Introducción

La inteligencia artificial (IA) puede parecer, de entrada, un tema de ciencia ficción que nada tiene que ver con la filosofía. Sin embargo, la ciencia ficción, cuando es valiosa, plantea grandes cuestiones filosóficas. En concreto, la inteligencia artificial ha sido tratada por los mejores novelistas del género, tales como Isaac Asimov, Philip K. Dick y William Gibson, como una excusa para reflexionar sobre muchas grandes cuestiones filosóficas, de las que aquí destacamos dos: qué es ser humano y las consecuencias distópicas de la liberación de la razón instrumental o subjetiva respecto de los imperativos éticos dictados por la razón axiológica u objetiva. Estos son los dos temas centrales de los que nos ocuparemos en el presente estudio. El primero lo abordaremos a través del examen de las condiciones de posibilidad técnicas de la IA. El segundo, haciendo lo mismo con sus condiciones de posibilidad sociales.

Toda tecnología, entendiendo esta en un sentido amplio que abarca tanto la técnica como los instrumentos producidos por ella, ha de satisfacer dos tipos de condiciones de posibilidad, es decir, de requisitos para ser posible. Por un lado están las condiciones puramente técnicas, que refieren a la posibilidad material de manipular la naturaleza para obtener la tecnología en cuestión. Así, por ejemplo, las computadoras electrónicas son técnicamente posibles, pues existen en efecto, y todo lo efectivo es, cuando menos, posible. Sin embargo, hay grupos sociales, como los *amish*, que las excluyen de su vida ordinaria porque no se adecúan a sus intereses, regidos estos por causas de diversa índole. Tal rechazo apunta al otro tipo de condiciones de posibilidad que debe cumplir toda tecnología: las sociales, que refieren a su compatibilidad con el espíritu, es decir, con lo humano que trasciende la mera naturaleza. Naturaleza y espíritu, por tanto, imponen sus respectivas exigencias.

Las condiciones de posibilidad técnicas de la IA las examinaremos en el capítulo séptimo, y las sociales, en el octavo. Pero antes será necesario un extenso trabajo preliminar que nos ocupará el resto de los capítulos: desde el

primero al sexto. Empezaremos por describir la noción vulgar de IA, que es la que se maneja a pie de calle, y que a grandes rasgos es la de una máquina con unas destrezas intelectuales similares a las de un ser humano. En sentido estricto, atendiendo al significado de las siglas *IA*, una máquina que imitase el pensamiento de un animal inferior también sería una inteligencia artificial. Sin embargo, el tema de este estudio no serán las condiciones de posibilidad de la duplicación de cualquier inteligencia, sino solo de la inteligencia humana, que es la que coincide con la noción vulgar de IA. Así definida, en términos antropocéntricos, lo cierto es que la IA es una tecnología que todavía no existe. Hay máquinas que hacen cosas asombrosas, como jugar al ajedrez mejor que el mejor ajedrecista humano, pero no se puede decir, en rigor, que sean auténticas inteligencias artificiales. Su inferioridad respecto a nosotros se debe, principalmente, a que carecen de dos habilidades: en el mundo social, del lenguaje, y en el mundo físico, de nuestra versatilidad para hacer un intento pasable en casi cualquier cosa.

La IA es, por tanto, una tecnología que todavía no existe y, en consecuencia, su noción vulgar no puede proceder del contacto mundano con ella. A nuestro juicio, su presencia en el imaginario popular es resultado de la acción de tres agentes: la mercadotecnia, los investigadores de la IA y la ciencia ficción. La mercadotecnia lleva décadas pregonando que podemos adquirir bienes materiales cada vez más inteligentes. Televisores, teléfonos, automóviles e incluso objetos no mecánicos como los tejidos: la inteligencia es un atributo que abunda en todos ellos, y que se predica en un grado proporcional a la capacidad del objeto para comprender y satisfacer eficazmente las órdenes o deseos de su propietario. Los investigadores de la IA, por su parte, llevan mucho tiempo prometiendo que están cerca de duplicar la inteligencia humana. Desde la fundación de su disciplina, allá por los años 50, no han cesado de hacer previsiones que jamás se cumplen, y no solo predicciones, sino de afirmar contra la evidencia que sus máquinas son capaces de hacer cosas que, sencillamente, no hacen. Su estrategia, en muchos casos, es la de repetir la misma mentira una y otra vez con la esperanza de que los demás terminen creyéndola. Se pueden postular muchas causas de tan deshonesta conducta, pero la fundamental es que se comportan así para conseguir financiación. Por encima de su condición de científicos, son seres humanos con ambiciones profesionales y con necesidades materiales que cubrir. Por último, respecto a la ciencia ficción, es quizás, en retroalimentación con los otros dos, el agente que más ha contribuido a configurar la noción vulgar de IA. En los peores casos, a través de obras literarias y audiovisuales elaboradas con la única finalidad de recoger beneficios, y en lo mejores, con obras que han utilizado el tema, como decíamos al comienzo, para abordar asuntos de gran importancia filosófica.

