Introducción

¿Es posible construir máquinas inteligentes? ¿Es el cerebro una máquina? Estas son dos preguntas que han obsesionado a grandes pensadores durante siglos. El desarrollo de la inteligencia artificial (IA) ha acercado ambas cuestiones e incluso para muchos investigadores las ha unificado en el sentido de que se están usando los mismos conceptos, técnicas y experimentos en los intentos de diseñar máquinas inteligentes y en investigar la naturaleza de la mente. Actualmente, todavía sabemos poco acerca del cerebro; sin embargo, estamos siguiendo un camino que pasa por considerarlo un sistema computacional y hemos empezado a explorar el espacio de posibles modelos computacionales que permitan emular su funcionamiento.

El objetivo último de la IA —lograr que una máquina tenga una inteligencia de tipo *general* similar a la humana— es de los más ambiciosos que se ha planteado la ciencia. Por su dificultad, es comparable a otros grandes objetivos científicos como explicar el origen de la vida, el origen del universo o conocer la estructura de la materia. A lo largo de los últimos siglos, este afán por construir máquinas inteligentes nos ha conducido a inventar modelos o metáforas del cerebro humano. Por ejemplo, en el siglo XVII, Descartes se preguntó si un complejo sistema mecánico compuesto de engranajes, poleas

y tubos podría, en principio, emular el pensamiento. Dos siglos después, la metáfora fueron los sistemas telefónicos, ya que parecía que sus conexiones se podían asimilar a una red de neuronas. Actualmente, el modelo dominante es el computacional, basado en el ordenador digital y, por consiguiente, es el modelo que se contempla en este libro.

Unas consideraciones sobre el concepto de 'inteligencia'

La inteligencia no es una característica exclusiva de los humanos. En la naturaleza existen muchos animales que exhiben un comportamiento inteligente, en el sentido de que planifican, son capaces de prever las consecuencias de sus acciones y emplean útiles o herramientas para conseguir sus propósitos. Algunos animales tienen también capacidades, aunque muy limitadas, para procesar el lenguaje. En definitiva, hay una larga lista de observaciones científicas que avalan manifestaciones de inteligencia en chimpancés, delfines, elefantes y otros animales. Por este motivo es más correcto hablar de *inteligencias* que de *inteligencia*, y no sería absurdo pensar que la IA pueda llegar a constituir un nuevo tipo de inteligencia, aunque, como veremos, distinta de la de animales y humanos.

Centrándonos en la inteligencia humana, que es el referente principal en IA, en este capítulo introducimos brevemente los modelos computacionales más importantes, empezando por la distinción entre *IA débil* e *IA fuerte*, dos visiones que se corresponden, respectivamente, con los dos siguientes intentos de definición:

1. La IA es la ciencia e ingeniería que permite diseñar y programar ordenadores de forma que realicen tareas que requieren inteligencia.
2. La IA es la ciencia e ingeniería que permitirá replicar la inteligencia humana mediante máquinas.

La hipótesis del sistema de símbolos físicos y la inteligencia artificial: IA débil frente a IA fuerte

