26 Fundamentos filosóficos

En este capítulo examinamos lo que significa pensar y si los artefactos podrían y deberían alguna vez llegar a hacerlo.

Como se mencionó en el Capítulo 1, los filosófos existían mucho antes que los computadores y llevaban tiempo intentando solucionar algunas cuestiones relacionadas con la IA: ¿cómo trabaja la mente? ¿Es posible que las máquinas actúen de forma inteligente, igual que las personas? Y si así fuera, ¿tendrían mentes? ¿Cuáles son las implicaciones éticas de las máquinas inteligentes? A lo largo de los primeros 25 capítulos de este libro, hemos estudiado cuestiones de IA en sí misma, pero en este capitulo estudiaremos la agenda del filósofo.

HIPÓTESIS DE LA IA DÉBIL

HIPÓTESIS DE LA IA FUERTE

En primer lugar, observemos la terminología: los filósofos definen la **hipótesis de la IA débil** como la afirmación de que es posible que las máquinas actúen con inteligencia (o quizá mejor, *como si* fueran inteligentes); de la misma manera, la **hipótesis de la IA fuerte** consiste en la afirmación de que las máquinas sí piensan *realmente* (opuesto al pensamiento *simulado*).

La mayoría de los investigadores de IA dan por sentado la hipótesis de la IA débil, y no se preocupan por la hipótesis de la IA fuerte, con tal de que funcione su programa no les interesa si se llama simulación de inteligencia o inteligencia real. Sin embargo, todos deberían preocuparse por las implicaciones éticas de su trabajo.

26.1 IA débil: ¿pueden las máquinas actuar con inteligencia?

Algunos filósofos han intentado demostrar que la IA es imposible; que las máquinas no tendrán la posibilidad de actuar inteligentemente. Algunos han utilizado argumentos que tratan de dar el alto a la investigación en IA:

> La Inteligencia Artificial *abordada desde dentro del culto al computacionalismo* no tendrá ni siquiera un atisbo de fantasma de posibilidad de producir resultados duraderos... Es hora de desviar los esfuerzos de los investigadores en IA, y la gran cantidad de dinero disponible para su soporte, y dirigirse a caminos distintos del enfoque computacional (Sayre, 1993).

Obviamente, si la IA es imposible o no lo es, dependerá de cómo se defina. En esencia, la IA consiste en la búsqueda del mejor programa agente en una arquitectura dada. Con esta formulación, la IA es posible por definición: para cualquier arquitectura digital de k bits de almacenamiento existirán exactamente 2^k programas agente y todo lo que habrá que hacer para encontrar el mejor es enumerarlos y probar todos ellos. Esto podría no ser viable para una k grande, pero los filósofos abordan más la teoría que la práctica.

Nuestra definición de IA funciona bien para el problema de encontrar un buen agente, dependiendo de la arquitectura. Por tanto, nos sentimos tentados a acabar esta sección aquí mismo, respondiendo afirmativamente a la pregunta formulada en el título. Sin embargo, los filósofos están interesados en el problema de comparar dos arquitecturas, la humana y la de la máquina. Además, ellos por tradición han formulado la pregunta de la siguiente manera: **«¿Pueden pensar las máquinas?»** Desgraciadamente, esta cuestión no está bien definida. Para ver por qué, consideremos las dos cuestiones siguientes:

¿PUEDEN PENSAR LAS MÁQUINAS?

- ¿Pueden volar las máquinas?
- ¿Pueden nadar las máquinas?

La mayoría de las personas están de acuerdo en que la respuesta a la primera cuestión es sí, que los aviones pueden volar, pero la respuesta a la segunda es no; los barcos y los submarinos se mueven por el agua, pero eso no es nadar. Sin embargo, ni las preguntas ni sus respuestas afectan en absoluto a las vidas laborales de los ingenieros aeronáuticos ni navales, ni a las de los usuarios de sus productos. Las respuestas no tienen mucho que ver con el diseño o con las características de los aviones o de los submarinos, y sin embargo sí tienen que ver mucho más con la forma en que se han elegido utilizar las palabras. La palabra nadar («swim» en inglés) ha llegado a tener el significado de «moverse por el agua mediante el movimiento de las partes del cuerpo», mientras que la palabra «fly» (volar) no tiene dicha limitación en un medio de locomoción[1]. La posibilidad práctica de las «máquinas pensantes» lleva viviendo con nosotros durante sólo 50 años o así, tiempo insuficiente para que los angloparlantes se decidan a dar un significado a la palabra «pensar».

