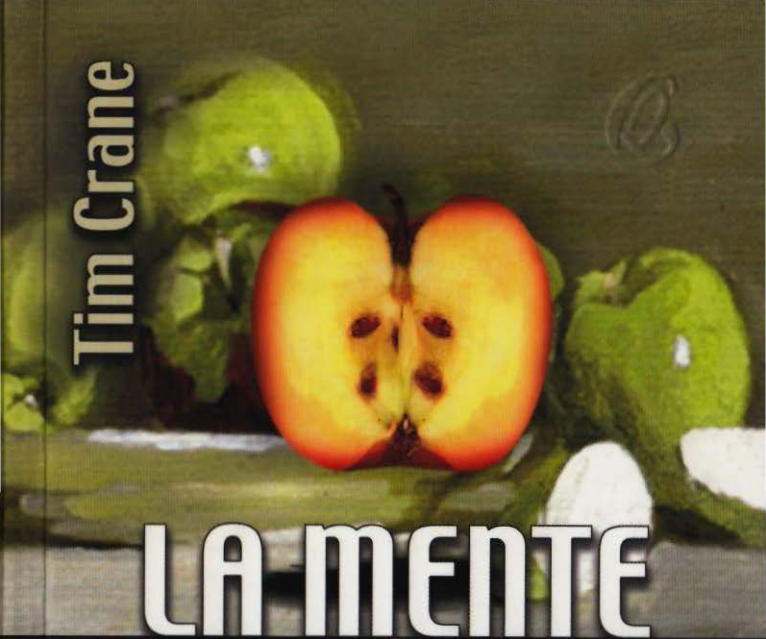


Tim Crane

LA MENTE MECÁNICA

Introducción filosófica a mentes, máquinas y representación mental

BREVIARIOS

BREVIARIOS

del

FONDO DE CULTURA ECONÓMICA

559

LA MENTE MECÁNICA

Traducción de

Juan Almela

La mente mecánica

Introducción filosófica a mentes, máquinas y representación mental

por Tim Crane

FONDO DE CULTURA ECONÓMICA

Primera edición en inglés, 1995
Segunda edición en inglés, 2003
Primera edición en español, 2008

Crane, Tim
La mente mecánica. Introducción filosófica a mentes, máquinas y representación mental / Tim Crane ; trad. de Juan Almela. — México : FCE, 2008
375 p. : ilus. ; 17 ×11 cm — (Colec. Breviarios ; 559)
Título original The Mechanical Mind. A philosophical introduction to minds, machines and mental representation
ISBN 978-968-16-8351-1

1. Filosofía de la mente I. Almela, Juan, tr. II. Ser. III t.

LC BD418.3 Dewey 082.1 B846 V. 559

Distribución mundial

Comentarios y sugerencias: editorial@fondodeculturaeconomica.com
www.fondodeculturaeconomica.com
Tel. (55) 5227-4672 Fax (55) 5227-4694

Empresa certificada ISO 9001: 2000

Título original: *The Mechanical Mind. A Philosophical Introduction to Minds, Machines and Mental Representation*
© 1995, 2003, Tim Crane
Primera edición en inglés publicada por Penguin Books
Segunda edición en inglés publicada por Routledge
11 New Fetter Lane, Londres, EC4P 4EE
Publicado simultáneamente en EUA y Canadá por Routledge
29 West Street 35, Nueva York, 10001

Traducción autorizada de la edición en lengua inglesa publicada por Routledge, integrante de Taylor & Francis Group

Diseño de portada: Teresa Guzmán Romero

D. R. © 2008, Fondo de Cultura Económica
Carretera Picacho-Ajusco, 227; 14738 México, D. F.

Se prohíbe la reproducción total o parcial de esta obra —incluido el diseño tipográfico y de portada—, sea cual fuere el medio, electrónico o mecánico, sin el consentimiento por escrito del editor.

ISBN 978-968-16-8351-1

Impreso en México • *Printed in Mexico*

ÍNDICE

Prefacio a la primera edición 13
Prefacio a la segunda edición 17
Introducción 21
La imagen mecánica del mundo 22
La mente 26

I. *El rompecabezas de la representación* 31
La idea de representación 35
Imágenes y parecido 38
Representación lingüística 47
Representación mental 52
Pensamiento y conciencia 57
Intencionalidad 63
La tesis de Brentano 73
Conclusión: de la representación a la mente 78
Lecturas adicionales 79

II. *Cómo entender a los pensadores y sus pensamientos* 81
El problema mente-cuerpo 83
Cómo entender otras mentes 88
La imagen causal de los pensamientos 100
Psicología del sentido común 112
La ciencia del pensamiento: ¿eliminación o vindicación? 123
Teoría y simulación 133
Conclusión: de la representación a la computación 138
Lecturas adicionales 140

III. *Computadoras y pensamiento* 142
Preguntar lo adecuado 143
Computación, funciones y algoritmos 145
Máquinas de Turing 154
Codificación y símbolos 166
Ejemplificación y computación de una función 169
Algoritmos automáticos 172
¿Computadoras pensantes? 180
Inteligencia artificial 187
¿Puede el pensamiento ser capturado por reglas y representaciones? 194
El cuarto chino 202
Conclusión: ¿puede pensar una computadora? . 208
Lecturas adicionales 209

IV. *Los mecanismos del pensamiento* 211
Cognición, computación y funcionalismo 213
El lenguaje del pensamiento 218
Sintaxis y semántica 222
El argumento del lenguaje del pensamiento ... 226
La modularidad de la mente 237
Problemas para el lenguaje del pensamiento ... 246
Computadoras "sesudas" 254
Conclusión: ¿explica la computación la representación? 265
Lecturas adicionales 266

V. *Explicando la representación mental* 268
Reducción y definición 269
Definiciones conceptuales y naturalistas 273
Teorías causales de la representación mental ... 277
El problema del error 281
Representación mental y éxito en la acción 292
Representación mental y función biológica 298

La evolución y la mente 305
Contra la reducción y la definición 315
Conclusión: ¿puede la representación ser explicada reductivamente? 327
Lecturas adicionales 328

VI. *La conciencia y la mente mecánica* 330
La historia hasta aquí 330
La conciencia, "lo parecido" y los qualia 336
Conciencia y fisicalismo 342
Los límites del conocimiento científico 353
Conclusión: ¿qué nos enseñan los problemas de la conciencia acerca de la mente mecánica? . 358
Lecturas adicionales 360

Glosario 363
Cronología 369
Índice de figuras 375

Pero ¿cómo, y por cuál arte, lee el alma que tal imagen o trazo en la materia... significa tal objeto? ¿Aprendemos semejante Alfabeto en nuestro estado Embrionario? Y ¿cómo es que no tengamos noción de ningunas aprehensiones congénitas tales?... Que por diversidad de movimientos conjuremos imágenes, distancias, magnitudes, colores, cosas a las que no se parecen, lo atribuimos a algunas secretas deducciones.

JOSEPH GLANVILL,
The Vanity of Dogmatizing, 1661

PREFACIO A LA PRIMERA EDICIÓN

Este libro es una introducción a algunas de las principales preocupaciones de la filosofía contemporánea de la mente. Hay muchas maneras de escribir un libro de introducción. En lugar de dar una descripción justa de todas las teorías filosóficas recientes de la mente, decidí seguir una línea de pensamiento que captura la esencia de lo que me parece el más interesante de los debates contemporáneos. El centro de esta línea de pensamiento es el problema de la representación mental: ¿cómo puede la mente representar el mundo? Este problema es el hilo que ata los capítulos y en torno a este hilo se trenzan los otros temas fundamentales del libro: la naturaleza de la explicación psicológica cotidiana, la naturaleza causal de la mente, la mente como una computadora y la reducción del contenido mental.

Aunque hay una línea argumentativa continua, he tratado de construir el libro de tal manera que (en cierta medida) los capítulos pueden ser leídos independientemente uno del otro. Así, el capítulo I introduce el problema de la representación y discute la representación pictórica, lingüística y mental. El capítulo II trata de la naturaleza de la psicología del sentido común (la llamada "popular") y la naturaleza causal de los pensamientos. El capítulo III se ocupa de la cuestión de si las computadoras pueden pensar, y el capítulo IV se pregunta si nuestras mentes son computadoras en algún sentido. El capítulo final discute las teorías de la representación mental y el breve epílogo suscita algunas

dudas escépticas acerca de las limitaciones del punto de vista mecánico de la mente. Así, quienes se interesan en la cuestión de si la mente es una computadora pueden leer los capítulos III y IV independientemente del resto del libro. Y quienes se interesan más en los problemas puramente "filosóficos" podrían preferir leer los capítulos I y II separadamente. He tratado de indicar dónde la discusión se vuelve más complicada, y qué secciones preferiría eliminar un principiante. En general, sin embargo, los capítulos IV y V son más difíciles de estudiar que los capítulos I a III.

Al final de cada capítulo he proporcionado sugerencias para mayor lectura. En las notas finales se ofrecen referencias más detalladas, destinadas únicamente al estudiante que trata de seguir el debate; nadie necesita leer las notas finales con objeto de entender el libro.

He presentado la mayoría del material de este libro en conferencias y seminarios en el University College de Londres durante los últimos años, y agradezco mucho a mis discípulos por sus reacciones. Agradezco también a los auditorios de las universidades de Bristol, Kent y Nottingham, donde las versiones iniciales de los capítulos III y IV fueron presentadas como conferencias. Quisiera agradecer a Stefan McGrath por su inapreciable consejo editorial, a Caroline Cox, Stephen Cox, Virginia Cox, Petr Kolář, Ondrej Majer, Michael Ratledge y Vladimir Svoboda por sus útiles comentarios acerca de versiones anteriores de algunos capítulos, a Roger Bowdler por sus dibujos y a Ted Honderich por su generoso estímulo en una etapa inicial. Tengo además una deuda especial con mis colegas Mike Martin, Greg McCulloch, Scott Sturgeon y Jonathan Wolff por sus detallados y agudos comentarios a la penúltima versión del libro completo, que resultaron una revisión sustancial, y me aho-

rraron muchos errores. Esta penúltima versión fue escrita en Praga, mientras era yo huésped del Departamento de Lógica de la Academia Checa de Ciencias. Mi más cálido agradecimiento se dirige a los miembros del Departamento —Petr Kolář, Pavel Materna, Ondrej Majer y Vladimir Svoboda, así como Marie Duži—, por su amable hospitalidad.

University College de Londres,
noviembre de 1994

PREFACIO A LA SEGUNDA EDICIÓN

Los principales cambios que he hecho para esta segunda edición son sustituir el epílogo por un nuevo capítulo acerca de la conciencia, la adición de nuevas secciones acerca de la modularidad y la psicología evolucionista en los capítulos IV y V, y la adición del glosario y la cronología al final del libro. También he corregido muchos errores estilísticos y filosóficos y puesto al corriente la sección de otras lecturas. Mi manera de ver la intencionalidad ha cambiado de ciertos modos desde que escribí este libro. Ahora adopto un enfoque intencionalista ante todos los fenómenos mentales, según se esbozó en mi libro de 2001, *Elements of Mind* (Oxford University Press). Sin embargo, he resistido la tentación de modificar de modo significativo la exposición del capítulo I, excepto cuando dicha exposición implicaba auténticos errores.

Agradezco mucho a Tony Bruce por su entusiasta apoyo para la nueva edición de este libro, a numerosos informes anónimos de lectores de Routledge por su excelente consejo y a Ned Block, Katalin Farkas, Hugh Mellor y Huw Price por sus comentarios críticos detallados acerca de la primera edición.

University College de Londres,
agosto de 2002

A mis padres

INTRODUCCIÓN

Un amigo observó que llamar a este libro *La mente mecánica* es un poco como llamar a una novela de misterio *Fue el mayordomo*. Sería una vergüenza si el título tuviese esta connotación, porque la meta del libro es esencialmente suscitar y examinar problemas, más que resolverlos. A grandes rasgos, trato de hacer dos cosas con este libro: primero, explicar el problema filosófico de la representación mental; segundo, examinar las cuestiones acerca de la mente que surgen cuando se trata de resolver este problema a la luz de supuestos filosóficos dominantes. Fundamental para estos supuestos es la idea que llamo "mente mecánica". A grandes rasgos, éste es el punto de vista de que la mente debe ser considerada como un mecanismo causal, un fenómeno natural que se comporta de una manera regular, sistemática, como el hígado o el corazón.

En el capítulo 1 introduzco el problema filosófico de la representación mental. Este problema es fácil de enunciar: ¿cómo puede la mente representar algo? Mi creencia, por ejemplo, de que Nixon visitó China trata de Nixon y de China, pero ¿cómo puede un estado de mi mente "ocuparse" de Nixon y de China?, ¿cómo puede mi estado mental dirigirse a Nixon o a China?, ¿qué es para una mente representar algo?, ¿qué es para *algo* (ya sea una mente o no) representar otra cosa?

El problema, que algunos filósofos contemporáneos llaman "el problema de la intencionalidad", tiene orígenes re-

motos. Recientes adelantos en la filosofía de la mente —junto con desarrollos en las disciplinas afines, la lingüística, la psicología y la inteligencia artificial— han suscitado el viejo problema de una manera nueva. Así, por ejemplo, la cuestión de si una computadora puede pensar es reconocida ahora como estrechamente ligada al problema de la intencionalidad. Y lo mismo ocurre con la cuestión de si podrá haber una "ciencia del pensamiento": ¿puede la mente ser explicada por la ciencia o requiere su propio modo no científico de explicación? Una respuesta completa a esta cuestión depende, como veremos, de la naturaleza de la representación mental.

Como fundamento de los más recientes intentos de responder tales cuestiones está lo que he llamado el punto de vista mecánico de la mente. La representación se considera un problema porque es difícil entender cómo un simple mecanismo puede representar el mundo, cómo estados del mecanismo pueden "alcanzar el exterior" y dirigirse al mundo. El propósito de esta introducción es dar mayor idea de lo que quiero decir al hablar acerca de la mente mecánica, esbozando los orígenes de la idea.

La imagen mecánica del mundo

La idea de que la mente es un mecanismo natural deriva de pensar en la naturaleza misma como una especie de mecanismo. Así, comprender esta manera de ver la mente requiere comprender —en términos muy generales— esta manera de contemplar la naturaleza.

El punto de vista occidental moderno del mundo nos conduce a la "revolución científica" del siglo XVII y las ideas

de Galileo, Francis Bacon, Descartes y Newton. En la Edad Media y en el Renacimiento el mundo se había considerado en términos orgánicos. La tierra misma era juzgada una especie de organismo, según lo ilustra de manera colorida este pasaje de Leonardo da Vinci: "Podemos decir que la tierra tiene un alma vegetativa y que su carne es la tierra, sus huesos son las estructuras de las rocas... su sangre son los depósitos de agua... su respiración y sus pulsos son el flujo y el reflujo del mar".[1]

Esta imagen orgánica del mundo, como podemos llamarla, debía mucho a las obras de Aristóteles, el filósofo que tuvo, con mucho, la mayor influencia durante la Edad Media y el Renacimiento. (De hecho, su influencia era tan grande que a menudo simplemente era llamado "el Filósofo".) En el sistema del mundo de Aristóteles, todo debía tener su "lugar" o condición natural, y las cosas hacían lo que hacían porque estaba en su naturaleza alcanzar su condición natural. Esto se aplicaba a las cosas inorgánicas tanto como a las orgánicas: las piedras caen al suelo a causa de que su lugar natural está en el suelo, el fuego se eleva a su lugar natural en el cielo, y así sucesivamente. Todo en el universo era visto como si tuviese su carácter final, punto de vista que estaba plenamente en armonía con una concepción del universo cuya fuerza impulsora es Dios.

En el siglo XVII todo esto empezó a desmoronarse. Un importante cambio fue que el método aristotélico de explicación —en términos de fines y "naturalezas"— fue remplazado por un método mecánico de explicación, en términos del comportamiento determinista regular de la materia

[1] Citado por Peter Burke, *The Italian Renaissance* (Cambridge, Polity Press, 1986), p. 201.

en movimiento. Y el modo de encontrar algo acerca del mundo no era estudiando e interpretando las obras de Aristóteles, sino observando y experimentando, así como midiendo matemáticamente las magnitudes e interacciones de la naturaleza. El uso de la medida matemática en la comprensión científica del mundo era uno de los elementos clave de la nueva "imagen mecánica del mundo". Galileo tuvo fama de hablar acerca de

> este gran libro del universo que... no puede ser comprendido a menos de que uno empiece por comprender el lenguaje y leer el alfabeto en el cual está compuesto. Está escrito en el lenguaje de las matemáticas, y sus caracteres son triángulos, círculos y otras figuras geométricas, sin las cuales es humanamente imposible entender ni una palabra de él.[2]

La idea de que el comportamiento del mundo podía medirse y entenderse en términos de ecuaciones matemáticas precisas o leyes de la naturaleza, estuvo en el corazón del desarrollo de la ciencia de la física, tal como hoy la conocemos. Hablando muy toscamente podemos decir que según la imagen mecánica del mundo, las cosas hacen lo que hacen no a causa de que traten de alcanzar su lugar natural o final, o porque obedezcan la voluntad de Dios, sino, más bien, porque se tienen que mover según ciertos modos de acuerdo con las leyes de la naturaleza.

En los términos más generales, esto es lo que entiendo por un punto de vista mecánico de la naturaleza. Por su-

[2] Galileo, "The Assayer", en Stillman Drake, *Discoveries and Opinions of Galileo* (Nueva York, Doubleday, 1957), pp. 237-238.

puesto, el término "mecánico" significaba —y a veces sigue significando— algo mucho más específico. Se consideró que los sistemas mecánicos sólo interactuaban en contacto y deterministamente, por ejemplo. Progresos posteriores de la ciencia —por ejemplo la física de Newton, con su postulación de fuerzas gravitacionales que aparentemente actúan a distancia, o el descubrimiento de que los procesos físicos fundamentales no son deterministas— refutaron la imagen mecánica del mundo en este sentido específico. Estos descubrimientos, por supuesto, no socavan la imagen general de un mundo de causas que actúa de acuerdo con leyes o regularidades naturales; y esta idea más general es a la que me referiré como "mecánica" en este libro.

En la imagen "orgánica" del mundo de la Edad Media y el Renacimiento se concibieron las cosas de acuerdo con las líneas de las cosas orgánicas. Todo tenía su lugar natural, ajustado al funcionamiento armonioso del "animal" que es el mundo. Pero con la imagen mecánica del mundo la situación se invirtió: las cosas orgánicas fueron juzgadas siguiendo las líneas de las cosas inorgánicas. Todo, orgánico e inorgánico, hacía lo que hacía porque estaba causado por algo más, de acuerdo con principios que podían ser formulados precisamente, matemáticamente. René Descartes (1596-1650) fue famoso por sostener que los animales no humanos son máquinas que carecen de cualquier conciencia o mentalidad: pensó que el comportamiento de los animales podía ser explicado de modo enteramente mecánico. Y conforme se desarrolló la imagen mecánica del mundo, el reloj, más bien que el animal, se tornó una metáfora dominante. Como escribió Julien de La Mettrie, un precursor dieciochesco del punto de vista mecánico de la mente, "el

cuerpo es sólo un reloj... el hombre sólo es una colección de resortes que se dan cuerda mutuamente".[3]

Así, no es sorprendente que, hasta mediados de este siglo, un gran misterio para la imagen mecánica del mundo fuese la naturaleza de la vida misma. Muchos supusieron que había en principio una explicación mecánica de la vida por descubrir —Thomas Hobbes había afirmado confiadamente en 1651 que "la vida es sólo un movimiento de los miembros"—;[4] el único problema era encontrarla. Gradualmente, se descubrió más y más acerca de cómo la vida era un proceso puramente mecánico, lo que culminó en el descubrimiento de la estructura del ADN por Watson y Crick en 1953. Ahora, según parece, la capacidad de los organismos para reproducirse puede ser explicada, en principio, en términos químicos. Lo orgánico puede ser explicado en términos de lo inorgánico.

La mente

¿Dónde deja esto a la mente? Aunque estaba perfectamente dispuesto a considerar a los animales como puras máquinas, Descartes no hizo lo mismo con respecto a la mente humana: aunque pensó que la mente (o el alma) tiene efectos sobre el mundo físico, la colocó fuera del universo mecánico. Sin embargo, muchos filósofos mecanisistas, en siglos posteriores, no podían aceptar este punto de vista particular de Descartes, y por lo tanto se enfrentaban al mayor reto al ex-

[3] J. de la Mettrie, *Man, the Machine* (1748, traducido por G. Bussey; Chicago, Open Court, 1912).

[4] Hobbes, *Leviathan* (1651), introducción, p. 1. [Ed. en el FCE, *Leviatan,* México, 2003.]

plicar el lugar de la mente en la naturaleza. El misterio que permanecía en la imagen mecánica del mundo era la explicación de la mente en términos mecánicos.

Igual que con la explicación mecánica de la vida, muchos supusieron que habría una explicación igual de la mente. Ejemplos particularmente buenos de este punto de vista se encuentran entre las muletillas de los materialistas de los siglos XVIII y XIX: la espléndida observación de La Mettrie de que "el cerebro tiene músculos para pensar, tal como las piernas tienen músculos para andar", o el lema del fisiólogo Karl Vogt de que "el cerebro segrega el pensamiento precisamente igual que el hígado segrega la bilis".[5] Éstos son, por supuesto, manifiestos materialistas más bien que teorías.

De modo que ¿cómo sería una explicación de la mente? Una idea influyente en la filosofía de los últimos 40 años es que explicar la mente significaría mostrar que es en realidad, ni más ni menos, materia. Los estados mentales en realidad son, ni más ni menos, estados químicos del cerebro. Esta noción "fisicalista" depende normalmente de suponer que explicar algo plenamente es explicarlo en última instancia en términos de ciencia (diremos más acerca de este modo de ver en el capítulo VI). Esto es, las ciencias distintas de la física deben tener su legitimidad vindicada por la física; todas las ciencias deben ser *reducibles* a la física. Lo normal es que esto signifique que el contenido de las ciencias distintas de la física sea deducible o derivable de la física (más "puentes" que enlazan los conceptos físicos con los no físicos) y que, por lo tanto, todo lo explicable por cualquier ciencia sea explicable en términos de la física. Éste es el

[5] La cita de La Mettrie es de *Man, the Machine*. La cita de Vogt es de John Passmore, *A Hundred Years of Philosophy* (Harmondsworth, Penguin, 1968), p. 36.

punto de vista —conocido en ocasiones como "reduccionismo"— que yace detrás de la broma memorable de Rutherford de que "hay física; y hay colección de timbres".[6]

Este reduccionismo extremo es verdaderamente muy implausible, y es muy dudoso si la práctica científica realmente se ajusta a él. Muy pocas ciencias no físicas han sido reducidas realmente a la física en este sentido, y parece haber pocas perspectivas de que la ciencia del porvenir aspire a reducir todas las ciencias a la física. Si acaso, la ciencia parece estarse volviendo más diversificada y más unificada. Por esta razón (y otras) pienso que podemos distinguir entre la idea general de que la mente puede ser explicada mecánicamente (o causalmente explicada en términos de alguna ciencia por otra) y la más extrema tesis reduccionista. Podría creerse que puede haber una ciencia de la mente sin creer que esta ciencia tiene que reducirse a la física. Éste será un supuesto conductor de este libro, aunque no pretendo haber presentado aquí argumentos en su favor.[7]

Mi personal modo de ver, que trato de defender en este libro, es que una explicación mecánica de la mente debe

[6] Citado por Christopher Longuet-Higgins, "The Failure of Reductionism", en C. Longuet Higgins *et al.*, *The Nature of Mind* (Edimburgo, Edinburgh University Press), p. 16. Véase también David Charles y Kathleen Lennon (eds.), *Reduction, Explanation and Realism* (Oxford, Oxford University Press, 1991). El término "reduccionismo" ha significado muchas cosas en filosofía; para una exposición más detallada, véase mis *Elements of Mind* (Oxford, Oxford University Press, 2001), § 15.

[7] Para argumentos en defensa de esta pretensión, véase Tim Crane y D. H. Mellor, "There is no Question of Physicalism", *Mind,* 99 (1990), reimpreso en D. H. Mellor, *Matters of Metaphysics* (Cambridge, Cambridge University Press, 1991), y Tim Crane, "Against Physicalism", en Samuel Guttenplan (ed.), *A Companion to the Philosophy of Mind* (Oxford, Blackwell, 1994).

demostrar (por lo menos) cómo la mente es parte del mundo de las causas y los efectos, parte de lo que los filósofos llamarían el "orden causal" del mundo. Otra cosa que debe hacer una explicación mecánica de la mente es dar los detalles de generalizaciones que describen regularidades causales en la mente. En otras palabras, una explicación mecánica de la mente está destinada a la existencia de *leyes naturales* de la psicología. Precisamente como la física informa acerca de las leyes que gobiernan el mundo no mental, así la psicología informa acerca de las leyes que gobiernan la mente: puede haber una ciencia natural de la mente.

Sin embargo, mientras este punto de vista es asumido por la mayoría de los filósofos de la mente en sus rasgos principales, su aplicación a muchos de los fenómenos de la mente es sumamente problemática. Dos clases de fenómenos sobresalen como obstáculos al punto de vista mecánico de la mente: el fenómeno de la conciencia y el fenómeno del pensamiento. De ahí la reciente preocupación de la filosofía con dos cuestiones: primera, ¿cómo puede un simple mecanismo ser consciente?; y, segunda, ¿cómo puede un simple mecanismo pensar en cosas y representarlas? El tema central de este libro es generado por la segunda cuestión: el problema del pensamiento y la representación mental. Así, los capítulos I a V se ocupan en gran medida de este problema. Pero un tratamiento pleno de la mente mecánica también requiere decir algo acerca del problema de la conciencia: ninguna teoría mecánica de la mente que no consiguiera enfrentarse a este fenómeno mental tan esencial podría considerarse como una teoría completa de la mente. Éste es el tema del capítulo VI.

I. EL ROMPECABEZAS DE LA REPRESENTACIÓN

Cuando la nasa envió el *Pioneer 10* al espacio para explorar el sistema solar en 1972, se puso a bordo una placa de metal grabada con varias imágenes y signos. Parte de la placa era un diagrama de un átomo de hidrógeno, en tanto que otra era un diagrama de los tamaños relativos de los planetas en nuestro sistema solar, indicando el planeta del cual procedía el *Pioneer 10*. La imagen mayor en esta placa era la silueta de un hombre y una mujer desnudos, con el hombre alzando la mano para saludar. La idea que había detrás de esto era que cuando el *Pioneer 10* acabase por abandonar el sistema solar, podría continuar un camino cualquiera a través del espacio, tal vez para ser descubierto en millones de años por alguna forma vital ajena. Y tal vez estos alienígenas serían inteligentes y conseguirían comprender los diagramas, reconocer el grado de nuestro conocimiento científico y darse cuenta de que nuestras intenciones hacia ellos, quienes fueran, eran pacíficas.

Me parece que hay algo muy humorístico en esta historia. Supóngase que el *Pioneer 10* alcanzase alguna estrella distante. Y supóngase que la estrella tuviera un planeta con condiciones susceptibles de sostener la vida. Y finalmente supóngase que alguna forma de vida de este planeta fuera inteligente y tuviera alguna especie de órganos sensoriales con los cuales pudiera percibir la lámina volante. Todo esto es sumamente improbable. Pero incluso al admitir estas

improbables suposiciones, ¿no parecería aún *más* imposible que los alienígenas consiguieran comprender qué significan los símbolos de la placa?

Piénsese en algunas de las cosas que deberían comprender. Tendrían que comprender que los símbolos sobre la placa *eran* símbolos, que aspiraban a representar cosas y no eran sencillamente raspaduras en la placa, o simple decorado. Una vez que los alienígenas supieran que eran símbolos, tendrían que comprender de qué clase de símbolos se trataba: por ejemplo, que el diagrama del átomo de hidrógeno era un diagrama científico y no una simple estampa. Entonces deberían tener alguna idea de qué clases de cosas representaban los símbolos: que el dibujo del hombre y la mujer simbolizaban formas vitales en lugar de elementos químicos, que el diagrama del sistema solar simbolizaba nuestra parte del universo antes que la forma de quienes diseñaron el vehículo espacial. Y —quizá lo más absurdo de todo—, aun cuando se hiciesen una idea de lo que el hombre y la mujer eran, tendrían que reconocer que la mano alzada era un signo de saludo pacífico y no de impaciencia, de agresión o de desdén, o sencillamente que era la posición normal de esta parte del cuerpo.

Cuando se considera todo esto, ¿no parece aún más improbable que los alienígenas comprendieran los símbolos que la llegada del vehículo espacial a un planeta con vida inteligente, para empezar?

Una cosa que ilustra esta historia, en mi opinión, es algo acerca del problema filosófico o el rompecabezas de la representación. Los dibujos y símbolos de la placa representan cosas —átomos, seres humanos, el sistema solar—, pero la historia sugiere que hay algo desconcertante acerca del modo como lo logran. Porque cuando nos imaginamos en

la posición de los alienígenas, nos damos cuenta de que no podemos decir qué representan estos símbolos con sólo mirarlos. Ningún grado de escrutinio de las marcas en la placa puede revelar que estas marcas representan a un hombre y aquéllas representan a una mujer, y que estas otras marcas representan un átomo de hidrógeno. Las marcas de la placa pueden ser entendidas de muchas maneras, pero parece que nada en las marcas *mismas* revela cómo entenderlas. Ludwig Wittgenstein, cuya filosofía estaba dominada por cuestiones acerca de la representación, expresó brevemente: "Cada signo *por sí mismo* parece muerto; *¿qué* le da vida?"[1]

El rompecabezas filosófico acerca de la representación puede ser planteado sencillamente preguntando cómo es posible que una cosa represente otra. Planteada así, la cuestión puede parecer un poco oscura, y puede ser difícil ver exactamente qué es desconcertante acerca de ella. Una razón para esto es que la representación sea un hecho tan familiar en nuestras vidas. Palabras dichas y escritas, imágenes, símbolos, gestos, expresiones faciales, pueden todas verse como representaciones a partir del tejido de nuestra vida cotidiana. Sólo cuando iniciamos la reflexión acerca de cosas como la historia del *Pioneer 10* empezamos a ver cuán desconcertantes son realmente. Nuestras palabras, imágenes, expresiones, y así por el estilo, representan, significan o revelan cosas, pero ¿cómo?