En una ponencia, con motivo de la recepción del prestigioso Premio Turing en 1975, Allen Newell y Herbert Simon formularon la hipótesis del sistema de símbolos físicos (SSF), según la cual "todo sistema de símbolos físicos posee los medios necesarios y suficientes para llevar a cabo acciones inteligentes". Por otra parte, dado que los seres humanos somos capaces de mostrar conductas inteligentes en el sentido general, entonces, de acuerdo con la hipótesis, nosotros también somos sistemas de símbolos físicos. Conviene aclarar a lo que se refieren Newell y Simon: un SSF consiste en un conjunto de entidades denominadas símbolos que, mediante relaciones, pueden combinarse formando estructuras más grandes —como los átomos que forman moléculas— y que pueden ser transformados aplicando un conjunto de procesos. Estos pueden crear nuevos símbolos, crear y modificar relaciones entre símbolos, almacenarlos, comparar si dos símbolos son iguales o distintos, etc. Estos símbolos son físicos en tanto que tienen un sustrato físico-electrónico (en el caso de los ordenadores) o físico-biológico (en el caso de los seres humanos). Efectivamente, en el caso de los ordenadores, los símbolos se realizan mediante circuitos electrónicos digitales, y en el caso de los seres humanos, mediante redes de neuronas. En definitiva, de acuerdo con la hipótesis SSF, la naturaleza del sustrato (circuitos electrónicos o redes de neuronas) carece de importancia siempre y cuando dicho sustrato permita procesar símbolos. No olvidemos que se trata de una hipótesis y por lo tanto no debe ser ni aceptada ni rechazada *a priori*. En cualquier caso, su validez o refutación se deberá verificar, de acuerdo con el método científico, con ensayos experimentales. La IA es precisamente el campo científico dedicado a intentar comprobar esta hipótesis en el contexto de los ordenadores digitales, es decir, si un ordenador convenientemente programado es capaz o no de tener conducta inteligente de tipo general.

Es importante el matiz de que debería tratarse de inteligencia de tipo *general* y no *especializada*, ya que la inteligencia de los seres humanos es de tipo general. Exhibir inteligencia específica es otra cosa bien distinta. Por ejemplo, los programas que juegan al ajedrez a nivel de Gran Maestro son incapaces de jugar a las damas a pesar de ser un juego mucho más sencillo. Se requiere diseñar y ejecutar un programa distinto e independiente del que le permite jugar al ajedrez para que el mismo ordenador juegue también a las damas. En el caso de los seres humanos no es así, ya que cualquier jugador de ajedrez puede aprovechar sus conocimientos sobre este juego para, en cuestión de segundos, jugar a las damas bien. El diseño y realización de inteligencias artificiales que únicamente muestran comportamiento inteligente en un ámbito muy especializado es lo que se conoce por IA débil en contraposición con la IA fuerte, a la que se referían Newell y Simon. Aunque estrictamente la hipótesis SSF se formuló en 1975, ya estaba implícita en las ideas de los pioneros de la IA, e incluso en las ideas de Alan Turing en sus escritos sobre máquinas inteligentes (Turing, 1950).

Quien introdujo esta distinción entre IA débil e IA fuerte fue el filósofo John Searle en un artículo crítico con la IA, publicado en 1980, que provocó —y sigue provocando— mucha polémica (Searle, 1980). La IA fuerte implicaría que un ordenador convenientemente programado no simula una mente, sino que *es una mente* y por consiguiente debería ser capaz de pensar igual que un ser humano. Searle intenta demostrar que la IA fuerte es imposible. Conviene aclarar que no es lo mismo IA general que IA fuerte. Existe obviamente una conexión, pero solamente en un sentido: toda IA fuerte será necesariamente general, pero puede haber IA generales que no sean fuertes, es decir, que simulen la capacidad de exhibir inteligencia general de la mente pero sin ser mentes.

La IA débil consistiría, según Searle, en construir programas que ayudan al ser humano en sus actividades mentales en lugar de duplicarlas. La capacidad de los ordenadores para realizar tareas específicas mejor que las personas ya se

ha demostrado. En ciertos dominios los avances de la IA débil superan en mucho la pericia humana, como, por ejemplo, buscar soluciones a fórmulas lógicas con muchas variables o jugar al ajedrez. También se asocia con la IA débil el hecho de formular y probar hipótesis acerca de aspectos relacionados con la mente (por ejemplo, la capacidad de razonar deductivamente, de aprender inductivamente, etc.) mediante la construcción de programas que llevan a cabo dichas funciones, aunque sea mediante procesos completamente distintos a los que lleva a cabo el cerebro. Todos los avances logrados hasta ahora en el campo de la IA son manifestaciones de la IA débil. A lo largo de este libro hablaremos sobre todo de la IA débil y veremos numerosos ejemplos de dichos avances.

