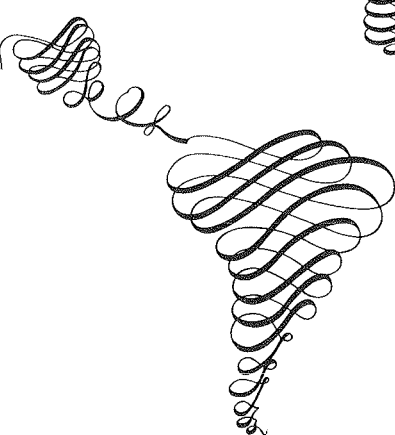




La polémica de la inteligencia artificial

Los últimos avances de las ciencias física y matemática ha suscitado en algunos científicos la cuestión de una posible máquina pensante. En esta controversia reciente sobre la viabilidad de una inteligencia artificial marcó un hito Alan Turing en 1950 con su afirmación rotunda, siendo rebatido un decenio después por J. R. Lucas con su negativa absoluta. Posteriormente, también los filósofos se han sumado a la polémica. En la actualidad, los defensores y detractores del ordenador pensante se apoyan respectivamente en el carácter algorítmico o no del pensamiento.

Antonio F. Rañada *

¿Llegarán a pensar las máquinas?

EN el estudio de la materia se usa a menudo un enfoque reduccionista que supone que todas las propiedades

* Departamento de Física Teórica. Universidad Complutense de Madrid.

de un sistema complejo pueden ser explicadas basándose únicamente en las de sus componentes simples. La primera manifestación de este tipo se debe a los griegos Leucipo y Demócrito quienes afirmaron en el siglo V a. de J. C. que no hay más que átomos y espacio vacío y que las propiedades de esos átomos y su disposición mutua permiten explicar las de todos los cuerpos —así por ejemplo la mayor pesadez del plomo sería debida a que sus átomos están más apretados. Nótese que, aunque el énfasis atomista se pone en el elemento constitutivo, los átomos, aparece ya la estructura al hablar de la disposición mutua.

Desde entonces, la aplicación del reduccionismo a las relaciones entre la biología por un lado y la física y la química por el otro se ha visto envuelta en multitud de malentendidos y confusiones, cargados muchas veces de confrontación ideológica, que incluso han contribuido a retrasar la aceptación general de la idea de átomo hasta ya entrado el siglo XX (1). Ocurrió así en buena medida por la mezcla de distintos aspectos del reduccionismo que pueden calificarse como *ontológicos* (¿están los seres vivos compuestos de los mismos elementos que la materia inanimada?), *metodológicos* (¿deben estudiarse los seres vivos investigando los procesos físico-químicos subyacentes o, más bien, las estructuras complejas como la célula, el organismo o el grupo social?) y *epistemológicos*, es decir explicativos (¿se puede deducir las leyes de la biología de las de la física y la química?).

En el terreno ontológico cabe decir que, aunque hace ya tiempo que se llegó a comprender que los procesos de la vida son explicables sin recurrir a sustancias inmateriales, eso no significa que un ser vivo no sea más que átomos y moléculas. Suponer, porque un sistema esté compuesto de ciertos elementos, que no sea más que esos elementos es una falacia conocida como «del nada más que». Un automóvil es mucho más que una cierta cantidad de acero y plástico y no se puede decir que un ordenador sea sólo un conjunto de semiconductores, cables e interruptores. Muy al contrario, la organización de todos sus componentes en una estructura con subsistemas de complejidad creciente y las interacciones entre sus niveles y partes son precisamente su característica más propia. O, dicho de otro modo: en un sistema complejo no es cierto que el todo sea igual a la suma de sus partes.

(1) Ver por ejemplo *Studies in the Philosophy of Biology*, editado por F. J. Ayala y Th. Dobzhansky, MacMillan, Londres, 197.

El reduccionismo epistemológico lleva a la idea de *propiedades emergentes* que son las que tiene un sistema pero no se pueden predicar de sus partes o componentes. Aunque los átomos están formados por electrones y nucleones que están sometidos a las leyes de la física, sus propiedades químicas son emergentes porque no se pueden predicar de esos elementos básicos. Así ocurre con el enlace químico (por ejemplo, dos átomos de hidrógeno se unen a uno de oxígeno para formar una molécula de agua) pues, aunque los electrones intervienen en ese enlace, no se pueden enlazar ellos mismos. De igual manera, las propiedades de los seres vivos emergen de la estructura compleja en que se organizan sus constituyentes, que no gozan de ellas, como ocurre con un hígado que tiene muchas propiedades que no lo son ni sus células una a una ni sus componentes químicos. Desde esta perspectiva, tanto metodológica como epistemológica, hay una jerarquía de ciencias *Física* → *Química* → *Biología* → *Neurología* → *Sociología*, ordenadas en grado de complejidad creciente y de fundamentalidad decreciente.