Por un lado, la representación acude naturalmente a nosotros. Cuando hablamos uno con otro, o miramos una imagen, lo que representa suele ser inmediato, y no algo que tengamos que averiguar. Por otra parte, palabras e imá-

[1] Wittgenstein, *Philosophical Investigations* (Oxford, Blackwell, 1953), § 432.

genes no son sino pautas físicas: vibraciones en el aire, marcas en papel, piedra, plástico, película o (como en el *Pioneer 10*) placas metálicas. Tómese el ejemplo de las palabras. Es evidente que no hay nada en las pautas físicas de las palabras mismas que las hagan representar como lo hacen. Los niños a menudo se familiarizan con este hecho cuando repiten palabras para ellos mismos una y otra vez, hasta que parecen "perder" su sentido. Quienquiera que haya aprendido una lengua extranjera reconocerá que, por natural que parezca en el caso de nuestro propio lenguaje, las palabras no tienen significado *en sí y por sí mismas.* O, como dicen los filósofos, no tienen significado "intrínseco".

Por un lado, pues, la representación parece natural, espontánea y sin problemas. Por otra parte, la representación parece innatural, enrevesada y misteriosa. Igual que con los conceptos de tiempo, verdad y existencia (por ejemplo), el concepto de representación plantea un rompecabezas característico de la filosofía: lo que parece natural y evidente en nuestras vidas se torna, al reflexionar, hondamente misterioso.

El problema filosófico de la representación es un tema fundamental de este libro. Es uno de los problemas centrales de la filosofía contemporánea de la mente. Y otros muchos puntos filosóficos se acumulan alrededor de este problema: el lugar de la mente en la naturaleza, la relación entre pensamiento y lenguaje, la naturaleza de nuestra comprensión entre uno y otro, el problema de la conciencia y la posibilidad de máquinas pensantes. Todos estos puntos serán tocados aquí. La meta de este capítulo es aguzar nuestra comprensión del problema de la representación mostrando cómo ciertas soluciones en apariencia evidentes al respecto sólo conducen a nuevos problemas.

La idea de representación

Comenzaré afirmando algunas cosas muy generales acerca de la idea de representación. No hay que alarmarse de enunciar lo evidente: una representación es algo que representa a otra cosa. No digo que una representación sea algo que representa algo *más,* porque una representación puede representarse a sí misma. (Hay un ejemplo famoso de la "Paradoja del Mentiroso": "Esta oración es falsa" representa la oración misma.) El caso normal es donde una cosa —la representación misma— representa otra cosa, que podríamos llamar *objeto* de la representación. Por lo tanto podemos plantear dos cuestiones: una acerca de la naturaleza de las representaciones y una acerca de la naturaleza de los objetos de la representación.

¿Qué clase de cosas pueden ser las representaciones? Ya he mencionado las palabras y las imágenes, que son quizá los ejemplos más evidentes. Por supuesto, hay otros muchos casos. El diagrama de la placa del átomo de hidrógeno en el *Pioneer 10* no es un manojo de palabras ni una imagen, sino que representa el átomo de hidrógeno. Los numerales, como 15, 23, 1 001, etc., representan números. Los números pueden representar otras cosas también: por ejemplo, un numeral puede representar la longitud de un objeto (en metros o en pies) y un triple número puede representar un matiz particular de color representando su grado de matiz, saturación y claridad. Las estructuras de datos de una computadora pueden representar texto o números o imágenes. Los anillos de un árbol pueden representar su edad. Una bandera puede representar una nación. Una demostración política puede representar agresión. Un fragmento de músi-

ca puede representar un talante de insoportable melancolía. Las flores pueden representar la pena. Una ojeada o una expresión facial puede representar irritación. Y, como veremos, un estado mental —una creencia, una esperanza, un deseo o un anhelo— puede representar casi cualquier cosa.

Hay tantas clases de cosas que pueden ser representaciones, que costaría más de un libro discutirlas todas. Y, por supuesto, no pretenderé hacer esto. Enfocaré ejemplos sencillos de representación en el lenguaje y en el pensamiento. Por ejemplo, hablaré acerca de cómo es que puedo utilizar una palabra para representar una persona particular o cómo puedo preguntar (digamos) acerca de un perro. Me fijaré estos sencillos ejemplos porque los problemas filosóficos acerca de la representación surgen incluso en los casos más sencillos. Introducir casos más complejos —tal como de qué manera una pieza de música puede representar una actitud— en esta etapa sólo haría el punto más difícil y confundiría más la mente que antes. Ignorar estos casos complejos no significa que considere que no son importantes o que no tienen interés.[2]

Ahora nuestra segunda cuestión: ¿qué clase de cosas pueden ser objetos de representación? La respuesta es, evidentemente, casi cualquier cosa. Las palabras y las imágenes pueden representar un objeto físico, como una persona o una casa. Pueden representar un rasgo o una propiedad de un objeto físico; por ejemplo, la forma de una persona o el color de una casa. Oraciones como "Alguien está en mi casa" pueden representar lo que podríamos llamar hechos, situaciones o estados de cosas: en este caso, el hecho de que

[2] Acerca de la cuestión, por ejemplo, de cómo puede la música expresar emoción, véase Malcolm Budd, *Music and the Emotions* (Londres, Routledge, 1986).

alguien esté en mi casa. Los objetos no físicos pueden ser representados también: si hay números, no son evidentemente objetos físicos (¿cuándo en el mundo físico se da el número 3?). Representaciones —tales como palabras, imágenes, música y expresiones faciales— pueden representar talantes, sentimientos y emociones. Y las representaciones pueden representar cosas que no existen. Puedo pensar —o sea representar— en unicornios, dragones y el mayor número primo. Ninguna de estas cosas existe; pero pueden todas ser "objetos" de representación.

Este último ejemplo indica un rasgo curioso de la representación. Frente a ella, la expresión *"X* representa *Y"* sugiere que la representación es una *relación* entre dos cosas. Sin embargo, una relación entre dos cosas normalmente implica que estas dos cosas existen. Tómese la relación de *besar:* si beso a Santa Claus, entonces Santa Claus y yo debemos existir ambos. Y el hecho de que Santa Claus no exista explica por qué no puedo besarlo.

Pero esto no es cierto acerca de la representación: si pienso en Santa Claus, y por lo tanto lo represento, no se sigue que Santa Claus exista. La no existencia de Santa Claus no es obstáculo para mi representación de él, como sí lo es para besarlo. De esta manera, la representación parece muy diferente de otras relaciones. Como veremos más adelante, muchos filósofos han tomado este aspecto de la representación por centro de su naturaleza.

Así, hay muchas clases de representaciones, y muchas clases de cosas que pueden ser objetos de representación. ¿Cómo podemos hacer cualquier progreso en la comprensión de la representación? Hay dos clases de preguntas que podemos hacer.

Primero, podemos preguntar *cómo* alguna clase de re-

presentación —imágenes, palabras o lo que sea— consigue representar. Lo que queremos conocer *es* qué pasa con esta clase de representación que hace que tenga su papel representativo. (Como ejemplo considero más adelante la idea de que las imágenes podrían representar cosas *pareciéndose* a ellas.) Evidentemente, no supondremos que la historia de una forma de representación se aplicará por necesidad a todas las demás formas: el modo como las imágenes representan no será el mismo que la manera como la música representa, por ejemplo.

En segundo lugar, podemos preguntar si alguna forma particular de representación es más *básica* o *fundamental* que las otras. Es decir, si podemos explicar algunas clases de representación en términos de otras. Por ejemplo, una cuestión de la filosofía contemporánea es si podemos explicar el modo como el lenguaje representa en términos de capacidades representativas de estados mentales, o si necesitamos explicar la representación mental en términos de lenguaje. Si hay una clase de representación que sea más fundamental que las otras clases, entonces estamos claramente en el camino de comprender la representación en conjunto.

Mi propio punto de vista es que la representación mental —la representación del mundo por estados mentales— es la forma más fundamental de representación. Para ver cómo éste puede ser un punto de vista razonable, necesitamos mirar brevemente una representación pictórica y lingüística.

Imágenes y parecido

Tal como se presentan las cosas, la manera como las imágenes representan parece ser más rectilínea que otras formas

de representación. Pues, en tanto que nada hay de intrínseco en la palabra *perro* que haga que represente perros, de seguro hay algo intrínseco en un cuadro de un perro que hace que parezca un perro, esto es, *aquello a lo cual se asemeja la imagen.* Las imágenes de perros se ven aproximadamente como perros, parecen perros de alguna manera, y lo hacen a causa de sus rasgos intrínsecos: su forma, su color, y así sucesivamente. Tal vez, entonces, una imagen representa lo que representa por parecerse a dicha cosa.

La idea de que una imagen representa pareciéndose sería una respuesta a la primera clase de cuestión mencionada anteriormente: ¿cómo una clase particular de representación consigue representar? La respuesta es: las imágenes representan cosas pareciéndose a dichas cosas. (Esta respuesta puede entonces usarse como fundamento para una respuesta a la segunda cuestión; la sugerencia será que todas las formas de representación pueden ser explicadas en términos de representación pictórica. Pero, como veremos más adelante, esta idea no tiene esperanza.) Llamemos a esta idea "teoría del parecido para la representación pictórica", o "teoría del parecido", por usar pocas palabras. Discutir la teoría del parecido de manera más precisa requiere un poco de terminología filosófica básica.

Los filósofos distinguen entre dos maneras de que la verdad de una opinión dependa de la verdad de otra. Llaman a estas dos maneras condiciones "necesaria" y "suficiente". Decir que una pretensión particular *A* es una condición *necesaria* para alguna otra pretensión *B,* es decir esto: *B* es verdad sólo si *A* es verdad también. Intuitivamente, *B* no será verdad sin que *A* lo sea, de suerte que la verdad de *A* es *necesaria* (es decir, necesitada, requerida) para la verdad de *B.*

Decir que *A* es una condición *suficiente* para *B* es decir

esto: si *A* es verdadera, entonces *B* es verdadera también. Intuitivamente, la verdad de *A* asegura la verdad de *B* o, en otras palabras, la verdad de *A basta* para la verdad de *B*. Decir que *A* es una condición necesaria y suficiente para la verdad de *B* es decir esto: si *A* es verdad, *B* es verdad, y si *B* es verdad, *A* es verdad. (Esto es expresado a veces diciendo que "*A* es verdad si y sólo si *B* es verdad" y "si y sólo si es abreviado en ocasiones como 'sic'".)

Ilustremos esta distinción con un ejemplo. Si estoy en Londres, entonces estoy en Inglaterra. Así, estar en Inglaterra es una *condición necesaria* para estar en Londres: no puedo precisamente estar en Londres sin estar en Inglaterra. Análogamente, estar en Londres es una *condición suficiente* para estar en Inglaterra: estar en Londres bastará para estar en Inglaterra. Pero estar en Londres claramente no es una condición necesaria para estar en Inglaterra, ya que hay muchas maneras de estar en Inglaterra sin estar en Londres. Por la misma razón, estar en Inglaterra no es una condición suficiente para estar en Londres.

La teoría del parecido toma la representación pictórica como dependiente del parecido entre la imagen y lo que representa. Expresemos esta dependencia más precisamente en términos de condiciones necesarias y suficientes: una imagen (llámese *P*) representa algo (llámese *X*) si y sólo si *P* se parece a *X*. O sea que un parecido entre *P* y *X* es tanto necesario como suficiente para que *P* represente *X*.

Esta manera de plantear la teoría de la semejanza es ciertamente más precisa que nuestra inicial formulación vaga. Pero, desgraciadamente, expresarla de esta manera más precisa sólo muestra sus problemas. Tomemos primero la idea de que el parecido podría ser una condición suficiente. Y ni siquiera sólo parecido pictórico.

Decir que el parecido es suficiente para la representación es decir esto: si *X* se parece a *Y,* entonces *X* representa *Y.* La primera cosa que debe llamarnos la atención es que "se parece" es algo vago. Pues, en un sentido, casi todo se parece a todo lo demás. Éste es el sentido en el cual el algo debe precisamente tener algún rasgo en común con dicha cosa. Así, en este sentido, no sólo se parecen mi padre y mi madre, porque yo me parezco a ellos, sino que también se parecen a mi mesa —mi mesa y yo somos ambos objetos— y al número 3: son ambos objetos de una clase o de otra. Pero no soy una representación de ninguna de estas cosas.

Tal vez necesitemos angostar los modos o aspectos según los cuales algo se parece a algo más si queremos que el parecido sea base de la representación. Pero nótese que no ayuda si decimos que si *X* se parece a *Y en algún aspecto* entonces representa *Y.* Pues me parezco a mi padre según ciertos aspectos —digamos rasgos de carácter—, pero esto no me convierte en una representación suya y, evidentemente, no queremos añadir que *X* debe parecerse a *Y* y a esos aspectos según los cuales *X representa Y,* ya que esto haría la teoría del parecido circular y poco informativa: si *X* se parece a *Y* y a estos segundos aspectos en los cuales *X* representa *Y,* entonces *X* representa *Y.* Esto puede ser cierto, pero escasamente será un análisis de la noción de representación.

Hay otro problema en que el parecido sea una condición suficiente. Supóngase que especificamos algunos aspectos según los cuales algo se parece a otra cosa: un retrato de Napoleón, por ejemplo, podría parecerse a Napoleón en la expresión facial, en las proporciones del cuerpo, la posición característica del brazo, y así sucesivamente. Sin embargo, parece ser un hecho evidente acerca del parecido el que si *X* se parece a *Y* entonces *Y* se parece a *X.* (Los filó-

sofos plantean esto diciendo que el parecido es una relación *simétrica.*) Si me parezco a mi padre en algunos aspectos, entonces mi padre se parece a mí en ciertos aspectos. Esto no conduce a representación. Si la imagen se parece a Napoleón entonces Napoleón se parece a la imagen. Napoleón no representa la imagen. Así que no puede ser suficiente el parecido para la representación pictórica si hemos de evitar hacer que cualquier objeto sea él mismo una representación pictórica de su imagen.

Finalmente, debemos considerar el hecho evidente de que todo se parece a sí mismo. (Los filósofos plantean esto diciendo que el parecido es *reflexivo.*) Si se supone que el parecido es suficiente condición para la representación, entonces se sigue que todo se representa a sí mismo. Esto es absurdo. No es satisfactoria una teoría de la representación pictórica que convierta *todo* en imagen de sí misma. Esto trivializa por completo la idea de representación pictórica.

Así, la idea de que el parecido sería una condición suficiente de representación pictórica es algo sin esperanza.[3] ¿Significa esto que la teoría del parecido falle? Todavía no, pues la teoría del parecido podría decir que el parecido no es una condición suficiente, sino una condición necesaria. Esto es, que si una imagen *P* representa *X,* entonces *P* se parecerá a *X* en ciertos aspectos, aunque no viceversa. ¿Qué haremos con esta sugerencia?

Tal como se presenta, parece muy plausible. Si un retrato representa a la reina, entonces de seguro debe de parecérsele en algún aspecto. Después de todo, puede ser que sea una buena imagen para ser un "buen parecido". Sin embar-

[3] Véase Nelson Goodman, *Languages of Art* (Indianapolis, Hackett, 1976), cap. I.

go, hay problemas con esta idea también, pues una imagen puede ciertamente representar a algo sin parecérsele mucho. Abundante arte del siglo xx es representacional; pero esto no quiere decir que se base en parecido (considérense los cuadros cubistas). Las caricaturas y los dibujos esquemáticos, al igual que figuras de palos, tienen muy poco parecido en común con las cosas que representan. Sin embargo no acostumbramos preocuparnos reconociendo lo que representan. Una caricatura de la reina puede parecérsele mucho menos que un dibujo detallado de alguien diferente. Sin embargo, la caricatura sigue siendo una imagen de la reina.[4]

O sea que ¿cuánta semejanza se requiere para satisfacer la condición necesaria de la representación? Tal vez podría responderse que todo lo que se necesita es que haya *alguna* semejanza, por escasa que sea, entre la imagen y lo que representa. Tal vez el parecido puede tomarse con tanta soltura, que incorpore la representación implicada en los cuadros cubistas. Esto está muy bien, pero ahora la idea de parecido no funciona tanto en la teoría como antes. Si una imagen esquemática (digamos de las usadas en algunas corporaciones en sus logos) necesita parecerse a la cosa que representa sólo de una manera mínima, entonces es difícil ver cuánto es explicado diciendo que "si una imagen representa *X,* debe parecerse a *X*". Así, incluso cuando una imagen se parece a aquello que representa, debe haber factores diferentes del parecido que entran en la representación y la hacen posible.

[4] Como dice Wittgenstein: "No es la semejanza lo que hace de la imagen un retrato (pudo ser notablemente parecido a una persona, y sin embargo ser un retrato de alguien a quien se parece menos)". *Philosophical Grammar* (Oxford, Blackwell, 1974), § v.

No estoy negando que las imágenes a menudo se parecen a lo que representan. Evidentemente lo hacen, y esto puede ser parte de lo que las convierte en imágenes (en oposición a las oraciones, las gráficas o los diagramas). Lo único que estoy poniendo en tela de juicio es si la idea de parecido puede *explicar* mucho acerca de cómo representan las imágenes. La idea de que el parecido es una condición necesaria de la representación pictórica bien puede ser cierta, pero la cuestión es "*¿qué más* hace que una imagen represente como lo hace?"[5]

Un punto que requiere ser subrayado aquí es que las imágenes a menudo necesitan interpretación. Por ejemplo, en el *Juicio Final* de Miguel Ángel, en la Capilla Sixtina, vemos a las almas en el infierno luchando angustiadas al ver su final, con la figura monumental del Cristo encima, alzando la mano para juzgar. ¿Por qué no vemos las almas bienvenidas de las profundidades por el Cristo benévolo, con la mano alzada estimulando amistosamente: "hey, vengan, que aquí está más fresco"? (Recuérdese la imagen de la placa metálica del *Pioneer 10* con su mano alzada.) Pues bien, podríamos, pero no es así. La razón es que vemos la imagen a la luz de ciertos supuestos que hacemos al respecto, a los que podríamos vagamente llamar el "contexto" de la imagen. Sabemos que la imagen es una imagen del Juicio Final, y que en este juicio algunas almas son sentenciadas a la condenación eterna, con el Cristo como juez, y así sucesivamente. Esto es parte de por qué vemos la imagen de la manera como la vemos: la interpretamos.

Podemos situar las cosas mediante un ejemplo de Witt-

[5] Aunque Goodman arguye que no es ni siquiera necesario; véase *Languages of Art*, cap. I.

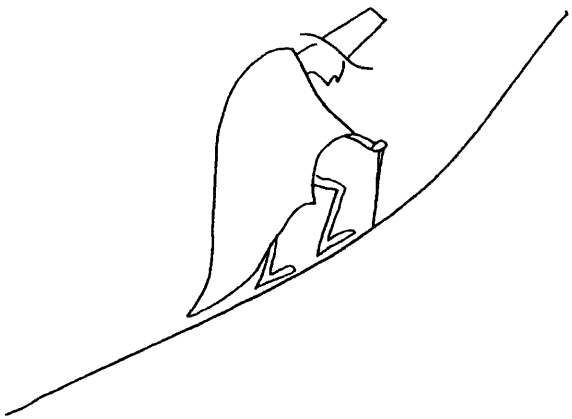


FIGURA 1.1. *Viejo con un bastón*

genstein.[6] Imaginemos la imagen de un hombre con un bastón subiendo una cuesta (véase la figura 1.1). ¿Qué hace que ésta sea la representación de un hombre subiendo por una cuesta, más bien que un hombre deslizándose suavemente cuesta abajo? Nada que haya en la imagen. Es porque estamos acostumbrados a nuestra experiencia cotidiana, y por la clase de contexto en la cual estamos acostumbrados a ver semejantes imágenes, por lo que vemos la imagen de un modo más que de otro. Tenemos que interpretar la imagen a la luz del contexto: la imagen no se interpreta sola.

No voy a perseguir la teoría del parecido o la interpretación de las imágenes en mayor detalle. La menciono aquí para ilustrar cuán poco la idea de parecido nos habla de representación pictórica. Lo que quiero hacer ahora es considerar brevemente la segunda cuestión que planteé al final

[6] *Philosophical Investigations*, p. 54.

de la última sección y aplicarla a la representación pictórica. Podríamos plantear la cuestión así: supongamos que tenemos una teoría completa de la representación pictórica. ¿Sería entonces posible para todas las demás formas de representación ser explicadas en términos de representación pictórica?

La respuesta es "no", por múltiples razones. Una razón a la cual ya hemos echado una ojeada: las imágenes a menudo requieren ser interpretadas y no servirá el decir que la interpretación debiera ser otra imagen, ya que ésta podría requerir interpretación también. Sin embargo, aunque la respuesta sea negativa, podemos aprender algo acerca de la naturaleza de la representación aprendiendo acerca de las limitaciones de la representación pictórica.

Un ejemplo sencillo puede ilustrar el punto. Supóngase que le digo a usted: "Si no llueve esta tarde, saldremos a pasear". Esto es una oración bastante sencilla, una representación lingüística. Pero supóngase que queremos explicar *toda* la representación en términos de representación pictórica; requeriríamos poder expresar la representación lingüística en términos de imágenes. ¿Cómo lograríamos esto?

Bien, quizá pudiéramos dibujar una imagen de una escena no lluviosa con usted y yo paseando en ella. ¿Cómo representamos la idea de "esta tarde"? No podemos poner un reloj en la imagen, pues hay que recordar que estamos tratando de reducir toda representación a imágenes, y un reloj no representa el tiempo pintándolo. (La idea de "pintar" el tiempo en realidad tiene poco sentido.)

Y hay la razón adicional de por qué esta primera imagen no podría ser cierta: es sólo una imagen de usted y yo caminando en una zona sin lluvia. Lo que quisimos expresar era una combinación particular y la relación entre dos ideas:

primero, *no* está lloviendo y, segundo, usted y yo salimos a pasear. Así, tal vez debiéramos trazar dos imágenes: una de la escena libre de lluvia y usted y yo caminando. Esto no puede ser tampoco correcto, pues ¿cómo podría este par de imágenes expresar la idea de que *si* no llueve, *entonces* saldremos a pasear?, ¿por qué no debieran tomarse las imágenes como representación sencillamente de una escena sin lluvia *y* usted y yo saliendo a pasear? O ¿por qué no representa la idea de que *o bien* saldremos de paseo *o* no lloverá? Cuando tratamos de representar la diferencia entre ... *y* ..., *si*... *entonces*..., y... *o bien*... en imágenes, dibujamos un blanco completo. Parece simplemente no haber otra manera de hacerlo.

Una cuestión importante que las imágenes no pueden hacer, pues, es representar ciertas clases de relaciones entre ideas. No pueden representar, por ejemplo, las relaciones que expresamos utilizando las palabras *si*..., *entonces*..., ... *y*..., *o bien*... *o*... y *no*. (¿Por qué *no*? Pues bien, la imagen de la escena sin lluvia puede igualmente ser una imagen de una escena soleada, ¿cómo podemos expresar pictóricamente la idea de que la escena es una escena *sin* lluvia? Tal vez dibujando lluvia y poniendo una cruz —como en un signo de "No fumar"— pero otra vez estamos usando algo que no es una imagen: la cruz.) Por esta razón cuando menos, es imposible explicar o reducir otras formas de representación a la representación pictórica.

Representación lingüística

Una imagen a veces puede valer por mil palabras, pero mil imágenes no pueden representar algunas de las cosas que

podemos representar usando palabras y oraciones. Así, ¿cómo podemos representar cosas usando palabras y oraciones?

Una idea natural es ésta: "Las palabras no representan cosas de ninguna manera natural; más bien representan por *convención.* Hay una convención entre los hablantes de un lenguaje de que las palabras que usan significarán la misma cosa para el uno y el otro; cuando los hablantes convengan o converjan en sus convenciones, lograrán comunicarse; cuando no, no lo harán".[7]

Es difícil negar que lo que las palabras representan es al menos en parte una cuestión de convención. ¿Cuál es la convención, exactamente? Considérese la palabra "perro". ¿Es la idea de que hay una convención entre los hablantes de nuestro idioma para usar la palabra "perro" para representar perros, y sólo perros (mientras se trate de hablar literalmente y hablar con verdad)? Entonces es difícil ver cómo la convención puede *explicar* la representación, según afirmamos la convención como "una convención de usar la palabra 'perro' para *representar* perros". Como la convención se enuncia usando la idea de representación, la da por descontado: no puede explicarla. (De nuevo, mi punto en cuestión no es que la convención no esté implicada en la representación lingüística; la cuestión es si el recurso a la convención puede explicarla de por sí.)

Un pensamiento igualmente natural es que las palabras representan por estar enlazadas convencionalmente a las

[7] Esto evidentemente es una manera muy sencilla de plantear la cuestión. Para más acerca de la convención, véase David Lewis, *Convention* (Oxford, Blackwell, 1969). Sobre el escepticismo acerca del papel de la convención en el lenguaje, véase Donald Davidson, "Communication and Convention", *Inquiries into Truth and Interpretation* (Oxford, Oxford University Press, 1984).

ideas que los pensadores intentan expresar usando estas palabras. La palabra "perro" expresa la idea de un perro por medio de una convención que enlaza la palabra a la idea. Esta teoría tiene una historia filosófica especial: algo parecido a ella se remonta al menos a Thomas Hobbes (1588-1679), y especialmente John Locke (1632-1704), que resumieron el punto de vista diciendo que las palabras son "marcas sensibles de ideas".[8]

¿Qué son ideas? Algunos filósofos han sostenido que son algo como las imágenes mentales: imágenes en la mente. Así, cuando uso la palabra "perro", esto se correlaciona con una imagen mental, en mi mente, de un perro. Una convención asocia la palabra "perro" con la idea que está en mi mente, y es en virtud de esta asociación como la palabra representa perros.

Esta teoría plantea muchos problemas. Sin ir más lejos, ¿es la imagen de mi mente una imagen de un perro particular, digamos Firuláis? En tal caso, ¿suponemos que la palabra "perro" significa *perro*, más bien que *Firuláis*? Además, es difícil suponer qué imagen de "perrería" en general sería.[9] Y aun si la teoría de las imágenes mentales de ideas de alguna manera puede dar razón de este problema, encontrará el problema mencionado al final de la última sección. Aunque muchas palabras pueden asociarse con imágenes mentales, otras muchas no: éste era el problema que teníamos al tratar de explicar *y, o, no* y *sí* en términos de imágenes.

Sin embargo, tal vez no todas las ideas son imágenes mentales —a menudo pensamos en palabras, por ejemplo,

[8] John Locke, *An Essay Concerning Human Understanding* (1689), libro III, cap. 2, § 1.

[9] Véase la crítica por George Berkeley de la doctrina de Locke de las ideas abstractas en *Principles of Human Knowledge* (1710).

y no en absoluto en imágenes—. En tal caso, las críticas a los dos últimos párrafos no dan en el blanco. De modo que pongamos a un lado la teoría de que las ideas son imágenes mentales, y nada más consideremos la pretensión de que las palabras representan expresando ideas, cualesquiera ideas que sean.

Esta teoría no recurre a una "convención para *representar* perros", de modo que no es vulnerable a la misma crítica que la teoría anterior. Pero no puede, por supuesto, explicar la representación, porque atrae ideas, y ¿qué son las ideas sino otra forma de representación? Una idea de perro representa perros tanto como la palabra "perro"; de modo que estamos en efecto recurriendo a una clase de representación (la idea) para explicar otra clase (la palabra). Esto está muy bien, pero si queremos explicar la representación en general, entonces también necesitamos explicar cómo representan las *ideas*.

Tal vez se piense que esto es preguntar demasiado. Tal vez no necesitamos explicar cómo las ideas representan. Si explicamos cómo las palabras representan, asociándolas con ideas, y explicamos también cómo las imágenes pueden ser interpretadas en términos de las ideas que la gente asocia con ellas en sus mentes, quizás podamos detenernos ahí. En resumidas cuentas, no podemos explicarlo todo y tenemos que dar algo por sabido. Así que ¿por qué no tomar las capacidades representacionales de las ideas como cosa establecida?

Pienso que esto es insatisfactorio. Si estamos contentos con dar por sabidos los poderes representacionales de la mente, entonces ¿por qué no dar un paso atrás y tener por sabidos los poderes del lenguaje? Porque no es tan bien entendida la mente como el lenguaje, de hecho en filosofía lo

contrario es lo probablemente cierto. Las ideas, los pensamientos y los fenómenos mentales generalmente parecen hasta más misteriosos que las palabras y las imágenes. De modo que, si acaso, esto sugeriría que debiéramos explicar las ideas en términos del lenguaje, antes que a la inversa. Sin embargo, no pienso que podamos hacer esto. Así es que necesitamos explicar la naturaleza representacional de las ideas.

Antes de pasar a discutir las ideas y la representación mental, quisiera dejar en claro lo que estoy diciendo acerca de la representación lingüística. No estoy diciendo que las nociones que mencioné —de convención, o de palabras que expresan ideas— sean las únicas opciones para una teoría del lenguaje. No, en absoluto. Las introduje sólo como ejemplos de cómo una teoría de la representación lingüística deberá, a fin de cuentas, echar mano de una teoría de la representación mental. Algunas teorías del lenguaje negarán esto, pero ignoraremos estas teorías aquí.[10]

La conclusión de esta discusión es que las palabras, igual que las imágenes, no representan en sí mismas ("intrínsecamente"). Necesitan interpretación, necesitan una interpretación que se les asigne de alguna manera. ¿Cómo podemos explicar esto? La respuesta natural, pienso, es que la interpretación es algo que la *mente* concede a las palabras. Palabras e imágenes ganan las interpretaciones que tienen, y por lo tanto representan lo que representan, en virtud de los estados mentales de quienes las usan. Estos estados menta-

[10] Véase, por ejemplo, el intento de Davidson de dilucidar el significado lingüístico en términos de verdad: *Inquiries into Truth and Interpretation* (Oxford, Oxford University Press, 1984). Vista de conjunto, Barry C. Smith, "Understanding Language", *Proceedings of the Aristotelian Society*, 92 (1992).

les son representacionales también. De manera que para comprender cabalmente la representación lingüística y pictórica tenemos que comprender la representación mental.

Representación mental

Así que ¿cómo representa la mente algo? Hagamos la cuestión un poco más fácil de manejar preguntando cómo los *estados* individuales de la mente representan algo. Por un "estado de la mente", o "estado mental", quiero decir aquí algo como una creencia, un deseo, una esperanza, un miedo, una conjetura, una expectación, una intención, una percepción, y así por el estilo. Pienso que todos estos son estados de la mente que representan el mundo de alguna manera. Esto requerirá un poco de explicación.