Ahora bien, el éxito de la IA como gancho comercial y como argumento para justificar partidas de los presupuestos públicos destinados a investigación debe tener una causa. Este asunto, el origen de la atracción por las máquinas pensantes, lo abordaremos en el *capítulo segundo* en un doble sentido: antropológico e histórico. El origen en sentido antropológico lo descubriremos examinando el significado de los mitos de la recreación del hombre por el hombre. De la mano del filósofo especialista en cibernética André Robinet apreciaremos similitudes y diferencias de estos mitos en las religiones monoteístas y politeístas. A continuación, rastrearemos el origen histórico de los primeros autómatas del pensamiento. Fue en el siglo XVII, recién iniciada la Modernidad, cuando, por causas técnicas y sociales, pudieron acometerse los primeros intentos plausibles de construir autómatas del pensamiento. En concreto, autómatas capaces de efectuar cálculos matemáticos. Entre los que se lanzaron a la aventura de intentar construir estos artefactos figuran nombres tan ilustres como los de Pascal y Leibniz. Sin embargo, el primero que lo conseguiría fue un humilde matemático anónimo, el alemán Wilhelm Shickard. Desde entonces, 1623, hubieron de pasar doscientos años hasta que un inventor inglés, Charles Babbage, llevase la cuestión un paso más allá. Con su máquina analítica Babbage planteó el diseño de una máquina casi tan versátil como las computadoras electrónicas actuales. Por desgracia, solo tuvo tiempo y dinero suficientes para construir la mitad.

Volviendo al origen histórico de las máquinas pensantes, en el siglo XVII la posibilidad de construir semejantes artefactos, capaces de realizar operaciones mentales, o cuando menos de imitar la conducta producida por ellas, fue recibida de manera desigual por la filosofía de la época. En la última parte del capítulo segundo contrastaremos la diferencia de posturas al respecto entre Descartes y Hobbes. El primero, padre del racionalismo moderno, se basó en su metafísica del dualismo de sustancias y en sus convicciones cristianas para argumentar en contra de la posibilidad técnica de duplicar de manera perfecta el pensamiento humano. A su entender, había dos facultades de nuestro intelecto que no podían ser producidas por combinaciones mecánicas, y son justo las dos que mencionamos en el capítulo anterior: en el mundo social, el lenguaje, y en el mundo físico, la flexibilidad para acometer casi cualquier tarea. Descartes demostró así una agudeza excepcional, al adelantarse en más de trescientos años al diagnóstico de los dos mayores obstáculos de la IA.

En cambio, Hobbes, empirista y padre de la psicología mecanicista moderna, no solo no apreciaba ningún impedimento técnico en la empresa de los autómatas del pensamiento, sino que, de haberse logrado construirlos, se habrían confirmado sus tesis materialistas. Apuntemos también que la elección de estos dos autores, Descartes y Hobbes, como exponentes del contexto fi-

losófico en el que surgieron los primeros autómatas del pensamiento, obedece no solo a su representatividad respectiva dentro de las corrientes racionalista y empirista, sino también a que el dualismo de sustancias cartesiano persiste actualmente, en cierta manera, en la corriente de la IA denominada IA simbólica, mientras que el corporalismo de Hobbes tiene su continuación en la otra gran corriente de esta disciplina: la IA subsimbólica.