Si los animales no son sino máquinas, idea propuesta la primera vez por Descartes, aunque sean muy complejas y con muchos niveles de organización, es posible en principio construir mecanismos artificiales que realicen las mismas funciones o una parte de ellas al menos. Las máquinas que se desarrollaron desde el principio de la Revolución Industrial (telares, bombas de agua, locomotoras) hacían precisamente eso. Su éxito y el desarrollo de la química llevaron a suponer que incluso el sustrato material de que están compuestos los seres vivos podría llegar a ser fabricado en el laboratorio, especialmente tras la síntesis de la urea por el químico alemán Wöhler en 1828. Actualmente es posible fabricar muchos de los componentes químicos elementales de la materia viva, aunque no ha sido posible, a pesar de los muchos esfuerzos realizados, ir más allá del nivel molecular y producir protoplasma artificialmente, menos aún una célula completa o alguno de sus órganos.

Es una idea aceptada hoy que las propiedades de la materia viva se pueden entender a partir de la física y la química. Por ejemplo, cuando una persona se apoya en su pierna, se producen en su fémur tensiones y deformaciones similares a las de una viga de un puente por el que pasa un camión, el movimiento de la sangre sigue las leyes de la hidrodinámica como lo hace el agua en las cañerías, y la naturaleza química de muchos procesos vitales se conoce muy bien. Esto no tiene solamente interés teórico, sino que permite aplicaciones terapéuticas, por ejemplo, al susti-

tuir un trozo defectuoso de ese hueso por una prótesis de titanio o porcelana, cuando se trata una malformación cardíaca o al curar a un diabético. Aunque hay aún muchos problemas por resolver, nadie duda que la aplicación de las ciencias físico-químicas al entendimiento de la materia viva seguirá avanzando a un paso acelerado y en ello se basa la extraordinaria explosión de la medicina de hoy.

Y naturalmente, puesto que el cerebro es materia viva, en último término un montón de neuronas, surge la idea de que un dispositivo artificial, el ordenador, pueda llegar a realizar sus mismas funciones. Nace así un campo nuevo de investigación, la Inteligencia Artificial (IA), que atrae a matemáticos, físicos, informáticos y filósofos que pretenden e intentan demostrar que un cerebro no es más que un ordenador complejo, construyendo para ello una máquina que sea capaz de razonar tanto e incluso más que una persona inteligente. Esto supone que el pensamiento se puede separar de su base material y reproducir fuera de ella, porque no es más que una combinación de indiferencias lógicas simples, que son inmateriales, lo mismo que un trozo de hierro se puede sintetizar a partir de átomos.

Surge así la pregunta: ¿es posible que un ordenador tenga mente? ¿Llegarán a pensar las máquinas?

Dos opiniones

LA nueva ciencia tuvo en seguida sus profetas y sus detractores que mantienen una polémica dura y apasionada, en la que no faltan las descalificaciones más graves. El desarrollo de la tecnología de los ordenadores (2) desde la Segunda Guerra Mundial ha sido espectacular. Esas máquinas son capaces hoy de realizar cálculos de asombrosa complejidad que exigen muchos millones de operaciones elementales por segundo, algo claramente imposible para ningún ser humano. Pueden almacenar ingentes cantidades de datos y ordenarlos, manejarlos o clasificarlos de muchas maneras con una rapidez y una precisión impensable tan sólo hace pocas décadas. Ya juegan al ajedrez mejor que la mayoría de las personas y una ha ganado una vez a un gran maestro internacional. Sin ellas sería imposible haber llegado a la Luna, diseñar un

(2) En este artículo se usa la palabra ordenador con preferencia sobre su sinónima computador o computadora, de uso también muy general.

avión o un gran puente, predecir el tiempo o llevar la contabilidad de una gran empresa. No cabe duda de que son capaces de algún tipo de actividad de las que se suelen asociar con la inteligencia, pero ¿piensan? ¿sienten? ¿son conscientes? o, si no lo hacen aún hoy ¿llegarán a hacerlo?

Para la cultura de masas de hoy, la idea de ordenador inteligente no es extraña. Por el contrario, el cine los ha hecho familiares, aliados a menudo antes que siervos, enemigos a veces que se rebelan contra sus creadores como ya lo hiciera el monstruo de Frankenstein. Se ha trivializado así una pregunta que no se puede abordar directamente sin considerar antes muchas otras como ¿qué es pensar? ¿qué relación tienen el sentido común y la intuición con el razonamiento lógico? ¿se puede pensar sin ser consciente?... Por ello, incluso si la IA fuese imposible porque ninguna máquina pueda nunca llegar a pensar, vale la pena reflexionar sobre ella porque ayuda a comprender cómo funciona la mente.

La cuestión empezó a tomar cuerpo en 1950 con un famoso artículo del matemático inglés Alan Turing (3) en el que se declara convencido de que las máquinas llegarán a tener un comportamiento que será considerado por todos como inteligente y rechaza varias objeciones en contra. Afirma, en conclusión, que «podrán competir con el hombre en todos los campos puramente intelectuales». En el mismo artículo propuso también el famoso *test de Turing*, discutido en la siguiente sección, para investigar el nivel de inteligencia de un ordenador.

De Turing es también una de las ideas más importantes para esta cuestión, la de *máquina universal*. Para entender este concepto cabe recordar que en un ordenador hay una base material, el hardware, y una base lógica (4), el software. El hardware es la materia y las piezas de que está construido; el software consiste en una colección de programas que son los conjuntos ordenados de instrucciones necesarias para la realización de las tareas. En ello se basa la enorme versatilidad de esas máquinas porque, cambiando los programas, se puede resolver una enorme variedad de problemas distintos. Turing mostró que todos los ordenadores son (idealmente) equivalentes en el siguiente sentido: cada ordenador puede comportarse exactamente igual y hacer lo mismo que cualquier otro, es decir, puede imitarlo perfectamente, si se le introduce el programa adecua-

(3) «Computing machinery and intelligence», *Mind*, 59, n.º 236, 1950; reproducido en *Controversia sobre mentes y máquinas*, editado por A. R. Anderson, Tusquets Editores, Barcelona, 1984.