Alan Turing, en su famoso artículo «Computing Machinery and Intelligence» (Turing, 1950), sugirió que en vez de preguntar si las máquinas pueden pensar, deberíamos preguntar si las máquinas pueden aprobar un test de inteligencia conductiva (de comportamiento), conocido como el Test de Turing. La prueba se realiza para que el programa mantenga una conversación durante cinco minutos (mediante mensajes escritos en línea, *online*) con un interrogador (interlocutor). Éste tiene que averiguar si la conversación se está llevando a cabo con un programa o con una persona; si el programa engaña al interlocutor un 30 por ciento del tiempo, este pasará la prueba. Turing conjeturó que, hacia

[1] En ruso, el equivalente de «nadar» *sí* se aplica a los barcos.

el año 2000, un computador con un almacenamiento de 10^9 unidades podría llegar a programarse lo suficientemente bien como para pasar esta prueba, pero no estaba en lo cierto. Algunas personas *han* sido engañadas durante cinco minutos; por ejemplo, el programa ELIZA y el chatbot en Internet llamado MGONZ han engañado a personas ignorantes que no se daban cuenta de que estaban hablando con un programa; el programa ALICE engañó a un juez en la competición del Loebner Prize en el año 2001. Sin embargo, ningún programa se ha acercado al criterio del 30 por ciento frente a jueces con conocimiento, y el campo en su conjunto de la IA no ha prestado mucha atención a los tests de Turing.

Turing también examinó una gran gama de posibles objeciones ante la posibilidad de las máquinas inteligentes, incluyendo virtualmente aquellas que han aparecido medio siglo después de que apareciera este artículo. Examinaremos algunas de ellas.

El argumento de incapacidad

El «argumento de incapacidad» afirma que «una máquina nunca puede hacer *X*». Como ejemplos de *X*, Turing enumera las siguientes acciones:

> Ser amable, tener recursos, ser guapo, simpático, tener iniciativas, tener sentido del humor, distinguir lo correcto de lo erróneo, cometer errores, enamorarse, disfrutar con las fresas con nata, hacer que otra persona también se enamore, aprender de la experiencia, utilizar palabras de forma adecuada, ser el tema de su propio pensamiento, tener tanta diversidad de comportamientos como el hombre, hacer algo realmente nuevo.

Turing tuvo que utilizar su intuición para adivinar aquello que en un futuro sería posible, pero nosotros tenemos el privilegio de poder mirar hacia atrás y ver qué es lo que ya pueden hacer los computadores. Es innegable que los computadores actualmente hacen muchas cosas que anteriormente eran sólo del dominio humano. Los programas juegan a la ajedrez, a las damas y a otros juegos, inspeccionan piezas de las líneas de producción, comprueban la ortografía en los documentos de los procesadores de texto, conducen coches y helicópteros, diagnostican enfermedades, y hacen otros cientos de tareas tan bien o mejor que los hombres. Los computadores han hecho pequeños pero significativos descubrimientos, en Astronomía, Matemáticas, Química, Mineralogía, Biología, Informática y otros campos que necesitan rendimiento a nivel de experto.

Debido a lo que conocemos actualmente acerca de los computadores, no es sorprendente que sean también buenas en problemas combinatorios tales como los del juego del ajedrez. Sin embargo, los algoritmos también funcionan a nivel humano en tareas que aparentemente se relacionan con el juicio humano, o como apunta Turing, «aprender a partir de la experiencia» y la capacidad de «distinguir lo que es correcto de lo incorrecto». Ya en el año 1955, Paul Meehl (*véase* también Grove y Meehl, 1996) estudió los procesos de la toma de decisiones de expertos formados en tareas subjetivas como predecir el éxito de un alumno en un programa de formación, o la reincidencia de un delincuente. De 20 estudios que Meehl examinó, en 19 de ellos encontró que sencillos algoritmos de aprendizaje estadístico (tal como la regresión lineal y Bayes simple) predicen mejor que los expertos. Desde el año 1999, el Educational Testing Service (Servicio de Exámenes Educativo) ha utilizado un programa automatizado para calificar millones de

preguntas de redacciones en el examen GMAT. Este programa concuerda con los examinadores en un 97 por ciento, aproximadamente al mismo nivel de concordancia entre dos personas (Burstein *et al.,* 2001).