En el *capítulo tercero* describiremos las características de las computadoras, imprescindibles para elucidar las condiciones de posibilidad técnicas de la IA en tanto que son las máquinas con las que los científicos pretenden construirla. Lo haremos distinguiendo tres dimensiones en ellas: formal, material y pedagógica. A nivel formal las computadoras electrónicas son sistemas formales, es decir, conjuntos de símbolos sobre los que se aplican reglas para formar y transformar expresiones. Para ilustrar el potencial de los sistemas formales, y también alguna limitación, describiremos el funcionamiento de las máquinas de Turing, artefactos ideales que hacen exactamente lo mismo que cualquier computadora real, esto es: ejecutar algoritmos, conjuntos finitos de instrucciones rutinarias cuya ejecución arroja un resultado deseado. El repaso de la historia de los autómatas matemáticos realizado en el capítulo anterior cobrará un nuevo sentido en este cuando descubramos que Alan Turing concibió sus máquinas con la intención de definir formalmente las tareas de computación que hasta el momento venían siendo efectuadas por redes de computadores humanos dotados de una inteligencia no especialmente brillante. Respecto a la dimensión material de las computadoras electrónicas, hablaremos sobre la suma de componentes y la miniaturización. Esta última es una técnica que durante décadas ha permitido mejorar las prestaciones de estas máquinas de manera exponencial pero que, sin embargo, está a punto de toparse con límites físicos infranqueables. Y, por último, desvelaremos la condición pedagógica inherente a toda técnica en general, y en concreto cómo afecta la pedagogía de las computadoras al alcance de lo que se puede hacer con ellas. En un texto tan antiguo como el *Fedro* de Platón encontraremos limitaciones que pesan decisivamente sobre el diseño de programas informáticos y que no pueden ser superadas ni con la más moderna tecnología.

Como ya hemos mencionado, la IA se divide en dos grandes corrientes o programas de investigación: IA simbólica e IA subsimbólica. La primera pretende duplicar la mente, y la segunda, el cerebro. Ahora bien, mente y cerebro puede ser concebidos de muchas maneras posibles. Exponer con detalle los modelos de la mente y del cerebro que pretenden ser imitados, respectivamente, por la IA simbólica y la IA subsimbólica será el objetivo del *capítulo cuarto*. Para ello comenzaremos dedicando una sección a exponer una serie de conceptos de filosofía de la ciencia que nos harán falta, tales como el de pa-

radigma de Thomas Kuhn y el de programa de investigación de Imre Lakatos, así como para abordar tres distinciones: realismo e instrumentalismo, racionalismo y relativismo, explicar y comprender. En la primera demostraremos que la ciencia, siempre en el sentido de la ciencia moderna, es un mero instrumento para la dominación de la naturaleza que no puede albergar pretensiones de verdad. En la segunda argumentaremos en favor de la tesis de que el método científico, entendido como procedimiento algorítmico que garantiza la obtención de conocimiento, es un mito. En su lugar, la actividad científica real procede aplicando estrategias lógicas y psicológicas tanto en el contexto de descubrimiento como en el de justificación. La existencia de estrategias psicológicas en el contexto de justificación constituye una prueba en favor del enfoque relativista de la ciencia frente al racionalista. Y, por último, en la tercera distinción abordaremos las diferencias entre explicar y comprender. Las ciencias de la naturaleza emplean un método explicativo, mientras que el de las ciencias sociales es comprensivo.

El método explicativo se fundamenta en un enfoque molecular que va de abajo a arriba, intentando reducir los fenómenos más complejos a fenómenos simples o atómicos. En cambio, el método compresivo tiene un enfoque molar que va de arriba a abajo, para captar el significado de los fenómenos más simples en el contexto total en el que se dan. La diferencia entre ambos métodos afecta de manera singular a la psicología, dado que, por ocuparse esta del estudio de la mente, y por ser la mente producto de la interacción de factores naturales y sociales, la psicología se ve obligada a intentar integrar la explicación y la comprensión, la molecularidad con la molaridad. Ante la imposibilidad de realizar semejante síntesis, los diversos paradigmas de la psicología han optado por privilegiar uno de los dos métodos. En el caso de la psicología cognitiva, o cognitivismo, que es el paradigma que suministra a la IA simbólica su modelo de la mente, su opción ha sido adoptar el enfoque explicativo, el propio de las ciencias de la naturaleza.