(4) Se emplea a veces la palabra *logicial*.

do. Por ello podemos desarrollar la discusión en términos de una sola máquina, calificada de universal, modelo ideal del que todos los ordenadores son realizaciones particulares. Nótese que se prescinde aquí de limitaciones de tipo práctico, suponiendo que todos ellos tienen suficiente memoria y capacidad, lo que está bien, pues uno que pretenda ser inteligente debe ser muy poderoso; esto no significa que un pequeño PC haga las mismas cosas, en su uso diario, que un enorme mainframe, sino que operan sobre los mismos principios. Pues bien, el problema de la IA puede ser planteado como la pregunta de si el cerebro humano es o no es una realización de la máquina universal de Turing. Es razonable admitir que en el cerebro hay también una dualidad hardware-software, pero el quid de la cuestión está en saber cómo funciona. Si lo hace como los ordenadores, es decir algorítmicamente, la respuesta sería positiva (ver sección 6). Si lo hace de otra manera, habría que contestar que no.

Pero unos años más tarde un artículo de J. R. Lucas (5) que defendía la opinión contraria a la teoría de Turing causó gran impacto. Su autor se basaba en el teorema de Gödel, que afirma que en todo sistema formal coherente hay proposiciones que son ciertas, pero no se pueden demostrar sin salirse del sistema (ver sección 5). Desde entonces, estas dos posturas se han ido depurando y afilando sus argumentos, de manera que existen hoy dos bandos:

1. Algunos creen que nuestro cerebro no es más que un ordenador hecho de carne y que la diferencia que hoy percibimos, en cuanto a capacidad de razonar, entre los seres humanos y los ordenadores es puramente cuantitativa. Hoy por hoy nuestro cerebro es más complejo que los ordenadores más potentes, pero es posible y se llegará a construir uno que sea tan inteligente como cualquier ser humano. Más aún, podrán calificarse como mucho más inteligentes porque su mayor rapidez hará que resuelvan los mismos problemas en menos tiempo. Esta afirmación, conocida como *hipótesis fuerte de la Inteligencia Artificial*, lleva a sus últimas consecuencias las ideas de Descartes de que los animales son máquinas y que todo el pensamiento es analítico.

2. Otros creen, por el contrario, que tal diferencia es cualitativa y que cualquier producto del arte humano, lo mismo un ordenador que un avión o un radar, es y será siempre incapaz de facultades mentales. Los

(5) J. R. Lucas, «Minds, machines and Gödel», *Philosophy*, 36, 120, 1961; reproducido en *Controversia sobre mentes y máquinas*, editado por A. R. Anderson, Tusquets editores, Barcelona, 1984.

ordenadores pueden realizar tareas que exijan mucha rapidez, reiteraciones y capacidad de manejo y almacenamiento de muchos datos, pero nunca llegarán a entender, porque el pensamiento es mucho más que cadenas de inferencias lógicas elementales. Cerebro y ordenadores operan siguiendo principios distintos.

Los defensores de la IA se basan en la idea, sostenida ya por Descartes, de que el pensamiento consiste solamente en razonamientos analíticos y puede, por tanto, reducirse siempre a cadenas lineales de operaciones lógicas simples, mediante las que cualquier problema tiene siempre solución. Los pensamientos que parecen muy complejos se diferencian de los simples únicamente porque en ellos intervienen muchas de esas operaciones o manejan una cantidad de información muy alta. La hipótesis fuerte de la IA fue interpretada como la base de una nueva y radical revolución científica y vivida con excitación en varios centros académicos de Estados Unidos, en MIT con Marvin Minsky, en Carnegie-Mellon con Allen Newell, Herbert Simon y Cliff Shaw, en Stanford con Edward Feigenbaum, llegando a asegurarse que en veinte años (es decir, hacia la mitad de los ochenta) se habría llegado a construir y programar un ordenador al que todos se verían obligados a considerar inteligente. Está claro que se trata de una predicción incumplida, pero los profetas de la IA lo atribuyen a que la dificultad del problema es mayor de lo esperado, como ocurre y ha ocurrido con otras muchas predicciones científicas (la energía barata por fusión de plasmas o la vacuna contra el SIDA, por ejemplo). Quizás haya que esperar otros veinte años o quizás cien, pero para la humanidad eso no es gran cosa al fin y al cabo.

En el libro *La quinta generación* (6), el profeta de la IA Edward Feigenbaum lo expresa claramente:

En la clase de sistema inteligente concebido por los diseñadores de la quinta generación, la velocidad y el poder del procesado se incrementarán dramáticamente; pero, más importante aún, las máquinas tendrán el poder de razonar: ordenarán automáticamente vastas cantidades de conocimiento para servir cualquier propósito propuesto por el hombre, desde el diagnóstico médico al diseño de productos, desde las decisiones de dirección a la enseñanza.

