26 Fundamentos filosóficos  
HIPÓTESIS DE LA IA DÉBILHIPÓTESIS DE LA IA FUERTE  
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 2k 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 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 locomoción1. 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

1076 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
1 En ruso, el equivalente de «nadar» sí se aplica a los barcos.  
¿PUEDEN PENSAR LAS MÁQUINAS?  
el año 2000, un computador con un almacenamiento de 109 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 incapacidadEl «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  
FUNDAMENTOS FILOSÓFICOS 1077  
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 matemáticaEs 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 má-  
quinas 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.

1078 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
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 informalidadUna 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 (véaseCapítulo10).  
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: ¿QuéFUNDAMENTOS FILOSÓFICOS 1079  
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:

1080 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
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 autó-  
nomamente 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 má-  
quinas 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

FUNDAMENTOS FILOSÓFICOS 1081  
pensar. De nuevo esta objeción fue prevista por Turing, y cita unas palabras del Profesor Geoffrey Jefferson (1949):  
Hasta que una máquina pueda escribir un soneto o componer un concierto porque sienta los pensamientos y las emociones, y no porque haya una lluvia de símbolos, podría  
reconocer que la máquina iguala al cerebro, es decir, no sólo escribirlo sino que sepa que  
lo ha hecho.  
Esto es lo que Turing llama el argumento de la consciencia, la máquina tiene que ser  
consciente de sus propias acciones y estados mentales. Aunque la consciencia sea un tema  
importante, el punto de vista clave de Jefferson se relaciona realmente con la fenomenología, o el estudio de la experiencia directa, es decir, la máquina tiene que sentir emociones realmente. Otros se centran en la intencionalidad, esto es, en la cuestión de si  
las creencias, deseos y otras representaciones supuestas de la máquina son de verdad algo  
que pertenece al mundo real.  
La respuesta de Turing a esta objeción es interesante. Podría haber presentado  
razones para demostrar que las máquinas pueden de hecho ser conscientes (o tener fenomenología, o tener intenciones). En cambio, Turing mantiene que la cuestión no está  
bien definida al decir, «¿Pueden pensar las máquinas?» Además, por qué deberíamos  
insistir en un estándar más alto para las máquinas que el usado para los humanos. Después de todo, en la vida ordinaria no tenemos nunca una evidencia directa sobre los  
estados mentales internos de otras personas. No obstante, Turing dice que «En vez de  
argumentar constantemente sobre este punto de vista, es usual mantener la convencióneducada de que todos pensamos».  
Turing argumenta que Jefferson estaría dispuesto a ampliar la convención educada  
a las maquinas si al menos tuviera experiencia con las que actúan de forma inteligente,  
y cita el diálogo siguiente el cual ha llegado a formar una parte muy significativa de la  
tradición oral de la IA y que debemos incluir:  
HOMBRE: En la primera línea de tu soneto que dice «te compararé con un día de verano»,  
¿no sería mejor decir «un día de primavera»?  
MÁQUINA: No rimaría.  
HOMBRE: Y qué tal un «día de invierno». Este sí qué rimaría bien.  
MÁQUINA: Sí, pero nadie quiere una comparación con un día de invierno.  
HOMBRE: ¿Podrías decir que Mr. Pickwick te recuerda a las navidades?  
MÁQUINA: En cierto modo, sí.  
HOMBRE: Sin embargo, las navidades representan un día de invierno, y no creo que a Mr.  
Pickwick le importara la comparación.  
MÁQUINA: No, creo que estés hablando en serio. Por un día de invierno se entiende un tí-  
pico día de invierno, y no un día especial de Navidad.  
Turing reconoce que la cuestión de la conciencia (consciencia) es difícil, pero niega que  
sea relevante para la práctica de la IA: «No quiero dar la impresión de que pienso que  
no hay misterio en torno a la conciencia… Sin embargo no creo que estos misterios tengan necesariamente que resolverse antes de la respuesta a la cuestión que estamos tratando en este trabajo». Coincidimos con Turing en que nos interesa crear programas que  
se comporten de forma inteligente y no en si alguien los declara reales o simulados. Por  
otro lado, muchos filósofos están especialmente interesados en esta cuestión. Como

1082 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
CONVENCIÓN EDUCADA  
ayuda para entenderlo tendremos en cuenta la cuestión de si otros artefactos se consideran reales.  
En 1848, Frederick Wöhler sintetizó urea artificial por primera vez. Este fue un hecho  
importante porque probó que la química orgánica y la inorgánica se podían unir, cuestión discutida muy fuertemente. Una vez que se consiguió la síntesis, los químicos reconocieron que la urea artificial era urea, porque tenía todas las propiedades físicas  
adecuadas. Igualmente, los edulcorantes artificiales son innegablemente edulcorantes,  
y la inseminación artificial (la otra IA) es innegablemente inseminación. Por otro lado,  
las flores artificiales no son flores, y Daniel Dennett señala que el vino artificial Chateau Latour no sería vino Chateau Latour, aunque no se pudiera distinguir químicamente,  
simplemente porque no se fabricó en el lugar adecuado ni de la forma adecuada. Ni  
tampoco un Picasso artificial sería un Picasso, independientemente del aspecto que tenga el cuadro.  
Podemos concluir diciendo que en algunos casos el comportamiento de un artefacto es importante, aunque en otros sea el pedigrí del artefacto lo que importa. Lo importante en cada caso parece ser una cuestión de convención. Sin embargo para las mentes  
artificiales, no existe una convención, y tenemos que depender de las intuiciones. El  
filósofo John Searle (1980) tiene una convención muy fuerte:  
Nadie piensa que la simulación por computador de una tormenta nos va a mojar… Y, ¿cual  
es la razón de que cualquier persona en su sano juicio suponga que la simulación por computador de los procesos mentales tendrían realmente procesos mentales? (pp. 37-38.)  
Aunque sea fácil reconocer que las simulaciones por computador de las tormentas no  
nos van a mojar, no está claro cómo aplicar esta analogía a las simulaciones por computador de los procesos mentales. Después de todo, la simulación de una tormenta en  
Hollywood utilizando aspersores y máquinas de viento sí que moja a los actores. La mayoría de las personas se sienten cómodas diciendo que la simulación por computador de  
la suma es la suma, y la de un juego de ajedrez un juego de ajedrez. ¿Son los procesos  
mentales parecidos a las tormentas, o son más parecidos a la suma o al ajedrez, o se parecen al Chateau Latour y al Picasso..., o a la urea? Todo depende de la teoría de los estados y los procesos mentales.  
La teoría del funcionalismo dice que un estado mental es cualquier condición causal  
inmediata entre la entrada y la salida. Bajo la teoría funcionalista, dos sistemas con  
procesos causales isomórficos tendrían los mismos estados mentales. Por tanto, un  
programa informático podría tener los mismos estados mentales que una persona. Desde  
luego, todavía no hemos dicho lo que significa realmente «isomórficos», pero la suposición es que existe algún nivel de abstracción por debajo del cual no importa una  
implementación específica; siempre que los procesos sean isomórficos hasta este nivel,  
tendrán lugar los mismos estados mentales.  
En contraste, la teoría del naturalismo biológico dice que los estados mentales son  
características emergentes de alto nivel originadas por procesos neurológicos de bajo  
nivel en las neuronas, y lo que importa son las propiedades (no especificadas) de las  
neuronas. Así pues, los estados mentales no se pueden duplicar justo en la base de algún  
programa que tiene la misma estructura funcional con el mismo comportamiento de entrada y salida; necesitaríamos que el programa se ejecutara en una arquitectura con la