Una vez concluida la digresión sobre conceptos de filosofía de la ciencia, expondremos los rasgos esenciales del cognitivismo, que son cinco. Dos son supuestos nucleares, y los otros tres son rasgos metodológicos. Los supuestos nucleares son la tesis internalista y la tesis del procesamiento de información. La tesis internalista es la que habíamos anticipado al final del capítulo segundo, que sostiene el dualismo de sustancias cartesiano. No de manera ontológica, es decir, no afirma en plena era espacial que la mente y el cuerpo sean sustancias distintas, pero sí de manera metodológica, en tanto que sostiene que para explicar el funcionamiento de un sistema intencional, como es la mente humana, es necesario postular la existencia de un nivel mental de representaciones independiente de los procesos biológicos de los que surge, esto es: que la

mente es explicable con independencia del cuerpo. En cuanto a la tesis del procesamiento de información, también conocida como metáfora computacional, caracteriza a la mente como un procesador de información similar a una computadora electrónica. La consecuencia de tal postulado es, como se puede apreciar, 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. En este círculo se han movido los investigadores de la IA durante décadas, y de él han tomado la confianza optimista que les ha empujado a afirmar que las máquinas pensantes se conseguirían en poco tiempo.

Una vez descritos los demás rasgos del cognitivismo, pasaremos a caracterizar la ciencia que proporciona a la IA subsimbólica el modelo del cerebro en el que se basa. Esta es la neurociencia. A diferencia de la psicología, en la que abunda la variedad de escuelas contrapuestas, la neurociencia lleva mucho tiempo progresando acumulativamente en torno a un solo paradigma reconocido por toda la comunidad científica: la modularidad. No obstante, no siempre fue así. Antes de alcanzar el consenso sobre la modularidad, la neurociencia se debatió durante siglos entre el holismo y el localizacionismo. La modularidad, como veremos, sostiene un cierto localizacionismo, pero moderado y conciliador con el holismo, en tanto que sostiene la localización precisa solo de funciones muy elementales, en lugar de las funciones complejas que Joseph Gall creyó haber localizado a finales del siglo XVIII.

El capítulo cuarto lo finalizaremos contrastando el cognitivismo con el reduccionismo materialista. En ambos casos se trata de enfoques eliminativistas: el primero pretende explicar la mente con independencia del cuerpo, y el segundo explicar la conducta sin prestar atención a la mente. Frente a cualquier tipo de eliminativismo, nuestra postura es la del emergentismo, teoría según la cual la mente emerge del cerebro pero no actúa sobre él, sino que es un mero epifenómeno. Aún a pesar de las dificultades que plantea el estudio de la mente a causa de la mencionada exigencia de sintetizar los enfoques explicativo y comprensivo por la constitución biosocial de la mente, a nuestro juicio es una tarea imprescindible, en tanto que el reduccionismo materialista, como demostraremos mediante argumentos de Hilary Putnam, incurre en el error de presuponer la transitividad de las explicaciones.

Habiendo descrito las características generales de los paradigmas de la mente y del cerebro en los que se basan la IA simbólica y la IA subsimbólica, en el *capítulo quinto* aumentaremos el nivel de concreción. Lo haremos exponiendo un modelo de la inteligencia a nivel cerebral elaborado por un ingeniero informático, Jeff Hawkins, y un modelo cognitivista de la inteligencia a

nivel mental elaborado por Roger Schank, un informático investigador de la IA que en los últimos tiempos ha pasado a dedicarse a la psicología. El modelo cognitivista de la inteligencia de Schank nos servirá para apreciar desde dentro las dificultades insuperables con las que se topa la IA simbólica en su intento por utilizar sistemas formales, que es lo que en el fondo son todas las computadoras electrónicas, para reproducir la versatilidad distintiva de la inteligencia humana. Si el psicologismo epistemológico-lógico, como por ejemplo el de Stuart Mill, contra el que luchaba Husserl, pretendía reducir la lógica a psicología, el cognitivismo pretende justo lo contrario: reducir la psicología a sistemas de lógica formal.

Hacia el final de la exposición del modelo de la inteligencia de Schank trazaremos un hilo conector con el tema del mito del método científico tratado en el capítulo previo. Nuestra tesis es que lo que subyace a la IA simbólica y al mito del método en el sentido antes precisado es la misma pretensión positivista: encontrar algoritmos generadores de teorías, ya sean científicas en el caso del método científico o precientíficas en el caso de la IA simbólica. En ambos casos se pretende automatizar la producción de conocimiento, una empresa imposible que, sin embargo, es signo de nuestro tiempo, y que está en continuidad con los propósitos para los que se crearon las redes de calculadores humanos y, más tarde, las computadoras electrónicas.