Los principales modelos en inteligencia artificial: simbólico, conexionista, evolutivo y corpóreo

El modelo dominante en IA ha sido el *simbólico*, que tiene sus raíces en la hipótesis SSF. De hecho, sigue siendo muy importante y actualmente se considera el modelo clásico en IA (también denominado GOFAI, de Good Old Fashioned AI). Es un modelo *top-down* que se basa en el razonamiento lógico y la búsqueda heurística como pilares para la resolución de problemas, sin que el sistema inteligente necesite formar parte de un cuerpo ni estar situado en un entorno real. Es decir, la IA simbólica opera con representaciones abstractas del mundo real que se modelan mediante lenguajes de representación basados principalmente en la lógica matemática y sus extensiones. Por este motivo los primeros sistemas inteligentes resolvían principalmente problemas que no requieren interactuar directamente con el entorno como, por ejemplo, demostrar teoremas o jugar al ajedrez (los programas que juegan al ajedrez no necesitan la percepción visual para ver las piezas en el tablero ni actuadores para mover las piezas). Ello no significa que la IA simbólica no pueda ser usada para, por ejemplo, programar el módulo de razonamiento de un

robot físico situado en un entorno real, pero, en los primeros años, los pioneros de la IA no disponían de lenguajes de representación del conocimiento ni de programación que permitieran hacerlo de forma eficiente y, por este motivo, los primeros sistemas inteligentes se limitaron a resolver problemas que no requerían interacción directa con el mundo real. Actualmente, la IA simbólica se sigue usando para demostrar teoremas o jugar al ajedrez, pero también para aplicaciones que requieren percibir el entorno y actuar sobre él como, por ejemplo, el aprendizaje y la toma de decisiones en robots autónomos, tal como veremos más adelante.

Simultáneamente con la IA simbólica también empezó a desarrollarse una IA bioinspirada llamada *conexionista*. Los sistemas conexionistas no son incompatibles con la hipótesis SSF, pero, contrariamente a la IA simbólica, se trata de una modelización *bottom-up*, ya que se basan en la hipótesis de que la inteligencia emerge a partir de la actividad distribuida de un gran número de unidades interconectadas que procesan información paralelamente. En la IA conexionista estas unidades son modelos aproximados de la actividad eléctrica de las neuronas biológicas.

Ya en 1943, McCulloch y Pitts propusieron un modelo simplificado de neurona biológica basándose en la idea de que una neurona es esencialmente una unidad lógica. Este modelo es una abstracción matemática con entradas (dendritas) y salidas (axones). El valor de la salida se calcula en función del resultado de una suma ponderada de las entradas, de forma que si dicha suma supera un umbral preestablecido entonces la salida es un 1, en caso contrario es 0. Conectando la salida de cada neurona con las entradas de otras neuronas se forma una red neuronal artificial. Respecto a lo que ya se sabía sobre el reforzamiento de las sinapsis entre neuronas biológicas, se vio que estas redes neuronales artificiales se podían entrenar para aprender funciones que relacionaran las entradas con las salidas mediante el ajuste de los pesos que sirven para ponderar las conexiones entre neuronas. Por este motivo, se pensó que serían mejores modelos para el aprendizaje, la cognición

y la memoria que los modelos basados en la IA simbólica. Sin embargo, los sistemas inteligentes basados en el conexionismo tampoco necesitan formar parte de un cuerpo ni estar situados en un entorno real y, desde este punto de vista, tienen las mismas limitaciones que los sistemas simbólicos. Por otra parte, las neuronas reales poseen complejas arborizaciones dendríticas con propiedades no solo eléctricas, sino también químicas nada triviales. Pueden contener conductancias iónicas que producen efectos no lineales. Pueden recibir decenas de millares de sinapsis variando en posición, polaridad y magnitud. Además, hoy día sabemos que en el cerebro hay unas células llamadas gliales que regulan el funcionamiento de las neuronas, siendo incluso más numerosas que estas. No existe ningún modelo conexionista que incluya a dichas células, por lo que en el mejor de los casos estos modelos son incompletos. En definitiva, toda la enorme complejidad del cerebro queda muy lejos de los modelos actuales y plantea dudas sobre la utilidad de grandes iniciativas, como el proyecto Human Brain de la Unión Europea. Esta inmensa complejidad del cerebro también conduce a pensar que la llamada *singularidad*, es decir, futuras superinteligencias artificiales que basadas en réplicas artificiales del cerebro superaran con mucho la inteligencia humana en un plazo de unos 20 años, es una predicción con muy poco fundamento.