Cuando digo que esperanzas, creencias, deseos y demás representan el mundo, quiero decir que toda esperanza, creencia o deseo *se dirige a* algo. Si se espera, hay que esperar *algo;* si se cree se debe creer en *algo;* si se desea debe desearse *algo.* No tiene sentido suponer que una persona podría simplemente esperar, sin esperar algo; creer sin creer algo; o desear sin desear algo. Lo que usted crea o desee es lo que es representado por su creencia o deseo.

Necesitaremos un término general conveniente para estados de la mente que representen el mundo, o un aspecto del mundo. Usaré el término "pensamiento", ya que parece ser el término más general y neutro perteneciente al vocabulario mental cotidiano. De aquí en adelante en este libro usaré el término "pensamiento" para aludir a todos los estados mentales representacionales. Así, los estados de creencia, deseo, esperanza, amor y demás son todos pensamientos en

mi sentido, ya que todos representan cosas. (Cuándo todos los estados mentales son pensamientos en este sentido, es cuestión que dejaré hasta el final del capítulo.)

¿Qué podemos decir en general acerca de cómo los pensamientos representan? Comenzaré con pensamientos que son de particular interés; esos pensamientos que representan *situaciones* (o tratan de ellas). Cuando espero que haya sopa de mariscos en el menú en mi restaurante esta noche, estoy pensando en diversas cosas: la sopa, el menú, mi restaurante favorito, esta noche. Pero no estoy pensando sólo en estas cosas de un modo casual o desconectado: estoy pensando en cierto hecho o *situación* posibles: la situación en la cual la sopa está en el menú de mi restaurante favorito esta noche. Es una variante inocua de esto el decir que mi estado de esperanza *representa* esta situación.

Sin embargo, considérese un pensamiento diferente que podría tener: la *creencia* de que hay sopa de mariscos en el menú esta noche. Este estado mental no representa la situación del mismo modo como la esperanza lo hace. Cuando creo que hay sopa de mariscos en el menú esta noche (porque he pasado por el restaurante y leído el menú), tomo la situación en cuestión como un hecho: tomo como un hecho acerca del mundo que hay sopa de mariscos en el menú esta noche. Pero cuando espero, no lo esperaré como un hecho acerca del mundo; más bien me gustaría que fuera un hecho que hubiese sopa de mariscos en el menú esta noche.

Así que hay dos aspectos de estos pensamientos: hay la "situación" representada y hay lo que podríamos llamar (a falta de una palabra mejor) la *actitud* que asumimos ante la situación. La idea de diferentes actitudes ante situaciones es ilustrada mejor mediante ejemplos.

Considérese la situación de que yo visite Budapest. Pue-

do esperar que visitaré Budapest, y puedo creer que he visitado Budapest. Todos estos pensamientos tratan o representan la misma situación —mi visita a Budapest— pero las actitudes asumidas ante esta situación son muy diferentes. La cuestión por lo tanto surge acerca de lo que hace diferentes a estas distintas actitudes; pero por el momento sólo me importa distinguir la situación representada a partir de la actitud asumida ante ella.

Así como la misma situación puede ser tema de diferentes actitudes, así el mismo género de actitud puede concernir a muchas situaciones diferentes. Realmente creo que visitaré Budapest pronto, y también creo que mi restaurante favorito no tiene sopa de mariscos en el menú de esta noche, y creo otras innumerables cosas. Creencias, esperanzas y pensamientos de este género pueden ser captados únicamente especificando: *a)* la actitud en cuestión (creencia, esperanza, expectación, etc.) y *b)* la situación representada. (Debe también notarse, de paso, que muchas actitudes llegan en grados: uno puede querer algo más o menos intensamente; y creer algo con más o menos convicción; pero esta complicación no afecta al cuadro general.) En general podemos describir estas clases de pensamientos esquemáticamente como sigue. Si *"A"* es la persona que se halla en el estado mental, "ψ" representa la actitud (la letra psi, por "psicológico") y *"S"* ocupa el lugar de la situación representada, la mejor descripción será de la siguiente forma:

Un ψ que S

Por ejemplo, Vladimir (*A*) cree (ψ) que está lloviendo (*S*); Renata (*A*) espera (ψ) que visitará Rumania (*S*), y así sucesivamente.

Bertrand Russell (1872-1970) llamó a los pensamientos que pueden ser pescados de esta manera "actitudes proposicionales", y esta denominación ha corrido con suerte.[11] Aunque puede resultar más bien oscuro al primer golpe de vista, el término "actitud proposicional" describe la estructura de estos estados mentales sumamente bien. Ya he explicado el término "actitud". Lo que Russell quiere decir con "proposición" es algo parecido a lo que estoy llamando "situación": es aquello hacia lo cual se dirige la atención (así, una proposición en este sentido no es un fragmento del lenguaje). Una actitud proposicional es por lo tanto cualquier estado mental que pueda describirse en el estilo "*A* ψ que *S*".

Otro fragmento de terminología que ha sido casi universalmente adoptado es el término "contenido", usado donde Russell usó "proposición". De acuerdo con esta terminología, cuando creo que hay cerveza en el refrigerador, el *contenido* de mi creencia es que *hay cerveza en el refrigerador*. Y análogamente con deseos, esperanzas y así sucesivamente, éstas son actitudes diferentes, pero todas tienen "contenido". Qué es exactamente "contenido" y qué es para un estado mental tener "contenido" (o "contenido representacional") son cuestiones que retornarán durante el resto de este libro, especialmente en el capítulo v. En la filosofía del momento, el problema de la representación mental se expresa a menudo como: "¿Qué es para un estado mental tener contenido?" Por el momento, podemos pensar en el contenido de un estado mental como aquello que distingue estados que implican la misma actitud el uno para el otro.

[11] Russell usó la expresión en *The Analysis of Mind* (Londres, George Allen and Unwin, 1921), cap. 12. Para una colección de lecturas, véase Nathan Salmon y Scott Soames (eds.), *Propositions and Attitudes* (Oxford, Oxford University Press, 1988).

Diferentes creencias se distinguen una de otra (o, en terminología filosófica, son "individuadas") por sus diferentes contenidos. Así son los deseos; y así con todas las actitudes.

Me he concentrado en la idea de una actitud proposicional, porque los pensamientos de esta forma se volverán muy importantes en el capítulo siguiente. Y aunque todas las actitudes proposicionales sean pensamientos (por definición), es importante recalcar que no todos los pensamientos (en mi sentido) son actitudes proposicionales, esto es, no todos los estados mentales representacionales pueden caracterizarse en términos de actitudes ante situaciones. Tómese el amor, por ejemplo. El amor es un estado mental representacional: no se puede amar sin amar algo o a alguien. Sin embargo, el amor no es (siempre) una actitud ante una situación; el amor puede ser una actitud hacia una persona, un lugar o una cosa. El amor no puede ser descrito en el estilo de "*A* ψ que *S*" (inténtese y véase). En mi terminología, entonces, el amor es una clase de pensamiento, pero no una actitud proposicional.[12]

Otro interesante ejemplo es el deseo. ¿Es éste una actitud hacia una situación? Vistas las cosas, no lo es. Supóngase que deseo una taza de café; mi deseo es por una cosa, una taza de café, no por ninguna situación. En la superficie, pues, el deseo se asemeja al amor. Muchos filósofos piensan que esto es engañoso, y que describe poco de un deseo para tratarlo como una actitud hacia una cosa. La razón es que una descripción más exacta del deseo es que es un deseo de que se dé una situación determinada: la situación en la cual *yo tengo una taza de café.* Todos los deseos, según se sostie-

[12] Para más acerca de este tema, véase Tim Crane, *Elements of Mind* (Oxford, Oxford University Press, 2001), § 34.

ne, son realmente deseos de *que esto y lo otro,* donde "esto y lo otro" es una especificación de una situación. El deseo, a diferencia del amor, es una actitud proposicional.

Ahora, denominando "pensamientos" a los estados mentales representacionales, no quiero dar a entender que estos estados sean necesariamente conscientes. Supóngase que Edipo realmente no desea matar a su padre ni casarse con su madre. Entonces, según el criterio esbozado antes ($A\ \psi$ que S), estos deseos cuentan como actitudes proposicionales y por lo tanto pensamientos. No obstante, no son pensamientos conscientes.

Podría parecer extraño distinguir entre pensamiento y conciencia de esta manera. Para justificar la distinción, necesitamos una breve digresión preliminar sobre el tema turbio de la conciencia; un tratamiento cabal de este asunto habrá de esperar hasta el capítulo VI.

Pensamiento y conciencia

La conciencia es lo que hace que nuestras vidas despiertas parezcan ser como son, y puede sostenerse que también es la fuente última de todo valor del mundo: "Sin esta iluminación interna —dijo Einstein al filósofo Hebert Feigl— "el universo no sería sino un montón de tierra".[13] Sin embargo, a pesar de la importancia de la conciencia, quiero distinguir ciertas cuestiones acerca del pensamiento a partir de preguntas sobre la conciencia. En cierta medida, estas cuestiones son independientes una de otra.

[13] Citado en H. Feigl, *The "Mental" and the "Physical"* (Minneapolis, University of Minnesota, 1967), p. 138.

Como digo, esto puede parecer un poco extraño. Después de todo, para mucha gente, las expresiones "pensamiento" y "conciencia" son prácticamente sinónimas. De seguro pensar es tener conciencia del mundo, tener conciencia de cosas en uno mismo y fuera, ¿cómo pues podemos comprender el pensamiento sin entender también la conciencia? (Algunas personas incluso creen que las expresiones "consciente" y "mental" son sinónimas; para ellos el punto es incluso más evidente.)

La razón para distinguir el pensamiento y la conciencia es muy sencilla. Muchos de los pensamientos son conscientes, pero no todos ellos lo son. Algunas de las cosas que pensamos son inconscientes. Así, si el pensamiento puede aun ser *pensado* no siendo consciente, entonces no puede *en general* ser esencial para que algo sea un pensamiento que sea consciente. Debe por lo tanto ser posible explicar lo que hace del pensamiento lo que es sin tener que explicar la conciencia.

¿Qué quiero decir cuando digo que algún pensamiento es inconsciente? Simplemente esto: hay cosas que pensamos, pero no tenemos *conciencia* de que las pensamos. Permítaseme dar unos cuantos ejemplos, algunos más controvertidos que otros.

Yo aceptaría apostar que usted cree que el presidente de los Estados Unidos normalmente lleva calcetines. Si yo le preguntara "¿Lleva el presidente de los Estados Unidos normalmente calcetines?", creo que usted contestaría que sí. Y lo que la gente dice es testimonio muy aceptable de lo que piensa: así, yo tomaría la respuesta de usted como buen testimonio del hecho de que cree usted que el presidente de los Estados Unidos normalmente usa calcetines. También conjeturaría yo que las palabras "El presidente de los Estados Unidos normalmente usa calcetines" nunca se presen-

taron a la mente consciente de usted. Es muy probable que la cuestión del calzado del presidente no haya surgido *conscientemente* para usted; nunca ha tenido *conciencia* de pensar en ello. Y sin embargo, al preguntarle, parece usted revelar que piensa que es cierto. ¿Sólo empezó usted a pensar en ello cuando le pregunté? ¿Puede realmente ser cierto decir que no tenía usted opinión a este respecto antes de que yo le preguntara? ("Hum, éste es un asunto interesante, nunca le había yo concedido ningún pensamiento antes, me pregunto cuál será la respuesta...") ¿Tiene más sentido decir que el pensamiento inconsciente estuvo presente sin cesar?

Este ejemplo podría parecer no poco trivial, de manera que ensayemos otro, más significativo (y controvertido). En el *Menón,* diálogo platónico, Sócrates trata de defender su teoría de que todo conocimiento es reminiscencia de verdades conocidas en la vida previa del alma. Para persuadir a su interlocutor (Menón) de esto, Sócrates interroga a uno de los esclavos de Menón acerca de un fragmento sencillo de geometría: si el área de un cuadrado con lados de N unidades de largo es cierto número de unidades, ¿cuál es el área de un cuadrado con lados de $2 \times N$ unidades de largo? Interrogado sencillamente (sin dejar pasar nada), el esclavo de Menón casi siempre logra la respuesta correcta. El diálogo continúa:

Sócrates: ¿Qué piensas, Menón? ¿Ha contestado con alguna opinión que no fuera propia?
Menón: No, fueron todas suyas.
Sócrates: Sin embargo no sabía, según convinimos hace unos minutos.
Menón: Cierto.

Sócrates: Pero estas opiniones estuvieron en algún lugar de él, ¿no es así?
Menón: Sí.[14]

Sócrates, por lo tanto, arguye que el conocimiento es recuerdo, pero éste no es el punto de vista que me interesa aquí. Lo que me interesa es la idea de que uno puede tener una especie de "conocimiento" o (digamos) ciertos principios matemáticos "en algún lugar" de uno, sin ser explícitamente consciente de ellos. Esta clase de conocimiento puede ser "recuperado" (por usar la palabra de Sócrates) y hecha explícita, pero puede también yacer dentro de la mente de alguien sin ser recuperado nunca. El conocimiento implica el pensamiento de algo; es una clase de pensamiento. Así, si puede haber conocimiento inconsciente, puede haber pensamiento inconsciente.

Hay algunas dificultades terminológicas para hablar acerca de "pensamientos inconscientes". Para algunas personas, los pensamientos son episodios en la mente consciente, de modo que deben ser conscientes por definición. Ciertamente, muchos filósofos han pensado que la conciencia era esencial para todos los estados mentales, y por lo tanto para los pensamientos. Descartes fue uno: para él la idea de pensamiento inconsciente habría sido una contradicción en los términos. Y algunos siguen concordando con él.[15]

Sin embargo, creo que estos días muchos más filósofos (y no filósofos también) están preparados para tomar muy seriamente la idea del pensamiento inconsciente. Una in-

[14] *Menón* en Hamilton y Cairns (eds.), *Plato: Collected Dialogues* (Princeton, Princeton University Press, 1961), p. 370.

[15] Véase John R. Searle, *The Rediscovery of the Mind* (Cambridge, MIT Press, 1992), cap. 7.

fluencia aquí es la contribución de Freud a la concepción moderna de la mente. Freud reconoció que de muchas de las cosas que hacemos no puede darse plena razón por nuestras mentes conscientes. Lo que explica estas acciones son nuestras creencias y deseos *inconscientes*, muchos de los cuales están "enterrados" tan profundamente en nuestras mentes que necesitamos cierto tipo de terapia —el psicoanálisis— para sacarlos a la luz.[16]

Nótese que podemos aceptar esta pretensión freudiana sin aceptar detalles específicos sobre la teoría de Freud. Podemos aceptar la idea de que nuestras acciones pueden a menudo ser gobernadas por creencias y deseos inconscientes, sin aceptar muchas de las ideas (asociadas popularmente al nombre de Freud) acerca de cuáles son estas creencias y deseos, y cuál es su causa, por ejemplo el complejo de Edipo o "envidia del pene". De hecho, la idea esencial es muy cercana a nuestra manera ordinaria de pensar acerca de la mente de otra gente. Todos conocemos gente de quien pensamos que no "está en sus cabales" o que se engaña acerca de algo. ¿Cómo podrían no tener noción de sus propios pensamientos, si los pensamientos son esencialmente conscientes?

En todo caso, por todas estas razones, creo que hay pensamientos inconscientes, y también pienso que no necesitamos comprender la conciencia a fin de entender el pensamiento. Esto no significa que esté negando que hay cosa tal como el pensamiento consciente. Los ejemplos discutidos eran ejemplo de pensamientos *traídos* a la conciencia: se tra-

[16] La idea del inconsciente en este sentido es más antigua que Freud; para una interesante discusión, véase Neil Manson, "'A Tumbling Ground for Whimsies?' The History and Contemporary Relevance of the Conscious/Unconscious Contrast", en Tim Crane y Sarah Patterson (eds.), *History of the Mind-Body Problem* (Londres, Routledge, 2000).

jo a la conciencia de usted el pensamiento de que el presidente de los Estados Unidos usa calcetines, el esclavo de Menón llevó a su mente consciente el conocimiento geométrico que no tenía noción de poseer, y los pacientes del psicoanálisis traen a su mente consciente pensamientos y sentimientos que no saben que tienen. Y muchos de los ejemplos que emplearé a través del libro serán pensamientos conscientes. Empero, en lo que estoy interesado es en lo que los hace *pensamientos,* no en lo que los hace *conscientes.*

En su libro bien conocido, *The Emperor's New Mind,* el matemático y físico Roger Penrose supone que "la verdadera inteligencia requiere conciencia".[17] Puede parecer como si no conviniera yo con esta observación; no obstante, en realidad no es ése el caso. Decir que la verdadera inteligencia (o pensamiento) requiere conciencia no significa que para comprender la naturaleza del pensamiento tengamos que comprender la naturaleza de la conciencia. Significa nada más que cualquier cosa que podamos pensar debe también ser consciente. Una analogía acaso ayude: puede ser cierto que cualquier cosa que piensa, o es inteligente, debe estar viva. Quizá. En tal caso, entonces la "auténtica inteligencia requiere vida". No obstante, esto no significaría *por sí mismo* que a fin de comprender el pensamiento tengamos que comprender la vida. Tendríamos sencillamente que presuponer que las cosas que piensan están también vivas. Nuestra explicación del pensamiento no sería también una explicación de la vida. Y similarmente con la conciencia. Así, no estoy discrepando de la observación de Penrose. No obstante tampoco es que convenga

[17] Roger Penrose, *The Emperor's New Mind* (Londres, Vintage, 1990), p. 526. [Ed. en el FCE: *La mente nueva del emperador,* México, 1996.]

con él. Estoy manteniéndome neutral en esta cuestión, porque no sé si podría haber algún ser que tuviese pensamientos que fuesen totalmente inconscientes. Sin embargo, por fortuna, no necesito responder esta difícil cuestión a fin de llevar adelante los temas de este libro.

Hasta aquí, pues, la idea de que muchos pensamientos son inconscientes. Ya es tiempo de regresar a la idea de la representación mental. ¿Qué hemos aprendido acerca de la representación mental? Hasta aquí, no gran cosa. Sin embargo, describiendo en términos muy generales la noción de un *pensamiento* y articulando la distinción entre *actitud* y *contenido* (o *situación*), hemos logrado un principio. Ahora cuando menos tenemos algunas categorías básicas con qué trabajar, al plantear nuestra pregunta acerca de la naturaleza de la representación mental. En el siguiente apartado enlazaré la discusión hecha hasta aquí con algunas ideas importantes de la tradición filosófica.

Intencionalidad

Los filósofos tienen una palabra técnica para la naturaleza representacional de estados de la mente: hablan de la "intencionalidad". Los estados mentales que exhiben intencionalidad —los que representan— son a veces, por lo tanto, llamados "estados intencionales". Esta terminología puede ser confusa, especialmente porque no todos los filósofos usan los términos del mismo modo. Pero es necesario considerar el concepto de intencionalidad, ya que constituye el punto de partida de la mayoría de los intentos de los filósofos para vérselas con el rompecabezas de la representación.

El término "intencionalidad" deriva de los filósofos es-

colásticos de la Edad Media, que estaban muy interesados en cosas acerca de la representación. Estos filósofos usaban el término *intentio* para aludir al concepto, y el término *esse intentionale* (existencia intencional) —por ejemplo, santo Tomás de Aquino (*ca.* 1225-1274)— para el modo como las cosas pueden ser representadas conceptualmente en la mente. El término "existencia intencional" (o "inexistencia") fue revivido por el filósofo alemán Franz Brentano (1838-1917). En su libro *Psicología desde un punto de vista empírico* (1874), Brentano sostuvo que los fenómenos mentales son caracterizados

> por lo que los escolásticos de la Edad Media llamaron intencional... inexistencia del objeto, y que nosotros, aunque con expresiones no del todo sin ambigüedad, llamaríamos relación con un contenido, dirección hacia un objeto (que no está aquí para ser entendido como realidad) u objetividad inmanente.[18]

Las cosas son más sencillas aquí de lo que parecerían al golpe de vista. Las frases "inexistencia intencional", "relación con un contenido" y "objetividad inmanente", a despecho de diferencias superficiales entre ellas, son todas diferentes maneras de expresar la misma idea: que los fenómenos mentales implican representación o presentación del mundo. "Inexistencia" se toma para expresar la idea de que el objeto de un pensamiento —de qué trata el pensamiento— existe

[18] Franz Brentano, *Psychology from an Empirical Standpoint* (traducido por Rancurello, Terrell y McAlister; Londres, Routledge and Kegan Paul, 1973), p. 88. Para más acerca de los orígenes del término "intencionalidad", véase mi artículo "Intentionality", *Routledge Encyclopedia of Philosophy* (Londres, Routledge, 1998).

en el acto de pensar, en sí mismo. Esto no es decir que cuando pienso en mi perro haya un perro "en" mi mente. Más bien, es precisamente la idea de que mi perro es *intrínseco* a mi pensamiento, en el sentido de que lo que lo hace el pensamiento que es, es el hecho de que tenga a mi perro como objeto.

Empezaré entendiendo la idea de intencionalidad tan sencillamente como se pueda, como *direccionalidad acerca de algo.* Los filósofos contemporáneos usan a medida el término "acercaidad" como un sinónimo de "intencionalidad": los pensamientos tienen "acercaidad" porque son *acerca* de cosas. (Prefiero el término "direccionalidad", por razones que no tardarán en aparecer.) La esencia de la pretensión de Brentano es que lo que distingue fenómenos mentales de fenómenos físicos es que mientras todos los fenómenos mentales exhiben esta direccionalidad ningún fenómeno físico la exhibe. Esta pretensión, de que la intencionalidad es la "marca de lo mental", se llama a veces *tesis de Brentano.*

Antes de considerar si la tesis de Brentano es verdadera, necesitamos aclarar un par de posibles confusiones acerca de la expresión "intencionalidad". La primera es que el mundo se ve como si pudiera tener algo que ver con las ideas ordinarias de *intención,* y actuar *intencionalmente.* Hay evidentemente un enlace entre la idea filosófica de la intencionalidad y la idea de intención. Por un lado, si pretendo ejecutar alguna acción, *A,* es entonces natural pensar que represente a *A* (en algún sentido) para mí mismo. De modo que las intenciones pueden ser estados representacionales (y por ende "intencionales").

Empero, aparte de estas conexiones, no hay enlace filosófico sustancial entre el concepto de intencionalidad y el concepto ordinario de intención. Las intenciones en sentido or-

dinario son estados intencionales, pero la mayoría de los estados intencionales poco tienen que hacer con las intenciones.

La segunda confusión posible es algo más técnica. Los principiantes quizá prefieran pasar directamente al apartado siguiente, "La tesis de Brentano" (véase p. 73).

Esta segunda confusión es entre la intencionalidad (en el sentido que estoy usando aquí) y la *intensionalidad,* rasgo de algunos contextos lógicos y lingüísticos. Las palabras "intensionalidad" e "intencionalidad" se pronuncian de la misma manera, lo que aumenta la confusión y conduce a autores cuidadosos, como John Searle, a especificar si están hablando de "intencionalidad-con-c" o "intensionalidad-con-s".[19] Searle tiene razón: la intencionalidad y la intensionalidad son cosas diferentes y es importante mantenerlas separadas en nuestra mente.

Para ver por qué, necesitamos introducir algún vocabulario técnico de la lógica y la filosofía del lenguaje. Un contexto lingüístico o lógico (esto es, una parte del mismo lenguaje o cálculo lógico) es intensional cuando no es *extensional.* Un contexto extensional es uno acerca del cual son verdaderos los siguientes principios:

(A) El principio de intersustitutividad de las expresiones correferentes.

(B) El principio de generalización existencial.

Los títulos de estos principios parecen bastante formidables, pero las ideas lógicas que hay detrás son bien sencillas. Me explicaré.

[19] Véase John R. Searle, *Intentionality* (Cambridge, Cambridge University Press, 1983).

El principio (A) de intersustitutividad de las expresiones correferentes es un título bastante complicado para una idea muy sencilla. La idea es, ni más ni menos, que si un objeto tiene dos nombres, *N* y *M,* y se dice algo verdadero acerca de ello usando *M,* no puede convertirse esta verdad en una falsedad sustituyendo *M* por *N.* Por ejemplo, el nombre original de George Orwell era Eric Arthur Blair (tomó el nombre de Orwell del Río Orwell, en Suffolk). Como ambos nombres se refieren al mismo hombre, no puede cambiarse el enunciado verdadero:

George Orwell escribió *Animal Farm*

en una falsedad sustituyendo el nombre George Orwell por el nombre Eric Arthur Blair. Porque el enunciado:

Eric Arthur Blair escribió *Animal Farm*

es igualmente cierto. (Asimismo, sustituir George Orwell por Eric Arthur Blair no puede convertir una falsedad en una verdad; por ejemplo "George Orwell escribió *La guerra y la paz*".) La idea que hay detrás de esto es muy sencilla: porque la persona de quien está usted hablando es la misma en ambos casos, no importa a la verdad de lo que usted dice qué palabras use para hablar de él.

Los términos "George Orwell" y "Eric Arthur Blair" son "términos correferentes": esto es, se refieren al mismo objeto. El principio (A) dice que estos términos pueden ser sustituidos uno por otro sin cambiar la verdad o falsedad de la oración en la cual se presentan (por lo tanto es llamado a veces principio de "sustitutividad *salva veritate*", literalmente, "salvando la verdad").

¿Qué puede ser más sencillo? Desgraciadamente no tenemos que buscar mucho en pos de casos en que este simple principio es violado. Considérese a alguien —llamémoslo Vladimir— que cree que George Orwell escribió *Animal Farm* pero ignora el nombre original de Orwell. Entonces el enunciado:

Vladimir cree que George Orwell escribió *Animal Farm*

es cierto, en tanto que el enunciado:

Vladimir cree que Eric Arthur Blair escribió *Animal Farm*

es falso. La sustitución de los términos correferentes no preserva, en este caso, la verdad. Nuestro principio aparentemente evidente de la sustitutividad de los términos correferentes ha fracasado. ¿Cómo puede fallar este principio? Parece evidente por sí mismo.

Por qué este principio falla en determinados casos —notablemente en oraciones acerca de creencias y otros determinados estados mentales— es un problema de la filosofía del lenguaje. Sin embargo, no necesitamos insistir en las razones del fracaso aquí; sólo deseo señalarlo con el propósito de definir el concepto de intensionalidad. El fracaso del principio (A) es una de las marcas de la no extensionalidad, o intensionalidad.

La otra marca es el fracaso del principio (B), "generalización existencial". Este principio dice que podemos inferir que algo existe a partir de un enunciado hecho acerca de ello. Por ejemplo, a partir del enunciado:

Orwell escribió *Animal Farm*

podemos inferir que:

existe alguien que escribió *Animal Farm.*

Esto es, si el primer enunciado es verdadero, el segundo es verdadero también.

Otra vez, un ejemplo prominente de dónde puede fallar la generalización existencial, son los enunciados acerca de creencias. El enunciado

Vladimir cree que Santa Claus vive en el Polo Norte

puede ser cierto, en tanto que el siguiente enunciado es sin duda falso:

Existe alguien que Vladimir cree que vive en el Polo Norte.

En vista de que el primero de estos dos enunciados puede ser cierto en tanto que el segundo es falso, el segundo no puede seguirse lógicamente a partir del primero. Éste es un ejemplo del fracaso de la generalización existencial.

Resumiendo: la intensionalidad es un rasgo de las oraciones y cuestiones lingüísticas; una oración es intensional cuando no es extensional; es no extensional cuando uno o dos de sus principios (A) y (B) puede no ser aplicable. Nótese que digo que los principios *pueden* fallar en su aplicación, no que deban. Por supuesto, hay muchos casos en los que podemos sustituir expresiones correferentes en oraciones de creencia; y hay muchos casos en los cuales puede concluirse que algo existe a partir de una oración de creencia que trata de esa cosa. Empero, la cuestión es que no tenemos *garantía* de que estos principios se mantengan

para todas las oraciones de creencia y otros "contextos intensionales".

¿Qué tiene que ver esta intensionalidad con nuestro asunto, la intencionalidad? A primera vista hay una evidente conexión. Los ejemplos que usamos de oraciones que exhiben intensionalidad eran oraciones acerca de creencias. Es natural suponer que el principio de sustitutividad de los términos correferentes fracasa aquí porque el que una oración de creencia sea verdadera depende no nada más del *objeto representado* por el creyente, sino del *modo* como el objeto es representado. Vladimir representa a Orwell *como Orwell,* y no *como Blair.* Así, la intensionalidad parece ser un resultado de la naturaleza de la representación que interviene en una creencia. Tal vez, entonces, la intensionalidad de las *oraciones* de creencia es una consecuencia de la intencionalidad de las creencias mismas.

Igualmente con el fracaso de la generalización existencial. El fracaso de este principio en el caso de las oraciones de creencia es tal vez una consecuencia natural del hecho (mencionado antes) de que las representaciones pueden representar "cosas" que no existen. El hecho de que podamos pensar acerca de cosas que no existen parece ser una de las características definitorias de la intencionalidad. Así, una vez más, tal vez, la intensionalidad de (por ejemplo) las *oraciones* de creencia es una consecuencia de la intencionalidad de las creencias mismas.[20]

Sin embargo, esto es lo más lejos que podemos llegar asociando las nociones de intensionalidad e intencionalidad. Hay dos razones por las cuales no podemos ligar las dos nociones más estrechamente:

[20] Para más acerca de la distinción entre intencionalidad e intensionalidad, véase *Elements of Mind,* §§ 4 y 35.

1. *Puede haber intensionalidad sin intencionalidad (representación).* Esto es, puede haber oraciones que sean intensionales pero no tengan nada que ver con la representación mental. Los ejemplos mejor conocidos son oraciones que implican las nociones de *posibilidad* y *necesidad.* Decir que algo es necesariamente así, en este sentido, es decir que no podría haber sido de otro modo. De las dos oraciones verdaderas:

   Nueve es necesariamente mayor que cinco
   El número de planetas es nueve

no podemos inferir que:

   El número de planetas es necesariamente mayor que cinco

   puesto que no es necesariamente verdad que haya nueve planetas. Podía haber sólo cuatro planetas, o ninguno. Así que el principio de sustitutividad de los términos correferentes ("nueve" y "el número de planetas") falla, pero no por nada que tenga que ver con la representación mental.[21]
2. *Puede haber descripciones de intencionalidad que no exhiban intensionalidad.* Un ejemplo es dado por oraciones de la forma "*X* ve a *Y*". Ver es un caso de intencionalidad, o representación mental. No obstante, si Vladimir ve a Orwell, entonces de fijo ve también a Blair, y al autor de *The Road to Wigan Pier,* y así sucesivamente. El

[21] Véase W. V. Quine, "Reference and Modality" y "Quantifiers and Propositional Attitudes", en L. Linsky (ed.), *Reference and Modality* (Oxford, Oxford University Press, 1971).

principio (A) parece aplicarse a "*X* ve a *Y*". Además, si Vladimir ve a Orwell, entonces con seguridad hay alguien a quien ve. Así el principio (B) se aplica a oraciones de la forma "*X* ve a *Y*".[22] No todas las descripciones de la intencionalidad son intensionales; de modo que la intensionalidad en la descripción no es necesaria para la intencionalidad para ser descrita.