FUNDAMENTOS FILOSÓFICOS 1083FUNCIONALISMO NATURALISMO BIOLÓGICO  
misma potencia causal que las neuronas. La teoría no dice por qué las neuronas tienen  
esta potencia causal, ni tampoco qué otras instanciaciones físicas podrían tenerla o no.  
Para investigar estos dos puntos de vista examinaremos uno de los problemas más  
antiguos de la filosofía de la mente, y retomaremos tres experimentos pensados.  
El problema de mente-cuerpoEl problema mente-cuerpo cuestiona cómo se relacionan los estados y los procesos mentales con los estados y los procesos (específicamente del cerebro) del cuerpo. Como si  
no fuera bastante difícil, vamos a generalizar el problema como un problema de «arquitectura-mente», para que nos permita hablar sobre la posibilidad de que las máquinas tengan mentes.  
¿Por qué es un problema el problema mente-cuerpo? Para la primera dificultad nos  
remontaremos a René Descartes, quien abordó el tema de cómo un alma inmortal interactúa con un cuerpo mortal, y concluyó diciendo que el alma y el cuerpo son dos tipos  
de cosas diferentes, una teoría dualista. La teoría monista, frecuentemente llamada materialismo, mantiene que no existen cosas tales como almas inmateriales, sino sólo objetos materiales. Como consecuencia, los estados mentales, tales como sentir dolor,  
saber que alguien está montando a caballo, o creer que la capital de Austria es Viena,  
son estados del cerebro. John Searle, de forma concisa, resume esta idea con el slogan,  
«Los cerebros producen las mentes».  
El materialista se debe enfrentar por lo menos con dos obstáculos serios. El primer  
problema es el de la libertad de elección: ¿Cómo puede ser que una mente puramente  
física, cuyas transformaciones están regidas por las leyes de la física, conserve todavía  
el libre albedrío? La mayoría de los filósofos consideran que este problema necesita una  
reconstitución cuidadosa de nuestra noción ingenua de libertad de elección, en vez de  
presentar una amenaza para el materialismo. El segundo problema tiene que ver con el  
tema general de la conciencia (y cuestiones de entendimiento y de autoconocimientorelacionadas, aunque no idénticas). Para simplificar, por qué se siente algo cuando se tienen ciertos estados cerebrales, mientras que probablemente no se siente nada al tener otros  
estados físicos (por ejemplo, al ser una roca).  
Para empezar a responder estas cuestiones, necesitamos formas de hablar sobre los  
estados del cerebro a niveles más abstractos que las configuraciones específicas de todos los átomos del cerebro de una persona en particular en un momento concreto. Por  
ejemplo, cuando pienso en la capital de Austria, mi cerebro sufre miríadas de cambios  
diminutos de un picosegundo al otro, pero estos no constituyen un cambio cualitativoen el estado del cerebro. Para tenerlo en cuenta, necesitamos una noción de tipos de estados del cerebro bajo la cual podamos juzgar si dos estados del cerebro pertenecen al  
mismo tipo o a un tipo diferente. Existen diferentes opiniones sobre lo que quiere decir  
tipo en este caso. Casi todo el mundo cree que si tomamos un cerebro y sustituimos los  
átomos de carbono por un nuevo conjunto de átomos de carbono2, el estado mental no

1084 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
2 Quizás incluso los átomos de un isótopo de carbón diferente, como se hace algunas veces en los experimentos de exploración cerebral.  
PROBLEMA MENTECUERPO DUALISTA MONISTA MATERIALISMO LIBERTAD DE ELECCIÓN CONCIENCIA  
se verá afectado. Esto es bueno porque los cerebros reales están sustituyendo continuamente sus átomos mediante procesos metabólicos, y sin embargo esto en sí no parece  
causar perturbaciones mentales importantes.  
Ahora vamos a examinar una clase en particular de estado mental: las actitudesproposicionales (tratadas por primera vez en el Capítulo 10), conocidas también como  
estados intencionales. Estos son estados tales como creer, conocer, desear, temer, y otros  
más que se relacionan con algunos aspectos del mundo exterior. Por ejemplo, la creencia de que Viena es la capital de Austria es una creencia sobre una ciudad en particular  
y su estado. Nos vamos a cuestionar si es posible que los computadores tengan estados  
intencionales que nos ayuden a entender cómo caracterizar dichos estados. Por ejemplo,  
se podría decir que el estado mental de desear una hamburguesa difiere del estado de  
desear una pizza porque en el mundo real una hamburguesa y una pizza son cosas diferentes. Es decir, los estados intencionales tienen una conexión necesaria con otros  
objetos del mundo externo. Por otro lado, anteriormente, hemos argumentado que los  
estados mentales son estados del cerebro, y de aquí que los estados mentales de identidad o no-identidad se deberían determinar permaneciendo completamente «dentro de la  
cabeza», sin hacer referencia al mundo real. Para examinar este dilema retomaremos el  
experimento del pensamiento que intenta separar a los estados intencionales de sus objetos externos.  
El experimento del «cerebro en una cubeta»Imagínese que al nacer le extraen el cerebro de su cuerpo y lo colocan en una cubeta con  
una ingeniería maravillosa. Esta cubeta mantiene su cerebro y le permite crecer y  
desarrollarse. Al mismo tiempo, su cerebro recibe unas señales electrónicas de un  
simulador informático que pertenece a un mundo totalmente ficticio, y las señales motoras de su cerebro son interceptadas y utilizadas para modificar la simulación cuando  
sea adecuado3. A continuación, el estado del cerebro podría tener el estado mental MueroPor (Yo, Hamburguesa) aunque no tenga un cuerpo con el que sentir hambre ni sentido del gusto para experimentarlo, y puede que tampoco haya hamburguesas en el mundo  
real. En ese caso, ¿sería el mismo estado mental que el del cerebro en un cuerpo?  
Una forma de resolver el dilema es decir que el contenido de los estados mentales  
puede ser interpretado desde dos puntos de vista diferentes. La visión del «contenidoextenso» interpreta el dilema desde el punto de vista de un observador omnisciente desde  
fuera con acceso a la situación completa, y que puede distinguir las diferencias del mundo. De esta manera, bajo el contenido extenso, las ideas del cerebro en una cubeta son  
diferentes de las de una persona «normal». El contenido estrecho sólo tiene en cuenta  
el punto de vista subjetivo interno, y bajo este punto de vista todas las creencias serían  
las mismas.  
El pensamiento de que una hamburguesa es maravillosa tiene una cierta naturaleza  
intrínseca, esto es que existe algo que es como tener esa creencia. Ahora vamos a entrar  
en la esfera de qualia, o de las experiencias intrínsecas (de la palabra latina que signi