Es evidente que los computadores pueden hacer muchas cosas tan bien o mejor que el ser humano, incluso cosas que las personas creen que requieren mucha intuición y entendimiento humano. Por supuesto, esto no significa que los computadores utilicen la intuición y el entendimiento para realizar estas tareas, las cuales no forman parte del *comportamiento*, y afrontamos dichas cuestiones en otro sitio, sino que la cuestión es que la primera conjetura sobre los procesos mentales que se requieren para producir un comportamiento dado suele ser equivocada. También es cierto, desde luego, que existen todavía muchas tareas en donde los computadores no sobresalen (por no decirlo más bruscamente), incluida la tarea de Turing de mantener una conversación abierta.

La objeción matemática

Es bien conocido, a través de los trabajos de Turing (1936) y Gödel (1931), que ciertas cuestiones matemáticas, en principio, no pueden ser respondidas por sistemas formales concretos. El teorema de la incompletitud de Gödel (*véase* el Apartado 9.5) es el ejemplo más conocido en este respecto. En resumen, para cualquier sistema axiomático formal F lo suficientemente potente como para hacer aritmética, es posible construir una «sentencia Gödel» $G(F)$ con las propiedades siguientes:

- G(F) es una sentencia de F, pero no se puede probar dentro de F.
- Si F es consistente, entonces $G(F)$ es verdadero.

Filósofos como J. R. Lucas (1961) han afirmado que este teorema demuestra que las máquinas son mentalmente inferiores a los hombres, porque las máquinas son sistemas formales limitados por el teorema de la incompletitud, es decir no pueden establecer la verdad de su propia sentencia Gödel, mientras que los hombres no tienen dicha limitación. Esta afirmación ha provocado mucha controversia durante décadas, generando muchos libros entre los que se incluyen dos libros del matemático Sir Roger Penrose (1989, 1994) quien repite esta afirmación con nuevos giros (como por ejemplo, la hipótesis de que los hombres son diferentes porque sus cerebros operan por la gravedad cuántica). Examinemos solamente tres de los problemas de esta afirmación.

En primer lugar, el teorema de la incompletitud de Gödel se aplica sólo a sistemas formales que son lo suficientemente potentes como para realizar aritmética. Aquí se incluyen las máquinas Turing, y la afirmación de Lucas en parte se basa en la afirmación de que los computadores son máquinas de Turing. Esta es una buena aproximación, pero no es del todo verdadera. Aunque los computadores son finitos, las máquinas de Turing son infinitas, y cualquier computador por tanto se puede describir como un sistema (muy grande) en la lógica proposicional, la cual no está sujeta al teorema de incompletitud de Gödel.

En segundo lugar, un agente no debería avergonzarse de no poder establecer la verdad de una sentencia aunque otros agentes sí puedan. Consideremos la sentencia siguiente

> J. R. Lucas no puede consecuentemente afirmar que esta sentencia es verdadera.

Si Lucas afirmara esta sentencia, entonces se estaría contradiciendo a sí mismo, por tanto Lucas no puede afirmarla consistentemente, y de aquí que esta sentencia sea verdadera. (La sentencia no puede ser falsa, porque si lo fuera Lucas entonces no podría afirmarla consecuentemente, por tanto sería verdadera.) Así pues, hemos demostrado que existe una sentencia que Lucas no puede afirmar consecuentemente mientras que otras personas (y máquinas) sí pueden. Sin embargo, esto no hace que cambiemos de idea respecto a Lucas. Por dar otro ejemplo, ninguna persona podría calcular la suma de 10 billones de números de 10 dígitos en su vida, en cambio un computador podría hacerlo en segundos. Sin embargo, no vemos esto como una limitación fundamental en la habilidad de pensar del hombre. Durante miles de años los hombres se han comportado de forma inteligente antes de que se inventaran las máquinas, de manera que no es improbable que el razonamiento matemático no tenga más que una función secundaria en lo que implica ser inteligente.