En el capítulo quinto expondremos también la teoría de las inteligencias múltiples (teoría IM) de Howard Gardner, la cual, frente a las teorías de la inteligencia de Hawkins y Schank, es la más acertada a nuestro juicio, por dos razones principales. La primera es su caracterización relativista, en el mejor sentido, de la inteligencia como una facultad múltiple, distribuida y contextualizada, a diferencia de los enfoques unitarios, solipsistas y etnocéntricos que predominan en la psicometría, y de los cuales también hablaremos sobre la marcha. La otra razón de nuestro adhesión a la teoría de las inteligencias múltiples de Gardner es que los criterios que propone para identificar una inteligencia abarcan factores muy diversos, desde biológicos y culturales hasta históricos. Semejante amplitud refleja el carácter complejo de la inteligencia, en lugar de eludirlo como hacen aquellos psicólogos que definen la inteligencia en términos operacionales como aquella facultad que se mide con los test de inteligencia.

Por supuesto, en una reflexión filosófica sobre IA y la naturaleza de la inteligencia no podía faltar un examen del test de Turing. A él dedicaremos la última sección del capítulo quinto. Lo expondremos, así como la refutación formulada contra él por John Searle con su famoso argumento de la habitación china, y también las objeciones que se le han planteado al argumento de Searle y que han sido recopiladas por él mismo, para proseguir con las respuestas de

Searle a dichas objeciones, y finalmente exponer nuestras propias objeciones a esas respuestas. Nos sumergiremos, por tanto, en un diálogo de varias capas que culminará con una defensa por nuestra parte del criterio conductista de la inteligencia en el que se basa el test de Turing. Además evaluaremos la importancia que el matemático inglés otorgaba al lenguaje. Turing coincidía con Descartes, y nosotros con ellos dos, en que el lenguaje es una condición necesaria, y en cierto sentido suficiente, de la inteligencia humana.

En el *capítulo sexto* repasaremos la historia de la IA desde su fundación oficial en la conferencia de Dartmouth, celebrada en 1956, hasta el presente. En sus inicios la IA se adscribió al paradigma cognitivista, fundado oficialmente en un simposio del MIT que tuvo lugar en ese mismo año un mes después, y en el que participaron también dos de los asistentes que estuvieron en Dartmouth: Allen Newell y Herbert Simon, los autores de la primera IA. O, más correctamente, el primer intento de IA, ya que el *Logic Theorist*, que así se llamaba el programa informático, no era una verdadera IA. De hecho, como veremos a lo largo del capítulo, en toda la historia de esta disciplina jamás se ha logrado crear una IA en el sentido fuerte que coincide con la noción vulgar de IA. Newell y Simon, junto con otros destacados investigadores como Marvin Minsky, se aferraron desde el principio al programa de investigación de la IA simbólica, que es el de la duplicación de la mente entendida en términos cognitivistas, al tiempo que empleaban su poder para hundir académicamente a aquellos que, como Frank Rosenblatt, se atrevían a incursionar en el enfoque de la IA subsimbólica, que es, recordemos, el de la duplicación de las redes de neuronas o, por lo menos, la simulación de su funcionamiento. La situación cambió hacia los años 80, y desde entonces ambos puntos de vista, simbólico y subsimbólico, coexisten e incluso se complementan. No obstante, ninguno de ellos ha conseguido construir una máquina tan inteligente como un ser humano. Cada programa de investigación presenta sus propios problemas, y en este capítulo los examinaremos. El principal de la IA simbólica es el de la limitación de dominio, es decir, la falta de versatilidad, que es, como venimos señalando, una de las dos características distintivas de nuestro intelecto. La otra, el lenguaje natural, tampoco ha podido ser duplicada.

Aprovechando el hilo conductor de la historia de la IA, expondremos también algunas de las estrategias, técnicas y arquitecturas más relevantes inventadas por los investigadores en su intento de crear máquinas pensantes. Esta exposición dará lugar a la presentación de alternativas opuestas tales como la consabida de IA simbólica vs subsimbólica, IA humana vs ajena, IA fuerte vs débil, IA abstracta vs situada, y métodos fuertes vs débiles. Se trata de conceptos que, aunque han sido acuñados con esos nombres por ingenieros y matemáticos, nos remiten en muchos casos a problemas filosóficos. Por ejemplo,