(6) Edward Feigenbaum y Pamela McCorduck, *The fifth generation*, Addison-Wesley, 1983. El título alude a las generaciones de los ordenadores. La quinta es la de la inteligencia artificial, estando basadas las cuatro primeras en las lámparas de vacío, los transistores, los circuitos integrados y los circuitos de alta integración.

Pero, al comprobar que el problema es más difícil de lo previsto y ante el alargamiento de los plazos, el mismo Edward Feigenbaum lanza la idea, menos ambiciosa pero más segura, de *sistema experto* que algunos interpretan como un primer paso hacia la IA, otros como el abandono de las radicales pretensiones primeras.

Un sistema experto es un programa de ordenador que intenta imitar e incluso superar a un experto en un ámbito concreto de la actividad humana, por ejemplo a un médico que diagnostica una enfermedad o a un economista que toma una decisión sobre una empresa. Ya no se trata de reproducir el pensamiento humano, sino simplemente la pericia de un profesional competente. Esta pretensión es más sencilla porque en algunos campos reducidos los expertos trabajan siguiendo reglas, aunque a menudo sin ser plenamente conscientes de ello. En esos dominios la capacidad del ordenador de examinar muchas alternativas y aplicarles reglas claras puede resultar ventajosa, ya que, si el campo es estrecho, puede ser innecesario el sentido común. Un sistema experto consta de dos elementos: *la base de conocimiento*, una serie de reglas, y *la máquina de inferencias*, un programa que aplica las reglas al caso que se le presenta. Las reglas pueden ser de dos clases: reglas lógicas del tipo si-entonces (por ejemplo, si el número de glóbulos blancos supera una cierta cantidad, entonces buscar una infección) o reglas heurísticas basadas en la práctica de un experto, codificadas tras entrevistar a varios de ellos.

Al abandonar la intuición y el sentido común, los sistemas expertos han conseguido éxitos, tanto mayores cuanto más reducido sea el campo de operación. Por ejemplo, hay sistemas para prospección minera (el descubrimiento de dos yacimientos desconocidos en Oregon se presentó como un triunfo resonante hace algunos años), para análisis químicos, para diagnóstico médico, elección de antibióticos, ayuda al estudio, jugar al ajedrez, controlar el tráfico aéreo o el de una gran ciudad... Hoy se usan en muchos campos de manera habitual como ayuda a la toma de decisiones, aunque siempre la última responsabilidad está todavía en manos de una persona.

Pero frente a esa línea de entusiastas que presentan una visión triunfalista de la IA, la primera crítica de Lucas fue seguida por una serie de científicos y filósofos que dan varios tipos de argumentos en contra de la posibilidad de que los ordenadores puedan ser inteligentes (7).

(7) R. Penrose: *El vestido nuevo del emperador*. Mondadori, 1992; este libro, que toma partido contra la hipótesis fuerte de la IA, ha

El test de Turing y la habitación china

EN su artículo de 1950, Alan Turing propuso medir la inteligencia de un ordenador sometiéndole a una conversación con humanos de la siguiente manera. Un ordenador y un voluntario se encierran en una habitación ocultos de otra persona que los interroga por turno, pasándoles las preguntas por una ventanilla, a las que responden de manera impersonal en textos escritos a máquinas o mediante una pantalla. El voluntario contesta las preguntas intentando convencer al interrogador de que el humano es él, pero el ordenador está programado para mentir y simular las respuestas que daría un hombre o una mujer. Por ejemplo, si le piden un cálculo matemático complicado y responde muy deprisa, mucho más que lo haría una persona, sería descubierto en el momento, por lo que tarda en contestar o se equivoca. El juego consiste en que el interrogador averigüe quién es el ordenador y quién el humano, basándose en las respuestas. Para ello podría tender trampas dialécticas haciendo preguntas absurdas a las que sea difícil de contestar siguiendo reglas fijas, como «Ayer vi a una jirafa conduciendo una moto sin casco. ¿Qué le parece?» Si, tras muchas preguntas, el interrogador no consiguiese descubrir cuál de los dos es el ordenador, éste habría pasado el test y, según Turing, debería ser considerado inteligente.

Naturalmente hay muchas discusiones sobre la realización práctica del test de Turing, aunque en principio puede admitirse como un buen método. Se ha conseguido que algunos aparatos lo pasen si las preguntas son de tipos especiales. Como ejemplo, hay programas que simulan entender historias de una cierta clase como «Un hombre entró en un restaurante y pidió una hamburguesa. Cuando se la trajeron estaba quemada y se marchó enfadado sin dar propina» o «Un hombre entró en un restaurante y pidió una hamburguesa. Estaba muy buena y dio una propina», respondiendo acertadamente si el hombre se comió la hamburguesa en cada caso, aunque no se lo habían dicho. Pero, sin duda son versiones muy simples y toscas del test.

El filósofo de Berkeley John Searle publicó en 1980 (8) un argumento

causado mucho impacto por la personalidad de su autor. Es muy recomendable para los interesados en el tema, aunque su lectura no sea fácil sin una base en física o matemáticas.