En tercer lugar, y de manera mucho más importante, aunque reconozcamos que los computadores tienen limitaciones sobre lo que pueden demostrar, no existen evidencias de que los hombres sean inmunes ante esas limitaciones. Es realmente sencillo demostrar con rigor que un sistema formal no puede hacer *X*, y afirmar entonces que los hombres *pueden* hacer *X* utilizando sus propios métodos informales, sin dar ninguna evidencia de esta afirmación. En efecto, es imposible demostrar que los hombres no están sujetos al teorema de incompletitud de Gödel, porque cualquier prueba rigurosa contendría una formalización del talento humano declarado como no formalizable. De manera que nos quedamos con el llamamiento a la intuición de que los hombres, de alguna forma, pueden realizar hazañas superhumanas de comprensión matemática. Esta atracción se expresa con argumentos como «debemos asumir nuestra propia consistencia, si el pensamiento puede ser posible» (Lucas, 1976). Sin embargo ciertamente se sabe que los hombres son inconsistentes. Esto es absolutamente verdadero para el razonamiento diario, pero también es verdadero para un pensamiento matemático cuidadoso. Un ejemplo muy conocido es el problema del mapa de cuatro colores. En 1879, Alfred Kempe publicó una prueba que tuvo una gran acogida y contribuyó a que le eligieran Fellow de Royal Society. Sin embargo, en 1890, Percy Heawood apuntó que existía un error y el teorema quedó sin demostrar hasta el año 1977.

El argumento de la informalidad

Una de las críticas más persistentes e influyentes de la IA como empresa la realizó Turing mediante su «argumento de la informalidad del comportamiento». En esencia, esta afirmación consiste en que el comportamiento humano es demasiado complejo para poder captarse mediante un simple juego de reglas y que debido a que los computadores no pueden nada más que seguir un conjunto (juego) de reglas, no pueden generar un comportamiento tan inteligente como el de los hombres. En IA la incapacidad de capturarlo todo en un conjunto de reglas lógicas se denomina **problema de cualificación** (*véase* Capítulo10).

El filósofo que ha propuesto principalmente este punto de vista ha sido Hubert Dreyfus, quien elaboró una serie de críticas influyentes a la Inteligencia Artificial: *¿Qué*

es lo que no pueden hacer los computadores? (1972), *¿Qué es lo no pueden hacer todavía los computadores?* (1992). Junto con su hermano elaboró también *Mind Over Machine* (1986).

La postura que critican se vino a llamar «Good Old-Fashioned AI» (IA muy anticuada), GOFAI, término que empezó a utilizar Haugeland (1985). Se supone que este término afirma que todo comportamiento inteligente puede ser capturado por un sistema que razona lógicamente a partir de un conjunto de hechos y reglas, los cuales describen el dominio. Por tanto, se corresponde con el agente lógico más simple que se describió en el Capítulo7. Dreyfus está en lo cierto cuando dice que los agentes son vulnerables al problema de la cualificación. Como se vio en el Capítulo 13, los sistemas de razonamiento probabilístico son más adecuados para dominios abiertos. La crítica de Dreyfus por lo tanto no va en contra de los computadores *per se*, sino en contra de una forma en particular de programarlos. Sin embargo, sería razonable suponer que un libro llamado *Lo que no pueden hacer los sistemas lógicos de primer orden basados en reglas sin aprender* podría haber tenido menos impacto.

Bajo el punto de vista de Dreyfus, la pericia del hombre incluye el conocimiento de algunas reglas, pero solamente como un «contexto holístico» o «conocimiento base» *(background)* dentro del que operan los hombres. Proporciona como ejemplo el comportamiento social adecuado al dar o recibir regalos: «Normalmente se responde simplemente en las circunstancias adecuadas y dando el regalo adecuado». Al parecer hay que «tener un sentido directo de cómo hay que hacer las cosas y qué esperar». Esta misma afirmación se realiza en el contexto del juego del ajedrez: «Un maestro de ajedrez tendría que averiguar simplemente qué hacer, pero un buen maestro simplemente observa el tablero como exigiendo un cierto movimiento... y obtiene la respuesta apropiada rápidamente en su cabeza». Es cierto que gran parte de los procesos del pensamiento de una persona que da un regalo o de un gran maestro en ajedrez se llevan a cabo a un nivel que no está abierto a la introspección por la mente consciente. Sin embargo, esto no significa que no existan los procesos de pensamiento. Una cuestión importante que Dreyfus no responde es *cómo* aparece el movimiento de ajedrez adecuado en la cabeza del gran maestro. Esto nos lleva a pensar en un comentario de Dennett (1984):

> Es como si los filósofos se fueran a proclamar expertos en explicar los métodos de los magos en el escenario, y entonces cuando preguntamos cómo hace el truco del serrucho para partir en dos a una mujer, ellos dan la explicación de que es totalmente evidente: el mago no parte en dos a la mujer con la sierra, simplemente parece que lo hace. «Pero, ¿Cómo lo hace?», y los filósofos responden «No es de nuestra incumbencia».