El último argumento *(2)* es en realidad bastante cuestión de controversia, pero realmente no lo necesitamos para diferenciar la intencionalidad de la intensionalidad. El primer argumento será el que para nosotros se encargará de ello por su cuenta: en la terminología de las condiciones necesarias y suficientes antes introducida, podemos decir que la intensionalidad no es suficiente para la intencionalidad, y puede ni siquiera ser necesaria. Esto es, puesto que es posible tener intensionalidad sin mencionar para nada la intencionalidad, la intensionalidad no es suficiente para la presencia de intencionalidad. Esto es suficiente para mostrar que éstos son conceptos muy diferentes, y que no podemos usar la intensionalidad como criterio de intencionalidad.[23]

Dejemos ahora atrás la intensionalidad, y volvamos a nuestro tema principal: la intencionalidad. Nuestra última tarea en este capítulo será considerar la tesis de Brentano de que la intencionalidad es la "marca' de lo mental.

[22] Véase Fred Dretske, *Seeing and Knowing* (Londres, Routledge and Kegan Paul, 1969), cap. 1.

[23] Estas observaciones van dirigidas contra Quine: véase *Word and Object* (Cambridge, MIT Press, 1960), especialmente pp. 219-221.

La tesis de Brentano

Según se observó antes, Brentano pensó que todos y sólo los fenómenos mentales exhiben intencionalidad. Esta idea, la tesis de Brentano, ha sido muy influyente en la filosofía reciente. No obstante, ¿es verdadera?

Dividamos la cuestión en dos subcuestiones:

1. ¿Exhiben todos los estados mentales intencionalidad?
2. ¿Sólo los términos mentales exhiben intencionalidad?

Otra vez es útil la terminología de las condiciones necesarias y suficientes. La primera subcuestión puede ser reformulada: ¿es la mentalidad suficiente para la intencionalidad? Y la segunda: ¿es la mentalidad necesaria para la intencionalidad?

Es tentador pensar que la respuesta a la primera subcuestión es que no. Decir que todos los estados mentales exhiben intencionalidad es decir que todos los estados mentales son representacionales. Sin embargo —sigue diciendo esta línea de pensamiento—, podemos saber por introspección que muchos estados mentales no son representacionales. Supóngase que tengo un intenso dolor en la base de mi espinazo. Este dolor es un estado mental: es el género de estado en el cual sólo un ser consciente podría estar. No obstante, los dolores no parecen ser representacionales del modo como lo son los pensamientos: los dolores son sólo sentimientos, no tratan de nada ni están "dirigidos" a ello. Otro ejemplo, supóngase que tiene usted un género de depresión generalizada o malestar. Puede que esté usted deprimido sin conseguir decir qué es aquello que lo deprime. ¿Es éste otro

ejemplo de un estado intencional sin direccionalidad hacia un objeto?

Tomemos primero el caso del dolor. En primer lugar, debemos dejar en claro lo que queremos decir al afirmar que el dolor es un estado mental. A veces llamamos a un dolor "físico", para diferenciarlo del dolor "mental" de (digamos) la pérdida de un ser amado. Son evidentemente clases muy diferentes de estado metal y es erróneo pensar que tienen gran cosa en común, precisamente porque llamamos a los dos "dolor". Este hecho no hace que el dolor de (digamos) muelas sea algo *menos* mental. Pues el dolor es un estado de conciencia: nada podría tener un dolor a menos de que fuese consciente, y nada podría ser consciente a menos de que tuviera una mente.

¿Refuta la existencia de sensaciones la primera parte de la tesis de Brentano, de que la mentalidad es suficiente para la intencionalidad? Sólo si es verdadero que estamos completamente despojados de cualquier intencionalidad. Y esto no parece ser cierto.[24] Aunque no diríamos que mi dolor de cintura es "acerca" de nada, tiene algún carácter representacional, en la medida en que lo siento en mi espalda. Pude tener un dolor que se siente exactamente lo mismo, "a modo de dolor", pero está en lo alto de mi espinazo más bien que en la base del espinazo. La diferencia de cómo los dos dolores se sienten sería puramente cuestión de *dónde* se siente que son. Para aclarar esto más vívidamente: pudiera tener dos dolores, uno en cada mano, que se sintieran exac-

[24] Véanse D. M. Armstrong, *A Materialist Theory of the Mind* (Londres, Routledge and Kegan Paul, 1968; reimpreso en 1993), cap. 14; y M. G. F. Martin, "Bodily Awareness: A Sense of Ownership", en J. Bermúdez y N. Eilan (eds.), *The Body and the Self* (Cambridge, MIT Press, 1995).

tamente igual, salvo que uno lo sentiría en la mano derecha y el otro lo sentiría en la mano izquierda. Esta localización sentida es plausiblemente una diferencia de intencionalidad —en que el estado mental es "dirigido"— de manera que no es verdad que los dolores (al menos) no tengan intencionalidad en absoluto.

Por supuesto, esto no significa que los dolores sean actitudes proposicionales en el sentido de Russell. Pues no se dirigen a situaciones. Una atribución de dolor —"Oswaldo siente dolor"— no se ajusta a la forma "A ψ que S", forma que consideré como criterio para la atribución de actitudes proposicionales. Sin embargo, el hecho de que un estado mental no sea una actitud proposicional no quiere decir que no sea intencional porque, según hemos visto ya, no todos los pensamientos o estados intencionales de la mente son actitudes proposicionales (el amor fue nuestro anterior ejemplo). Y si entendemos la idea del "carácter representacional" o intencionalidad del modo como lo estoy haciendo de manera general aquí, es difícil negar que los dolores tengan carácter representacional.

¿Qué ocurre con el otro ejemplo, de depresión no dirigida o malestar? Pues bien, como es natural, hay tal cosa como la depresión en la cual la persona que sufre de ella no puede identificar qué es aquello que la deprime. Esto por sí mismo no significa que tal depresión carezca de objeto, que no tenga direccionalidad. Por un lado, no puede haber un criterio para que algo sea un estado intencional que el sujeto deba ser capaz de identificar; de otra manera algunas formas de autoengaño serían imposibles. Aún más importante, la descripción de esta clase de emoción como no dirigida hacia algo, lo describe mal. Pues la depresión de cualquier clase es típicamente una visión totalmente negativa del

mundo externo, según la frase económica de Lewis Wolpert.[25] Esto ocurre tanto con la depresión que "no es acerca de nada en particular" como con la depresión que tiene un objeto definido, fácil de identificar. La depresión generalizada es un modo de experimentar el mundo *en general,* todo parece mal, nada merece ser hecho, el mundo de la persona deprimida "se encoge". Esto es, la depresión generalizada es una manera como la mente de uno es dirigida hacia el mundo —y por lo tanto es intencional— ya que el mundo "en general" puede aún ser objeto de un estado mental.

No es evidente, pues, que haya ningún estado mental que sea plenamente no intencional. Sin embargo, puede seguir habiendo *propiedades* o *rasgos* mentales que son no intencionales: por ejemplo, aunque mi dolor de muelas tiene una direccionalidad intencional sobre mi muela, puede tener una calidad distintiva de *tenacidad* que no es intencional en absoluto: la tenacidad no está dirigida hacia nada, sino que sencillamente *está* ahí. Estas propiedades aparentes son a veces conocidas como *qualia.* Si una sensación como el dolor tiene estas propiedades, entonces puede haber un *elemento* residual en la sensación que no sea intencional, aun cuando la sensación considerada como un estado mental en conjunto sea intencional. Así, aun si la primera parte de la tesis de Brentano es cierta para estados mentales plenos —son todos intencionales—, puede aun haber un elemento no intencional en la vida mental. Esto sería algo como una victoria pírrica de la tesis de Brentano.

[25] Lewis Wolpert, *Malignant Sadness: The Anatomy of Depression* (Londres, Faber, 1999). Esto es semejante a la descripción de la depresión o melancolía dada por Sartre en su *Sketch for a Theory of the Emotions* (Londres, Methuen, 1971; publicado originalmente en 1939); véase especialmente pp. 68-69.

Hasta aquí, pues, la idea de que la mentalidad es suficiente para la intencionalidad. No obstante, ¿es necesaria la mentalidad para la intencionalidad? Esto es, ¿es cierto que si algo exhibe intencionalidad, entonces ese algo debe (o tiene que ser) una mente? Esto es más arduo. Para sostener que las mentes no son las únicas cosas que tienen intencionalidad, debemos dar un ejemplo de algo que tenga intencionalidad pero no tenga mente. Y parece que los ejemplos abundan. Tómense los libros. Este libro contiene muchas oraciones, todas las cuales tienen sentido, representan cosas y por lo tanto tienen intencionalidad en algún grado. Sin embargo, el libro no tiene mente.

La réplica natural a esto es emplear la línea de pensamiento que usé al discutir antes la representación lingüística. Esto es, debemos decir que las oraciones del libro no tienen intencionalidad *intrínsecamente,* sino que únicamente la tienen a causa de que se *interpretan* por los lectores del libro. Las interpretaciones proporcionadas por los estados mentales del lector, sin embargo, tienen una intencionalidad intrínseca.

Los filósofos a veces marcan la distinción entre libros y mentes a este respecto hablando de intencionalidad "original" y "derivada". La intencionalidad presente en un libro es meramente *derivada* intencionalmente: es derivada de los pensamientos de aquellos que escribirán y leerán el libro. No obstante, nuestras mentes tienen intencionalidad *original:* su intencionalidad no depende de la intencionalidad de algo más, ni deriva de ella.[26]

De manera que podemos reformular nuestras cuestio-

[26] Para esta distinción, véase John Haugeland, "The Intentionality All-stars", en J. Tomberlin (ed.), *Philosophical Perspectives 4: Action Theory and the Philosophy of Mind* (Atascadero, Ridgeview, 1990), pp. 385 y 420

nes como sigue: ¿puede algo distinto de las mentes tener intencionalidad original? Esta cuestión es muy desconcertante. Un problema que plantea es que si hubiésemos de encontrar algo que exhibiese intencionalidad original, es difícil ver cómo podría ser una cuestión *adicional* que semejante cosa tuviese una mente. Así pues, ¿queremos decir que sólo las mentes, tal como las conocemos, pueden exhibir intencionalidad original? La dificultad aquí es que comienza a tener el aire de una simple estipulación: si, por ejemplo, descubriéramos que las computadoras son capaces de intencionalidad original, bien pudiésemos decir: ¡Cuán asombroso! ¡Una computadora puede tener mente!, o pudiésemos decidir usar los términos diferentemente, y decir: ¡Cuán asombroso! ¡Algo puede tener intencionalidad original sin tener mente! La diferencia entre las dos reacciones puede parecer en gran medida cuestión de terminología. En el capítulo III habré de decir más acerca de esta cuestión.

La segunda parte de la tesis de Brentano —que la mentalidad es una condición necesaria para la intencionalidad— introduce algunas cuestiones desconcertantes, pero parece muy plausible a rasgos generales. Sin embargo, debemos reservar el juicio sobre ella hasta que descubramos un poco más acerca de lo que es tener una mente.

Conclusión: de la representación a la mente

El ejemplo de la "carta" interestelar del *Pioneer 10* recalcó la naturaleza de rompecabezas de la representación. Después

nota 6. Véase también John Searle, *Intentionality,* p. 27, para una distinción parecida.

de esto, consideré la representación pictórica, y la teoría del parecido de la representación pictórica, pues esta clase de representación parecía, a primera vista, ser más sencilla que otras clases. Empero, esta apariencia era engañosa. No sólo el parecido presenta una base frágil sobre la cual fundar la representación, sino que también las imágenes necesitan interpretación; la interpretación parece necesaria para la representación lingüística también. Y entonces sugerí que la interpretación deriva de la representación mental, o intencionalidad. Para comprender la representación, necesitamos entender estados representacionales de la mente. Éste es el tema del siguiente capítulo.

LECTURAS ADICIONALES

El capítulo 1 de Nelson Goodman, *Languages of Art* (Indianapolis, Hackett, 1976) es una discusión importante de la representación pictórica. *Why Does Language Matter to Philosophy?* de Ian Hacking (Cambridge, Cambridge University Press, 1975) es un relato muy legible semihistórico de la relación entre ideas y representación lingüística. Una buena introducción a la filosofía del lenguaje es la de Alex Miller, *Philosophy of Language* (Londres, UCL Press, 1997). Más adelantados son Richard Larson y Gabriel Segal, *Knowledge of Meaning: An Introduction to Semantic Theory* (Cambridge, MIT Press, 1995), que integra ideas de la filosofía reciente del lenguaje y la lingüística. Una excelente colección de lecturas esenciales acerca de esta área de la filosofía del lenguaje está en A. W. Moore (ed.), *Meaning and Reference* (Oxford, Oxford University Press, 1993). Para más acerca de la intencionalidad, véase el

capítulo 1 de mis *Elements of Mind* (Oxford, Oxford University Press, 2001). Una discusión importante está en Robert Stalnaker, *Inquiry* (Cambridge, MIT Press, 1984), capítulos 1 y 2. La *Intentionality* de John Searle (Cambridge, Cambridge University Press, 1983) es un libro accesible acerca de los fenómenos de la intencionalidad. Una colección útil de ensayos, muchos de ellos bastante técnicos, acerca de la idea de la "actitud proposicional", es la de Nathan Salmon y Scott Soames (eds.), *Propositions and Attitudes* (Oxford, Oxford University Press, 1988). La mejor colección en un volumen de lecturas sobre la filosofía de la mente en general sigue siendo la de David Rosenthal (ed.), *The Nature of Mind* (Oxford, Oxford University Press, 1990). Para mayor lectura sobre la conciencia, véase más adelante el capítulo VI, p. 330.

II. CÓMO ENTENDER A LOS PENSADORES Y SUS PENSAMIENTOS

He dicho que para comprender la representación tenemos que comprender el pensamiento. Sin embargo ¿cuánto realmente conocemos acerca del pensamiento? o, puestas así las cosas, ¿cuánto sabemos acerca de la mente en general?

Sería tentador pensar que ésta es una cuestión que sólo puede ser contestada realmente por la ciencia del cerebro. Si tal fuese el caso, la mayoría de la gente sabría muy poco acerca del pensamiento y de la mente. Después de todo la mayor parte de las personas no ha estudiado el cerebro, y hasta para los expertos algunos aspectos del cerebro siguen siendo del todo misteriosos. De esta manera, si tuviéramos que comprender los detalles de la función cerebral a fin de comprender las mentes, muy pocos de nosotros sabríamos realmente algo acerca de ellas.

No obstante, de seguro hay un sentido en el cual sabemos mucho acerca de las mentes. De hecho, las mentes son tan familiares para nosotros que este hecho puede escapar inadvertido al principio. Lo que quiero decir es que sabemos que tenemos pensamientos, experiencias, recuerdos, sueños, sensaciones y emociones, y sabemos que otras personas tienen lo mismo. Tenemos muy plena conciencia de distinciones sutiles entre clases de estado mental, entre esperanza y expectación, por ejemplo, o pena y remordimiento. Este conocimiento de las mentes se aplica para comprender a

otra gente. Mucho de nuestra vida cotidiana depende de nuestro conocimiento de lo que piensan otras personas, y a menudo somos no poco acertados para saberlo. Sabemos lo que otras personas están pensando observándolas, escuchándolas, hablándoles y logrando conocer sus caracteres. Con frecuencia este conocimiento de la gente nos permite predecir qué harán, a menudo con una precisión que pondría en vergüenza al servicio meteorológico.

Lo que tengo presente aquí son casos muy ordinarios de "predicción". Por ejemplo, supóngase que llama usted a una amiga y decide verla mañana a la hora de comer. Yo apostaría que (dependiendo de quién sea la amiga) muchos de nosotros confiaríamos más en que aparecerá una amiga de lo que confiamos en la predicción del clima. Sin embargo, al hacer esta "predicción" estamos contando con nuestro conocimiento de la mente: que ella *comprende* las palabras que se le dicen, que ella *sabe* dónde está el restaurante, que ella *quiere* encontrarlo a usted para comer, y así sucesivamente.

Así, en este sentido cuando menos, todos somos expertos en la mente. Sin embargo, nótese que esto no significa, por sí mismo, que la mente sea algo diferente del cerebro. Pues es perfectamente coherente con el hecho de que sepamos mucho acerca de la mente el sostener que esos estados mentales (como el deseo, la comprensión, etc.) son en última instancia nada más estados bioquímicos del cerebro. Si éste fuera el caso, entonces nuestro conocimiento de las mentes sería *también* el conocimiento de los cerebros, aunque podría no parecernos así.

Afortunadamente no tenemos que establecer la cuestión de si la mente es el cerebro a fin de mostrar lo que sabemos acerca de la mente. Para explicar por qué no, necesito decir

una que otra cosa acerca del notorio "problema mente-cuerpo".

El problema mente-cuerpo

El problema mente-cuerpo es la cuestión de cómo la mente y el cuerpo están conectados entre sí. Sabemos que *están* conectados, por supuesto; considero que cuando el cerebro de una persona está dañado, su capacidad de pensar es transformada. Todos sabemos que cuando las personas toman narcóticos, o beben demasiado alcohol, estas actividades corporales afectan al cerebro, que a su vez afecta los pensamientos que tienen. Nuestras mentes y la materia que constituye nuestros cuerpos están evidentemente relacionados, pero ¿cómo?

Una razón de que éste sea un problema es porque, por un lado, parece evidente que *debemos* precisamente estar hechos de materia y, por otra parte, parece evidente que *no podemos* estar nada más hechos de materia; debemos ser algo más. Pensamos que debemos ser nada más materia, por ejemplo, porque creemos que los seres humanos se han desenvuelto a partir de formas inferiores de vida, que a su vez estaban hechos enteramente de materia (cuando las mentes evolucionaron por vez primera, la materia prima a partir de la cual evolucionaron era simplemente materia compleja). Y es plausible creer que estamos enteramente hechos de materia, por ejemplo: si toda mi materia fuese suprimida, gradualmente, no quedaría nada de mí.

Sin embargo, parece muy difícil creer que somos, por debajo de todo, sólo materia, sólo unos cuantos dólares de carbono, agua y algunos minerales. Es fácil para cualquiera que haya experimentado el menor trastorno de su cuerpo al-

canzar el sentido de que es en todo caso *increíble* que esta materia frágil y confusa constituya su naturaleza como agente pensante consciente. Igualmente, aunque la gente habla a veces de la "química" que ocurre entre personas enamoradas, este uso es evidentemente metafórico; la idea de que el amor mismo es literalmente "nada sino una reacción química compleja" parece simplemente absurda.

Oí una vez una historia (probablemente apócrifa) que ilustra este sentimiento.[1] Según esta historia, algunos investigadores médicos de los años cuarenta descubrieron que las gatas que estaban desprovistas de magnesio en su dieta dejaban de ocuparse de su descendencia. Esto fue informado en un periódico bajo el encabezado: "El amor maternal es magnesio". Si la historia es verdad, es algo que no importa; lo que importa es por qué la encontramos divertida. Pensar en nuestras vidas mentales conscientes como "realmente" interacciones físicas complejas entre productos químicos parece ser tan absurdo como pensar que el amor maternal es "realmente" magnesio.

¿O no? Los científicos encuentran correlaciones más y más detalladas entre los trastornos psicológicos y sustancias químicas específicas en el cerebro.[2] ¿Hay un límite de lo que pueden descubrir acerca de tales correlaciones? Parece un recurso final desesperado, desde una posición de ignorancia casi total, decir que *debe* haber un límite. Pues la verdad es que no lo sabemos. Tal vez la verdad no es tan sen-

[1] Oí la historia de P. J. Fitzpatrick (por desgracia no he logrado hallar la fuente).

[2] Para una introducción muy clara y legible, véase la primera mitad de Mark Solms y Oliver Turnbull, *The Brain and the Inner World: An Introduction to the Neuroscience of Subjective Experience* (Nueva York, Other Press, 2002).

cilla como "el amor maternal es magnesio", pero ¿podría no estar demasiado lejos de ello?

Así, somos arrastrados primero de una manera, y luego de otra. Por supuesto, pensamos que sólo somos materia organizada de una manera compleja; pero entonces, al pensarlo, parece imposible que seamos materia nada más; debe haber más de nosotros que esto. Esto, en los términos más escuetos, es una manera de expresar el problema mente-cuerpo. Ha demostrado ser uno de los problemas más intratables de la filosofía, tanto que algunos filósofos han pensado que es imposible resolverlo. El filósofo inglés del siglo XVII Joseph Glanvill (1636-1680) expresó esta idea de modo muy convincente: "Cómo el espíritu más puro se une a esta pella es un nudo demasiado duro para que lo deshaga la humanidad caída".

Otros son más optimistas, y han ofrecido soluciones a este problema. Algunos —*materialistas* o *fisicalistas*— piensan que, a pesar de nuestros sentimientos en contra, es posible demostrar que la mente es sólo materia compleja: la mente es precisamente la materia del cerebro organizada según cierta manera compleja. Otros piensan que la mente no puede ser sólo materia, sino que debe ser algo más, algún otro tipo de cosa. Quienes creen, por ejemplo, que tenemos almas "inmateriales", que sobreviven a la muerte de nuestros cuerpos, deben negar que nuestras mentes sean las mismas cosas que nuestros cuerpos. Pues si nuestras mentes fueran lo mismo que nuestros cuerpos, ¿cómo podrían sobrevivir a la aniquilación de tales cuerpos? Estos filósofos son *dualistas,* ya que piensan que hay *dos* tipos principales de cosa: la material y la mental. (Una solución menos común estos días es suponer que todo es a fin de cuentas mental: éste es el *idealismo.*)

El materialismo, en una de sus muchas variedades, tiende a ser el enfoque ortodoxo al problema mente-cuerpo estos días. El dualismo es menos común, pero aun así es defendido vigorosamente por sus proponentes.[3] En un apartado del capítulo VI ("Conciencia y fisicalismo") retornaré a este problema y trataré de hacerlo más preciso y de esbozar lo que está en tela de juicio entre el dualismo y el materialismo. Por el momento podemos poner el problema mente-cuerpo a un lado mientras investigamos el problema de la representación mental. Me explicaré.

El problema acerca de la representación mental puede expresarse muy sencillamente: ¿cómo puede la mente representar cualquier cosa? Supóngase por el momento que el materialismo es cierto: la mente no es sino el cerebro. ¿Cómo ayuda esto al problema de la representación mental? Podemos hablar simplemente de otro modo, preguntando: ¿cómo puede el *cerebro* representar cualquier cosa? Esto parece tan difícil de comprender como la cuestión acerca de la mente. Con toda su complejidad, el cerebro es nada más un trozo de materia, y cómo un trozo de materia puede representar otra cosa parece precisamente tan desconcertante como que una mente pueda representar algo, sea o no la mente un trozo de materia.

Supóngase por un momento que el materialismo es cierto, y pensemos en lo que lleva usted dentro de la cabeza. Hay alrededor de 100 000 millones de células cerebrales. Éstas forman una sustancia de consistencia gris y blanca,

[3] Para una crítica estándar del dualismo, véase Peter Smith y O. R. Jones, *The Philosophy of Mind* (Cambridge, Cambridge University Press, 1986), caps. 1-3. Para el dualismo contemporáneo, véase W. D. Hart, *The Engines of the Soul* (Cambridge, Cambridge University Press, 1988), y John Foster, *The Immaterial Self* (Londres, Routledge, 1991).

acuosa, parecida al yogurt. Aproximadamente un kilogramo de esta materia constituye el cerebro de usted. Si el materialismo es verdad, entonces esta sustancia como yogurt, sola, lo ayuda a pensar, en usted, su vida y el mundo. Le permite a usted razonar acerca de lo que le conviene hacer. Le permite tener experiencias, recuerdos, emociones y sensaciones. Pero ¿cómo? ¿Cómo puede esta sustancia acuosa parecida al yogurt —esta "pella"— constituir los pensamientos de usted?

Por otra parte, supongamos que el dualismo es cierto: la mente no es el cerebro sino otra cosa, distinta del cerebro, como un "alma inmaterial". Entonces parece que podemos plantear la misma cuestión acerca del alma inmaterial: ¿cómo puede un alma inmaterial representar cualquier cosa? Descartes creía que la mente y el cuerpo eran cosas distintas: la mente era, para Descartes, un alma inmaterial. Creía también que la esencia de esta alma era pensar. No obstante, decir que la esencia de la mente es pensar no contesta a la cuestión de *cómo* logra pensar el alma. En general, no es muy satisfactorio responder a la cuestión de "¿Cómo es que *esto* hace *eso*?" con la respuesta de "Bueno, es porque está en la naturaleza (o es la naturaleza) de esto hacer eso". Eso sería como cuando el famoso doctor de la pieza de Molière, *Le malade imaginaire* [El enfermo imaginario], respondía a la cuestión de cómo el opio lo hace a uno dormir diciendo que tiene una *virtus dormitiva,* una "virtud dormitiva", o sea que está en la esencia o naturaleza del opio hacer dormir a uno.

Tanto el materialismo como el dualismo, pues, necesitan una solución al problema de la representación. Contestar el problema mente-cuerpo con materialismo o dualismo no resuelve por sí mismo el problema de la representación,

pues éste permanecerá aun cuando hayamos convenido en el materialismo o el dualismo como respuesta al problema anterior. Si el materialismo es cierto, y todo es materia, necesitamos todavía conocer cuál es la diferencia entre materia pensante y materia no pensante. Y si el dualismo es cierto, entonces necesitamos todavía saber qué es en esta mente no material lo que le permite pensar.

(Por otra parte, si el idealismo es verdad, entonces hay un sentido en el cual todo es pensamiento, sea como sea, de modo que el problema no se presenta. Sin embargo, es mucho más difícil creer en el idealismo de este género —por decirlo brevemente— que en muchos puntos de vista filosóficos, así que parecería que traficara uno con un misterio en el lugar del otro.)

Esto significa que podemos discutir los principales puntos de este libro sin tener que decidir si el materialismo o el dualismo es la solución correcta al problema mente-cuerpo. La controversia materialismo/dualismo no es directamente pertinente para nuestros problemas. Para los propósitos de este capítulo, es buena cosa. Pues, aunque no sabemos con ningún detalle cuál es la relación entre la mente y el cerebro, en lo que estoy interesado aquí es en *lo* que sabemos acerca de las mentes en general y el pensamiento en particular. Éste es el tema del resto de este capítulo. Volveremos al problema mente-cuerpo en el capítulo VI.

Cómo entender otras mentes

¿Así que, qué sabemos acerca de la mente? Una manera de enfocar esta cuestión es preguntar: ¿cómo sabemos de la mente? Por supuesto, no es la misma cuestión. (Compáren-

se las preguntas "¿Qué sabemos acerca del agua?" y "¿Cómo averiguamos acerca del agua?") Como veremos, en el caso de la mente, preguntar *cómo* sabemos iluminará considerablemente lo que *sabemos*.

Una cosa que parece evidente es que estamos enterados acerca de las mentes de otros de una manera muy diferente del modo como conocemos nuestras propias mentes. Sabemos acerca de nuestras mentes en parte por introspección. Si trato de representar lo que pienso acerca de determinada cuestión, puedo concentrarme en el contenido de mi mente consciente hasta que doy con ello. Pero no puedo concentrarme del mismo modo acerca del contenido de la mente de usted al imaginar lo que piensa. Algunas veces, por supuesto, no puedo decir lo que realmente pienso, y tengo que consultar a otros —un amigo o un terapeuta, quizá— acerca de la significación de mis pensamientos y acciones, y lo que revelan acerca de mi mente. El punto es que averiguar acerca de la mente propia no es *siempre* como esto, en tanto que aprender acerca de las mentes de otros siempre lo es.

El modo como conocemos acerca de los estados de la mente de otros no es, por así decirlo, *simétrico* con el modo como conocemos nuestros propios estados mentales. Esta "asimetría" se relaciona con otra asimetría importante: las diferentes maneras que usamos para conocer acerca de la posición de nuestros propios cuerpos y los cuerpos de otros. A fin de conocer si tiene usted las piernas cruzadas, tengo que mirar, o usar alguna otra forma de observación o inspección (podría preguntarle). Sin embargo, no necesito ninguna clase de observación para saber si tengo las piernas cruzadas. Normalmente conozco esto inmediatamente sin observación. Asimismo, puedo decir típicamente qué pien-

so sin tener que observar mis palabras y vigilar mis acciones. Sin embargo, no puedo decir qué piensa usted sin observar sus palabras y sus acciones.

Cuando se trata de las mentes de otros, parece evidente que todo lo que tenemos que hacer es ver lo que dice y hace la gente: su comportamiento observable. Así ¿cómo podemos deducir del conocimiento de la gente el conocimiento de lo que piensa?

Cierta especie de escepticismo filosófico afirma que no podemos. Éste es el "escepticismo acerca de otras mentes", y el problema que suscita se conoce como "el problema de las otras mentes". Esto necesitará una breve digresión. De acuerdo con este punto de vista escéptico, todo lo que realmente sabemos acerca de otra persona son hechos acerca de su comportamiento observable. La gente *podría* comportarse como lo hace sin tener mente en absoluto. Por ejemplo, todas las personas que usted ve a su alrededor podrían ser robots programados por algún científico loco para comportarse como gente consciente, pensante: podría usted ser la única mente real entre las circundantes. Ésta es una hipótesis estrafalaria, por supuesto: pero parece compatible con la evidencia que tenemos acerca de otras mentes.