FUNDAMENTOS FILOSÓFICOS 10853 Esta situación puede resultar familiar para los que hayan visto la película de 1999, The Matrix.  
CONTENIDO EXTENSO ESTADOS INTENCIONALES QUALIA CONTENIDO ESTRECHO  
fica aproximadamente «tales cosas»). Suponga que mediante algún accidente del cableado  
retinal o neuronal la persona X experimenta en rojo el color que la persona Y percibe como  
verde, y viceversa. Entonces cuando ambas vean las luces del semáforo actuarán de la  
misma manera, pero la experiencia que tienen será de alguna manera diferente. Ambas  
pueden reconocer que el nombre de su experiencia es «la luz es roja», pero las experiencias se sienten de diferente manera. No está claro si eso significa que son los mismos estados mentales o que son estados mentales diferentes.  
A continuación retomaremos otro experimento del pensamiento que aborda la cuestión  
de si los objetos físicos diferentes a las neuronas humanas pueden tener estados mentales.  
El experimento de la prótesis cerebralEl experimento de la prótesis cerebral fue introducido por Clark Glymour a mediados  
de los años 70, y retocado posteriormente por John Searle (1980), aunque se asocia más  
comúnmente al trabajo de Hans Moravec (1988). Trata de lo siguiente: suponga que la  
neurofisiología ha evolucionado hasta tal punto que el comportamiento y las conexiones de entrada y salida de todas las neuronas del cerebro humano se entienden perfectamente. Además, suponga que podemos construir mecanismos electrónicos  
microscópicos que imitan este comportamiento y que pueden interconectarse con el tejido neuronal. Y finalmente, suponga que una técnica quirúrgica milagrosa puede sustituir las neuronas individuales con los mecanismos electrónicos sin interrumpir el  
funcionamiento del cerebro por completo. El experimento consiste en sustituir gradualmente todas las neuronas de la cabeza de alguien con mecanismos electrónicos y a continuación invertir el proceso para retornar al sujeto a su estado biológico normal.  
Nos preocupa tanto el comportamiento externo como la experiencia interna del sujeto, durante y después de la operación. Por definición del experimento, el comportamiento externo del sujeto no debe sufrir ningún cambio en comparación con lo que se  
observaría si la operación no se llevase a cabo4. Ahora bien, aunque la presencia o ausencia de consciencia no la pueda asegurar fácilmente un tercero, el sujeto del experimento debería por lo menos poder registrar cualquier cambio en su propia experiencia  
consciente. Aparentemente existe una confrontación directa de intuiciones de lo que podría llegar a ocurrir. Moravec, un investigador y funcionalista en robótica, está convencido de que su consciencia no se vería afectada. Searle, un filósofo y naturalista biólogo,  
también está convencido de que la consciencia desaparecería:  
Ante nuestro asombro encontramos que en efecto estamos perdiendo el control del comportamiento externo. Encontramos, por ejemplo, que cuando los doctores comprueban  
nuestra visión, dicen: «Hemos puesto un objeto rojo delante de usted; por favor, díganos  
lo que ve». Desearemos responder gritando «No puedo ver nada. Me he quedado absolutamente ciego». Sin embargo, hablaremos y diremos sin ningún control por nuestra parte: «Observo que hay un objeto rojo delante de mí»... Nuestra experiencia consciente se  
va reduciendo poco a poco a la nada, mientras que se puede observar externamente que  
nuestro comportamiento es el mismo. (Searle, 1992)

1086 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
4 Imaginemos que utilizamos un sujeto idéntico de «control» que sufre una operación placebo para poder  
comparar los dos comportamientos.  
Sin embargo, podemos hacer algo más que argumentar, a partir de la intuición. En  
primer lugar, hay que destacar que para que el comportamiento externo siga siendo el  
mismo, mientras que el individuo se va sumiendo gradualmente en la inconsciencia, debe  
darse el caso de que la propia voluntad del sujeto se elimine instantáneamente y totalmente; de otra manera, la reducción del conocimiento se vería reflejada en el comportamiento externo, con palabras tales como «Ayuda, mi conocimiento se está reduciendo»,  
u otras palabras con ese mismo efecto. Esta eliminación instantánea de la propia voluntad  
como resultado de la sustitución gradual de neuronas una a una parece ser una afirmación improbable.  
En segundo lugar, examinemos qué ocurre si formulamos al sujeto cuestiones relacionadas con su experiencia consciente durante el período en donde no existen neuronas reales. Por las condiciones del experimento, obtendremos respuestas del tipo, «Me  
encuentro bien. Debo decir que estoy algo sorprendido porque creía en el argumento de  
Searle». O de lo contrario, podríamos atizar al sujeto con un bastón acabado en punta y  
observaríamos la respuesta, «¡Ay, cómo duele!». Ahora bien, durante el curso normal  
de sucesos, el escéptico puede desechar tales salidas de programas de IA tales como simples artimañas. Ciertamente, es bastante fácil utilizar una regla tal como «Si el sensor  
12 identifica “Alto” entonces la salida será “Ay”». Sin embargo lo importante es que, debido a que hemos duplicado las propiedades funcionales de un cerebro humano normal,  
supondremos que el cerebro electrónico no contiene dichas artimañas. Debemos entonces tener una explicación de las manifestaciones de la consciencia producidas por el cerebro electrónico que atañe sólo a las propiedades funcionales de las neuronas. Y además,esta explicación debe aplicarse también al cerebro real, el cual tiene las mismas propiedades funcionales. Parece entonces que existen sólo dos conclusiones posibles:  
1. Los mecanismos causales de la consciencia que generan estas clases de salidas  
en los cerebros normales todavía están funcionando en la versión electrónica,  
que, por tanto, es consciente.  
2. Los sucesos mentales conscientes en el cerebro normal no tienen conexión  
causal con el comportamiento, y quedan fuera del cerebro electrónico, que no  
es, por tanto, consciente.  
Aunque no podamos descartar la segunda posibilidad, esto reduce la consciencia a lo que  
los filósofos llaman rol epifenomenal, algo que ocurre, pero que no deja pistas como si  
lo fuera en el mundo observable. Además, si la consciencia es en efecto epifenomenal, entonces el cerebro debe tener un segundo mecanismo inconsciente responsable del «Ay».  
En tercer lugar, examinemos la situación en que la operación se ha invertido y el sujeto tiene un cerebro normal. Una vez más, el comportamiento externo del sujeto debe  
ser, por definición, como si la operación no hubiera ocurrido. En particular, deberíamos  
poder preguntar, «¿Cómo estaba durante la operación?, ¿recuerda el bastón acabado en  
punta?». El sujeto debe tener recuerdos precisos de la naturaleza real de sus experiencias conscientes, incluyendo las qualia, a pesar del hecho de que según Searle no ha habido  
tales experiencias.  
Searle podría responder que no hemos definido el experimento adecuadamente. Por  
ejemplo, si las neuronas reales quedan suspendidas entre el momento que se extraen y  
el tiempo en que se reemplazan en el cerebro, por supuesto entonces no «recordarán»