(8) J. Searle: «Minds, brains and programs», en *The behavioral and brain sciences*, vol. 3, Cambridge University Press, reproducido en *The mind's I*, editado por D. R. Hofstadter, Basic Books, Penguin, 1981.

contrario a la hipótesis fuerte de la IA con el que pretende demostrar que ser capaz de respuestas aceptables sobre un tema no significa *entender*. Searle imagina a una persona que no habla el chino, encerrado en una habitación con diccionarios y tratados sobre un tema del que es completamente ignorante, tal como magnetohidrodinámica o bioquímica. Alguien le presenta preguntas escritas en chino para las que, usando los diccionarios y los tratados, construye respuestas también en chino que serían indistinguibles de las de un chino nativo conocedor del tema, aunque la persona no entienda nada sobre lo que significan. Desde el punto de vista práctico, el proceso sería terriblemente lento, pero esto no importa porque aquí se trata de una cuestión de principio. Para acelerarlo se puede suponer que en la habitación hay ediciones bilingües de los tratados y que muchas personas se ocupan de las respuestas. Según Searle, es posible mantener una conversación inteligente con la habitación, la cual pasarían así con éxito el test de Turing, a pesar de no entender nada, lo que prueba que puede haber mucha distancia entre un comportamiento que parece inteligente y el verdadero entendimiento.

Algunos han criticado el argumento, diciendo que, aunque la persona o las personas no entiendan, la habitación es un ente superior con un nivel de organización más alto que sí lo hace. En alguna versión, se ha metido dentro a todos los habitantes de la India (menos los que entienden el chino, claro está) que serían algo así como las neuronas de nuestro cerebro, pues el entender es propio del conjunto de todas ellas y no de cada una en particular. Pero resulta difícil comprender qué significa que la India como nación entienda algo que ninguno de sus ciudadanos entiende. Por ello y en cualquier caso, parece que Searle señala un verdadero problema para la IA y que su habitación china muestra la cautela con que hay que hablar del comportamiento inteligente.

Las reglas y el sentido común

HUBERT Dreyfus, un filósofo de Berkeley que trabajó antes sobre IA en Harvard y en el MIT, es uno de los críticos más feroces de la hipótesis fuerte de la IA. En 1972 publicó un libro polémico y provocativo, *Lo que los ordenadores no pueden hacer* (9),

(9) H. L. Dreyfus: *What Computers can't do*. Harper & Row, 1972.

seguido en 1986 de *Mind over machine* (10), escrito con su hermano Stuart que es profesor de matemáticas en la Universidad de California y especialista en investigación operativa. Dreyfus reacciona fuertemente contra la idea de que los ordenadores y los humanos sean dos «razas» de una misma «especie», caracterizada por representar el mundo exterior mediante cadenas de símbolos. Su crítica se centra en la imposibilidad de programar el sentido común, pues «después de treinta años de esfuerzo nadie tiene la menor idea de cómo hacerlo», y en que el pensamiento no se puede reducir a la aplicación de reglas, que es lo más que pueden hacer los ordenadores. Según él, nuestros hábitos mentales están profundamente marcados por toda nuestra experiencia vital anterior y por nuestros cuerpos, lo que expresa en una frase redonda: «Nunca se podrá programar nuestro pensamiento porque no somos espíritus puros» (11). Hay que entender que los ordenadores sí lo son porque, al ser máquinas universales en el sentido de Turing, son absolutamente intercambiables con la programación adecuada, como cuestión de principio, y nada depende en ellos de la materia particular con que estén contruidos.

Para Dreyfus es muy importante comprender que el pensamiento consiste en algo más que en seguir reglas. Además, y en esto se apoya el filósofo Ludwig Wittgenstein, «las reglas no contienen las reglas de su propia aplicación» (12). La siguiente historia tomada de la página de sucesos de un periódico ilustra esta objeción. Un conductor de autobús, sancionado por no atender adecuadamente a un viajero que sufrió un ataque cardíaco, se defiende asegurando que siguió estrictamente la regla de su compañía que prohíbe a los conductores desviarse de la ruta prevista sin tener permiso para ello. La culpa, por tanto, no fue suya sino de las reglas que son necesarias para evitar el desorden.

Naturalmente lo que ocurre es que las reglas deben interpretarse según el contexto, pero tanto ese conductor del autobús como los ordenadores son incapaces de pensamiento contextual, es decir que, al no entender nada, no se dan cuenta de que, a veces, el contexto dice que se rompan las reglas o, al menos, que se cambie su interpretación, que es lo

(10) H. L. Dreyfus and S. E. Dreyfus: *Mind over machine: The power of human intuition and expertise in the era of the computer*. The Free Press, Mac Millan, New York, 1986.

(11) Entrevista con Guitta Pessis-Pasternak: *Dynasteurs*. 1 de noviembre de 1990, reproducida en Guitta Pessis-Pasternak, *Faut-il brûler Descartes?*, Editions La Découverte, Paris, 1991, pág. 213.

(12) H. Collins: «Will machines ever think?», *New Scientist*, 20 de junio de 1992, pág. 36.

que haría una persona con sentido común. Un partidario de la IA diría que es posible sustituir al conductor por un ordenador programándole adecuadamente, por ejemplo con una instrucción del tipo: «Seguir siempre la ruta prevista, excepto si un viajero sufre un ataque al corazón, en cuyo caso ir al hospital más próximo», con lo que sería por lo menos más inteligente que el conductor. Pero es imposible prever siempre todas las posibilidades porque enfrentarse a lo imprevisto no se puede reducir a reglas. Hay que fiarse del sentido común o de la intuición. Por ejemplo podría ser mejor ir a un hospital que no es el más próximo pero está mejor equipado o seguir el consejo de un médico que estuviese a bordo o el hospital podría estar cerrado por un incendio o el autobús podría haber sido secuestrado, etc.