Compárese el escepticismo acerca de otras mentes con el escepticismo acerca de la existencia del "mundo externo" (esto es, el mundo exterior a nuestras mentes). Esta clase de escepticismo afirma que, al formar las creencias de usted acerca de los objetos del mundo, todo lo que a usted realmente le queda es la evidencia de sus sentidos: sus creencias formaron el fundamento de experiencias. Estas experiencias y creencias podrían ser precisamente como son y sin embargo el mundo "externo" ser muy diferente del modo como se piensa que es. Por ejemplo, su cerebro podría ser

conservado en un recipiente de nutrientes, y sus nervios entrantes y salientes estimulados por un científico loco para hacer que parezca que usted experimenta el mundo de los objetos cotidianos. Ésta es también una hipótesis estrafalaria, pero asimismo parece compatible con la experiencia de usted.[4]

Estas versiones del escepticismo no pretenden ser posiciones filosóficamente sostenibles: ha habido pocos filósofos en la historia que hayan sostenido seriamente que otra persona carece de mente. Lo que hace el escepticismo es forzarnos a revelar lo que realmente sabemos, y forzarnos a justificar cómo lo sabemos. Para contestar al escepticismo, necesitamos hacer un relato de lo que se sabe de alguna manera, y por lo tanto dar razón de lo que sabemos "realmente". Así, los argumentos en pro y en contra del escepticismo pertenecen propiamente a la teoría del conocimiento (llamada *epistemología*) y caen fuera del alcance de este libro. Por esta razón, voy a dejar el escepticismo a un lado. Mi interés en este libro es lo que creemos que es verdad acerca de nuestras mentes. De hecho, todos creemos que sabemos mucho sobre la mente de otros, y pienso que estamos indudablemente en lo cierto en esta creencia. De manera que dejemos a los epistemólogos explicarnos qué es el conocimiento; sea como sea, más valdría tener en cuenta el hecho evidente de que conocemos mucho acerca de las mentes de los otros.

Nuestra cuestión es, pues, *cómo* logramos saber acerca de otras mentes, no *si* sabemos. Esto es, dado que sabemos tantas cosas acerca de las mentes de otros, ¿cómo sabemos ta-

[4] Esta última pretensión es rechazada por quienes sostienen un punto de vista "externalista" del pensamiento y la experiencia: véase, por ejemplo, John McDowell, "Singular Thought and the Extent of Inner Space",

les cosas? Un aspecto del argumento escéptico que parece arduo negar es éste: todo lo que tenemos que andar al comprender a otro sujeto es su comportamiento observable. ¿Cómo podría ser de otra manera? De seguro no percibimos los pensamientos o las experiencias de otros; percibimos sus palabras y acciones observables.[5] De modo que la cuestión es: ¿cómo pasamos del comportamiento observable al conocimiento de sus mentes? Una respuesta que se propuso seriamente en una ocasión es que el comportamiento observable es, en algún sentido, *todo lo* que implica tener una mente: por ejemplo, si realmente sentir dolor es un "comportamiento doloroso" (llorar, gemir quejándose, etc.). Este modo de ver es conocido como *behaviorismo,* y vale la pena iniciar nuestro examen de nuestro conocimiento de las mentes con un examen del behaviorismo.

Aunque parece muy implausible, durante un breve tiempo el behaviorismo fue, en el siglo xx, popular tanto en la psicología como en la filosofía de la mente.[6] Da una respuesta rectilínea a la cuestión de cómo conocemos las mentes de otros. Sin embargo, vuelve muy problemática la cuestión de cómo conocemos nuestras mentes, porque, según señalé antes, podemos conocer nuestras propias mentes sin

en P. Pettit y J. McDowell (eds.), *Subject, Thought and Context* (Oxford, Clarendon Press, 1986). Para la fantasía del "cerebro en una tina", véase Hilary Putnam, *Reason, Truth and History* (Cambridge, Cambridge University Press, 1980), cap. 1.

[5] Pero véase John McDowell, "On 'The Reality of the Past' ", en C. Hookway y P. Pettit (eds.), *Action and Interpretation* (Cambridge, Cambridge University Press, 1978), especialmente p. 136.

[6] Para alguna bibliografía behaviorista, véase W. G. Lycan (ed.), *Mind and Cognition* (Oxford, Blackwell, 1990), § 1; para una crítica del behaviorismo, véase Ned Block, "Psychologism and Behaviourism", *Philosophical Review,* 90 (1980).

observar nuestro comportamiento. (De ahí la broma filosófica popular, repetida *ad nauseam* a generaciones de estudiantes: dos behavioristas se encuentran en la calle; uno le dice al otro: "Te sientes muy bien hoy, ¿cómo me siento yo?") Este aspecto del behaviorismo va de la mano con su deliberado descuido (o incluso su rotunda negación) de la experiencia subjetiva, consciente; cómo se siente, desde dentro, tener una mente.

No quiero enfocar estos inconvenientes del behaviorismo, que se discuten en detalle en muchos otros libros acerca de la filosofía de la mente. Lo que quiero es concentrarme en la inadecuación *interna* del behaviorismo: el hecho de que, *incluso en sus propios términos,* no puede dar razón de los hechos acerca de la mente puramente en términos de comportamiento.[7]

Una objeción inicial evidente del behaviorismo es que tenemos muchos pensamientos que no son revelados para nada en el comportamiento. Por ejemplo, creo que Riga es la capital de Latvia, aunque nunca he expresado esa creencia con ningún comportamiento. Así, ¿el behaviorismo negaría que tengo esta creencia? No. El behaviorismo diría que la creencia no requiere un comportamiento *real,* sino una *disposición* a comportarse. Compararía la creencia con una disposición tal como la solubilidad de un trozo de azúcar. Un trozo de azúcar puede ser soluble aunque nunca sea colocado en agua; la solubilidad de la pieza reside en el hecho de que está *dispuesta* a disolverse al ser puesta en agua. Análogamente, creer que Riga es la capital de Latvia es estar dispuesto a comportarse de cierta manera.

[7] Véase R. M. Chisholm, *Perceiving: A Philosophical Study* (Ithaca, Cornell University Press, 1957), especialmente cap. 11, § 3.

Esto parece más plausible hasta que preguntamos cuál es esta "cierta manera'. ¿Cuál es el comportamiento que concierne a la creencia de que Riga es la capital de Latvia como la disolución del azúcar se relaciona con su solubilidad? Una posibilidad es que la conducta es verbal: diciendo "Riga es la capital de Latvia" al ser interrogado acerca de la cuestión "¿cuál es la capital de Latvia?" (Así, preguntar la cuestión sería análogo a poner el azúcar en el agua.)

Sencilla como es, esta sugerencia no puede ser correcta, pues yo responderé sólo "Riga es la capital de Latvia" a la cuestión "¿cuál es la capital de Latvia?" si, entre otras cosas, entiendo el español. Empero, entender el español no es una condición previa para creer que Riga es la capital de Latvia: numerosos latvios monoglosos tienen creencias verdaderas acerca de su capital. Así, entender español debe ser un estado mental distinto de creer que Riga es la capital de Latvia, y esto también debe ser explicado en términos behavioristas. Rebasemos la cuestión de si entender el español puede ser explicado en términos puramente behavioristas —a lo cual la respuesta es sin duda "no"—,[8] y continuemos con este ejemplo por un momento.

Supóngase que la explicación behaviorista de mi comprensión de que la oración "Riga es la capital de Latvia' está en términos de mi disposición a enunciar la oración. Esta disposición no puede, evidentemente, ser sólo la disposición a hacer los *sonidos* "Riga es la capital de Latvia": un loro podría tener esta disposición sin comprender la oración. Lo que necesitamos (cuando menos) es la idea de que

[8] Para una crítica del punto de vista behaviorista del lenguaje, que se ha vuelto clásico, véase la reseña que hizo Chomsky del libro del behaviorista B. F. Skinner, *Verbal Behaviour,* reimpreso en Ned Block (ed.), *Readings in the Philosophy of Psychology,* vol. II (Londres, Methuen, 1980).

los sonidos son enunciados con comprensión, es decir, ciertos enunciados de la oración, y ciertas maneras de responder al enunciado, son *apropiados* y otros no. ¿Cuándo es apropiado enunciar la oración? ¿Cuando creo que Riga es la capital de Latvia? No necesariamente, ya que puedo enunciar la oración entendiendo sin comprenderla. Tal vez enuncio la oración porque quiero que mis oyentes crean que Riga es la capital de Latvia, aunque yo (erróneamente) crea que lo es Vilnius.

No obstante, en cualquier caso, el behaviorista no puede recurrir a la *creencia* de que Riga es la capital de Latvia al explicar cuándo es correcto enunciar la oración, ya que al enunciar la oración se suponía que mostraba tener la creencia. Así, esta explicación daría vueltas. La lección general, aquí, es que el pensamiento no puede ser definido cabalmente en términos de comportamiento: se necesita mencionar otros pensamientos también. Cada vez que tratamos de asociar un pensamiento con un fragmento de comportamiento, descubrimos que esta asociación no valdrá a menos que estén en su sitio otros estados mentales. Y tratar de asociar cada uno de estos otros estados mentales con otros fragmentos de comportamiento conduce a los mismos problemas. El pensamiento individual de usted puede estar asociado con muchos fragmentos diferentes de comportamiento, *dependiendo de qué otros pensamientos tenga usted.*

Un ejemplo más sencillo facilitará la cuestión. Un hombre mira desde una ventana, va al armario y agarra un paraguas antes de abandonar su casa. ¿Qué está pensando? La respuesta evidente es que pensó que estaba lloviendo. Nótese que, aun si esto es cierto, este pensamiento no lo conduciría a agarrar el paraguas a menos que quiera también mantenerse seco *y* él crea que agarrar el paraguas lo ayudará

a estar seco *y* crea que este objeto es su paraguas. Esto podría parecer tan evidente que difícilmente habría que decirlo. Pero al reflexionar, es evidente que si no tuviese estos (indudablemente inconscientes) pensamientos, sería harto misterioso por qué habría de tomar su *paraguas* cuando pensara que estaba lloviendo. Adónde conduce este punto es, en mi opinión, claro: aprendemos acerca de los pensamientos de otros haciendo conjeturas razonadas acerca de lo que parece coherente con su comportamiento.

Sin embargo, según muestran nuestros pequeños ejemplos, hay muchas maneras de encontrar coherencia con un fragmento de comportamiento, atribuyendo a quien piensa pautas muy diferentes de pensamiento. ¿Cómo, entonces, escogemos entre todas las versiones competidoras posibles de lo que son los pensamientos de alguien? La respuesta, creo, es que hacemos esto empleando, o presuponiendo, varias hipótesis generales acerca de lo que es ser un pensador. Tómese el ejemplo del hombre y el paraguas. Podríamos hacer las siguientes conjeturas acerca de cuál es su estado mental:

> Él creyó que estaba lloviendo, y quiso mantenerse seco (y, no hay ni que añadir, pensó que el paraguas lo ayudaría a estar seco y pensó que se trataba de su paraguas, etcétera).

> Él pensó que hacía sol y quiso el paraguas para protegerse del calor del sol (y supuso que el paraguas lo protegería del sol y creyó que esto era su paraguas, etcétera).

> Él no tenía opinión acerca del clima, pero creyó que su paraguas tenía poderes mágicos y quiso agarrarlo para alejar los malos espíritus (y creyó que era su paraguas, etcétera).

Él planeaba matar a un enemigo y creyó que su paraguas contenía un arma (y creyó que era su paraguas, etcétera).

Todas éstas son explicaciones *posibles* para hacer lo que hizo, y podemos pensar en muchas más. Sin embargo, dado que realmente está lloviendo, y lo sabemos, la primera explicación es, con mucho, la más probable. ¿Por qué? Pues bien, es en parte porque creemos que puede ver lo que vemos (que está lloviendo) y en parte porque pensamos que es en general algo indeseable mojarse estando totalmente vestido, y que la gente, cuando es posible, evita las cosas indeseables cuando no cuesta mucho esfuerzo... y así sucesivamente. En pocas palabras, hacemos ciertos supuestos acerca de su manera de ver sus alrededores, sus facultades mentales y su grado de racionalidad, y le atribuimos los pensamientos que es razonable tener para él, dadas dichas facultades.

Se ha vuelto costumbre entre muchos filósofos de la mente (y algunos psicólogos también) describir los supuestos e hipótesis que adoptamos al comprender otras mentes como una especie de *teoría* de otras mentes. Llaman a esta teoría "psicología del sentido común" o "psicología popular". La idea es que, así como nuestro conocimiento de sentido común del mundo físico descansa en el conocimiento de algunos principios generales del comportamiento característico de los objetos ("física popular"), así nuestro conocimiento de sentido común de otras mentes descansa en el conocimiento de algunos principios generales del comportamiento característico de la gente ("psicología popular").

Convengo con la idea de que nuestro conocimiento de sentido común de otros pensadores es una especie de teoría. Prefiero denominarla "psicología del sentido común", en

lugar de "psicología popular". Éstas son sólo etiquetas, por supuesto, y en un sentido no importa gran cosa cuál usemos. Para mi oído, el término "psicología popular" porta la connotación de que los principios implicados son simplemente "sabiduría popular", tonterías populares tautológicas, del tipo de "muchas manos hacen el trabajo ligero". Así, en la medida en que la etiqueta "psicología popular" puede sugerir que el conocimiento que interviene es ingenuo y trivial, la etiqueta implica una actitud denigrante hacia la teoría. Como veremos, mucho gira en torno a la teoría, de manera que es mejor no prejuzgar las cosas demasiado al principio.[9]

Ya que entender por qué otros pensadores hacen lo que hacen deriva (más a menudo que no) del conocimiento de su comportamiento observable, al entendimiento dado por la psicología del sentido común se le suele llamar "la explicación del comportamiento". Así, los filósofos dicen a menudo que la función de la psicología del sentido común es la explicación del comportamiento. En un sentido esto es verdad: estamos explicando el comportamiento, *buscando el sentido* del comportamiento, atribuyéndole estados mentales. No obstante, de otra manera, la expresión "explicación del comportamiento" es equívoca, pues hace parecer como si nuestra ocupación principal fuera siempre lo que la gente está *haciendo,* antes que lo que está *pensando.* Evidentemente, a menudo queremos saber qué está pensando la demás gente a fin de averiguar qué hará, o para darle sentido a lo que habrá hecho, pero en ocasiones es pura curiosidad la que nos hace querer averiguar qué está pensando. Aquí nuestro

[9] Véase Kathleen Wilkes, "The Long Past and the Short History", en R. Bogdan (ed.), *Mind and Common-Sense* (Cambridge, Cambridge University Press, 1991), p. 155.

interés no está en su comportamiento como tal, sino en los hechos psicológicos que se organizan y "yacen detrás del comportamiento", esos hechos que dan sentido al comportamiento.

Los behavioristas, por supuesto, negarían que hubiese algo psicológico que yaciera detrás del comportamiento. Podrían aceptar, simplemente como hecho fundamental, que ciertas interpretaciones del comportamiento son más naturales para nosotros que otras. Así, en nuestro ejemplo del paraguas, el behaviorista puede aceptar que la razón de que el hombre agarre su paraguas es que creyó que iba a llover, y así sucesivamente. Ésta es la cosa natural que decir y el behaviorista podría convenir en ello. No obstante, en vista de que, de acuerdo con el behaviorismo, no hay verdadera sustancia en la idea de que algo podría estar *produciendo* el comportamiento o *generándolo,* no debemos tomar la descripción de cómo los pensamientos del hombre conducen a su comportamiento como si fuesen literalmente *verdaderos.* Estamos "a gusto" con ciertas explicaciones más que con otras; pero esto no significa que sean verdad. Son simplemente más naturales para nosotros.

Este punto de vista es muy insatisfactorio. De seguro, al entender a otros, queremos saber qué es cierto de ellos, y no nada más qué explicaciones encontramos más natural dar. Y esto requiere, me parece, que nos interese lo que *hace* a estas explicaciones verdad y, por lo tanto, nos da la justificación para encontrar una explicación más natural que otras. Esto es, estamos interesados en qué es aquello que produce el comportamiento o lo genera. Así, para entender más profundamente lo que está mal con este punto de vista behaviorista, necesitamos mirar más de cerca la idea de pensamientos que yacen detrás del comportamiento.

La imagen causal de los pensamientos

Un aspecto de esta idea es precisamente el punto de vista ordinario, antes mencionado, de que no podemos percibir directamente los pensamientos de otra gente. Vale la pena decir aquí que este hecho por sí mismo no hace que las mentes de otras personas sean peculiares o misteriosas. Hay muchas cosas que no podemos percibir directamente, y que no por esa razón son misteriosas. Los microbios, por ejemplo, son demasiado pequeños para ser percibidos directamente; los agujeros negros son demasiado densos incluso para dejar escapar de ellos la luz, de manera que no podemos percibirlos. Sin embargo, nuestra incapacidad de percibir directamente estas cosas no las hace en sí mismas peculiares o misteriosas. Los agujeros negros pueden ser misteriosos, pero no es precisamente porque no podamos verlos.

Sin embargo, cuando digo que los pensamientos "yacen detrás" del comportamiento, no quiero decir nada más que los pensamientos no son directamente perceptibles. Quiero decir también que el comportamiento es el *resultado* del pensamiento, que los pensamientos *producen* comportamiento. Es así como sabemos acerca de los pensamientos: sabemos acerca de ellos a través de sus efectos. Esto es, los pensamientos están entre las causas del comportamiento: la relación entre pensamiento y comportamiento es una relación causal.

¿Qué significa decir que los pensamientos son las *causas* del comportamiento? Las nociones de causa y efecto están entre las ideas básicas que usamos para entender nuestro mundo. Piénsese cuán a menudo usamos estas nociones en la vida cotidiana: pensamos que la política económica del

gobierno causa inflación o gran desempleo, que fumar causa cáncer, el VIH causa sida, el dióxido de carbono en la atmósfera causa calentamiento global, que a su vez causa la elevación del nivel del mar, y así sucesivamente. La causalidad es, en palabras de David Hume (1711-1776), el "cemento del universo".[10] Decir que los pensamientos son las causas del comportamiento es en parte decir que este "cemento" (sea lo que sea) es lo que mantiene los pensamientos ligados al comportamiento detrás del cual yacen. Si mi deseo de una bebida me causó ir al refrigerador, entonces la relación entre mi deseo y mi acción es *en algún sentido* fundamentalmente igual que la relación entre alguien fumando y el que tenga cáncer: la relación de causa y efecto. Esto es, en algún sentido mis pensamientos me hacen moverme. Llamaré a este supuesto de que los pensamientos y otros estados mentales son las causas del comportamiento la "imagen causal del pensamiento".

Ahora, aunque hablamos de causas y efectos constantemente, hay grave disputa entre los filósofos acerca de lo que es realmente la causalidad, o incluso si hay algo como la causalidad.[11] Así, para entender cabalmente lo que significa decir que los pensamientos son las causas del comportamiento, necesitamos saber un poco acerca de la causalidad. Aquí me limitaré a algunos rasgos, sin controversia, de la causalidad, y mostraré cómo estos rasgos pueden aplicarse a la relación entre pensamiento y comportamiento.

Primero, cuando decimos que *A* causó *B*, normalmente

[10] David Hume, resumen de *A Treatise of Human Nature,* L. A. Selby-Bigge (ed.), (Oxford, Oxford University Press, 1978), p. 662.

[11] El mejor lugar para comenzar un estudio de la causalidad es la colección editada por Ernest Sosa y Michael Tooley, *Causation* (Oxford, Oxford University Press, 1993).

nos entregamos a la idea de que si *A* no hubiese ocurrido, *B* no habría ocurrido. Cuando decimos, por ejemplo, que alguien fumando causó su cáncer, creemos normalmente que si no hubiese fumado, entonces no habría adquirido cáncer. Los filósofos plantean esto diciendo que la causalidad implica *contrafactuales:* verdades acerca de asuntos "contrarios a los hechos". Así, podríamos decir que si creemos que *A* causó *B,* nos comprometemos a la verdad de la pretensión contrafactual: "Si *A* no hubiese ocurrido, *B* no habría ocurrido".

Aplicada a la relación entre pensamientos y comportamiento, esta pretensión acerca de la relación entre contrafactuales y causalidad afirma esto: si determinado pensamiento —supongamos que un deseo de beber— tiene cierta acción —beber— como resultado, entonces si este pensamiento no hubiera estado presente, la acción no habría estado tampoco. Si yo no hubiese tenido el deseo, entonces no me habría servido la bebida.

Lo que aprendimos en la discusión del behaviorismo fue que los pensamientos generan comportamiento sólo en presencia de otros pensamientos. Así, mi deseo de beber me hará obtener una bebida sólo si también creo que soy realmente capaz de obtener una bebida, y así sucesivamente. Esto es lo mismo exactamente que en los casos no mentales de causalidad: por ejemplo, podemos decir que cierto tipo de bacteria causó una epidemia, pero sólo en presencia de otros factores, como una vacunación inadecuada, la ausencia de atención médica de emergencia y una limpieza pertinente, y así sucesivamente. Podemos sumar esto diciendo que en *las circunstancias,* si las bacterias no hubiesen estado ahí, entonces no habría habido una epidemia. Análogamente con el deseo: *en las circunstancias,* si mi deseo no hubiera estado

presente, yo no habría obtenido la bebida. Esto es parte de lo que convierte el deseo en causa de la acción.

El segundo rasgo de la causalidad que mencionaré es la relación entre causalidad y la idea de explicación. Explicar algo es responder a una pregunta del tipo "¿por qué?" al respecto. Preguntar "¿por qué ocurrió la primera Guerra Mundial?" y "explicar los orígenes de la primera Guerra Mundial" es preguntar en gran medida la misma cosa. Una manera como pueden ser respondidas las cuestiones de "¿por qué?" es citando la causa de lo que uno quiere explicar. Así, por ejemplo, una respuesta a la pregunta "¿por qué le dio cáncer?" podría ser "porque fumaba"; una respuesta a "¿por qué hubo un incendio?" puede ser "porque hubo un cortocircuito".

Es fácil ver cómo esto se aplica a la relación entre pensamientos y comportamiento, puesto que la hemos empleado en nuestros ejemplos hasta aquí. Cuando preguntamos "¿por qué el hombre agarró su paraguas?" y respondemos "porque pensó que estaba lloviendo, etc.", estamos (según la imagen causal) explicando la acción citando la causa, los pensamientos que hay detrás de ella.

El rasgo final de la causalidad que mencionaré es el nexo entre la causalidad y las regularidades del mundo. Como muchas cosas en la teoría contemporánea de la causalidad, la idea de que la causa y la regularidad están enlazadas procede de Hume. Éste dijo que una causa es un "objeto seguido de otro, y donde todos los objetos, similares al primero, van seguidos de objetos similares al segundo".[12] Así, por ejemplo, este cortocircuito causó este fuego, y entonces to-

[12] Hume, *An Enquiry Concerning Human Understanding*, Selby-Bigge (ed.), (Oxford, Oxford University Press, 1975), § 7.

dos los sucesos parecidos a este cortocircuito causarán sucesos parecidos al de este incendio. Tal vez no haya nunca dos sucesos *exactamente* iguales; pero todo lo que requiere la pretensión es que dos acontecimientos parecidos en algún aspecto específico causarán sucesos parecidos en algún aspecto específico.

Ciertamente esperamos que el mundo sea regular. Cuando tiramos una pelota al aire esperamos que caiga al suelo, generalmente a causa de que estamos acostumbrados a cosas como la acontecida. Y si arrojáramos una pelota al aire y no descendiera hasta el suelo, normalmente concluiríamos que algo había intervenido, esto es, que otra *causa* detuvo la pelota que caía al suelo. Esperamos que causas similares tengan efectos similares. La causalidad parece implicar un elemento de regularidad.

Sin embargo, algunas regularidades parecen más regulares que otras. Hay una regularidad en mi consumo de pizza: nunca he comido una pizza de más de 50 centímetros de diámetro. Es también una regularidad que los objetos no sujetos, aparte de globos, etc., caen al suelo. Estas dos regularidades parecen ser muy diferentes. Pues sólo la modestia me contiene de consumir una pizza mayor de 50 centímetros, pero es la naturaleza la que impide a los objetos no sostenidos escapar por el espacio. Por esta razón, los filósofos distinguen entre simples *regularidades accidentales,* como la primera, y *leyes de la naturaleza,* como la segunda.

Así, si hay un elemento de regularidad en la causalidad, entonces debe ser la regularidad en la relación entre pensamiento y comportamiento (si ésta es realmente una relación causal). Discutiré la idea de que hay tales regularidades, y de cómo serán, en el siguiente apartado.

Tracemos estas varias ideas acerca de la causalidad y el

pensamiento. Decir que los pensamientos causan comportamiento es decir al menos lo siguiente:

1. La relación entre pensamiento y comportamiento implica la verdad de un contrafactual, con el efecto de que, *dadas las circunstancias,* si el pensamiento no hubiese estado presente, entonces el comportamiento no se habría manifestado.
2. Citar un pensamiento, o manojo de pensamientos, como causa de un fragmento de comportamiento es *explicar* el comportamiento, ya que citar causas es una manera de explicar efectos.
3. Las causas implican, típicamente, *regularidades* o *leyes,* de modo que, si hay una relación causal entre pensamiento y comportamiento, podemos esperar que haya regularidades en la conexión entre pensamiento y comportamiento.

En ningún lugar hemos dicho que la causalidad tenga que ser una relación *física.* Puede ser mental o física, dependiendo de si sus relaciones (sus *relata*) son mentales o físicas. Así, la imagen causal de la mente no implica fisicalismo o materialismo. Con todo, la imagen causal del pensamiento es un elemento clave en lo que estoy llamando visión "mecánica" de la mente. Según este modo de ver, la mente es un mecanismo causal: una parte del orden causal de la naturaleza, así como el hígado y el corazón son parte del orden causal de la naturaleza. Y sabemos acerca de las mentes de otros precisamente del mismo modo en que sabemos acerca del resto de la naturaleza: por sus efectos. La mente es un mecanismo que tiene sus efectos en el comportamiento.

¿Por qué habríamos de creer que los estados mentales son causas del comportamiento? Después de todo, una cosa es negar el behaviorismo pero otra es aceptar que los estados mentales son *causas* de comportamiento. Ésta no es una hipótesis trivial, algo que cualquiera que entendiese el concepto de un estado mental aceptaría. De hecho, muchos filósofos lo niegan. Por ejemplo, el punto de vista de que los estados mentales son causas de comportamiento es negado por Wittgenstein y algunos de sus seguidores. Según su modo de ver, describir la mente en términos de causas y mecanismos es cometer el error de imponer un modelo de explicación que sólo es realmente apropiado para las cosas y los sucesos no mentales. "El error —escribe G. E. M. Anscombe, discípulo de Wittgenstein— es pensar que la relación de ser hecho como ejecución de cierta intención, o de ser hecho intencionalmente, es una relación causal entre acto e intención."[13]

¿Por qué alguien pensaría esto? ¿Cómo podría argüirse que los estados mentales no son las causas del comportamiento? Pues bien, considérese el ejemplo del fenómeno mental del *humor.* Podemos distinguir entre el estado mental (o, más precisamente, el hecho) de entretenerse y las manifestaciones de ese estado: reír, sonreír, y así sucesivamente. Necesitamos hacer esta distinción, por supuesto, porque alguien puede divertirse silenciosamente, y alguien puede pretender divertirse y convencer a los demás de que está genuinamente divirtiéndose. ¿Esta distinción significa

[13] G. E. M. Anscombe, "The Causation of Behaviour", en C. Ginet y S. Shoemaker (eds.), *Knowledge and Mind* (Cambridge, Cambridge University Press, 1983), p. 179. Para otra exposición influyente, no causal, de la relación entre razón y acción, véase A. Melden, *Free Action* (Londres, Routledge and Kegan Paul, 1961).

que tenemos que pensar en el estado interior de divertirse y *causar* las manifestaciones externas? Los oponentes del punto de vista causal de la mente dicen que no. Debemos, más bien, pensar en la risa (en un caso genuino de diversión) como la *expresión* de entretenimiento. Expresar diversión en este caso no debe considerarse como un efecto de un estado interior, sino, mejor, como *constituyente* parcial de lo que es divertirse. Pensar en el estado interno de causar la expresión externa sería entendido como equívoco, como pensar en algunos hechos ocultos que una imagen (o una pieza de música) explica. Como dice Wittgenstein, "el habla con y sin pensamiento debe compararse con la ejecución de una pieza de música con o sin pensamiento".[14]

Esto puede ayudar a dar alguna idea de por qué algunos filósofos rechazan la imagen causal del pensamiento. Dada esta oposición, necesitamos razones para creer en la imagen causal del pensamiento. ¿Qué razones se pueden dar? Aquí mencionaré dos razones que apoyan la imagen causal. El primer argumento deriva de ideas de Donald Davidson.[15] El segundo es un argumento más general e "ideológico", que depende de aceptar cierta imagen del mundo en lugar de aceptar que cierta conclusión se sigue decisivamente de cierto conjunto de premisas indiscutibles.

El mejor modo de introducir el primer argumento es con un ejemplo. Considérese alguien, llamémoslo Boleslav, que desea matar a su hermano. Supongamos que está celoso de

[14] Véase Ludwig Wittgenstein, *Philosophical Investigations,* § 341. Para una excelente introducción al pensamiento de Wittgenstein acerca de estas cuestiones, véase Marie McGinn, *Wittgenstein and the Philosophical Investigations* (Londres, Routledge, 1995).

[15] Véase "Actions, Reasons and Causes", en Davidson, *Essays on Actions and Events* (Oxford, Oxford University Press, 1980).

él, pues siente que está frustrando su progreso propio en la vida. Podemos decir que Boleslav tiene una *razón* para matar a su hermano: podemos no pensar que es una razón muy buena, o muy moral, pero no deja de ser una razón. Una razón (en este sentido) es únicamente una colección de pensamientos coherentes con un plan dado de acción. Ahora, supóngase que Boleslav está metido una noche en una riña de bar, por razones completamente desconectadas de su plan criminal, y accidentalmente mata a un hombre que, desconocido para él, es su hermano (tal vez su hermano está disfrazado). De manera que Boleslav tiene una razón para matar a su hermano, y mata a su hermano, pero no habrá matado a su hermano *por esa razón.*

Compárese esta otra historia: Boleslav quiere matar a su hermano, por la misma razón. Va al bar, reconoce a su hermano y lo mata de un tiro. En este caso, Boleslav tiene una razón para matar a su hermano, y mata a su hermano por esa razón.

¿Cuál es la diferencia entre los dos casos? O, para decirlo de otra manera, ¿qué implica realizar una acción *por* una razón? La imagen causal de los pensamientos da una respuesta: alguien ejecuta una acción por una razón cuando su razón es una *causa* de su acción. De manera que, en el primer caso, el plan fratricida de Boleslav no causa la muerte de su hermano, aun cuando tuvo una razón para proceder así, y realizó el acto. Sin embargo, en el segundo caso el plan fratricida de Boleslav fue la causa de su acción. Es la diferencia en la causalidad del comportamiento de Boleslav la que diferencia los dos casos.

¿Cuán plausible es decir que la razón de Boleslav (su manojo asesino de pensamientos) fue la causa del crimen en el segundo caso pero no en el primero? Bien, recuérdense

los rasgos de causalidad mencionados antes; apliquemos dos de ellos a este caso. (No tendré en cuenta la conexión entre causación mental y leyes; esto será discutido en el siguiente apartado.)