FUNDAMENTOS FILOSÓFICOS 1087EPIFENOMENAL  
las experiencias durante la operación. Para tratar esta eventualidad, necesitamos asegurarnos de que el estado de las neuronas se actualiza para reflejar el estado interno de las  
neuronas artificiales que están reemplazando. Si los supuestos aspectos «no funcionales» de las neuronas reales dan como resultado entonces un comportamiento funcionalmente diferente del que se observa con las neuronas artificiales todavía en su lugar,  
tenemos un reductio ad absurdum simple, debido a que eso significaría que las neuronas artificiales no son equivalentes funcionalmente a las neuronas reales. (Véase el  
Ejercicio 26.3 para una posible refutación de este argumento.)  
Patricia Churchland (1986) señala que los argumentos funcionalistas que operan a  
nivel de neurona también pueden funcionar al nivel de cualquier unidad funcional más  
grande, un grupo de neuronas, un módulo mental, un lóbulo, una hemisfera, o todo el  
cerebro. Eso significa que si se acepta la noción de que el experimento de prótesis del  
cerebro muestra que el cerebro de sustitución es consciente, deberíamos también creer  
que la consciencia se mantiene cuando el cerebro entero se sustituye por un circuito que  
hace corresponder la entrada con la salida mediante una tabla de búsqueda. Esto es desconcertante para muchas personas (incluyendo al mismo Turing), que tienen la intuición  
de que las tablas de búsqueda no son conscientes, o por lo menos, que las experiencias  
concientes generadas durante la búsqueda en tabla no son las mismas que las generadas  
durante la operación de un sistema que se podría describir (incluso en un sentido computacional ingenuo) como accediendo y generando ideas, introspecciones, metas, etc.  
Esto podría sugerir que el experimento de la prótesis del cerebro no puede utilizar un  
reemplazo de todo el cerebro en conjunto, si ha de ser efectivo al guiar las intuiciones,  
pero esto no significa que se debe utilizar el reemplazo de un átomo cada vez, tal como  
Searle nos hace creer.  
La habitación chinaNuestro experimento final del pensamiento es quizás el más famoso de todos. Este se  
debe a John Searle (1980), quien describe un sistema hipotético que claramente está ejecutando un programa y que pasa la prueba de Turing, pero que igualmente y de manera clara no entiende (según Searle) ninguna entrada ni salida. Su conclusión es que  
ejecutar el programa adecuado (es decir, tener las salidas adecuadas) no es una condición suficiente para ser una mente.  
El sistema se compone de un hombre, que solamente entiende inglés, y que está equipado con un libro de reglas escrito en inglés y varias pilas de papel, algunas en blanco  
y algunas con inscripciones indescifrables. (El hombre entonces hace el papel de la CPU,  
el libro de normas es el programa y los papeles son el dispositivo de almacenamiento.)  
El sistema se encuentra dentro de una habitación con una apertura al exterior. A través  
de la apertura se van deslizando los papeles con símbolos indescifrables. El hombre encuentra los símbolos correspondientes en el libro de reglas y sigue las instrucciones. Las  
instrucciones pueden incluir escritura de símbolos en los papeles nuevos que van saliendo,  
encontrar símbolos en las pilas de papeles, reorganizar las pilas, etc. Finalmente las instrucciones harán que un símbolo o más sean transcritos a un trozo de papel que se pasa  
otra vez al mundo exterior.

1088 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
Hasta ahora todo está bien. Pero desde fuera, se observa un sistema que está sacando la entrada en forma de sentencias chinas y generando respuestas chinas que obviamente son tan «inteligentes» como las de la conversación imaginada por Turing5.  
Entonces Searle argumenta lo siguiente: la persona que está en la habitación no entiende  
el chino (supuestamente). El libro de reglas y las pilas de papel, que son sólo trozos de  
papel no entienden el chino. Por consiguiente no está habiendo comprensión del chino.  
De aquí que, según dice Searle, ejecutar el programa adecuado no genera necesariamente entendimiento.  
Al igual que Turing, Searle examinó e intentó rechazar una serie de respuestas a este  
argumento. Algunos comentaristas, incluyendo a John McCarthy y a Robert Wilensky,  
propusieron lo que Searle llama la respuesta de los sistemas. La objeción es que aunque  
se pueda preguntar si el hombre de la habitación entiende el chino, esto es como preguntar si la CPU puede admitir raíces cúbicas. En ambos casos, la respuesta es no, y en  
ambos casos según la respuesta del sistema, el sistema completo sí tiene la capacidad  
en cuestión. Ciertamente, si se pregunta a la habitación china si entiende el chino, la respuesta sería afirmativa (en un chino fluido). Según la convención educada de Turing, esto  
debería ser suficiente. La respuesta de Searle es reiterar la cuestión de que el entendimiento no está en el hombre y no puede estar en el papel, de forma que no puede haber  
entendimiento. También sugiere algo más: que nos podríamos imaginar al hombre memorizando el libro de reglas y el contenido de todas las pilas de papel, de manera que  
no hubiera nada que tener que entender excepto al hombre; y una vez más, si se pregunta  
al hombre (en inglés), la respuesta será negativa.  
En este punto hemos topado con los temas reales. El cambio a la memorización desde el papel es una pista falsa, porque ambas formas son simplemente instanciaciones  
físicas de un programa en ejecución. La afirmación real que hace Searle se basa en los  
cuatro axiomas siguientes (Searle, 1990):  
1. Los programas informáticos son entidades sintácticas y formales.  
2. Las mentes tienen contenidos o semánticas mentales.  
3. La sintaxis por sí misma no es suficiente para la semántica.  
4. Los cerebros originan las mentes.  
A partir de los tres primeros axiomas, Searle concluye diciendo que los programas no  
son suficientes para las mentes. En otras palabras, un agente que ejecuta un programa  
podría ser una mente, pero no es necesariamente una mente sólo por virtud de ejecutar programas. A partir de los cuatro axiomas concluye que «Cualquier otro sistema  
capaz de originar mentes tendría que tener poderes causales (por lo menos) equivalentes  
a los de los cerebros». De aquí infiere que cualquier cerebro artificial tendría que duplicar los poderes causales, no sólo ejecutar un programa en particular, y que los  
cerebros humanos no producen fenómenos mentales sólo por virtud de ejecutar un  
programa.

FUNDAMENTOS FILOSÓFICOS 10895 El hecho de que las pilas de papel podrían ser mucho más grandes que todo el planeta y de que la generación de respuestas llevaría millones de años no redunda en la estructura lógica del argumento. Un propósito de la formación filosófica es desarrollar un sentido bien afinado de las objeciones que guardan relación y  
las que no.  
Las conclusiones de que los programas no son suficientes para las mentes sí que proceden de estos axiomas, si somos generosos en su interpretación. Pero la conclusión no  
es satisfactoria: todo lo que Searle ha demostrado es que si explícitamente negamos el  
funcionalismo (eso es lo que hace su axioma (3)) entonces no podemos concluir necesariamente diciendo que los no-cerebros son mentes. Esto es bastante razonable, de manera que el argumento completo se reduce si el axioma (3) puede ser aceptado. Según  
Searle, el punto de vista del argumento de la habitación china es proporcionar intuiciones al axioma (3). Pero la reacción de este argumento muestra que proporciona intuiciones sólo a los que ya se han inclinado a aceptar la idea de que meros programas no  
pueden generar un entendimiento de verdad.  
Reiterando, el objetivo del argumento de la habitación china es refutar la IA fuerte,  
es decir la afirmación de que ejecutar la clase adecuada de programa da como resultado necesariamente una mente. Esto se hace exhibiendo un sistema aparentemente inteligente que ejecuta la clase adecuada de programa que no es, según Searle,  
demostrablemente una mente. En este sentido, Searle apela a la intuición, no a la prueba: sólo hay que observar la habitación; ¿qué hay en ella que sea una mente? Sin embargo, se podría hacer el mismo argumento sobre el cerebro: observe el conjunto de  
células (o de átomos), que funcionan ciegamente según las leyes de la Bioquímica (o la  
Física): ¿Qué hay ahí que sea una mente? ¿Por qué un trozo de cerebro puede ser una  
mente mientras que un trozo de hígado no puede serlo?  
Además, cuando Searle admite que materiales diferentes a las neuronas podrían en  
principio ser una mente, debilita su argumento aún más por dos razones: en primer lugar sólo tenemos las intuiciones de Searle (o las nuestras propias) para decir que la habitación china no es una mente, y en segundo lugar, aunque determinemos que la  
habitación no es una mente, no nos dice nada sobre si un programa que se ejecuta en algún otro medio físico (incluyendo un computador) podría ser una mente.  
Searle permite la posibilidad lógica de que el cerebro esté implementando de verdad un programa de IA de la clase tradicional. Sin embargo, si el mismo programa se  
ejecuta en la clase inadecuada de máquina no sería una mente. Searle no cree que las  
«máquinas no puedan tener mentes», en cambio afirma que algunas máquinas sí que tienen mentes, los hombres son máquinas biológicas con mentes. No nos queda ninguna  
guía para ver qué tipos de máquinas califican o no califican.  
26.3 La ética y los riesgos de desarrollarla Inteligencia ArtificialHasta ahora nos hemos concentrado en si podemos desarrollar la IA, pero debemos también tener en cuenta si deberíamos hacerlo. Si es más probable que los efectos de la tecnología de la IA sean más negativos que positivos, sería responsabilidad moral de los  
trabajadores en su campo redirigir su investigación. Muchas de las nuevas tecnologías  
han tenido efectos negativos no intencionados: el motor de combustión trajo la polución  
ambiental y la pavimentación del paraíso; la fisión nuclear produjo el desastre de