Minsky y Feigenbaum (13) y los defensores de la IA argumentan que el sentido común se puede generar mediante un conjunto suficientemente vasto de reglas y muchos datos almacenados en la memoria de un ordenador. Para Dreyfus, en cambio, eso es imposible, aunque sí ocurre que en áreas reducidas las reglas pueden llegar a cubrir todas las posibilidades y, en esos terrenos, un ordenador puede ganar a una persona: eso es precisamente lo que hacen los sistemas expertos antes mencionados, aunque el argumento de la habitación china sugiere que no entienden nada.

Se ha hablado mucho de los programas de ajedrez que son sistemas expertos y que hoy pueden vencer a los aficionados de nivel medio. Desde 1950 han ido ganando en complejidad y eficacia, pudiendo examinar actualmente varias decenas de millones de posiciones finales tras cuatro jugadas, antes de decidir cada movimiento, por lo que cuando ganan lo hacen por la «fuerza bruta». Pero son completamente ajenos a la estrategia, como ilustra el caso del ajedrecista escocés David Levy, más o menos el jugador número mil en el escalafón mundial, que era gran partidario de la IA y estaba convencido de que un programa de ordenador llegaría a batir a los mejores campeones. Pero pudo ganar por cuatro a cero a un ordenador Cray con el programa Cray-Blitz, campeón del mundo de ordenadores en 1984, al darse cuenta que le convenía complicar el juego porque la fuerza bruta de la máquina se hacía insuficiente, mientras que el «saber cómo» y la intuición de un buen jugador le permite orientarse en

(13) Entrevista con Guitta Pessis-Pasternak, *Dynasteurs*, 1 de noviembre de 1990, reproducida en Guitta Pessis-Pasternak, *Faut-il brûler Descartes?*, Editions La Découverte, París, 1991, pág. 220.

una situación confusa sin seguir reglas lógicas. Tras ello llegó a decir que «el programa no entendía lo que estaba pasando» (14).

Realmente parece que la reflexión humana es más compleja que una aplicación reiterada de reglas, aunque a veces sea también eso. El pensamiento oscila siempre entre los polos del análisis y la síntesis, entre la deducción y la inducción, mientras que los ordenadores trabajan sólo en el modo analítico deductivo. Creo que se puede mantener que muchas de las mejores creaciones de la mente humana se han obtenido con una componente importante de razonamiento ni analítico ni deductivo. Esto es evidente en las obras de arte, como la Gioconda de Leonardo, el Quijote de Cervantes o un cuarteto de Beethoven, que no se han creado siguiendo reglas porque no obedecen a ningún esquema lineal lógico. Según un lugar común, las obras de los científicos son de otro tipo pues en ellas reina la lógica más estricta en todo su esplendor, pero, si tomamos las de cualquiera de los grandes, digamos un Einstein o un Darwin, veremos que, muy al contrario, la analogía juega en ellas un papel muy importante. Valga, si no, como ejemplo la descripción que Freeman Dyson hace del trabajo de su amigo el gran físico norteamericano Richard Feynman, Premio Nobel en 1965, y uno de los grandes del medio siglo (15), al hablar de las dificultades que tuvo para que se aceptasen sus primeras ideas:

La razón por la que su física era tan difícil de captar por los físicos ordinarios era que no usaba ecuaciones... Tenía una visión física de cómo ocurren las cosas y esa visión le daba las soluciones directamente con un mínimo de cálculo. No me sorprende que los que habían pasado sus vidas resolviendo ecuaciones estaban desconcertados por sus ideas. Sus mentes eran analíticas; la suya era pictórica.

Los algoritmos y el teorema de Gödel

SE decía más arriba que el problema de la inteligencia artificial consiste en saber si el cerebro se puede considerar o no como una realización particular de la máquina universal de Turing, para lo que es preciso discutir el concepto de algoritmo (16).

(14) *Los Angeles Times*, 12 de mayo de 1984, historia comentada en la obra citada de los hermanos Dreyfus.

(15) Freeman Dyson: *Disturbing the universe*. Harper and Row, New York, 1979.

(16) Penrose, obra citada.

Un algoritmo (17) es simplemente una sucesión de operaciones elementales, perfectamente especificadas y ordenadas, que sirven para hacer algo preciso. Es muy importante que esas operaciones estén perfectamente definidas, de manera que se pueda saber cuál es la primera, la décima, la número 34, etc. Por ejemplo, un algoritmo para freír un huevo sería: 1. Se pone una sartén en el fuego con aceite, 2. Se coge un huevo de la nevera, 3. Se casca, 4. Se echa en la sartén, 5. Se esperan dos minutos, 6. Se coge con una espumadera, 7. Se apaga el fuego. En otras palabras: un algoritmo es una receta. También son algoritmos las secuencias de operaciones de un obrero en una cadena de montaje, aunque se suele utilizar esta idea en un contexto matemático, donde las operaciones tienen sentido aritmético. Por ejemplo, para multiplicar o sumar dos números se emplean algoritmos, aunque no pensamos en ello al hacerlo. Lo más importante de esta idea es la descomposición de un trabajo en tareas elementales y su realización sucesiva en un orden completamente prefijado, de modo que, si se omite alguna o se altera el orden de las operaciones, el algoritmo no funciona. Por ejemplo, si al freír un huevo se omite el segundo paso, es imposible realizar la tercera operación (no se puede cascar un huevo que no se ha cogido) y si se intercambian la 5 y la 6 el huevo no se fría. Nótese además que un algoritmo es incapaz de dar respuesta a una situación imprevista, por ejemplo, si está roto el reloj y no se pueden medir los dos minutos, si el huevo no está en la nevera o si no hay espumadera.