En primer lugar, el rasgo contrafactual: parece correcto decir que, en el primer caso, siendo iguales otras cosas (es decir, manteniendo todas las otras circunstancias tan iguales como es posible), si Boleslav no hubiera tenido sus pensamientos fratricidas, entonces de todas maneras habría matado a su hermano. Matar a su hermano en la riña es independiente de sus pensamientos fratricidas. No obstante, en este segundo caso no es así.

Segundo, el rasgo explicativo de la causalidad. Cuando preguntamos "¿por qué Boleslav mató a su hermano?" en el primer caso, no es una buena respuesta decir que "porque estaba celoso de su hermano". Sus celos no *explican* por qué mató a su hermano en este caso; no mató a su hermano *a causa* de los deseos fratricidas que tenía. En el segundo caso, sin embargo, matar a su hermano se explica por los pensamientos fratricidas: debemos tratarlos como causa.

Lo que la argumentación sostiene es que *necesitamos* distinguir entre estos dos géneros de caso, y que *podemos* distinguirlos pensando en la relación entre razón y acción como una relación causal. Y esto nos da una respuesta a la cuestión: ¿qué es hacer algo por una razón, o qué es actuar según una razón? La respuesta es: actuar según una razón es tener esa razón como causa de la acción de uno.

Pienso que esta argumentación es convincente. Pero no se impone absolutamente. Pues el argumento mismo no excluye una explicación alternativa de lo que es actuar con una razón. La estructura del argumento es como sigue: aquí

hay dos situaciones que evidentemente difieren; necesitamos explicar la diferencia entre éstas; recurriendo a la causalidad se logra explicar la diferencia entre ellas. Esto puede ser correcto, pero nótese que no excluye la posibilidad de que haya alguna otra explicación aún *mejor* de lo que es actuar con una razón. Está abierto, por lo tanto, al oponente de la imagen causal del pensamiento responder a la argumentación ofreciendo otra posible explicación. Así, la primera argumentación no persuadirá a este oponente.

Sin embargo, es útil ver este argumento de Davidson en su contexto histórico. La argumentación es una entre muchas otras que surgieron en oposición al punto de vista anterior, que atribuí a Wittgenstein y a sus seguidores: el punto de vista de que es un error pensar en la mente en términos causales en absoluto. Estas otras argumentaciones apuntaban a mostrar que hay un componente causal esencial en muchos conceptos mentales. Por ejemplo, la *percepción* se analizaba como si implicara una relación causal entre quien percibe y el objeto percibido; la *memoria* se analizaba como si implicara una relación entre ésta y el hecho recordado; el conocimiento y la relación entre lenguaje y realidad se consideraban fundamentalmente basados en relaciones causales.[16] La argumentación de Davidson es parte de un movimiento que analizó muchos conceptos

[16] Acerca de la percepción, véase H. P. Grice, "The Causal Theory of Perception", en J. Dancy (ed.), *Perceptual Knowledge* (Oxford, Oxford University Press, 1988); acerca de la memoria, véase C. B. Martin y Max Deutscher, "Remembering", *Philosophical Review,* 75 (1966); para el conocimiento, véase Alvin Goldman, "A Causal Theory of Knowing", *Journal of Philosophy,* 64 (1967); para el lenguaje y la realidad, véase Dennis D. W. Stampe, "Toward a Causal Theory of Linguistic Representation", *Midwest Studies in Philosophy,* II (1977).

mentales en términos de causalidad. Frente a este fundamento puedo introducir mi segundo argumento en pro de la imagen causal del pensamiento.

La segunda argumentación es la que llamo argumento ideológico. La llamo así porque depende de aceptar determinada imagen del mundo: la imagen mecánica/causal. Esta imagen ve la naturaleza completa como si obedeciera a ciertas leyes generales causales —las leyes de la física, la química, la biología, etc.— y sostiene que la psicología también tiene sus leyes y que la mente se ajusta al orden causal de la naturaleza. A través de la naturaleza encontramos la causalidad, la sucesión regular de sucesos y la determinación de uno de éstos por otro. ¿Por qué estaría la mente exenta de esta clase de determinación?

Después de todo, la mayoría de la gente cree que los estados mentales pueden ser *afectados* por causas del mundo físico: los colores que ve usted, las cosas que huele, la comida que saborea, las cosas que usted escucha, todas estas experiencias son el resultado de ciertos procesos físicos puramente mecánicos fuera de la mente. Todos sabemos cómo nuestra mente puede ser afectada por productos químicos —estimulantes, antidepresivos, narcóticos, alcohol—, y en todos estos casos esperamos una conexión regular, como una ley, entre tomar la sustancia química y la naturaleza del pensamiento. Así, si los estados mentales pueden ser efectos, ¿cuáles se supone que son las razones de pensar que no pueden también ser causas?

Admito que esto cae muy lejos de ser una argumentación concluyente. Es difícil ver cómo se puede tener una argumentación filosófica *concluyente* para semejante punto de vista general, omniabarcante. Lo que voy a suponer aquí, en cualquier caso, es que, dada esta visión total del mundo

no mental, necesitamos razones harto positivas para creer que el mundo mental no funciona de la misma manera.

Psicología del sentido común

Hasta aquí, por el momento, la idea de que los estados mentales son las causas del comportamiento. Volvamos a la idea de la psicología del sentido común: la idea de que cuando entendemos la mente de otros, empleamos (en algún sentido) una especie de "teoría' que caracteriza o describe estados mentales. Adam Morton ha llamado a esta idea la "teoría teoría" de la psicología del sentido común —es decir, la *teoría* de que la psicología del sentido común es una *teoría*— y tomaré esta etiqueta de él.[17] Para entender esta "teoría teoría" necesitamos saber qué es una teoría y cómo la teoría del sentido común en la psicología se aplica a los estados mentales. Entonces debemos preguntarnos cómo se supone que esta teoría es empleada por los pensadores.

En los términos más generales, podemos pensar en una teoría como en un principio, o colección de principios, que es ideado para explicar ciertos fenómenos. Para que haya una teoría de los estados mentales, pues, tiene que haber una colección de principios que explique los fenómenos mentales. Donde interviene la psicología del sentido común, estos principios podrían ser tan sencillos como las tautologías de que, por ejemplo, "la gente generalmente trata de lograr el objeto de sus deseos (manteniendo iguales otras cosas)" o que "si una persona mira un objeto que tiene

[17] Adam Morton, *Frames of Mind* (Oxford, Oxford University Press, 1980), p. 7.

delante, con buena luz, creerá normalmente que el objeto está delante de él/ella (manteniendo iguales otras cosas)". (La aparente trivialidad de estas tautologías se discutirá más adelante.)

Sin embargo, del modo como se entiende normalmente, la pretensión de que la psicología del sentido común es una teoría no es sencillamente la pretensión de que hay principios que describen el comportamiento de estados mentales. Lo que se quiere decir, aparte de esto, es que los estados mentales son lo que los filósofos llaman "entidades teóricas".[18] Esto es, no son justamente estos estados mentales los describibles por una teoría, sino también la (verdadera, completa) teoría de los estados mentales *nos dice todo lo que hay que saber acerca de ellos.* Compárese la teoría del átomo. Si conociéramos una colección de principios generales que describieran la estructura y el comportamiento del átomo, nos dirían todo lo que necesitamos conocer acerca de los átomos en general, ya que todo lo que hay que saber acerca de los átomos está contenido en la teoría completa y verdadera del átomo. (Contrástese con los colores: puede argüirse la falsedad de que todo lo que sabemos acerca de los colores está contenido dentro de la teoría física de los colores. También sabemos *cómo son* los colores, lo cual no es algo que pueda darse teniendo conocimiento de la teoría de los colores.)[19] Los átomos son entidades teóricas, no sólo en

[18] Acerca de entidades teóricas, véase David Lewis, "How to Define Theoretical Terms", en sus *Philosophical Papers,* vol. 1 (Oxford, Oxford University Press, 1985). La idea deriva de F. P. Ramsey, "Theories", en sus *Philosophical Papers,* D. H. Mellor (ed.) (Cambridge, Cambridge University Press, 1991). Para una buena exposición de la pretensión de que los estados mentales son entidades teóricas, véase Stephen P. Stich, *From Folk Psychology to Cognitive Science* (Cambridge, MIT Press, 1983).

[19] Para un punto de vista contrario, véase J. J. C. Smart, *Philosophy and*

el sentido de que son planteamientos de una teoría, sino también porque su naturaleza se agota por la descripción de ellos que da la teoría. Igualmente, según la "teoría teoría" todo lo que sabemos acerca de, digamos, la *creencia,* está contenido en la teoría completa de la creencia.

Una analogía pudiera ayudar a aclarar el punto.[20] Piénsese en la teoría como si fuera una *historia.* Considérese una historia que dice: "Una vez había un hombre llamado Rey Lear, que tenía tres hijas, llamadas Goneril, Regan y Cordelia. Un día les dijo..." y así sucesivamente. Ahora bien, si se pregunta "¿quién era el Rey Lear?", una respuesta perfectamente correcta sería parafrasear alguna parte de la historia: "el Rey Lear es el hombre que dividió su reino, desheredó a su hija favorita, enloqueció y terminó entre las matas" y así sucesivamente. Empero, si uno pregunta: "¿el Rey Lear tenía un hijo?, ¿qué le ocurrió? o ¿qué clase de peinado usaba el Rey Lear?", la historia no tiene respuesta. Pero no es que haya algún hecho acerca del hijo de Lear o sobre su peinado que la historia deje sin mencionar; más bien es que todo lo que hay que saber acerca de Lear está contenido dentro de la historia. Pensar que podría haber más es errar en ella. Igualmente, pensar que hay más acerca de los átomos que lo contenido dentro de la teoría completa y verdadera de los átomos es (desde este punto de vista de las teorías) no conseguir apreciar que los átomos son entidades teóricas.

La analogía con la psicología del sentido común es ésta.

Scientific Realism (Londres, Routledge and Kegan Paul, 1963), y D. M. Armstrong, *A Materialist Theory of the Mind,* cap. 12.

[20] Oí a R. B. Braithwaite sugerir esta analogía en un programa de radio de D. H. Mellor acerca de la filosofía de F. P. Ramsey, "Better than the Stars", BBC Radio, 3 (27 de febrero de 1978).

La teoría de la creencia, por ejemplo, podría decir algo como: "Hay estos estados, creencias, que interactúan causalmente con deseos para causar acciones..." y así por el estilo, haciendo la lista de los hechos familiares acerca de las creencias y sus relaciones con otros estados mentales. Una vez que se ha establecido la lista de todos estos hechos familiares, la lista da una "definición teórica" del término "creencia". La naturaleza de las creencias será, vistas así las cosas, enteramente agotada por estas tautologías acerca de las creencias. No hay más en la teoría de las creencias que no esté contenido dentro de la teoría de la creencia; y lo mismo con otros tipos de pensamiento.[21]

Es importante distinguir, en principio, la idea de que la psicología del sentido común es una teoría de la imagen causal de los pensamientos como tales. Uno podría aceptar la imagen causal de los pensamientos —que, recuérdese, es sencillamente la pretensión de que los pensamientos tienen efectos sobre el comportamiento— sin aceptar la idea de la psicología del sentido común como una teoría (véase "Teoría y simulación", p. 133). Sería también posible negar la teoría causal de los pensamientos —negar que los pensamientos tengan efectos— aunque aceptando la concepción de la psicología del sentido común como teoría. Este modo de ver puede ser considerado por alguien, escéptico acerca de la causación, por ejemplo, aunque éste sería un modo muy desacostumbrado de ver.

Teniendo esto presente, necesitamos decir más acerca de cómo se supone que funciona la "teoría teoría" y qué dice la teoría acerca de lo que son los pensamientos. Tomemos

[21] Éste es el enfoque tomado por David Lewis en "Psychophysical and Theoretical Identification", en Ned Block (ed.), *Readings in the Philosophy of Psychology* (Londres, Methuen, 1980), vol. 1.

otro ejemplo sencillo y cotidiano. Supóngase que vemos a una mujer corriendo por una calle vacía, llevando diversas bolsas, mientras la alcanza un autobús, acercándose a una parada. ¿Qué está haciendo ella? La respuesta obvia es: corre para tomar el autobús. Las reflexiones anteriores de este capítulo deben hacernos conscientes de que hay diversas posibilidades aparte de la respuesta evidente: tal vez piensa que alguien está persiguiéndola, o tal vez simplemente quiere hacer ejercicio. Pero dado el hecho de que la calle esté, por lo demás, vacía, y que la gente no haga ejercicio llevando grandes bolsas, alcanzamos la conclusión evidente.

Como con nuestro ejemplo anterior, descartamos las interpretaciones más desacostumbradas porque no nos parecen razonables o racionales. Dando esta interpretación de su comportamiento suponemos cierto grado de racionalidad en la mente de la mujer: suponemos que está persiguiendo su meta inmediata (tomar el autobús), sin duda con objeto de alcanzar alguna meta a largo plazo (llegar a casa). Suponemos esto porque hay, a nuestro modo de ver, cosas razonables que hacer, y ella está usando modos razonables de intentar hacerlas (en oposición, digamos, a acostarse en mitad del suelo, delante del autobús, esperando que el chofer la recoja).

Decir esto no es negar la existencia de un comportamiento irracional y estrambótico. Claro que no. Si todo el comportamiento fuese irracional y estrambótico, no podríamos hacer estas hipótesis acerca de lo que está pasando por la mente de la gente. No sabríamos cómo elegir entre una hipótesis estrafalaria y otra. A fin de que la interpretación de otros pensadores sea posible en general, entonces tenemos que suponer que hay cierta regularidad en la conexión entre pensamiento y comportamiento. Y si la relación

entre los pensamientos de la gente y su conducta ha de ser suficientemente regular para permitir interpretaciones, entonces es natural esperar que la psicología del sentido común contendrá generalizaciones que detallan estas regularidades. De hecho, si la psicología del sentido común realmente es una teoría, esto es lo que debiéramos esperar en todo caso, pues una teoría es (por lo menos) una colección de principios o leyes generales.

Así, la siguiente pregunta es: ¿hay algunas generalizaciones psicológicas? El escepticismo acerca de tales generalizaciones puede proceder de múltiples fuentes. Un modo común de escepticismo se funda en la idea de que, si hubiera generalizaciones psicológicas, de seguro (como psicólogos del sentido común que somos) las conoceríamos. De hecho, lo hacemos muy mal si debemos traer a la mente cualesquiera generalizaciones plausibles. Como dice Adam Morton, "los principios como 'cualquiera que piensa que hay aquí un tigre dejará el cuarto' son... casi siempre falsos".[22] Y cuando realmente logramos traer a la mente algunas verdaderas generalizaciones, pueden ser objeto de desencanto; considérese nuestro ejemplo anterior: "La gente generalmente trata de alcanzar el objeto de sus deseos (siendo iguales otras cosas)". Nos inclinamos a decir: "¡Por supuesto! ¡Díganme algo que yo no conozca!" Aquí tenemos a Morton una vez más:

> Lo más notable acerca de la psicología del sentido común... es la combinación de un poder explicativo vigoroso y versátil con una gran ausencia de hipótesis poderosas o atrevi-

[22] Adam Morton, *Frames of Mind,* p. 37. Véase también Stephen Schiffer, *Remnants of Meaning* (Cambridge, MIT Press, 1987), pp. 28-31.

> das. Cuando uno trata de sacar a luz principios de explicación psicológica generalmente usados en la vida cotidiana, se encuentran únicamente tautologías sosas, y sin embargo, en casos particulares, se producen hipótesis interesantes, osadas y agudas, acerca de por qué una persona... actúa de determinada manera.[23]

Hay evidentemente algo cierto acerca de esta cuestión; pero tal vez es un poco exagerada. Después de todo, si la "teoría teoría' es correcta acerca de la psicología del sentido común, estamos empleando esta teoría todo el tiempo, cuando interpretamos el uno al otro. Así que difícilmente sorprenderá si encontramos "tautológicas" las generalizaciones que usamos. Serán tautológicas porque son familiares —pero esto no quiere decir que no sean poderosas—. Compárese nuestra teoría cotidiana de los objetos físicos: "la física popular". Sabemos que los objetos sólidos resisten la presión y la penetración por otros objetos. Esto es, en un sentido, una tautología, pero es una tautología de las que guían en el trato con el mundo de los objetos.

Otra manera en que el defensor de la "teoría teoría" puede responder es diciendo que es sólo mediante el supuesto de que tenemos *algún* conocimiento de la teoría psicológica de otras mentes como podemos explicar satisfactoriamente cómo conseguimos interpretar tan felizmente a otras personas. Sin embargo, este conocimiento no necesita ser explícitamente conocido por nosotros —esto es, no necesitamos traer este conocimiento a nuestras mentes conscientes—. No obstante, este conocimiento inconsciente, como el conocimiento matemático del esclavo de Menón que fue

[23] Adam Morton, *op. cit.*, p. 28.

discutido en el capítulo I (véase "Pensamiento y conciencia", p. 57) está aquí, con todo. Y explica cómo nos entendemos los unos a los otros justamente como (digamos) el conocimiento inconsciente o "tácito" de reglas lingüísticas explica cómo entendemos el lenguaje. (Volveremos a esta idea en el capítulo IV.)

Hasta aquí, pues, he sostenido que la psicología del sentido común opera suponiendo que la gente es en gran medida racional, y suponiendo la verdad de ciertas generalizaciones. Podríamos no conseguir enunciar todas estas generalizaciones. Pero dado que conocemos algunas de ellas —hasta las "tautologías sosas"— podemos ahora preguntar: ¿qué dicen las generalizaciones de la psicología del sentido común que son los pensamientos mismos?

Volvamos al ejemplo de la mujer que corre en pos del autobús. Si alguien preguntara por qué la interpretamos como corriendo tras el autobús, una cosa que podríamos decir es: "Pues bien, es evidente: el autobús llega". Sin embargo, cuando se piensa en ello, esto no es del todo exacto. Pues no es el hecho de que el autobús llegue lo que hace que haga lo que hace, sino el hecho de que ella *cree* que el autobús llega. Si el autobús estuviera llegando y ella no se diese cuenta, entonces no andaría corriendo detrás del autobús. Igualmente, si creyese que el autobús llegaba aunque en realidad no lo hiciera (tal vez al confundir el ruido de un camión con el ruido del autobús), de todas maneras correría.

En términos más generales, lo que la gente hace está determinado por cómo cree que es el mundo y cómo un pensador considera que el mundo no es siempre como es el mundo (todos nos equivocamos). Decir que un pensador "considera el mundo" de determinada manera es sólo otro

modo de decir que el pensador *representa* el mundo como si fuera de determinada manera. Así, lo que los pensadores hacen es determinado por cómo representan el mundo. Esto es, según la psicología del sentido común, los pensamientos que determinan el comportamiento son *representacionales.*

Nótese que es *como* las cosas son representadas en el pensamiento lo que importa para la psicología del sentido común, no nada más *qué* objetos están representados. Alguien que cree que el autobús llega debe representar el autobús *como un autobús* y no (por ejemplo) sólo como un *vehículo motorizado de alguna clase,* pues ¿por qué alguien correría en pos de un vehículo motorizado de alguna clase? O considérese a Boleslav: aunque mató a su hermano en la primera escena y se representó a su hermano de alguna manera, no se lo representó *como su hermano,* y por eso su deseo de matar a su hermano no es la causa del crimen. (Recuérdese el ejemplo de Orwell en el capítulo 1: "Intencionalidad".)

La otra parte central de la concepción de sentido común, por lo menos de acuerdo con la imagen causal de los pensamientos, es que éstos son las causas del comportamiento. La concepción de sentido común afirma que cuando damos una explicación del comportamiento de alguien en términos de creencias y deseos, la explicación cita las causas del comportamiento. Cuando decimos que la mujer corre en pos del autobús *porque* cree que el autobús está llegando y ella quiere volver a casa en él, este *porque* expresa causalidad, ni más ni menos que *porque* en "le dio cáncer *porque* fumaba" expresa causalidad.

Combinando la imagen causal del pensamiento con la "teoría teoría", llegamos a lo siguiente: la psicología del senti-

do común contiene generalizaciones que describen los efectos y efectos potenciales de tener ciertos pensamientos. Verbigracia: los sencillos casos que hemos discutido son ejemplos en los que alguien depende de lo que él o ella cree y lo que él o ella quiere o desea. Así, la imagen causal más la "teoría teoría" dirían que la psicología del sentido común contiene una generalización o manojo de generalizaciones acerca de cómo creencias y deseos interactúan para causar acciones. Un intento tosco de formular una generalización sería: "Las creencias se combinan con deseos para causar acciones que apuntan a la satisfacción o cumplimiento de estos deseos".[24]

Así, por ejemplo, si deseo un vaso de vino y creo que hay algo de vino en el refrigerador, y creo que el refrigerador está en la cocina, y creo que la cocina está aquí al lado, esto provocará que actúe yo de una manera que apunte a la satisfacción del deseo: por ejemplo podría moverme hacia el refrigerador. (Para más acerca de esto, véase en el capítulo v "Representación mental y éxito en la acción".)

Por supuesto, yo no podría hacerlo, aunque tuviera todas estas creencias y este deseo, si tuviera otro deseo, más fuerte, de conservar clara la cabeza, o si creyera que el vino pertenecía a alguien más y pensara que no debía tomarlo. Entonces puedo no actuar en mi deseo de un vaso de vino. Esto no socava la generalización, ya que ésta es compatible con cualquier número de deseos que interactúan para ocasionar mi acción. Si mi deseo de mantener clara la cabeza es más fuerte que mi deseo de tomar una bebida, entonces será causa de una acción diferente (evitar el refrigerador, ir a pasear por el campo, o cualquier cosa así).

[24] Véase Robert Stalnaker, *Inquiry* (Cambridge, MIT Press, 1984), capítulo 1.

Todo lo que la generalización dice es que uno actuará de una manera que se oriente a satisfacer los deseos de uno, cualesquiera que sean.

Vale la pena recalcar otra vez que los cursos de pensamiento como éste no se supone que crucen la mente consciente de uno. Alguien que quiere una bebida difícilmente pensará, de modo consciente: "quiero una bebida; la bebida está en el refrigerador; el refrigerador está ahí al lado; por lo tanto debería ir ahí", y así sucesivamente. (Si esto es lo que él o ella están pensando de manera consciente, entonces probablemente no es recomendable tomar otra bebida.) La idea es, más bien, que hay pensamientos inconscientes, con estos contenidos representacionales, que provocan el comportamiento de un pensador. Estos pensamientos son los "resortes" causales de las acciones del pensador, no necesariamente los ocupantes de sus mentes conscientes.

Al menos eso es lo que dice la versión causal de la "teoría teoría"; es tiempo ahora de valorar la "teoría teoría". Para valorarla necesitamos hacernos dos preguntas centrales. Primero, ¿da la "teoría teoría" una descripción correcta de nuestra comprensión psicológica cotidiana entre unos y otros? Es decir, ¿es correcto hablar acerca de psicología del sentido común como una clase de teoría a fin de cuentas, o debe entenderse de alguna otra manera? (Téngase en cuenta que rechazar la "teoría teoría" con estos fundamentos no es *ipso facto* rechazar la imagen causal de los pensamientos.)

La segunda pregunta es que aun si nuestra comprensión psicológica cotidiana entre unos y otros es una teoría, ¿es una *buena* teoría? Esto es, supóngase que la colección de principios y simplezas acerca de creencias y deseos que causan acciones (y demás), que estoy llamando psicología del sentido común, es en realidad una teoría de las mentes

humanas; ¿hay algunas razones para pensar que es una teoría verdadera de las mentes humanas? Esto podría parecer una pregunta rara, pero, según veremos, nuestra actitud ante ella puede afectar toda nuestra actitud hacia la mente.

Será más sencillo si tomamos estas preguntas en orden inverso.

La ciencia del pensamiento: ¿eliminación o vindicación?

Supongamos, entonces, que la psicología del sentido común es una teoría: la teoría de la creencia, el deseo, la imaginación, la esperanza, el miedo, el amor y demás estados psicológicos que nos atribuimos unos a otros. Al llamar a esta teoría psicología *del sentido común,* los filósofos implícitamente la contrastan con la disciplina científica de la psicología. La psicología del sentido común es una teoría cuyo dominio requiere nada más una mente bastante madura, una pizca de imaginación y alguna familiaridad con otra gente. En este sentido, todos somos psicólogos. La psicología científica, sin embargo, usa muchos conceptos técnicos y cuantitativos que sólo una pequeña proporción de "psicólogos del sentido común" entienden. Ambas teorías pretenden, a pesar de todo, ser teorías de la misma cosa: la mente. ¿De modo que cómo están relacionadas?

No bastaría sencillamente con suponer que de hecho la psicología científica y la psicología del sentido común son teorías de cosas diferentes; la psicología científica es la teoría del cerebro, mientras que la psicología del sentido común es la teoría de la mente o la persona. Hay cuando menos tres razones por las cuales esto no funcionará. En primer lugar, por todo lo que hemos dicho acerca de estas teorías hasta

aquí, la mente podría no *ser* sino el cerebro. Como dije en el capítulo 1, ésta es una cuestión que podemos dejar a un lado al discutir el pensamiento y la representación mental. Sin embargo, cualquiera que sea la conclusión que alcancemos en esto, ciertamente no debemos suponer que sólo porque tenemos dos teorías tenemos dos cosas. (Compárese: el sentido común dice que la mesa es de madera compacta; la física de partículas dice que la mesa es, principalmente, espacio vacío. Es una mala inferencia concluir que hay dos mesas sencillamente porque hay dos teorías.)[25]

En segundo lugar, la psicología científica habla acerca de una multitud de las mismas clases de estados mentales igual que hablamos de ellos en la psicología del sentido común. Los psicólogos científicos tratan de responder algunas de estas preguntas: ¿cómo funciona la memoria?, ¿cómo vemos objetos?, ¿por qué soñamos?, ¿qué son imágenes mentales? Todos estos estados y sucesos mentales —memoria, visión, soñar e imaginería mental— son familiares en la psicología del sentido común. No hay que tener pretensiones científicas para poder aplicar los conceptos de memoria o visión. Tanto la psicología científica como la del sentido común tienen cosas que decir acerca de estos fenómenos. No hay razón para suponer desde el principio que el fenómeno de la visión para un psicólogo científico es diferente de la visión para un "psicólogo" del sentido común.

Finalmente, abundante psicología científica actual es realizada sin referencia al auténtico funcionamiento del

[25] La inferencia, famosa, fue hecha, sin embargo: véase Arthur Eddington, *The Nature of the Physical World* (Cambridge, Cambridge University Press, 1929), pp. xi-xiv.

cerebro. Esto no es porque los psicólogos implicados sean normalmente dualistas cartesianos, sino más bien porque a menudo tiene más sentido ver cómo la mente funciona en gran escala, en términos macroscópicos —en términos de comportamiento ordinario— antes de mirar los detalles de su implementación neural. Así, la idea de que la psicología científica se ocupa sólo del cerebro no es cierta ni aun para la práctica real de la psicología.

Dado que la psicología científica y la psicología del sentido común se ocupan de la misma cosa —la mente—, la cuestión de la relación entre ellas se torna urgente. Hay muchas actitudes posibles con las cuales tomar esta relación, pero al final se reducen a dos: *vindicación* o *eliminación*. Veamos estos dos enfoques.

Según el enfoque de la vindicación, ya sabemos (o tenemos buena razón para creer) que las generalizaciones de la psicología del sentido común son en gran medida ciertas. Así, una de las cosas que podemos esperar de la psicología científica es una explicación de *cómo* o *por qué* son ciertas. Sabemos, por ejemplo, que si los perceptores normales miran un objeto con buena luz, con nada en el camino, creerán que el objeto está delante de ellos. Así, una de las metas de una psicología científica de la visión y la cognición es explicar por qué esta humilde verdad es, de hecho, verdadera: qué tiene que ver con nosotros, en relación con nuestros cerebros y nuestros ojos, y en relación con la luz, que hace posible para nosotros ver objetos y formar creencias acerca de ellos sobre la base de verlos. El enfoque de la vindicación podría usar una analogía con la física de sentido común. Antes de Newton la gente ya sabía que si un objeto es lanzado por el aire, con el tiempo volverá al suelo. Pero hizo falta la física de Newton para explicar *por qué* esta verdad

es, de hecho, verdadera. Y así es como las cosas serán con la psicología del sentido común.[26]

En contraste, el enfoque de la eliminación afirma que hay muchas razones para dudar de si la psicología del sentido común es verdadera. Y si no es verdadera, entonces debemos dejar que la ciencia de la mente o el cerebro se desarrollen sin tener que emplear las categorías de la psicología del sentido común. La psicología científica no está obligada a explicar por qué las generalizaciones del sentido común son verdaderas, ¡porque hay buenas razones para pensar que no lo son! Así, debiéramos esperar que la psicología científica acabaría por eliminar la del sentido común, antes que vindicarla. Este enfoque utiliza una analogía con teorías desacreditadas tales como la alquimia. Los alquimistas pensaban que había una "piedra filosofal" que podía convertir el plomo en oro. Sin embargo, la ciencia no mostró por qué esto era cierto: no era cierto y la alquimia acabó siendo eliminada. Y así es como las cosas serán con la psicología del sentido común.[27]

En vista de que los proponentes del enfoque de la eliminación son siempre materialistas, éste es conocido como *materialismo eliminativo.* Según uno de sus principales defensores, Paul Churchland:

> El materialismo eliminativo es la tesis de que nuestro concepto de sentido común de los fenómenos psicológicos constituye una teoría radicalmente falsa, una teoría tan

[26] El enfoque vindicatorio ha sido defendido por Jerry Fodor: véase *Psychosemantics* (Cambridge, MIT Press, 1987), cap. 1.

[27] Acerca del enfoque eliminativo, véase especialmente Paul M. Churchland, "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy,* 78 (1981), y Patricia S. Churchland, *Neurophilosophy* (Cambridge, MIT Press, 1986).

> fundamentalmente defectuosa que tanto sus principios como su ontología acabarán por ser desplazadas... por una neurociencia completada.

Por "la ontología de la teoría", Churchland quiere decir esas cosas que la teoría pretende que existen: deseos, intenciones y algo por el estilo. ("Ontología" es el estudio del ser, de lo que existe.) Así, decir que la ontología de la psicología del sentido común es defectuosa equivale a decir que la psicología del sentido común es errónea acerca de lo que está en la mente. De hecho, los materialistas eliminativos normalmente pretenden que ninguno de los estados mentales que postula la psicología del sentido común existe. Esto es, no hay creencias, deseos, intenciones, recuerdos, esperanzas, miedos ni nada por el estilo.

