAFÍA

Baltar, E. (2020): "El posthumanismo en la UCI de la realidad": https://telos.fundaciontelefonica.com/telos-114-analisis-baltar-el-posthumanismo-en-la-uci-de-la-realidad/

Carabantes López, M. (2016): *Inteligencia artificial: Una perspectiva filosófica*. Madrid: Escolar y Mayo.

Gardner, H. (2011): *Inteligencias múltiples. La teoría en la práctica*. Barcelona: Paidós.

Russell, S. y **Norvig, P.** (2011): *Inteligencia Artificial: Un Enfoque Moderno*. New Jersey: Prentice Hall (3ª ed.).

Searle, J. (2000): *Razones para actuar. Una teoría del libre albedrío*. Oviedo: Ediciones Nobel.

Turing, A. (1985): "¿Puede pensar una máquina?" En Turing, A., Putnam, H. y Davidson, D. *Mentes y máquinas*. Madrid: Tecnos.

Wilhelmsen, Frederick D. (2023): *La estructura paradójica de la existencia*. Madrid: Dykinson. Traducción, prólogo y notas de Ernesto Baltar.

Wooldridge, M. (2020): *The Road to Conscious Machines. The Story of AI*. Dublin: Penguin Books.

Zubiri, X. (1998): *Inteligencia sentiente, I. Inteligencia y realidad*. Madrid: Alianza Editorial.

NOTAS

1 Véase **Ernesto Baltar,** "El posthumanismo en la UCI de la realidad", *Telos*, Fundación Telefónica, 16 de octubre de 2020: https://telos.fundaciontelefonica.com/telos-114-analisis-baltar-el-poshumanismo-en-la-uci-de-la-realidad/

2 No deja de ser curioso en este sentido que algunos de los que están llevando más lejos los planteamientos de la IA fuerte propugnen precisamente la superación de esa distinción mediante la fusión de lo natural y lo artificial, de lo humano y lo tecnológico, por ejemplo en los llamados cíborgs o en los seres biónicos, seres que son medio artificiales medio naturales (por ejemplo, pueden ser personas que lleven dentro de su cuerpo determinados implantes o chips integrados). Eso, que hasta hace poco podía sonar a ciencia-ficción, lo están defendiendo actualmente distintas corrientes adscritas al posthumanismo o transhumanismo, supuestamente con el objetivo de mejorar las capacidades físicas y cognitivas del ser humano, hasta llegar a una superación de lo biológico.

3 Ya en 1962 Allen Newell y Herbert Simon desarrollaron el "Sistema de Resolución General de Problemas" (SRGA), donde compararon las etapas del proceso de razonamiento con las de los seres humanos ante los mismos problemas.

PALABRAS CLAVE

Inteligencia artificial ● Filosofía ● Máquinas ● Racionalidad ● Aprendizaje ● Conciencia ● Agente racional ● IA fuerte ● Ciencia ● Técnica ● Robótica

[4] Ese "como si" no es un elemento trivial, sino que nos parece la clave de la cuestión.

[5] En el marco del test de Turing, el ordenador debía tener cuatro capacidades fundamentalmente para poder engañar al evaluador: procesamiento del lenguaje natural (entender y hablar el inglés); representación del conocimiento (para almacenar lo que se conoce o siente); razonamiento automático (para utilizar la información acumulada para responder preguntas y extraer nuevas conclusiones), y aprendizaje automático (para adaptarse a las nuevas circunstancias y detectar y extrapolar patrones) (cfr. Turing 1985). Posteriormente se fue perfeccionando el test y la llamada "prueba global de Turing" incluye una señal de vídeo para que el evaluador pueda valorar la capacidad de percepción del ordenador y además pueda pasar objetos a través de una ventanita. Aquí ya el ordenador tendría que estar dotado de visión computacional (para percibir objetos) y robótica (para manipular y mover objetos). Por tanto, el desarrollo de esas seis capacidades (4+2) forman parte de la IA.

[6] Cuando en el contexto de la IA se habla de razonamiento y adopción de decisiones, se incluye la planificación, la programación, la búsqueda de soluciones y la optimización de los resultados. El primer paso es convertir los datos en conocimiento, creando un modelo; a continuación se extraen inferencias a través de reglas simbólicas, se busca la optimización entre todas las soluciones posibles a un problema y finalmente se toma una decisión sobre la acción que se debe llevar a cabo. Este grupo de técnicas incluye el aprendizaje automático, las redes neuronales, el aprendizaje profundo o los árboles de decisión, entre otros instrumentos que permiten aprender a resolver problemas que no se pueden describir mediante reglas de razonamiento simbólicas.

[7] La categoría de "robots" incluye, por ejemplo, manipuladores robóticos, vehículos autónomos, drones, robots humanoides, robots de limpieza, etc.

[8] Por ejemplo, la intencionalidad es un concepto que existe desde la filosofía medieval, que retoma Franz Brentano a finales del XIX, y que en el siglo XX van a estudiar en el ámbito de la teoría de la mente autores como Daniel Dennett.

[9] Esto lo estudia la neurociencia, pero también lo ha estudiado toda la teoría moderna del conocimiento. Para que nos podamos plantear, como hizo Alan Turing, si puede pensar una máquina, lo primero que tenemos que saber es a qué estamos llamando pensar. Además, aunque una máquina lograse pasar el test de Turing, eso no quiere decir que realmente esté pensando sino que simplemente estaría simulando esa acción de pensar.

[10] Sobre esta cuestión recomendamos el libro de John Searle, *Razones para actuar*. Ya desde las categorías aristotélicas de acción y pasión está planteada esa cuestión, y se ha estudiado en toda la historia de la filosofía. Se ha desarrollado asimismo en el campo de la economía la teoría de la decisión o la teoría de juegos.

[11] Toda la filosofía moderna no dejó de dar vueltas a esta cuestión: Descartes no lo resuelve, Malebranche y el ocasionalismo le dieron una solución teológica con el "recurso a Dios", Spinoza lo disuelve al unificar las sustancias en una sola, Leibniz aplica el principio de la armonía preestablecida, etc. En el ámbito del empirismo tenemos a Hume con su crítica a la idea del yo y del alma como sustancia, y de esa crítica de Hume saldrá en el siglo XX toda la filosofía analítica: la famosa crítica de Gilbert Ryle al dualismo cartesiano (el "dogma del fantasma en la máquina"), el fisicalismo del empirismo lógico (el Círculo de Viena), Quine, Wittgenstein, Dennet, Richard Rorty, etc.

[12] De hecho, como ya anunciaba Frederick D. Wilhelmsen hacia 1970 partiendo de las ideas de Marshall McLuhan sobre la revolución tecnológica de la informática, "en el futuro las élites educadas tendrán que imitar el acto mismo de existir: deben sintetizar en una unidad una increíble cantidad de información que abrumará a la humanidad y la sumirá en la angustia, a menos que el hombre emerja verdaderamente en una Era de la Metafísica, una era capaz de plantear la pregunta decisiva que no puede ser asumida ni contestada por la tecnología informática: ¿por qué?" (2013, 10).