Descartes en su *Discurso del método* de 1637 propuso una manera de razonar para resolver cualquier problema por complejo que sea. La geometría analítica, desarrollada por él mismo y en la que los problemas espaciales se reducen a operaciones aritméticas, le había producido una fuerte impresión, lo que le llevó al convencimiento de que un modo de proceder parecido es posible siempre. Por ello su método consiste en analizar cualquier cuestión en elementos básicos aislables y en reducir su solución a aplicar reglas precisas de manera sucesiva. De este modo, argumentaba, todo el conocimiento podría deducirse a partir de primeros principios, los cuales se podrían deducir a su vez preguntándose cómo debe haber organizado el mundo un Dios racional y bueno. Este punto de vista puede resumirse diciendo que, para Descartes, el pensamiento humano procede mediante algoritmos.

(17) Esta palabra viene del nombre del matemático persa del siglo IX Abu Ja'far al-Khowarizm, autor de un libro muy importante de Álgebra.

Esta primacía del razonamiento analítico dominó durante mucho tiempo la escena del pensamiento, en parte bajo el impulso del enorme desarrollo de las matemáticas. Un capítulo posterior de esta tradición que nos interesa aquí es el intento por el gran matemático alemán David Hilbert de desarrollar un método para la resolución de todos los problemas de matemáticas, inada menos que el sueño de Descartes! Al ser invitado a dar la conferencia de honor en el Congreso Internacional de Matemáticos de París en 1900, su tema elegido fue la resolución de problemas como motor del pensamiento matemático, proponiendo veinte de ellos, los llamados *Problemas de Hilbert*, como reto ante el siglo que empezaba, algunos de los cuales no se han resuelto todavía. Hilbert trabajaba entonces en un programa muy ambicioso para fundamentar la matemática sobre bases absolutamente firmes e indudables, a base de axiomas y reglas de inferencia aplicadas sucesivamente de modo algorítmico, que permitiese construir un método para probar la verdad o falsedad de cualquier proposición. Pero no lo había conseguido completamente, por lo que colocó en el décimo lugar de su lista de problemas el conocido como *Entscheidungsproblem* que consiste en probar que ese método existe realmente.

Tres décadas más tarde, en 1931, el joven matemático austriaco Kurt Gödel probó un teorema que echó por tierra las esperanzas de Hilbert. El llamado teorema de Gödel dice que todo sistema formal de axiomas y reglas de procedimiento, que contenga la aritmética elemental y que esté libre de contradicciones, incluye necesariamente afirmaciones que no se pueden probar ni refutar desde dentro del sistema. Se dice que la verdad de tales afirmaciones es *indecible*. La prueba del teorema es complicada, incluso su enunciado técnico lo es, pero para nuestro propósito basta con decir que implica que hay proposiciones que sabemos que son ciertas, pero que no se pueden probar. Como ejemplo tomemos la paradoja del mentiroso, ya estudiada por los griegos, que consiste en saber si la frase «Esta frase no es cierta» es verdadera o falsa. En los dos casos hay una contradicción, pues si suponemos que es verdadera, resulta que es falsa y al revés. Una frase parecida es «Esta proposición no se puede probar». Llamemos *P* a esta proposición que tiene algo de perturbador porque *P* es verdadera solamente si no se puede probar. Algunos dicen que esta contradicción no tiene importancia porque las proposiciones que no se pueden probar aunque sean ciertas son o muy complicadas o carecen de interés como puede ocurrir con *P*. Esta es una posición pragmática que sirve para salir del atolladero, pero que ignora que la fugitiva relación

entre verdad y prueba es una de las más importantes cuestiones del pensamiento y de la lógica, por lo que la sola existencia de tales proposiciones hace necesario replantear la idea de verdad. Lo que más nos importa aquí es que cualquier procedimiento algorítmico tiene límite de aplicación y no sirve para todas las afirmaciones contenidas en su sistema de axiomas y reglas.

La crítica de Lucas a la IA antes mencionada, que inició la corriente de opinión contraria, se basa precisamente en el teorema de Gödel que sugiere la necesidad de acudir a procedimientos no algorítmicos para decidir algunas cuestiones. Pero, la máquina universal de Turing trabaja siguiendo algoritmos, lo que significa que todos los ordenadores hacen lo mismo, afirmación conocida como *la tesis de Church-Turing*, lo que impone una limitación objetiva de las máquinas para comportarse de manera inteligente.