Esto podría llamar la atención como un punto de vista increíble. ¿Cómo puede cualquier persona razonable *pensar* que no hay *pensamientos*? ¿No es cosa que se refute sola como *decir* que no hay *palabras*? No obstante, antes de valorar el punto de vista, adviértase cuán suavemente parece desprenderse de la concepción de la psicología del sentido común como teoría y de los estados mentales como entidades teóricas mencionados en la sección anterior. Recuérdese que, según esta concepción, toda la naturaleza de los pensamientos es descrita por la teoría. La respuesta a la pregunta de "¿qué son los pensamientos?" es: "Los pensamientos son lo que la teoría de los pensamientos dice que son". Así, si la teoría de los pensamientos resulta ser falsa, es en gran medida verdadera, o es que no hay pensamientos en absoluto. (Compárese: los átomos son lo que la teoría de los átomos dice que son. No hay nada más en ser un átomo que lo que la teoría dice; así que si la teoría es falsa, no hay átomos.)

Los materialistas eliminativos adoptan el punto de vista de que la psicología del sentido común es una teoría, y entonces arguyen que la teoría es falsa.[28] ¿Por qué creen que la teoría es falsa? Una razón que dan es que (contrariamente al enfoque de la vindicación) la psicología del sentido común no explica realmente gran cosa: "La naturaleza y dinámica de la enfermedad mental, la facultad de la imaginación creativa... la naturaleza y función psicológica del sueño... la rica variedad de ilusiones perceptuales... el milagro de la memoria... la naturaleza del proceso de aprendizaje en sí mismo...",[29] todos estos fenómenos, según Churchland, son "completamente misteriosos" a la psicología del sentido común y probablemente sigan siéndolo. Una segunda razón para rechazar la psicología del sentido común es que está "estancada", ha mostrado escasa señal de desarrollo a través de su larga historia (cuya duración Churchland, no poco arbitrariamente, establece en 25 siglos).[30] Una tercera razón es que parece haber poca probabilidad de que las categorías de la psicología del sentido común (creencia, deseo y cosas así) se "reducirán" a categorías físicas, es decir, parece muy improbable que los científicos consigan decir, de modo detallado y sistemático, qué fenómenos físicos sustentan las creencias y los deseos. (Recuérdese el absurdo de "el amor materno es magnesio".) Si esto no puede hacerse, argumenta Churchland, hay poca probabilidad de hacer que la psicología del sentido común sea científicamente respetable.

[28] Para un planteamiento particularmente claro de esta línea de argumentación, véase especialmente Stephen Stich, *From Folk Psychology to Cognitive Science* (Cambridge, MIT Press, 1985).

[29] Churchland, "Eliminative Materialism and the Propositional Attitudes", p. 73.

[30] *Ibid.*, p. 76.

Antes de valorar estas razones debemos retornar a la cuestión que seguramente todavía le preocupa a usted: ¿cómo puede alguien creer que no hay creencias? De hecho, ¿cómo puede alguien incluso afirmar tal teoría? Pues afirmar algo es expresar una creencia en ello; pero si el materialismo eliminativo es correcto, entonces no hay creencias, de modo que nadie puede expresarlas. Así, ¿no son los materialistas eliminativos, por sus propias luces, sólo ondas sonoras que suenan y vibran con sonidos sin sentido? ¿No se refuta a sí misma su teoría?

Churchland ha respondido a este argumento trazando una analogía con la creencia del siglo XIX en el *vitalismo,* la tesis de que no es posible explicar la diferencia entre cosas vivas y no vivas en términos plenamente fisicoquímicos, sino sólo recurriendo a la presencia de un espíritu vital o "entelequia" que explica la presencia de la vida. Imagina a alguien arguyendo que la negación del vitalismo ("antivitalismo") se autorrefuta:

> Mi sabio amigo ha afirmado que no hay cosas tales como un espíritu vital. Mas la afirmación es incoherente. Pues si es cierta, entonces mi amigo no tiene espíritu vital, y por lo tanto debe estar *muerto.* Si está muerto, entonces su afirmación es sólo una cadena de ruidos, despojada de sentido o verdad. Evidentemente, el supuesto de que el antivitalismo es cierto ¡implica que no puede ser cierto! QED.[31]

El argumento parodiado es éste: los vitalistas sostenían que estaba en la naturaleza del estar vivo que el cuerpo de

[31] Paul M. Churchland, *Matter and Consciousness* (Cambridge, MIT Press, 1984), p. 48.

uno contuviera una entelequia vital, de modo que alguien que rechazara la existencia de entelequia vital pretendía en efecto que nada está vivo (incluyéndose ellos mismos). Éste es un mal argumento. Churchland pretende que la carga de la autorrefutación contra el materialismo eliminativo implica un argumento igualmente malo: que es afirmar algo, de acuerdo con la psicología del sentido común es expresar una creencia en ella; de manera que cualquiera que niegue la existencia de creencias cree, en efecto, que nadie afirma nada (incluyendo los materialistas eliminativos).

Ciertamente, el argumento a favor del vitalismo es malo. No obstante, la analogía no es muy persuasiva. Pues en tanto que podemos fácilmente encontrar sentido en la idea de que la vida podría no implicar una entelequia vital, es muy arduo encontrar sentido en la idea análoga de que la afirmación no implicaría la expresión de una creencia. La aserción misma es una noción de la psicología del sentido común: afirmar algo es pretender que es cierto. En este sentido, la afirmación es próxima a la idea de creencia: creer algo es sostenerlo como verdadero. Así, si la psicología del sentido común es eliminada, la aserción y la creencia tendrán que irse.[32]

Churchland puede responder que no debiéramos dejar que el desenvolvimiento futuro de la ciencia fuera dictado por lo que podemos o no imaginar o encontrar sentido en ello. En el siglo XIX hubo gente que no podía encontrar sentido en la idea de que la vida no consistiera en "una entelequia" vital; esta gente era víctima de las limitaciones de sus propias imaginaciones. Así, por supuesto, aunque es

[32] Véase Hilary Putnam, *Representation and Reality* (Cambridge, MIT Press, 1988).

una buena idea tener conciencia de nuestros límites cognitivos, semejante precaución por sí misma no nos acerca a la posición eliminativa.

No necesitamos poner en claro este punto acerca de la autorrefutación a fin de afirmar el materialismo eliminativo; pues, cuando se examina, los argumentos positivos en apoyo del punto de vista no son muy persuasivos que se diga. Los repasaré brevemente.

En primer lugar, tómese la idea de que la psicología del sentido común no ha explicado mucho. En vista de esto, el hecho de que la teoría que explica el comportamiento en términos de creencias y deseos no explica también por qué dormimos (ni las otras cosas mencionadas antes), no es *en sí misma* una razón para rechazar creencias y deseos. Pues ¿por qué tendría la teoría de las creencias y los deseos que explicar el sueño? Esta respuesta parece requerir demasiado del punto de vista de la vindicación.

En segundo lugar, consideremos el cargo de que la psicología del sentido común está "estancada". Esto es muy discutible. Un ejemplo notable de cómo la teoría del sentido común de la mente parece haber cambiado es el lugar que asigna a la conciencia (véase el capítulo 1). Se acepta ampliamente que, desde Freud, mucha gente de Occidente cree que hay sentido en suponer que algunos estados mentales (por ejemplo los deseos) no son conscientes. Éste es un cambio en la visión de la mente que puede plausiblemente ser considerado parte del sentido común.

En cualquier caso, aun si la psicología del sentido común no ha cambiado mucho con los siglos, esto no establecería en sí mismo gran cosa. El hecho de que una teoría no haya cambiado durante muchos años puede ser un signo de estancamiento de la teoría o de que está extremada-

mente *bien* establecida. Cuál de éstos sea el caso, depende de cuán buena sea la teoría para explicar los fenómenos, no de la ausencia de cambio como tal. (Compárese: la física del sentido común cree que los cuerpos sin apoyo caen al suelo, y no ha cambiado durante muchos siglos. ¿Concluiremos que esta creencia del sentido común está estancada?)

En tercer lugar, está el punto de si las categorías de la psicología popular pueden reducirse a categorías físicas (o neurofisiológicas). Aquí el supuesto es: a fin de que una teoría sea científicamente respetable, tiene que ser reducible a la física. Éste es un supuesto muy extremo y, según sugerí en la introducción, no tiene que aceptarse a fin de aceptar la idea de que la mente puede ser explicada por la ciencia. En este caso, el enfoque de la vindicación puede rechazar el reduccionismo sin rechazar la explicación científica de la mente.[33]

Así, al menos si no se autorrefutan a fin de cuentas, los argumentos del materialismo eliminativo no son muy convincentes. Las razones específicas que los materialistas eliminativos ofrecen en defensa de la teoría son muy discutibles. Sin embargo, muchos filósofos de la mente son perturbados por la simple posibilidad del materialismo eliminativo. La razón es que esta posibilidad (por remota que sea) es de las implícitas en la "teoría teoría". Pues si la psicología del sentido común realmente es una teoría empírica —esto es, una teoría que pretende ser verdadera en el mundo ordinario de la experiencia—, entonces, como cualquier teoría empírica,

[33] Para mayor discusión de estos puntos contra el materialismo eliminativo, véase T. Horgan y James Woodward, "Folk Psychology is Here to Stay", en W. G. Lycan (ed.), *Mind and Cognition,* y Colin McGinn, *Mental Content* (Oxford, Blackwell, 1989), cap. 2.

sus proponentes deben aceptar la posibilidad de que pueda un buen día ser falsificada. No importa cuánto creamos en las teorías de la evolución o la relatividad; debemos aceptar (por lo menos) la posibilidad de que un día puedan revelarse como falsas.

Un modo de evitar esta desventurada situación es rechazar la "teoría teoría" por completo como descripción de nuestro entendimiento ordinario de otras mentes. Este enfoque daría una respuesta negativa a la primera pregunta planteada al final del último apartado, "¿Da la 'teoría teoría' una noción adecuada de la psicología del sentido común?" Echemos una breve ojeada a esta actitud.

Teoría y simulación

Así, hay muchos filósofos que opinan que la "teoría teoría" tergiversa rotundamente lo que hacemos cuando aplicamos conceptos psicológicos a la comprensión de las mentes entre unos y otros. Su alternativa más que comprender las mentes de otros implica una especie de proyección imaginativa en sus mentes. Esta proyección la llaman diversamente "replicación" o "simulación".

La esencia de la idea es fácil de captar. Cuando tratamos y representamos lo que alguien más está haciendo, a menudo nos ponemos "en sus zapatos", tratando de ver las cosas desde su perspectiva. Esto es, imaginativamente "simulamos" o "replicamos" los pensamientos que podrían explicar su conducta. Al reflexionar sobre las acciones de otro, según Jane Heal, "lo que aspiro a hacer es replicar o recrear su pensamiento. Me pongo en lo que pienso que es su estado inicial, imaginando el mundo desde su punto de vista y

luego delibero, razono y reflexiono para ver qué decisión surge".[34]

Un punto de vista parecido fue expresado hace más de 40 años por W. V. Quine:

> Las actitudes proposicionales [...] pueden ser pensamientos suyos que implican algo como la cita de una respuesta verbal imaginada a una situación imaginada. Poniendo nuestras verdaderas individualidades en papeles irreales, generalmente no sabemos cuánta realidad mantener constante. Surgen aprietos. Pero a pesar de ellos nos encontramos atribuyendo creencias, deseos y empeños incluso a seres carentes del poder del habla; tal es nuestro virtuosismo dramático. Nos proyectamos aun en lo que según este comportamiento nos imaginamos que puede haber sido el estado mental de un ratón, y lo dramatizamos como una creencia, deseo o empeño, verbalizado según parezca pertinente y natural a nosotros en el estado así fingido.[35]

Los investigadores recientes han empezado a tomar muy en serio la observación de Quine, y hay varias opciones que surgen acerca de cómo llenar los detalles. Empero, común a todas ellas es la idea de que imaginar lo que alguien piensa no es ver su comportamiento y aplicarle una teoría. Más bien es algo más parecido a una *habilidad* que tenemos: la

[34] Jane Heal, "Replication and Functionalism", en Jeremy Butterfield (ed.), *Language, Mind and Logic* (Cambridge, Cambridge University Press, 1986). Véase Robert Gordon, "Folk Psychology as Simulation", *Mind & Language,* 1 (1986); Alvin Goldman, "Interpretation Psychologised", *Mind & Language,* 4 (1989), y un número especial de *Mind & Language,* 7, núms. 1 y 2 (1992).

[35] Quine, *Word and Object* (Cambridge, MIT Press, 1960), p. 219.

habilidad de imaginarnos en las mentes de otros y predecir y explicar su comportamiento como resultado.

Es fácil ver cómo esta "teoría de la simulación" de la psicología del sentido común puede evitar el tema de la eliminación de la mente. El argumento eliminativo materialista en el último apartado comenzó con los supuestos de que la psicología del sentido común era una teoría, que las cosas de que habla están plenamente definidas por la teoría, y de que está compitiendo con la psicología científica. La argumentación decía entonces que la psicología del sentido común no es muy buena teoría, y concluía que no hay buenas razones para pensar que existen estados mentales. Pero si la psicología del sentido común no es una teoría *en absoluto,* no está ni siquiera en competencia con la ciencia, y la argumentación no se alza del suelo.

Aunque adoptar la teoría de la simulación sería una manera de negar una premisa —la "teoría teoría"— en uno de los argumentos para el materialismo eliminativo, ésta no es una muy buena razón para creer en la teoría de la simulación. Pues, vista de otra manera, la teoría de la simulación podría estar perfectamente de acuerdo con los materialistas eliminativos: podría sostenerse que, si la psicología del sentido común no se presenta siquiera como ciencia, o como "protociencia", entonces no necesitamos pensar al respecto para nada. Así, uno podría adoptar la teoría de la simulación sin creer que las mentes existen realmente. (El supuesto aquí, claro está, es que las únicas pretensiones que nos dicen lo que hay en el mundo son las pretensiones hechas por las teorías científicas.)

Esta combinación de la teoría de la simulación y del materialismo eliminativo es, de hecho, sostenida por Quine. Contrástese la observación citada antes con la siguiente:

> El punto es [...] si en un resumen final ideal de todo [...] es eficaz presentar nuestro esquema conceptual para marcar una serie de entidades o unidades de un género mental, como se dice, además de las físicas. Mi hipótesis, adelantada en el espíritu de una hipótesis de ciencia natural, es que no es eficaz.[36]

En vista de que el materialismo eliminativo y la teoría de la simulación son compatibles de esta manera, evitar el materialismo eliminativo sería una muy mala motivación por sí misma para creer en la teoría de la simulación. Y, por supuesto, los teóricos de la simulación tienen varias razones independientes para creer en su teoría. Una razón ya ha sido mencionada en este capítulo (en el apartado "Psicología del sentido común"): nadie ha sido capaz de mostrar muchas generalizaciones poderosas o interesantes del sentido común. Recuérdese la observación de Adam Morton de que la mayoría de las generalizaciones de la psicología popular son "tautologías sosas". Esto no pretende ser un argumento fatal, pero (según afirman los teóricos de la simulación) debiera animarnos a buscar otra cosa que la "teoría teoría".

Por lo tanto, ¿qué haríamos con la teoría de la simulación? Ciertamente, muchos de nosotros reconocerán que así es, a menudo, como las cosas nos parecen cuando nos entendemos entre unos y otros. "Ver las cosas desde el punto de vista de alguien más", puede incluso ser prácticamente sinónimo de comprenderlas, y la incapacidad de ver las cosas desde los puntos de vista de otros es claramente un

[36] Quine, "On Mental Entities", *The Ways of Paradox* (Cambridge, Harvard University Press, 1976), p. 227.

fracaso en la capacidad de uno como psicólogo del sentido común. Mas si la simulación es tal parte evidente de nuestras vidas despiertos, ¿por qué alguien habría de negar que ocurre?, y si nadie (incluso un "teórico de la teoría") negara que acontece, ¿cómo se supone que la teoría de la simulación está en *conflicto* con la "teoría teoría"?, ¿por qué no podría un "teórico de la teoría" responder diciendo: "Estoy de acuerdo: así es como nos *parece* a nosotros entender otras mentes; pero no podría simularse a menos que tuviera usted conocimiento de alguna teoría subyacente cuya verdad hiciera posible la simulación. Esta teoría subyacente no necesita ser aplicada conscientemente; pero como todos sabemos, esto no significa que no esté ahí".

La respuesta depende de lo que entendamos cuando decimos que la psicología del sentido común es una teoría que se "aplica" a pensadores. En el apartado acerca de la psicología del sentido común citado, señalé que la "teoría teoría" podía decir que las generalizaciones psicológicas del sentido común eran conocidas de manera inconsciente por pensadores (una idea a la que volveremos en el capítulo IV). Sin embargo, vistas las cosas, parece como si este punto de vista no fuera directamente amenazado por la teoría de la simulación. Como la simulación se refiere a aquello de lo que estamos explícitamente conscientes en actos de interpretación, el hecho de que simulemos otros no muestra que no tengamos conocimiento tácito de las generalizaciones psicológicas del sentido común. Los teóricos de la simulación, por lo tanto, necesitan proporcionar argumentos independientes contra este punto de vista.

Es importante no lanzarse a conclusiones apresuradas de alguna clase. Son días relativamente tempranos de la teoría de la simulación, y muchos de los detalles no han sido

elaborados todavía. Sin embargo, parece que la "teoría teoría" puede defenderse si se le permite recurrir a la idea del conocimiento tácito; y la "teoría teoría" puede, según parece, aceptar el principal discernimiento de la teoría de la simulación, que a menudo interpreta a otros pensando en cosas desde su punto de vista, etc. De esta manera, podría ser posible sostener los mejores elementos de ambas actitudes para entender otras mentes. Tal vez no haya aquí ninguna disputa real, sólo una diferencia de hincapié.

Conclusión: de la representación a la computación

Y bien, ¿cómo podemos saber acerca de la mente? He considerado y apoyado una respuesta: aplicando conjeturas acerca de las mentes ajenas —o aplicando una teoría de la mente— para explicar su comportamiento. Examinar la teoría nos ayuda entonces a responder la otra pregunta: *¿qué* sabemos acerca de la mente? Esta pregunta puede ser respondida hallando qué dice la teoría acerca de las mentes. Tal como interpreto la psicología del sentido común, dice (por lo menos) que los pensamientos son estados de la mente que representan el mundo y que tienen efectos sobre el mundo. Así es como pasamos de una respuesta a la pregunta "¿cómo?" a una respuesta a la pregunta "¿qué?".

Hay varias maneras como podría seguirse indagando. La idea de un estado que representa el mundo y hace que su poseedor se comporte de determinada manera no es una idea que sea aplicable sólo a los seres humanos. Como nuestro conocimiento de los pensamientos deriva del comportamiento —y no necesariamente el verbal— es posible aplicar

los elementos básicos de la psicología del sentido común a otros animales también.

¿Qué tan abajo de la escala evolutiva conduce esta clase de explicación? ¿A qué clases de animales podemos aplicar esta explicación? Considérese este notable pasaje de C. R. Gallistel:

> En el monótono desierto tunecino, una hormiga de patas largas y movimiento rápido deja la protección del nido húmedo para buscar comida. Se mueve en curvas tortuosas, corriendo primero hacia aquí, luego hacia allá, pero gradualmente avanzando más y más desde la humedad sostenedora de vida del nido. Finalmente encuentra el cadáver de un escorpión, usa sus fuertes tenazas para arrancar un trozo casi de su propio tamaño; entonces da la vuelta para orientarse, a uno o dos grados de la línea recta que la une a la entrada del nido, un agujero de un milímetro de anchura, a 40 metros de distancia. Corre por una línea recta 43 metros, orientándose según el ángulo del sol. Tres metros más allá del punto en el cual debió haber encontrado la entrada, la hormiga emprende abruptamente una pauta de búsqueda mediante la cual la encuentra al fin. Un testigo de esta jornada hacia casa encuentra difícil resistir la inferencia de que la hormiga en su búsqueda de comida poseía en todo momento una representación de su posición relativa a la entrada del nido, una representación espacial que le permitía computar el ángulo solar y la distancia en la jornada de vuelta desde donde hubiese encontrado comida.[37]

[37] C. R. Gallistel, *The Organisation of Learning* (Cambridge, MIT Press, 1990), p. 1.

Aquí el comportamiento de la hormiga es explicado en términos de representaciones de localizaciones en su medio. Hay algo más, sin embargo: Gallistel habla de la "computación" del ángulo solar y de la distancia en el camino de retorno. ¿Cómo podemos hallar sentido a las representaciones "computarizadas" de una hormiga? ¿Por qué es "difícil de resistir" esta conclusión? Puestas así las cosas, ¿qué significa computar representaciones, al fin y al cabo? Ocurre, por supuesto, que lo que Gallistel considera verdad acerca de la hormiga, mucha gente lo ve como verdad de nuestras mentes, que conforme nos movemos por el mundo y pensamos acerca de él, computamos representaciones. Éste es el tema del próximo capítulo.

LECTURAS ADICIONALES

The Philosophy of Mind (Boulder, Westview, 1996), de Jaegwon Kim, es una de las mejores introducciones generales a la filosofía de la mente; también es buena *Philosophy of Mind and Congnition* (Oxford, Blackwell, 1996), de David Braddon-Mitchell y Frank Jackson. *Matters of the Mind* (Edimburgo, Edinburgh University Press, 2001), de William Lyons, es legible y accesible, con un enfoque novedoso en algunos puntos. El behaviorismo está representado adecuadamente por la parte 1 de W. G. Lycan (ed.), *Mind and Cognition* (Oxford, Blackwell, 1990; segunda edición, 1998); toda la antología contiene también lecturas esenciales acerca del materialismo eliminativo y de la psicología del sentido común o "popular". Para la idea de que los estados mentales son causas de comportamiento, véanse los ensayos de Donald Davidson reunidos en sus *Essays*

on Actions and Events (Oxford, Oxford University Press, 1980); Davidson también combina esta idea con una denegación de las leyes psicológicas (en "Sucesos mentales" y "La mente material"). Para la teoría causal de la mente, D. M. Armstrong, *A Materialist Theory of the Mind* (Londres, Routledge, 1968; reimpreso en 1993) es un clásico que bien vale la pena leer. Daniel C. Dennett ha desarrollado una posición particular acerca de las relaciones entre ciencia y psicología popular y entre la representación y la causalidad: véanse los ensayos en *The Intentional Stance* (Cambridge, MIT Press, 1987), especialmente "Auténticos creyentes" y "Tres clases de psicología intencional". Una versión interesante de la alternativa "simulación" a la "teoría teoría" es de Jane Heal, "Replicación y funcionalismo", en J. Butterfield (ed.), *Language, Mind and Logic* (Cambridge, Cambridge University Press, 1986). El debate simulación/"teoría teoría" está bien representado en los dos volúmenes de los que Martin Davies y Tony Stone son editores: *Folk Psychology: The Theory of Mind Debate* y *Mental Simulation: Evaluations and Applications* (ambos de Oxford, Blackwell, 1995).

III. COMPUTADORAS Y PENSAMIENTO

HASTA AQUÍ he tratado de explicar los problemas filosóficos de la naturaleza de la representación, y cómo se enlaza con nuestro entendimiento de otras mentes. Lo que la gente dice y hace es causado por lo que piensa —lo que cree, espera, quiere, desea y así por el estilo—, esto es, por sus estados mentales de representación o *pensamientos*. Lo que la gente hace es causado por las maneras como se representa el mundo. Si hemos de explicar el pensamiento, entonces tenemos que explicar cómo puede haber estados que puedan, al mismo tiempo, ser representaciones del mundo y causas de comportamiento.

Para comprender cómo cualquier cosa puede tener estos dos rasgos es útil introducir la idea de la mente como computadora. Muchos psicólogos y filósofos piensan que la mente es una especie de computadora. Hay muchas razones para pensar así, pero el nexo con nuestro tema del momento es éste: una computadora es un mecanismo causal que contiene representaciones. En este capítulo y en el siguiente explicaré esta idea y mostraré su importancia para los problemas que rodean el pensamiento y la representación.

La idea misma de que la mente es una computadora, o de que las computadoras pudieran pensar, inspira intensos sentimientos. Algunas personas encuentran esto excitante, otros lo encuentran absurdo, o incluso degradante para la naturaleza humana. Trataré y enfocaré este punto discutido

de una manera tan limpia como se pueda, valorando algunos de los argumentos principales en pro y en contra de las pretensiones de que las computadoras pueden pensar, y de que la mente es una computadora. No obstante, primero necesitamos entender estas pretensiones.

Preguntar lo adecuado

Es decisivo comenzar preguntando lo correcto. Por ejemplo, a veces se plantea la pregunta: ¿puede la mente humana ser modelada en una computadora? Pero incluso si la respuesta es afirmativa, ¿cómo podría eso mostrar que la mente es una computadora? El Tesoro de Inglaterra produce modelos de computadora de la economía, pero nadie piensa que esto muestre que la economía es una computadora. Este capítulo explicará cómo puede surgir esta confusión. Uno de los propósitos principales de este capítulo es distinguir entre dos cosas:

1. ¿Puede pensar una computadora?, o, más precisamente, ¿puede algo pensar sencillamente siendo una computadora?
2. ¿Es la mente humana una computadora?, o, más precisamente, ¿son algunos estados y procesos mentales reales de índole computacional?

Este capítulo se ocupará principalmente de la pregunta 1, y el capítulo IV de la pregunta 2. La distinción entre las dos puede no ser todavía clara, pero al final del capítulo sí debe serlo. Para entender estas dos cosas necesitamos saber por lo menos otras dos: primero, qué es una compu-

tadora, y, en segundo lugar, qué pasa con la mente que hace a las personas pensar que una computadora pudiera tener mente, o que la mente humana pudiera ser una computadora.

¿Qué es una computadora? Todos estamos familiarizados con las computadoras, muchos de nosotros las usamos cada día. Para muchos son un misterio y explicar cómo funcionan pudiera parecer una tarea muy difícil. Sin embargo, aunque los detalles de las computadoras modernas son pasmosamente complicados, los conceptos fundamentales que hay tras ellas son en realidad hermosamente sencillos. La dificultad de comprender las computadoras no es tanto captar los conceptos que implican, sino ver *cómo* estos conceptos son tan útiles.

Si usted está familiarizado con los conceptos fundamentales de las computadoras puede ser que desee saltarse los cinco apartados siguientes y pasar directamente al apartado de este capítulo que se llama "¿Computadoras pensantes?", en la p. 180. Si no se está familiarizado con estos conceptos entonces algo de la terminología que sigue acaso sea un poco intimidante. Puede usted querer leer los siguientes apartados bastante aprisa, y aquello de que tratan se volverá más claro después de haber leído el resto de este capítulo y del capítulo IV.

Para prepararse y comprender las computadoras, lo mejor es abandonar la mayoría de los supuestos que puede tener usted acerca de ellas. Las computadoras personales que usamos en nuestra vida cotidiana tienen normalmente un teclado como una máquina de escribir y una pantalla. Suelen estar hechas de una combinación de metal y plástico, y la mayoría de nosotros sabrá que tienen adentro cosas que se llaman "*chips* de silicio", que hace que funcionen de

alguna manera. Pónganse todas estas ideas a un lado por el momento: ninguno de estos rasgos de las computadoras es esencial para ellas. No es ni siquiera esencial para las computadoras ser electrónicas.

¿Qué es, pues, esencial para una computadora? La definición tosca a la que llegaré al final es ésta: *Una computadora es un dispositivo que procesa representaciones de una manera sistemática.* Esto es un poco vago hasta que entendemos más precisamente "procesos", "representaciones" y "sistemático". A fin de comprender estas nociones, hay dos ideas más que necesitamos comprender. La primera es la cuestión matemática bastante abstracta de una *computación.* La segunda es cómo las computaciones pueden ser *automatizadas.* Tomaré estas ideas sucesivamente.

Computación, funciones y algoritmos

La primera idea que necesitamos es la de una *función* matemática. Todos estamos familiarizados con esta idea, por la aritmética elemental. Algo de las primeras cosas que aprendemos en la escuela son las funciones aritméticas básicas: suma, resta, multiplicación y división. Entonces normalmente nos enteramos de otras funciones, tales como la función cuadrado (mediante la cual producimos el cuadrado de un número, x^2, multiplicando el número x por sí mismo), los logaritmos y así sucesivamente.

Según las aprendemos en la escuela, las funciones aritméticas no son números, sino cosas que se "hacen" a los números. Lo que aprendemos a hacer en la aritmética básica es tomar algunos números y aplicarles algunas funciones. Tómese la suma de dos números, 7 y 5. En efecto, toma-

mos estos dos números como la "entrada" a la función suma y obtenemos otro número, 12, a modo de "salida". Esta suma la representamos escribiendo: 7 + 5 = 12. Por supuesto, podemos poner cualesquiera otros dos números en los lugares ocupados por 7 y 5 (los lugares de entrada) y la función suma determinará un número único como salida. Hace falta adiestramiento para representar lo que será la salida para cualquier número que sea, pero el punto es que, de acuerdo con la función suma, hay exactamente un número que es la salida de la función para cualquier grupo dado de números de entrada.

Si tomamos el cálculo 7 + 5 = 12, y le quitamos los numerales 7, 5 y 12, obtenemos un símbolo complejo con tres "agujeros": _ + _ = _. En los primeros dos agujeros escribimos las entradas a la función suma, y en el tercer agujero escribimos la salida. La función misma podría entonces ser representada como _ + _, con los dos blancos indicando dónde deben insertarse los números de entrada. Estos blancos se indican por costumbre con letras cursivas *x, y, z,* y así sucesivamente; así, la función se escribiría por lo tanto $x + y$. Estas letras, llamadas "variables", son una manera útil de marcar los diferentes agujeros o *lugares* de la función.

Ahora, un poco de terminología. Las entradas de la función se llaman *argumentos* de la función, y la salida se llama *valor* de la función. Los argumentos en la ecuación $x + y = z$ son pares de números x y y tales que z es su valor. Esto es, el valor de la función suma es la suma de los argumentos de esa función. El valor de la función resta es el resultado de restar un número de otro (los argumentos). Y así sucesivamente.

Aunque la teoría matemática de las funciones es muy

complicada en sus detalles, la idea básica de una función puede ser explicada usando ejemplos sencillos como la suma. Y, aunque la presenté con un ejemplo matemático, la noción de una función es extremadamente general y puede extenderse a cosas que no son números. Por ejemplo, en vista de que cada quien tiene sólo un padre natural, podemos pensar en la expresión "el padre natural de *x*" como si describiera una función, que toma a las personas como argumentos y ofrece sus padres como valores. (Los familiarizados con la lógica elemental conocerán también que expresiones como "y" y "o" se conocen como funciones-*de-verdad*, por ejemplo: la proposición compleja P&Q implica una función que da el valor Verdad cuando ambos argumentos suyos son verdad, y el valor Falso de otra manera.)