1090 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
Chernobyl, la Isla de las Tres Millas y la amenaza de la destrucción mundial. Todos los  
científicos e ingenieros se enfrentan a consideraciones éticas de cómo deberían actuar  
en el trabajo, qué proyectos deberían o no deberían hacer, cómo los deberían abordar.  
Existe incluso un manual sobre la Ética de los Computadores (Ethics of Computing,Berleur y Brunnstein, 2001). Sin embargo, la IA parece que expone problemas nuevos  
yendo más allá de, por ejemplo, construir puentes que no se desmoronen:  
• Las personas podrían perder sus trabajos por la automatización.  
• Las personas podrían tener demasiado (o muy poco) tiempo de ocio.  
• Las personas podrían perder el sentido de ser únicos.  
• Las personas podrían perder algunos de sus derechos privados.  
• La utilización de los sistemas de IA podría llevar a la pérdida de responsabilidad.  
• El éxito de la IA podría significar el fin de la raza humana.  
Vamos a examinar en orden cada uno de estos temas.  
Las personas podrían perder sus trabajos por la automatización. La economía  
industrial moderna ha llegado a depender en general de los computadores, y selecciona  
programas de IA en particular. Por ejemplo, gran parte de la economía, especialmente  
en Estados Unidos, depende de la disponibilidad de créditos del consumidor. Las aplicaciones de las tarjetas de crédito, aceptaciones de cargos y detección de fraudes se  
realizan mediante programas de IA. Se podría decir que miles de trabajadores han sido  
desplazados por estos programas de IA, pero en efecto si eliminamos estos programas  
de IA estos trabajos no existirían, porque la mano de obra humana añadiría un coste adicional a las transacciones. Hasta ahora, la automatización por medio de la tecnología de  
la IA ha creado más trabajos de los que ha eliminado, y ha creado puestos de trabajo más  
interesantes y mejor pagados. Ahora que el programa IA canónico es un «agente inteligente» diseñado para ayudar a un hombre, la pérdida de trabajo preocupa menos que cuando la IA se centraba en los sistemas expertos diseñados para sustituir a los hombres.  
Las personas podrían tener demasiado (o muy poco) tiempo de ocio. Alvin Toffler  
en su libro Future Shock (1970) escribió, «La semana laboral se ha recortado el 50 por  
ciento a finales del siglo. No es absurdo decir que se volverá a reducir a la mitad hacia el  
año 2000». Arthur C. Clarke (1968b) escribió que las personas del año 2001 «podrían enfrentarse a un futuro de gran aburrimiento, en donde el problema principal de la vida es  
decidir cual de los cientos de canales de la televisión seleccionar». La única predicción  
que ha llegado a resultar cierta es el número de canales de televisión (Springsteen, 1992).  
En cambio, las personas que trabajan en las industrias muy relacionadas con el conocimiento han descubierto que forman parte de un sistema computerizado integrado que funciona 24 horas al día; para mantenerlo se han visto forzados a trabajar durante más horas.  
En una economía industrial, las recompensas son aproximadamente proporcionales al tiempo invertido; trabajar el 10 por ciento más llevaría a producir un incremento del 10 por  
ciento en los ingresos. En una economía de la información marcada por la comunicación  
de un ancho de banda alto y por una reproducción fácil de la propiedad intelectual (lo que  
Frank y Cook (1996) llaman «Winner-Take-All Society» (la Sociedad del ganador se hace  
con todo)), existe una gran recompensa por ser ligeramente mejor que la competencia;  
trabajar un 10 por ciento más podría significar un 100 por 100 de incremento en los ingresos. De forma que todos se sienten más presionados por trabajar más fuerte. La IA

FUNDAMENTOS FILOSÓFICOS 1091  
incrementa el ritmo de la innovación tecnológica y contribuye así a esta tendencia general, pero la IA también mantiene la promesa de permitirnos ahorrar tiempo y permitir que  
nuestros agentes automatizados hagan las cosas por un tiempo.  
Las personas podrían perder su sentido de ser únicas. En el libro de Weizenbaum  
(1976), Computer Power and Human Reason (El poder de los computadores y la razón  
humana), el autor del programa ELIZA, señala algunas de las posibles amenazas que supone la IA para la sociedad. Uno de los argumentos principales de Weizenbaum es que  
la investigación en IA hace posible la idea de que los hombres sean autómatas, una idea  
que produce pérdida de autonomía o incluso de humanidad. Podemos señalar que la idea  
ya existía mucho antes que la IA, nos podemos remontar a L’Homme Machine (La  
Mettrie, 1748). También destacaremos que la humanidad ha sobrevivido a otros contratiempos hacia nuestro sentido de la unicidad: De Revolutionibus Orbium Coelestium(Copernicus, 1543) desplazó a la tierra lejos del centro del sistema solar, y Descent ofMan (Darwin, 1871) puso al homo sapiens al mismo nivel que otras especies. La IA,  
aunque sea una materia de gran éxito, quizá sea por lo menos amenazante para las  
suposiciones morales de la sociedad del siglo XXI al igual que la teoría de la evolución  
lo fue para los del siglo XIX.  
Las personas podrían perder algo de sus derechos privados. Weizenbaum también  
señaló que la tecnología del reconocimiento de voz podría llevar a una intercepción  
extensa de cableados, y de aquí a la pérdida de las libertades civiles. No previó un mundo con amenazas terroristas que cambiarían el equilibrio de la vigilancia que estarían  
dispuestas a aceptar las personas, sino que reconocería correctamente que la IA tiene el  
potencial de producir una vigilancia en grandes cantidades. Esta predicción puede convertirse en realidad: el sistema clasificado Echelon del gobierno americano «consiste en  
una red de envíos de escucha, campos de antenas y estaciones de radar; el sistema está  
respaldado por computadores que utilizan traducción de lenguajes, reconocimiento de  
voz y palabras clave que buscan pasar por la criba automáticamente todo el tráfico de  
llamadas telefónicas, correos electrónicos, faxes y telex»6. Algunas personas reconocen  
que la computerización conduce a la pérdida de privacidad, por ejemplo, el consejero  
delegado de Sun Microsystems, Scott McNealy ha dicho que «De cualquier forma tenemos privacidad cero. Hay que superarlo». Otros no están de acuerdo. Por ejemplo, el  
juez Louis Brandeis en 1890 escribió que «La privacidad es el derecho más completo y  
extenso de todos... el derecho de la personalidad de uno mismo».  
La utilización de sistemas de IA podría llevar a la pérdida de responsabilidad. En  
la atmósfera de litigios que prevalece en Estados Unidos, la obligación legal se convierte  
en un tema importante. Cuando un médico depende del juicio de un sistema médico experto para hacer diagnóstico, ¿quién es el culpable si el diagnóstico es erróneo? Afortunadamente, debido en parte a la mayor influencia en medicina de métodos teóricos para  
las decisiones, se acepta que la negligencia no puede demostrarse si un médico lleva cabo  
procedimientos médicos que tienen una utilidad altamente esperada, incluso si el  
resultado real es catastrófico para el paciente. Por tanto la pregunta debería ser «¿Quién  
tiene la culpa si el diagnóstico no es razonable?». Hasta ahora, los juzgados han man-