A pesar de ello, los partidarios de la IA creen, siguiendo la tradición cartesiana, que el pensamiento es algorítmico, por lo que el cerebro puede imitarse completamente mediante un ordenador al que se programa para seguir el algoritmo adecuado en cada caso. Si a veces parece que nuestro razonar sigue un esquema más directo o más complejo, como cuando aplicamos el simple sentido común o cuando resolvemos un problema «de un chispazo», por intuición, esto se debe a que el algoritmo es muy complicado y no evidente. Feigenbaum y Minsky, entre otros, insisten en que el sentido común se puede llegar a programar introduciendo en un ordenador de gran memoria los datos que representen toda la experiencia acumulada de una persona y un conjunto muy grande de reglas simples. Para ellos, nuestro cerebro es una máquina de tratamiento de información muy poderosa y *nada más*.

Los detractores de la IA, por su parte, niegan el carácter algorítmico del pensamiento. Aceptan, desde luego, que en algunas ocasiones los hombres razonan mediante reglas, como a veces hace un médico al diagnosticar a un paciente a partir de sus análisis clínicos, pero dicen que muy a menudo la visión global u holística de un problema, no reducible a algoritmos, juega un papel muy importante. En otras palabras: si bien el pensamiento deductivo es algorítmico, el sentido común, la intuición, el pensamiento contextual, las analogías y la síntesis siguen procesos de tipo más complicado que nunca se pueden reducir a reglas. Este punto, el carácter algorítmico del pensamiento, es precisamente la clave del problema.

Tomemos el caso de un trabajador en una cadena de producción que debe colocar un circuito integrado en una lavadora. Normalmente se le da un conjunto de instrucciones que forman un algoritmo, pero, como el conductor de antes, debe estar preparado para variaciones sutiles. A veces el circuito no encaja bien y tiene que manejarlo de otra manera o trae adherida una pieza de metal que hay que quitar o le duele la mano derecha y debe usar la izquierda. El ideal algorítmico sería de una monotonía insoportable para cualquier persona (recordemos a Chaplin en *Tiempos modernos*) y es completamente distinto de cómo llevamos a cabo las tareas que debemos repetir, como lavarnos cada mañana o dar un golpe de revés al jugar al tenis, en las que tenemos que hacer pequeños ajustes para responder a lo imprevisto del contexto.

¿Una nueva física?

NO debe omitirse aquí la mención de una idea sorprendente propuesta hace poco por Roger Penrose en el libro antes citado. Si rechazamos la hipótesis fuerte de la IA, pero aceptamos que las leyes de la física se aplican a la materia viva, y el cerebro lo es sin duda, ¿por qué no se puede reproducir su funcionamiento con un mecanismo, sea un ordenador u otro distinto? Penrose cree que la física de la mente no se ha descubierto aún y que estará en la frontera entre la teoría cuántica y la gravedad, en lo que él llama la gravedad cuántica correcta que está aún por formular y que deberá tener un carácter no algorítmico, lo que sugiere que sería una teoría de una gran dificultad. Invirtiendo la creencia general, opina que para conseguirla no hay que hacer más cuántica a la gravedad sino más gravitatoria a la teoría cuántica. Es ésta una idea original, tan atractiva como imprecisa aún, difícil de entender y que quizás no sea muy aceptada, al menos por un tiempo. Pero, de ser correcta, abriría perspectiva sugerentes porque relacionaría la física de la mente con la del Universo.

Resumen

LA polémica de la inteligencia artificial es uno de los debates más interesantes del segundo medio siglo XX, pues toca cuestiones muy profundas de matemáticas, física, bioquímica, psico-

logía y filosofía. Entre los partidarios de lo que hemos llamado hipótesis fuerte de la IA, que afirma que el cerebro no es más que un ordenador hecho de carne, están todos aquellos que confían en que pronto se podrá reducir todo el comportamiento de la naturaleza entera a un esquema conceptual bien determinado y simple, en el sentido de que cualquier pregunta pueda hallar una respuesta no ambigua mediante la aplicación mecánica secuencial de una colección de reglas lógicas elementales y una cantidad muy grande de datos. Una importante consecuencia de la hipótesis fuerte es que la libertad humana no existiría, sería tan sólo una apariencia. Muchos creen que ese esquema es una consecuencia inevitable del desarrollo de la ciencia de hoy y que no tardará mucho. Su entusiasmo hace que se haya extendido la idea de que todos los científicos creen en la inteligencia artificial.

Sin embargo, esto no es así, pues son muchos los que consideran que el mundo, y especialmente el cerebro, es más complicado de lo que supone la IA y que el pensamiento no puede reducirse a reglas. Según éstos, a quienes los desarrollos de los últimos años parecen estar dando la razón, el sentido común, la intuición y las analogías no pueden reducirse a cadenas de reglas, por muy complicadas que sean. Por ello, aunque es posible fabricar sistemas expertos que funcionan bien en campos específicos y tienen importantes aplicaciones prácticas, nunca se podrá construir una máquina de la que se pueda decir que piensa. Un argumento importante en su favor lo constituye el teorema de Gödel que asestó un golpe muy duro a las esperanzas de poder codificar completamente el pensamiento, al afirmar que cualquier sistema lógico es necesariamente incompleto. Si este grupo tiene razón y yo creo que sí la tiene, para obtener un mecanismo artificial capaz de pensar como un cerebro, habría que construir inadménos que otro cerebro!