Dreyfus y Dreyfus (1986) proponen un proceso de adquisición de pericia en cinco etapas, comenzando con un procesamiento basado en reglas (del tipo propuesto en GOFAI) y terminando con la habilidad de seleccionar las respuestas correctas instantáneamente. Al realizar esta propuesta, Dreyfus y Dreyfus pasan en efecto de ser críticos a la IA a ser teóricos de IA, ya que proponen una arquitectura de redes neurales (neuronales) organizadas en una biblioteca de casos extensa, pero señalan algunos problemas. Afortunadamente, se han abordado todos sus problemas, algunos con éxito parcial y otros con éxito total. Entre estos problemas se incluyen los siguientes:

1. No se puede lograr una generalización buena de ejemplos sin un conocimiento básico. Afirman que no se sabe cómo incorporar el conocimiento básico en el proceso de aprendizaje de las redes neuronales. De hecho, en el Capítulo 19 vimos que existen técnicas para utilizar el conocimiento anterior en los algoritmos de aprendizaje. Sin embargo, esas técnicas dependen de la disponibilidad previa de conocimiento de forma explícita en los algoritmos de aprendizaje, algo que Dreyfus y Dreyfus niegan vigorosamente. Bajo nuestro punto de vista, esta es una buena razón para realizar un rediseño serio de los modelos actuales del procesamiento neuronal de forma que *puedan* sacar provecho del conocimiento aprendido anteriormente como lo hacen otros algoritmos de aprendizaje.
2. El aprendizaje de redes neuronales es una forma de aprendizaje supervisado (*véase* Capítulo18), que requiere la identificación anterior de las entradas relevantes y las salidas correctas. Por tanto, afirman que no puede funcionar autónomamente sin la ayuda de un entrenador humano. De hecho, el aprendizaje sin un profesor se puede conseguir mediante un **aprendizaje no supervisado** (Capítulo 20) y un **aprendizaje de refuerzo** (Capítulo 21).
3. Los algoritmos de aprendizaje no funcionan bien con muchas funciones, si seleccionamos un subgrupo de éstas, «no existe una forma conocida de añadir funciones nuevas, si el conjunto actual demuestra ser inadecuado para tener en cuenta los hechos aprendidos». De hecho, métodos nuevos tales como las máquinas vectoriales de soporte utilizan muy bien conjuntos grandes de funciones. Como vimos en el Capítulo 19, también existen formas importantes de generar funciones nuevas, aunque requiera más trabajo.
4. El cerebro es capaz de dirigir sus sensores para buscar la información relevante y procesarla para extraer aspectos relevantes para la situación actual. Sin embargo, afirman que «Actualmente, los detalles de este mecanismo ni se entienden y ni siquiera se hipotetizan para guiar la investigación en la IA». De hecho, el campo de la visión activa, respaldado por la teoría del valor de la información (Capítulo 16) tiene que ver exactamente con el problema de dirigir los sensores, y algunos robots ya han incorporado los resultados teóricos obtenidos.

En resumen, muchos de los temas que ha tratado Dreyfus, el conocimiento del sentido común básico, el problema de la cualificación, la incertidumbre, aprendizaje, formas compiladas de la toma de decisiones, la importancia de considerar agentes situados y no motores de interferencia incorpóreos, por ahora se han incorporado en el diseño estándar de agentes inteligentes. Bajo nuestro punto de vista, esta es una evidencia del progreso de la IA, y no de su imposibilidad.

26.2 IA fuerte: ¿pueden las máquinas pensar de verdad?

Muchos filósofos han afirmado que una máquina que pasa el Test de Turing no quiere decir que esté *realmente* pensando, sería solamente una *simulación* de la acción de