Otra modelización bioinspirada, también compatible con la hipótesis SSF y no corpórea, es la *computación evolutiva*. Los éxitos de la biología evolucionando organismos complejos hizo que a primeros de los años sesenta algunos investigadores se plantearan la posibilidad de imitar la evolución con el fin de que los programas de ordenador, mediante un proceso evolutivo, mejorasen automáticamente las soluciones a los problemas para los que habían sido programados. La idea es que estos programas, gracias a operadores de mutación y cruce de *cromosomas* que los modelan, crean nuevas generaciones de programas modificados, cuyas soluciones son mejores que las de las generaciones anteriores. Dado que podemos considerar que el objetivo de la IA es la búsqueda

de programas capaces de producir conductas inteligentes, se pensó que se podría usar la programación evolutiva para encontrarlos dentro del espacio de programas posibles. La realidad es mucho más compleja y esta aproximación tiene muchas limitaciones, aunque ha producido excelentes resultados en problemas de optimización y también en la invención de nuevos dispositivos (por ejemplo, un nuevo tipo de antena).

Una de las críticas más fuertes a estos modelos no corpóreos se basa en que un agente inteligente necesita un cuerpo para poder tener experiencias directas con su entorno (diríamos que el agente está *situado* en su entorno) en lugar de que un programador proporcione descripciones abstractas de dicho entorno codificadas mediante un lenguaje de representación del conocimiento. Sin un cuerpo, estas representaciones abstractas no tienen contenido semántico para la máquina. Sin embargo, mediante la interacción directa con el entorno, el agente puede relacionar las señales que percibe mediante sus sensores con representaciones simbólicas generadas a partir de lo percibido. Algunos expertos en IA incluso llegaron a afirmar que no era ni siquiera necesario generar dichas representaciones internas, es decir, que no es necesario que un agente deba tener una representación interna del mundo que le rodea, ya que el propio mundo es el mejor modelo posible de sí mismo, y que la mayor parte de las conductas inteligentes no requieren razonamiento, sino que emergen a partir de la interacción entre el agente y su entorno. Esta idea generó mucha polémica y uno de sus principales valedores, Rodney Brooks, unos años más tarde, admitió que hay muchas situaciones en las que una representación interna del mundo es necesaria para que el agente tome decisiones racionales.

La aproximación corpórea con representación interna ha ido ganando terreno en la IA y actualmente muchos la consideramos imprescindible para avanzar hacia inteligencias de tipo general. De hecho, basamos una gran parte de nuestra inteligencia en nuestra capacidad sensorial y motora. En otras palabras, *el cuerpo da forma a la inteligencia* (*the body shapes*

the way we think) y por lo tanto sin cuerpo no puede haber inteligencia de tipo general. Esto es así porque el *hardware* del cuerpo, en particular los mecanismos del sistema sensor y del sistema motor, determina el tipo de interacciones que un agente puede realizar. A su vez, estas interacciones conforman las habilidades cognitivas de los agentes, dando lugar a lo que se conoce como *cognición situada*. Es decir, se sitúa la máquina en entornos reales, como ocurre con los seres humanos, con el fin de que tengan experiencias interactivas que, eventualmente, les permitan llevar a cabo algo similar a lo que propone la teoría del desarrollo cognitivo de Piaget, según la cual un ser humano sigue un proceso de maduración mental por etapas y quizá los distintos pasos de este proceso podrían servir de guía para diseñar máquinas inteligentes. Estas ideas han dado lugar a una nueva subárea de la IA llamada *robótica del desarrollo* (*developmental robotics*). A lo largo del libro exploraremos con más detalle estos modelos.