1092 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
6 Véase «Eavesdropping in Europe» (Escucha a escondidas en Europa), Wired news, 30/09/1998, e informes citados por la UE.  
tenido que los sistemas médicos expertos desempeñan el mismo papel que los libros de  
texto médicos y los libros de referencias; los médicos son responsables de entender el  
razonamiento que soporta esta decisión y de utilizar su propio juicio a la ahora de decidir si se aceptan las recomendaciones del sistema. Por tanto, en el diseño de sistemas  
médicos expertos como agentes, no se debería considerar que las acciones afectan directamente a los pacientes, sino que influyen directamente en el comportamiento del mé-  
dico. Si los sistemas expertos se hacen más fiables y precisos que los hombres que hacen  
los diagnósticos, los médicos podrían tener obligaciones legales si no utilizan las recomendaciones de un sistema experto. Gawande (2002) explora esta premisa.  
Están empezando a aparecer temas similares en relación con la utilización de agentes inteligentes en Internet. Se han hecho progresos en la incorporación de limitaciones  
en los agentes inteligentes de forma que no pueden, por ejemplo, dañar los archivos de  
otros usuarios (Weld y Etzioni, 1994). El problema se magnifica cuando el dinero cambia de manos. Si las transacciones monetarias las realiza un agente inteligente en nombre de alguien, ¿está obligado por las deudas incurridas? ¿Sería posible que un agente  
inteligente tuviera activos o que realizara compras electrónicas en su propio nombre? Hasta ahora, parece que estas cuestiones no se entienden de forma clara. En nuestro conocimiento, ningún programa ha recibido ningún estado legal como individuo con fines  
financieros; por el momento, parece que no es razonable hacerlo. Los programas tampoco son considerados «conductores» para reforzar las regulaciones del tráfico en  
autovías reales. En las leyes californianas, por lo menos, no parece haber ninguna sanción legal que impida a un vehículo automatizado exceder los límites de velocidad, aunque  
el diseñador del mecanismo de control de vehículos sí tuviera obligación en caso de accidente. Al igual que con la tecnología para la reproducción humana, la ley tiene todavía que ponerse a la altura de los nuevos desarrollos.  
El éxito de la IA podría significar el fin de la raza humana. Casi cualquier tecnología tiene el potencial de hacer daño si se encuentra en las manos equivocadas, pero con  
la IA y la robótica, tenemos el problema nuevo de que las manos equivocadas podrían  
pertenecer a dicha tecnología. Incontables historias de ciencia ficción nos han alertado  
de los robots o de los «cyborgs» (robots-hombre) que se comportan de forma enajenada. Entre los primeros ejemplos se incluyen Frankenstein de Mary Shelly, o el ModernPrometheus (El Prometeo Moderno) (1818)7, y la obra de Karel Capek R.U.R. (1921),  
en la que los robots conquistan el mundo. En cine, también tenemos Terminador (TheTerminator) (1984), en donde se combinan clichés de robots que conquistan el mundo  
con el viajar en el tiempo, y Matrix (The Matrix) (1999), en donde se combinan robots  
que conquistan el mundo y el cerebro en una cubeta.  
Para una gran mayoría, parece que los robots son protagonistas de muchas historias  
de conquistas del mundo porque representan lo desconocido, de la misma manera que  
las brujas y los fantasmas de cuentos de otras décadas más antiguas. ¿Suponen una amenaza más creíble que las brujas y los fantasmas? Si los robots estuvieran diseñados adecuadamente como agentes que adoptan las metas de sus propietarios, probablemente no  
supondrían una amenaza: los robots que derivan de mayores avances sobre sus diseños  
actuales, sí que van a servir y no a conquistar. Los hombres utilizan su inteligencia de

FUNDAMENTOS FILOSÓFICOS 10937 Charles Babbage, cuando era joven, se vio influenciado por la lectura de Frankenstein.  
formas agresivas porque son innatas, por naturaleza. Sin embargo, las máquinas que construimos no tienen que ser innatamente agresivas, a menos que decidamos que así sean.  
Por un lado, es posible que los computadores logren una clase de conquista sirviendo y  
haciéndose indispensables, de la misma manera que los automóviles han conquistado en  
cierto sentido el mundo industrializado. Este escenario merece más atención. En 1965,  
I. J. Good escribió:  
Vamos a definir una máquina ultrainteligente como una máquina que puede sobrepasar  
con mucho todas las actividades intelectuales de cualquier hombre, por muy inteligente  
que sea. Puesto que el diseño de las máquinas es una de estas actividades intelectuales,  
una máquina ultrainteligente podría diseñar máquinas incluso mejores; entonces existiría incuestionablemente una «explosión de inteligencia», y la inteligencia del hombre quedaría bastante atrás. Así pues la primera máquina ultrainteligente es la última invención  
que haya hecho nunca el hombre, siempre que la máquina sea lo suficientemente dócil  
como para decirnos cómo controlarla.  
La «explosión de inteligencia» también ha sido llamada singularidad tecnológica por el  
profesor de Matemáticas y autor de ciencia ficción Vernor Vinge, (1993) quien escribió que,  
«Dentro de treinta años, tendremos el medio tecnológico de crear una inteligencia superhumana. Algún tiempo después, la era humana habrá acabado». Good y Vinge (y muchos  
otros) señalan correctamente que la curva del progreso tecnológico actualmente está creciendo de manera exponencial (tomemos en consideración, por ejemplo, la ley de Moore).  
Sin embargo, es un buen paso adelante extrapolar que la curva continuará hacia la singularidad de un crecimiento casi infinito. Hasta ahora, el resto de tecnologías han seguido una  
curva en forma de S, en donde el crecimiento exponencial finalmente disminuye.  
La preocupación y el miedo de Vinge radica en llegar a la singularidad, sin embargo  
otros científicos y futuristas gozan con esa idea. En el Robot, Mere Machine to TrascendentMind (Mera máquina hacia la mente trascendente), Hans Moravec predice que los robots  
se igualarán a la inteligencia humana en 50 años y a continuación la excederán:  
De manera bastante rápida podríamos quedar desplazados y fuera de la existencia. No  
estoy tan alarmado como muchos otros por esta última posibilidad, ya que considero que  
las máquinas del futuro son nuestra progenie, «hijos de mente» construidos a nuestra imagen y semejanza, es decir, nosotros mismos pero en una forma más potente. Al igual que  
los hijos biológicos de generaciones anteriores, representarán la mejor esperanza de la  
humanidad para un futuro a largo plazo. Nos corresponde a nosotros ofrecerles todas las  
ventajas, y cómo retirarnos cuando ya no podamos contribuir. (Moravec, 2000.)  
Ray Kurzweil, en The Age of Spiritual Machines (2000) predice que hacia el año 2099  
existirá «una fuerte tendencia hacia una fusión del pensamiento humano en el mundo de  
la inteligencia de la máquina que las especies humanas crearon inicialmente. Ya no existe una distinción clara entre los hombres y los computadores». Existe incluso una palabra nueva, trashumanismo, que se refiere al movimiento social real que ansía este futuro.  
Basta con decir que estos temas presentan un reto para la mayoría de los teóricos que consideran la preservación de la vida humana y de las especies como algo bueno.  
Finalmente, tomemos en consideración el punto de vista del robot. Si los robots  
adquieren consciencia, tratarlos entonces como meras «máquinas» (por ejemplo, tratarlos  
como algo aparte) podría ser inmoral. Los robots también deben actuar moralmente,

1094 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
SINGULARIDAD TECNOLÓGICA TRASHUMANISMO  
necesitaríamos programarlos con una teoría de lo que está bien y lo que está mal. Los  
escritores de ciencia ficción han abordado el tema de los derechos y responsabilidades  
de los robots, empezando por Isaac Asimov (1942). La famosa película titulada A.I.(Spielberg, 2001) se basó en una historia de Brian Aldiss sobre el robot inteligente que  
fue programado para creer que él era un humano y que no entiende el abandono final  
por parte de su madre-propietaria. La historia (y la película) convencen de la necesidad  
de los robots para moverse por sus derechos civiles.  
26.4 ResumenEste capítulo ha abordado los siguientes temas:  
• Los filósofos utilizan el término IA débil para la hipótesis de que las máquinas  
podrían comportarse posiblemente de forma inteligente, y el término IA fuerte para  
la hipótesis de que dichas máquinas contarían como si tuvieran mentes reales (en  
contraposición a las mentes simuladas).  
• Alan Turing rechazó la cuestión «¿Pueden pensar las máquinas?», y lo sustituyó  
por un test de comportamiento. Previeron muchas objeciones a la posibilidad de  
máquinas pensantes. Pocos investigadores en IA prestan atención al Test de  
Turing, y prefieren concentrarse en el rendimiento de sus sistemas en base a tareas prácticas, en vez de la habilidad de imitar a los humanos.  
• Actualmente existe el acuerdo general de que los estados mentales son los estados del cerebro.  
• Los argumentos a favor y en contra de la IA fuerte no son concluyentes. Unos pocos  
investigadores dominantes en la disciplina de la IA creen que cualquier cosa  
significativa gira en torno al resultado del debate.  
• La consciencia sigue siendo un misterio.  
• Se han identificado ocho amenazas potenciales para la sociedad que se exponen  
tanto ante la IA como ante una tecnología relacionada. Podemos concluir diciendo que algunas amenazas son improbables, pero merece la pena revisar dos de  
ellas en particular. La primera es que las máquinas ultrainteligentes podrían  
llevarnos a un futuro muy diferente del actual y puede que no sea de nuestro agrado. La segunda es que la tecnología de la robótica puede permitir a individuos con  
una psicopatía emplear armas de destrucción masiva. Concluimos diciendo que esto  
es más una amenaza de la biotecnología y nanotecnología que de la robótica.

NOTAS BIBLIOGRÁFICAS E HISTÓRICASLa naturaleza de la mente ha sido un tema estándar de la teorización filosófica desde tiempos inmemoriales hasta hoy en día. En el Phaedo, Platón consideró y rechazó de forma  
específica la idea de que la mente podría ser una «adaptación» o un modelo de organización de las partes del cuerpo, un punto de vista que se aproxima al punto de vista funFUNDAMENTOS FILOSÓFICOS 1095  
cionalista de la filosofía moderna de la mente. En cambio, determinó que la mente tenía que ser un alma inmortal, inmaterial, separable del cuerpo y diferente en sustancia,  
del punto de vista del dualismo. Aristóteles distinguió una variedad de almas (en griego υ) de los seres vivos, describiendo algunas, por lo menos, de forma funcionalista. (Véase Nussbaum (1978) para más información sobre el funcionalismo de Aristóteles.)  
Descartes destaca por su punto de vista dualista de la mente humana, aunque irónicamente su influencia histórica se dirigió hacia el mecanismo y el materialismo. Explí-  
citamente, concibió a los animales como autómatas, y se anticipó al Test de Turing,  
escribiendo que «no se puede concebir [que una máquina] elabore diferentes colocaciones  
de palabras para dar una respuesta con un significado adecuado para todo lo que se diga  
en su presencia, incluso de la misma forma que los hombres más aburridos cuando hablan» (Descartes, 1637). La ardiente defensa que hace Descartes del punto de vista de  
los animales como autómatas realmente tuvo el efecto de facilitar también la concepción de los hombres como autómatas, incluso aunque él mismo no diera ese paso. El libro L’Homme Machine o Man a Machine (La Mettrie, 1748) argumentó explícitamente  
que los hombres son autómatas.  
La filosofía analítica moderna normalmente ha aceptado el materialismo (a menudo mediante la teoría de la identidad del estado del cerebro (Place, 1956; Armstrong,  
1968), la cual afirma que los estados mentales son idénticos a los estados del cerebro)),  
sin embargo ha estado mucho más dividida en el funcionalismo, en la analogía de la  
máquina con le mente humana, y la cuestión de si las máquinas literalmente pueden pensar. Scriven (1953), por ejemplo, entre las primeras respuestas filosóficas a Turing  
(1950) «Computing Machinery and Intelligence», intentó negar que tuviera significado  
decir que las máquinas pudieran pensar, basándose en que dicha afirmación violaba el  
significado del mundo. En el apéndice de la reimpresión de este artículo se puede ver  
que, por lo menos, Scriven hacia el año 1963 se retractó de este punto de vista (Anderson, 1964). El científico informático Edsger Dijkstra dije que «La cuestión de si un computador puede pensar no es más interesante que la cuestión de si un submarino puede  
nadar». Ford y Hayes (1995) argumentan que el Test de Turing no es útil para la IA.  
El funcionalismo es la filosofía de la mente que más naturalmente sugiere la IA, y  
las críticas del funcionalismo a menudo toman forma a partir de las críticas sobre la IA  
(como en el caso de Searle). Siguiendo la clasificación utilizada por Block (1980), se  
pueden distinguir variedades del funcionalismo. La teoría de la especificación funcional (Lewis, 1966, 1980) es una variante de la teoría de la identidad del estado del cerebro la cual selecciona los estados del cerebro que se identifican con los estados mentales  
en su papel (rol) funcional. La teoría de la identidad del estado funcional (Putnam,  
1960, 1967) se basa más estrechamente en la analogía de la máquina. No identifica los  
estados mentales con los estados físicos del cerebro sino con los estados abstractos  
computacionales del cerebro concebidos expresamente como un dispositivo informático. Se supone que estos estados abstractos son independientes de la composición física  
específica del cerebro, lo que lleva a considerar que la teoría de la identidad del estado  
funcional es una forma de dualismo.  
Tanto la teoría de la identidad del estado del cerebro como las diferentes formas de  
funcionalismo han sido atacadas por autores afirmando que no tienen en consideración  
la qualia o el aspecto «a qué se asemejan» de los estados mentales (Nagel, 1974). En

1096 INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO  
cambio, Searle se ha centrado en la supuesta incapacidad del funcionalismo (Searle, 1980,  
1984, 1992). Churchland y Churchland (1982) rebaten ambas formas de crítica.  
El materialismo eliminativo (Rorty, 1965; Churchland, 1979) difiere de las otras  
teorías prominentes de la filosofía de la mente, en que no intenta dar cuenta de por qué  
son verdaderas las ideas de «la psicología popular» y del sentido común sobre la mente, sino que por el contrario las rechaza como si fueran falsas e intenta sustituirlas por  
una teoría puramente científica de la mente. En principio, esta teoría científica podría  
ser proporcionada por una IA clásica, pero en cambio, en la práctica, los materialistas  
eliminativos tienden a apoyarse en la investigación de la neuro ciencia y de las redes  
neuronales (Churchland, 1986), sin embargo, basándose en la IA clásica, especialmente en la investigación de la «representación del conocimiento» del tipo descrito en el  
Capítulo 10, tienden a confiar en la verdad de la psicología popular. Aunque el punto de  
vista del «estado intencional» (Dennett, 1971) pudiera interpretarse como funcionalista, probablemente debería considerarse como una forma de materialismo eliminativo, en  
el sentido de que la «postura intencional» no tiene que reflejar ninguna propiedad objetiva del agente hacia el que se toma esa postura. También se debería destacar que es  
posible ser un materialista eliminativo en algunos aspectos de la mentalidad a la vez que  
se analizan otras de forma diferente. Por ejemplo, Dennett (1978) es mucho más eliminativo en cuanto a la qualia que en cuanto a la intencionalidad.  
Ya se han proporcionado en este mismo capítulo otras fuentes de críticas importantes hacia la IA débil. Aunque estuvo muy en auge ridiculizar los enfoques simbólicos  
en la era de las redes post-neuronales, no todos los filósofos son críticos de GOFAI. De  
hecho, algunos son ardientes partidarios e incluso buenos profesionales. Zenon Pylyshyn  
(1984) ha argumentado que la cognición se puede entender mejor a través de un modelo computacional, no sólo en principio sino también en la forma de realizar una investigación contemporánea, y ha refutado específicamente las críticas de Dreyfus en cuanto  
al modelo computacional de la cognición humana (Pylyshyn, 1974). Gilbert Harman  
(1983), quien en el análisis de la revisión de las creencias ha realizado conexiones con  
la investigación de la IA en los sistemas del mantenimiento de verdad. Michael Bratman  
ha aplicado su modelo de «creencia-deseo-intención» de la psicología humana (Bratman,  
1987) a la investigación de la IA en planificación (Bratman, 1992). En el lado extremo  
de la IA fuerte, Aaron Sloman (1978, p. xiii) ha descrito incluso como «racista» el punto de vista de Joseph Weizenbaum (Weizenbaum, 1976) diciendo que las máquinas inteligentes hipotéticas no deberían considerarse personas.  
La literatura filosófica sobre las mentes, los cerebros, y otros temas relacionados es  
extensa y algunas veces difícil de leer sin una práctica adecuada en la terminología y en  
los métodos de argumentar empleados. La Enciclopedia de la Filosofía (Encyclopedia ofPhilosofy) (Edwards, 1967) es una ayuda sorprendentemente fidedigna y útil de este proceso. El Diccionario de Filosofía de Cambridge (The Cambridge Dictionary of Philosofy)  
(Audi, 1999) es un trabajo más reducido y más accesible, pero las entradas principales  
(tales como «filosofía de la mente») todavía abarcan 10 páginas o más. La Enciclopedia  
de la Ciencia Cognitiva del MIT (MIT Encyplopedia of Cognitive Sciencie) (Wilson y Keil,  
1999) aborda la filosofía de la mente así como la biología y la psicología de la mente.  
Otras colecciones de artículos sobre filosofía de la mente, incluyendo el funcionalismo  
y otros puntos de vista relacionados con la IA son El Materialismo y el Problema de Men

FUNDAMENTOS FILOSÓFICOS 1097MATERIALISMO ELIMINATIVO  
te-Cuerpo (Materialism and the Mind-Body Proble, Rosenthal, 1971) y Lecturas de la Filosofía de la Psicología (Readings in the Philosophy of Psycology, volumen 1, Block,  
1980). Biro y Shahan (1982) presentan una colección dedicada a los pros y los contras  
del funcionalismo. Entre las antologías que tratan más específicamente con la relación  
entre la Filosofía y la IA se incluyen Minds and Machines (Anderson, 1964), Philosophical Perspectives in Artificial Intelligence (Ringle, 1979), Mind Design (Haugeland,  
1981) y The Philosophy of Artificial Intelligence (Boden, 1990). Existen varias introducciones a la «cuestión IA» filosófica (Boden, 1977, 1990; Haugeland, 1985; Copeland,  
1993). The Behavioral and Brain Sciences, abreviado BBS, es una publicación importante  
dedicada a debates filosóficos y científicos sobre la IA y la neurociencia. En publicaciones como AI and Society, Law, Computers and Artificial Intelligence, y Artificial Intelligence and Law se tratan los temas de la ética y la responsabilidad en la IA.  
EJERCICIOS26.1 Repase la lista de presuntas incapacidades de las máquinas de Turing, identificando cuáles se han logrado, cuáles se pueden lograr en principio por cualquier programa, y cuáles son aún problemáticas porque requieren estados mentales conscientes.  
26.2 ¿La refutación de la habitación china demuestra necesariamente que los computadores con una programación adecuada tienen estados mentales? ¿La aceptación del argumento significa necesariamente que los computadores no pueden tener estados  
mentales?  
26.3 En el argumento de la prótesis del cerebro, es importante poder restaurar el cerebro de un sujeto a un estado normal, de tal manera que su comportamiento externo sea  
como si la operación no hubiera tenido lugar. ¿Puede un escéptico objetar razonablemente  
que esto requeriría actualizar las propiedades neuropsicológicas de las neuronas que se  
relacionan con la experiencia consciente, distintas de las que tienen que ver con el comportamiento funcional de las neuronas?  
26.4 Encuentre y analice un caso relacionado con los medios de comunicación populares sobre un argumento o más, en relación con la idea de que la IA es imposible.  
26.5 Intente definir los términos «inteligencia», «pensamiento» y «consciencia». Sugiera algunas posibles objeciones a sus definiciones.  
26.6 Analice las posibles amenazas a la sociedad de la tecnología de IA. ¿Cuáles son  
las amenazas más serias, y cómo se podrían combatir? ¿Cómo se podrían comparar con  
las posibles ventajas?  
26.7 ¿Cómo se pueden comparar las posibles amenazas de la tecnología de la IA con  
otras de otras tecnologías informáticas, y con las de las tecnologías nucleares, bionucleares  
y nanonucleares?  
26.8 Algunos críticos piensan que la IA es imposible, mientras que otros dicen que es  
demasiado posible, y que las máquinas ultrainteligentes suponen una amenaza. ¿Cuál  
de estas objeciones piensa que es más probable? ¿Sería una contradicción mantener ambas opiniones?