La idea de una función, pues, es sumamente general y nos apoyamos en ella implícitamente en nuestra vida cotidiana (cada vez que sumamos los precios de algo en el supermercado, por ejemplo). No obstante, una cosa es decir lo que es una función, en abstracto, y otra decir cómo la usamos. Para saber cómo emplear una función necesitamos un método a fin de obtener valor para un argumento o argumentos dados. Recuérdese lo que ocurre cuando se aprende aritmética elemental. Supóngase que se desea calcular el producto de dos números, 127 y 21. La manera normal de calcular esto es el método de la multiplicación larga:

$$\begin{array}{r} 127 \\ \times\ 21 \\ \hline 127 \\ +2540 \\ \hline 2667 \end{array}$$

Lo que se hace cuando se realiza la multiplicación larga es tan evidente que sería trivial analizarlo. De hecho, lo que se sabe cuando se sabe cómo hacer esto es algo increíblemente poderoso. Lo que se tiene es un método para calcular el producto de dos números *cualesquiera,* esto es, de calcular el valor de la función de multiplicación para cualesquiera dos argumentos. Este método es enteramente general: no se aplica a algunos números y a otros no. Y carece enteramente de ambigüedad: si se conoce el método, se conoce en cada etapa lo que se va a hacer entonces para producir la respuesta.

(Compárese un método como éste con los métodos que usamos para tratar con gente que acabamos de conocer. Tenemos algunas reglas toscas y dispuestas, que aplicamos: tal vez nos presentamos, sonreímos, damos la mano, preguntamos acerca de ellos, etc. Sin embargo, evidentemente estos métodos no dan respuestas "definidas"; en ocasiones nuestras sutilezas sociales disparan por la culata.)

Un método —tal como la multiplicación larga— para calcular el valor de una función se conoce como un *algoritmo.* Los algoritmos se llaman también "procedimientos eficaces", ya que son procedimientos que, si se aplican correctamente, resultan enteramente eficaces en dar sus resultados (a diferencia de los procedimientos que usamos al encontrarnos con personas). También se llaman "procedimientos mecánicos", pero yo preferiría no usar esta expresión, ya que en este libro estoy usando el término "mecánico" con un sentido menos preciso.

Es muy importante distinguir entre algoritmos y funciones. Un algoritmo es un *método* para encontrar el *valor* de una función. Una función puede tener más de un algoritmo para encontrar sus valores para cualesquiera argu-

mentos dados. Por ejemplo, multiplicamos 127 por 21 usando el método de la multiplicación larga. Sin embargo podíamos haber multiplicado sumando 127 a sí mismo 20 veces. Esto es, podríamos haber usado un algoritmo diferente.

Decir que hay un algoritmo para determinada función aritmética no es decir que una aplicación del algoritmo dé siempre un *número* como respuesta. Por ejemplo, puede quererse ver si determinado número se divide *exactamente* por otro número sin residuo. Cuando se aplica el algoritmo para la división, puede usted encontrar que no es así. De modo que la cuestión no es que el algoritmo dé un número como respuesta, sino que siempre dé un procedimiento para encontrar si hay una respuesta.

Cuando hay un algoritmo que da el valor de una función para cualquier argumento, entonces los matemáticos dicen que la función es *computable*. La teoría matemática de la computación es, en sus términos más generales, la teoría de las funciones computables, es decir, funciones para las cuales hay algoritmos.

Al igual que la noción de una función, la noción de un algoritmo es extremadamente general. Todo procedimiento eficaz para encontrar la solución de un problema puede llamarse algoritmo, mientras satisfaga las condiciones siguientes:

1. En cada etapa del procedimiento hay una cosa definida que hacer a continuación. Moverse paso a paso no requiere ninguna conjetura, discernimiento o inspiración especiales.

2. El procedimiento puede ser especificado en un número finito de pasos.

Así podemos pensar en un algoritmo como una regla, o un manojo de reglas, que den la solución de un problema dado. Estas reglas pueden entonces ser representadas como un "diagrama de flujo". Considérese, por ejemplo, un algoritmo muy sencillo para multiplicar dos números enteros, x y y, que funcione agregando y a sí mismo. Ayudará si se imagina el procedimiento realizado en tres trozos de papel, uno para el primer número (llámese a este papel X), otro para el segundo número (llámese a este trozo de papel Y) y uno para la respuesta (llámese a este trozo de papel la RESPUESTA). La figura III.1 muestra el diagrama de flujo; representa el cálculo mediante la siguiente serie de etapas:

Etapa *(i)*: Escríbase "0" en la RESPUESTA, y pásese al paso *(ii)*.

Etapa *(ii)*: ¿Es el número escrito en $X = 0$?
Si es sí, entonces pásese a la etapa *(v)*
Si es NO, entonces pásese a la etapa *(iii)*

Etapa *(iii)*: Réstese 1 del número escrito en X, escríbase el resultado en X, y pásese a la etapa *(iv)*

Etapa *(iv)*: Súmese el número escrito en Y a la RESPUESTA, y pásese a la etapa *(ii)*

Etapa *(v)*: ALTO

Apliquemos esto a un cálculo determinado, digamos 4 veces 5. (Si está usted familiarizado con esta clase de procedimiento, puede usted saltarse el ejemplo y pasar al siguiente párrafo.)

Empiécese escribiendo los números por multiplicar, 4 y 5, en los pedazos X y Y de papel respectivamente. Aplíquese la etapa *(i)* y escríbase 0 en la RESPUESTA. Entonces aplíquese la etapa *(ii)* y pregúntese si el número escrito en X es 0.

No lo es, es 4. De manera que pásese a la etapa *(iii)*, y réstese el número 1 del número escrito en X. Esto lo deja a uno con 3, de modo que debe escribirse esto en X y pasar a la etapa *(iv)*. Súmese el número escrito en Y (es decir, 5) a la RESPUESTA, que hace que la RESPUESTA diga 5. Muévase a la etapa *(ii)*, y pregúntese otra vez si el número en X es 0. No lo es, es 3. Así pásese a la etapa *(iii)*, réstese 1 del número escrito en X, escríbase 2 en X y pásese a la etapa *(iv)*. Súmese el número escrito en Y a la RESPUESTA, lo cual hace que la RESPUESTA se lea 10. Pregúntese de nuevo si el número escrito en X es 0. No lo es, es 2. Así pásese a la etapa *(iii)*, réstese 1 del número escrito en X, escríbase 1 en X y pásese a la etapa *(iv)*. Súmese el número escrito en Y a la RESPUESTA, lo cual hace que ésta se lea 15. Pregúntese de nuevo si el número

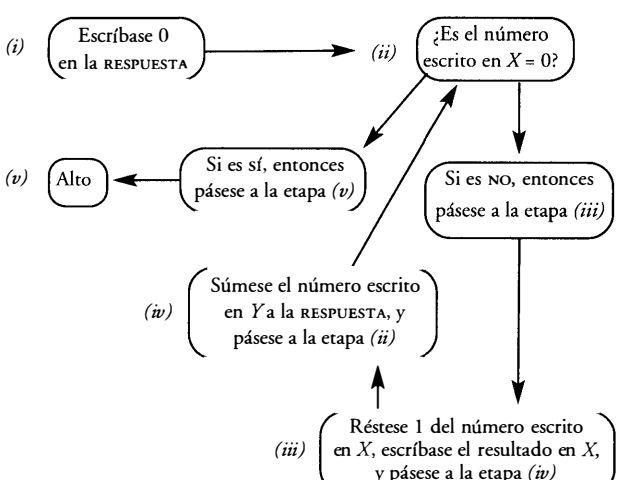


FIGURA III.1. *Diagrama de flujo para el algoritmo de multiplicación*

escrito en X es 0; no lo es, es 1. Así que pásese a la etapa *(iii),* réstese 1 del número escrito en X, escríbase 0 en X y pásese a la etapa *(iv).* Súmese el número escrito en Y a la RESPUESTA, que hace que ésta se lea 20. Pásese a la etapa *(ii)* y pregúntese si el número escrito en X es 0. Esta vez sí lo es, así que pásese a la etapa *(v),* y deténgase el procedimiento. El número escrito en la RESPUESTA es 20, que es el resultado de multiplicar 4 por 5.[1]

Éste es un modo harto laborioso de multiplicar 4 por 5. Sin embargo, el motivo de la ilustración no es que sea éste un *buen* procedimiento para que lo usemos. La cuestión es más bien que es un procedimiento enteramente *eficaz* en cada etapa, es completamente claro qué se hace a continuación, y el procedimiento termina en un número finito de etapas. El número de etapas podrá ser muy grande; pero dado cualquier par de números finitos, seguirá habiendo un número finito de etapas.

Las etapas *(iii)* y *(iv)* del ejemplo ilustran un rasgo importante de los algoritmos. Aplicando este algoritmo a la multiplicación, empleamos otras operaciones aritméticas: resta en la etapa *(iii),* suma en la etapa *(iv).* No tiene nada de malo hacer esto, mientras haya algoritmos para las operaciones de resta y suma también, como es el caso, por supuesto. De hecho, la mayoría de los algoritmos usarán otros algoritmos en alguna etapa. Piénsese en la multiplicación larga: usa la suma para reunir los resultados de las multiplicaciones "cortas". Por lo tanto, usará usted algún algoritmo

[1] El ejemplo es de Ned Block, "The Computer Model of the Mind", en Daniel N. Osherson *et al.* (eds.), *An Invitation to Cognitive Science,* vol. 3, *Thinking* (Cambridge, MIT Press, 1990). Éste es un excelente artículo de introducción que cubre terreno no cubierto en este capítulo, por ejemplo la prueba de Turing (véase más adelante).

para la suma cuando esté haciendo una multiplicación larga. De manera que nuestro laborioso algoritmo de multiplicación puede ser dividido en etapas que dependen únicamente de otros (tal vez más sencillos) algoritmos y simples "movimientos" de paso en paso. Esta idea es muy importante para comprender las computadoras, como veremos.

El hecho de que los algoritmos puedan ser representados por diagramas de flujo indica la generalidad del concepto de algoritmo. Tal como podemos escribir diagramas de flujo para toda suerte de procedimientos, así podemos escribir algoritmos para toda clase de cosas. Algunas recetas, por ejemplo, pueden ser representadas como diagramas de flujo. Considérese este algoritmo para cocer un huevo.

1. Enciéndase la estufa.
2. Llénese el recipiente con agua.
3. Póngase el recipiente sobre la estufa.
4. Cuando hierva el agua, añádase un huevo, y póngase el despertador.
5. Cuando suene el despertador, apáguese el gas.
6. Sáquese el huevo del agua.
7. Resultado: un huevo cocido.

Éste es un proceso que puede ser completado en un número finito de etapas, y a cada etapa la sigue una cosa definida, no ambigua, que hacer a continuación. No hace falta inspiración o conjetura. Así, en un sentido, cocer un huevo puede describirse como un procedimiento algorítmico (véase la figura III.2).

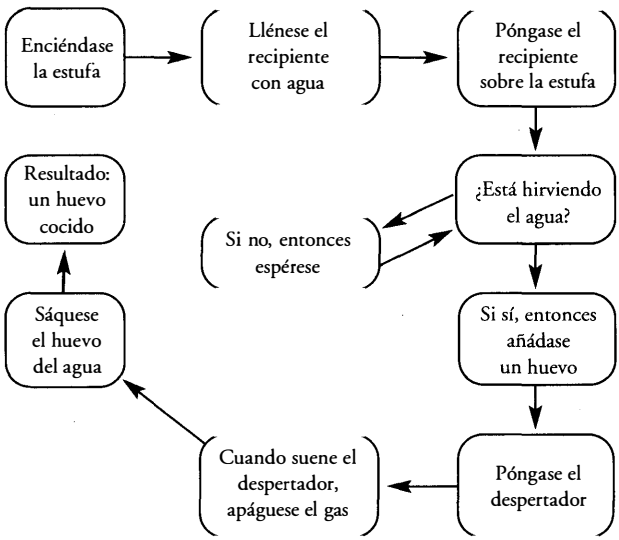


FIGURA III.2. *Un diagrama de flujo para cocer un huevo*

MÁQUINAS DE TURING

El uso de algoritmos para computar los valores de las funciones es al menos tan antiguo como las viejas matemáticas griegas. Pero no fue sino recientemente (de hecho, en los años treinta del siglo XX) cuando la idea fue objeto de escrutinio y los matemáticos trataron de dar un sentido preciso al concepto de algoritmo. Desde fines del siglo XIX se había dado un intenso interés en los *fundamentos* de las matemáticas. ¿Qué hace verdaderos a los enunciados matemáticos? ¿Cómo pueden las matemáticas ser puestas sobre un fundamento firme? Una cuestión que se volvió particularmente

urgente era: ¿qué *determina* si el método dado de cálculo es adecuado para la tarea en cuestión? Conocemos en particular casos en que un algoritmo es adecuado, pero ¿hay un método general que nos diga, para cualquier método propuesto de cálculo, si es un algoritmo o no?

La cuestión tiene honda importancia teórica para los matemáticos, porque los algoritmos residen en el meollo de la práctica matemática, pero si podemos decir qué son, no podemos realmente decir qué son las matemáticas. Una respuesta a la cuestión fue dada por el brillante matemático inglés Alan Turing en 1937. Aparte de ser un matemático de genio, Turing (1912-1954) fue, podría sostenerse, una de las personas más influyentes del siglo xx, de manera indirecta. Como veremos, desarrolló los conceptos fundamentales de los cuales derivaron las computadoras digitales modernas y todas sus consecuencias. También fue famoso por descifrar el código Enigma de los nazis durante la segunda Guerra Mundial. Esta clave fue utilizada para comunicarse con submarinos, que por entonces diezmaban la marina británica, y puede decirse que descifrar la clave fue uno de los principales factores que evitó a Gran Bretaña la derrota en aquel momento de la guerra.[2]

Turing respondió la cuestión acerca de la naturaleza de la computación de un modo vívido y original. En efecto, preguntó: ¿cuál es el dispositivo más sencillo que podía realizar cualquier computación sin importar cuán complicada? Procedió entonces a describir semejante dispositivo, que ahora es llamado (muy naturalmente) "máquina de Turing".

Una máquina de Turing no es una máquina en el sentido

[2] Para una descripción de la vida de Turing, véase la biografía de Alan Hodges, *Alan Turing: The Enigma* (Nueva York, Simon & Schuster, 1983).

ordinario de la palabra. O sea que no es una máquina física, sino más bien una especificación abstracta, teórica, de una máquina posible. Aun cuando se han construido máquinas con estas especificaciones, la cuestión con ellas no es (en primer lugar) ser construidas, sino ilustrar algunas propiedades muy generales de los algoritmos y las computaciones.

Puede haber muchas clases de máquina de Turing para diferentes tipos de computación. No obstante, todas tienen los siguientes rasgos en común: una cinta dividida en cuadrados y un dispositivo que puede escribir símbolos en la cinta y luego leer tales símbolos.[3] El dispositivo está también en ciertos "estados internos" (más al respecto, adelante), y puede mover la cinta a la derecha o a la izquierda, un cuadrado a la vez. Supongamos, para simplificar, que hay sólo dos clases de símbolo que pueden ser escritos en la cinta: "1" y "0". Cada símbolo ocupa precisamente un cuadrado de cinta —de manera que la máquina sólo puede leer un cuadrado a la vez—. (No tenemos que cuidarnos todavía de lo que estos símbolos "significan": considérense nada más como *marcas* en la cinta.)

De manera que el dispositivo puede hacer sólo cuatro cosas:

1. Puede mover la cinta un cuadrado a la vez, de izquierda a derecha o de derecha a izquierda.
2. Puede leer un símbolo en la cinta.
3. Puede escribir un símbolo en la cinta, ya sea escribiendo en un cuadrado en blanco o encima de otro símbolo.
4. Puede cambiar su "estado interno".

[3] De hecho, la cinta de la máquina necesita ser infinitamente larga. Para una explicación, véase, por ejemplo, Penrose, *The Emperor's New Mind,* cap. 2.

Las operaciones posibles de una máquina particular pueden representarse por su "tabla de la máquina". La tabla de la máquina es, en efecto, una serie de instrucciones de la forma "si la máquina está en el estado X y leyendo el símbolo S, entonces realizará determinada operación (por ejemplo, escribir o borrar un símbolo, moviendo la cinta) y pasará al estado Y (o seguirá en el mismo estado) y moverá la cinta a la derecha/izquierda". Si uno quiere, se puede pensar en la tabla de la máquina como "programa" de la máquina: dice a la máquina qué hacer. Al especificar una posición particular en la tabla de la máquina, necesitamos saber dos cosas: la *entrada* del momento a la máquina y su *estado* del momento. Lo que la máquina hace es *enteramente fijado* por estas dos cosas.

Todo esto parecerá muy abstracto, de modo que consideremos un ejemplo específico de una máquina de Turing: una que realiza una operación matemática sencilla, la de sumar 1 a un número.[4] A fin de obtener una máquina que realice una operación particular, necesitamos interpretar los símbolos en la cinta, esto es, hacer que representen algo. Supongamos que todos nuestros unos en la cinta representen números: 1 representa el número 1, de manera harto obvia. Pero necesitamos maneras de representar números diferentes de 1, así que usemos un método sencillo: más o menos como un prisionero puede representar sus días de prisión mediante hileras de arañazos en la pared, una línea o "cadena" de *n* representará el número *n*. Así, 111 representa 3, 11111 representa 5, y así sucesivamente.

[4] Véase Penrose, *The Emperor's New Mind*, p. 54. Véanse también los caps. 2 y 3 de Joseph Weizenbaum, *Computer Power and Human Reason* (Harmondsworth, Penguin, 1976).

Para permitir que dos o más números sean escritos en una cinta, podemos separar los números usando uno o más ceros. Los ceros sencillamente funcionan para marcar espacios entre los números: son la única "puntuación" en esta sencilla notación. Así por ejemplo, la cinta,

... 000011100111111000100 ...

representa la sucesión de números 3, 6, 1. En esta notación, el número de ceros no viene al caso para el número escrito. Las marcas ... (puntos suspensivos) indican que la cinta en blanco continúa indefinidamente en ambas direcciones.

Necesitamos también una especificación de los "estados internos" de la máquina; resulta que la simple máquina de que nos estamos ocupando sólo necesita dos estados internos, que podemos igualmente llamar estado A (el estado inicial) y estado B. La máquina particular de Turing que estamos considerando tiene un comportamiento especificado por las siguientes instrucciones:

1. Si la máquina está en el estado A, y lee un 0, entonces permanece en el estado A, escribe un 0 y se mueve un cuadrado a la derecha.
2. Si la máquina está en el estado A, y lee un 1, entonces cambia al estado B, escribe un 1 y se mueve un cuadrado hacia la derecha.
3. Si la máquina está en el estado B, y lee un 0, entonces cambia al estado A, escribe un 1 y se detiene.
4. Si la máquina está en el estado B, y lee un 1, entonces permanece en el estado B, escribe un 1 y se mueve un cuadrado a la derecha.

		ENTRADA 1	ENTRADA 0
ESTADO DE LA MÁQUINA	A	Cambiar a B; Escribir un 1; Mover la cinta a la derecha	Estar en A; Escribir un 0; Mover la cinta a la derecha
	B	Estar en B; Escribir un 1; Mover la cinta a la derecha	Cambiar a A; Escribir un 1; ALTO

FIGURA III.3. *Una tabla de la máquina para una máquina sencilla de Turing*

La tabla de la máquina para esta máquina se verá como la figura III.3.

Ahora imaginemos presentar a la máquina parte de una cinta que puede verse así:

0	0	0	1	1	0	0	0

La cinta representa el número 2. (Recuérdese, los ceros sirven nada más como "puntuación", no representan ningún número en esta notación.) Lo que queremos que haga la máquina es sumar 1 a este número, aplicando las reglas de la tabla de la máquina.

Lo hace así. Supóngase que comienza en el estado inicial, el estado A, leyendo el cuadrado de cinta a la extrema derecha. Entonces sigue las instrucciones de la tabla. La cinta se "verá" como esto durante este proceso (el cuadrado de la cinta que está siendo ahora leído por la máquina está subrayado):

(i) 0 0 0 1 1 0 0 0 . . .
(ii) . . 0 0 1 1 0 0 0 . . .
(iii). . 0 0 0 1 1 0 0 0 . . .
(iv) . . . 0 0 0 1 1 0 0 0 . . .
(v) . . . . 0 0 0 1 1 0 0 0 . . .
(vi) . . . . . 0 0 0 1 1 0 0 0 . . .
(vii). . . . . 0 0 1 1 1 0 0 0 . . .

En la línea *(vi)* la máquina está en el estado B, lee un 0, de manera que escribe un 1, pasa al estado A y se detiene. La "salida' está en la línea *(vii):* esto representa el número 3, de manera que la máquina ha logrado su tarea de sumar 1 a su entrada.

Puede preguntarse: ¿ha hecho la máquina realmente algo?, ¿cuál es el *objeto* de todo este tejemaneje tedioso en torno a una cinta imaginaria? Como nuestro ejemplo de un algoritmo para la multiplicación, antes, parece un modo laborioso de hacer algo enteramente trivial. Pero, como con nuestro algoritmo, la cosa no es trivial. Lo que la máquina ha hecho es *computar una función.* Ha computado la función $x + 1$ para el argumento 2. Ha computado esta función usando únicamente las "acciones" más sencillas, las "acciones" representadas por los cuatro cuadrados de la tabla de la máquina. Éstas son sólo combinaciones de los pasos muy sencillos que fueron parte de la definición de todo lo que una máquina de Turing puede hacer (leer, escribir, cambiar de estado, mover la cinta). Explicaré la lección de esto dentro de un momento.

Puede uno preguntarse acerca del papel de los "estados internos" en todo esto. ¿No se está metiendo de contrabando algo en la descripción de este dispositivo tan sencillo, tomando de sus "estados internos"? ¿Tal vez *están* haciendo

el cálculo? Creo que esta preocupación es muy natural; pero está fuera de lugar. Los estados internos de la máquina no son más o menos lo que la tabla de la máquina dice que son. El estado interno, B, es, por definición, el estado tal que si la máquina obtiene un 1 como entrada, la máquina hace esto y lo otro; y tal que, si recibe un 0 como entrada, la máquina hace esto y lo otro. Esto es todo lo que hay en estos estados.[5] ("Interno" puede por lo tanto ser equívoco, ya que sugiere que los estados tienen una "naturaleza oculta".)

Para designar una máquina de Turing que realice operaciones más complejas (tales como nuestro algoritmo de multiplicación de la sección anterior) necesitamos una tabla de la máquina más compleja, más estados internos, más cinta y una notación más complicada. *Sin embargo, no necesitamos operaciones básicas más rebuscadas.* No hay necesidad para nosotros de entrar en los detalles de máquinas de Turing, ya que los puntos básicos pueden ser ilustrados por nuestra sencilla culebra. Sin embargo, es importante detenerse en la cuestión de la notación.

Nuestra notación de marcas para números de nuestro prisionero tiene diversos inconvenientes que son obvios. Uno es que no puede representar el 0, un grave reparo. Otro es que los números muy grandes tomarán siglos para computarlos, ya que la máquina sólo puede leer un cuadrado a la vez. (Sumar 1 al número 7 000 000 requeriría una cinta con más cuadrados que habitantes hay en Londres.) Un sistema más eficiente es el sistema binario o base binaria, en que todos los números naturales son representados por combinaciones de unos y ceros. Recuérdese que, en la notación binaria, la columna ocupada por múltiplos de 10

[5] Véase Weizenbaum, *Computer Power and Human Reason,* pp. 51-53.

en el sistema de dos ingredientes que es común (base 10) está ocupada por múltiplos de 2. Esto nos da la siguiente traducción de denario a binario:

1 = 1
2 = 10
3 = 11
4 = 100
5 = 101
6 = 110
7 = 111
8 = 1000

Y así sucesivamente. Es evidente que los números de codificación en binario nos dan la posibilidad de representar números mucho mayores con más eficacia de lo que logran las marcas de nuestro prisionero.

Una ventaja de usar la notación binaria es que podemos diseñar máquinas de Turing de gran complejidad sin tener que añadir más símbolos al repertorio básico. Comenzamos con dos clases de símbolos, 0 y 1. En nuestra notación de marcas del prisionero, los ceros simplemente sirvieron para dividir los números unos de otros. Con la base 2, los ceros sirven como numerales, permitiéndonos escribir cualquier número como una cadena de unos y ceros. Empero, nótese que la máquina sigue necesitando sólo el mismo número de operaciones básicas: léase 1, escríbase un 1, léase 0, escríbase un 0, muévase la cinta. Así, usar la base 2 nos da la posibilidad de representar muchos más números mucho más eficientemente, sin tener que añadir más operaciones básicas a la máquina. (Es claro que necesitamos puntuación también para mostrar dónde cesa una instrucción o frag-

mento de entrada, y comienza otra. No obstante, con suficiente ingenio podemos codificar esto mediante unos y ceros también.)

Ahora estamos al borde de un descubrimiento muy emocionante. Con una notación adecuada, como la binaria, no sólo la *entrada* a una máquina de Turing (la cinta inicial), sino la *tabla de la máquina misma* puede codificarse como números en la notación. Para hacer esto necesitamos una manera de señalar las distintas operaciones de la máquina (leer, escribir, etc.), y los "estados internos" de la máquina, mediante números. Usamos las marcas "A" y "B" para los estados internos de nuestra máquina. Esto era puramente arbitrario: podíamos haber usado cualesquiera símbolos para estos estados: %, @, *, o lo que sea. Así, pudimos haber usado también números para representar estos estados. Y si usamos la base 2 podemos codificar estos estados internos y "acciones" como unos y ceros en una cinta de la máquina de Turing.

En vista de que la máquina de Turing está completamente definida por su tabla de la máquina, y cualquier tabla de la máquina de Turing puede ser numéricamente codificada, se sigue evidentemente que cualquier máquina de Turing puede ser numéricamente codificada. Así la máquina puede ser codificada en binario y escrita en la cinta de otra máquina de Turing. Así la otra máquina de Turing puede tomar la cinta de la primera máquina de Turing como entrada: puede *leer* la primera máquina de Turing. Todo lo que necesita es un método para convertir las operaciones descritas en la cinta de la primera máquina de Turing —el programa— en sus propias operaciones. Esto sólo será otra tabla de máquina, que puede ella misma ser codificada. Por ejemplo, supóngase que codificamos nuestra máquina "sú-

mese 1" al binario. Entonces podría ser representado en una cinta como una cadena de unos y ceros. Si añadimos algunos unos y ceros que representen un número (digamos 127) a la cinta, entonces éstos, más la codificación de nuestro máquina "sumar 1", puede ser la entrada de otra máquina de Turing. Esta máquina a su vez tendría un programa que interpreta nuestra máquina de "sumar 1". Puede entonces hacer exactamente lo que hace nuestra máquina de "sumar 1": puede sumar 1 al número alimentado, 127. Haría esto "simulando" el comportamiento de nuestra máquina original de "sumar 1".

Ahora, el descubrimiento emocionante es éste: hay una máquina de Turing que puede imitar el comportamiento de cualquier otra máquina de Turing. Como cualquier máquina de Turing puede ser codificada numéricamente, puede ser alimentada como entrada de otra máquina de Turing, mientras esa máquina tenga manera de leer su cinta. Turing demostró a partir de esto que, para realizar todas las operaciones que pueden realizar las máquinas de Turing, no necesitamos una máquina separada para cada operación. Necesitamos solamente *una* máquina que sea capaz de imitar cualquier otra máquina. Ésta se llama *máquina de Turing universal*. Y es la idea de una máquina de Turing universal la que reside detrás de las computadoras digitales modernas de propósito general. De hecho, no es una exageración decir que la idea de una máquina de Turing universal ha afectado probablemente el carácter de toda nuestra vida.

Sin embargo, decir que una máquina de Turing universal puede hacer cualquier cosa que cualquier máquina de Turing particular puede hacer, suscita la pregunta: ¿qué *pueden* hacer las máquinas de Turing?, ¿qué clase de opera-

ciones pueden realizar, aparte de la perfectamente trivial que he ilustrado?

Turing sostuvo que cualquier función computable puede en principio ser computada en una máquina de Turing, con sólo dar suficiente cinta y suficiente tiempo. Esto es, cualquier algoritmo podía ser ejecutado por una máquina de Turing. La mayoría de los lógicos y matemáticos aceptan ahora la pretensión de que ser un algoritmo *es sencillamente* ser capaz de ejecución en alguna máquina de Turing. Esto es, *ser capaz de ejecución en una máquina de Turing* en algún sentido nos dice lo que es un algoritmo. Esta pretensión se llama tesis de Church, del estadunidense Alonzo Church (n. 1903), quien independientemente llegó a conclusiones muy parecidas a las de Turing (a veces se habla de tesis de Church-Turing).[6] La idea básica de la tesis es, en efecto, dar un sentido preciso a la noción de algoritmo, enseñarnos qué es un algoritmo.

Puede aún quererse preguntar: *¿cómo* nos ha enseñado la idea de una máquina de Turing lo que es un algoritmo?, ¿cómo nos ha ayudado recurrir a estas interminables "cintas" y las tediosas cadenas de unos y ceros escritas en ellas? La respuesta de Turing podría ponerse como sigue: lo que hemos hecho es reducir cualquier cosa que reconocemos naturalmente como un procedimiento eficaz a una serie de etapas sencillas realizadas por un dispositivo muy sencillo. Estos pasos son tan sencillos que no es posible para nadie verlos como misteriosos. Lo que hemos hecho, pues, es quitarle el misterio a la idea de un procedimiento eficaz.

[6] Para una exposición muy clara de la tesis de Church-Turing, véase Clark Glymour, *Thinking Things Through* (Cambridge, MIT Press, 1992), pp. 313-315.

Codificación y símbolos

Una máquina de Turing es cierto tipo de dispositivo de *entrada-salida.* Se pone cierta cosa "en" la máquina —una cinta que contiene una cadena de unos y ceros— y se obtiene otra cosa: una cinta que contiene otra cadena de unos y ceros. Mientras tanto, la máquina realiza ciertas cosas a la entrada —las cosas determinadas por su tabla de la máquina o instrucciones— para volverla la salida.

Una cosa que podría haberle preocupado a usted, sin embargo, no es la definición de la máquina de Turing, sino la idea de que semejante máquina pueda ejecutar *cualquier* género de algoritmo. Es fácil ver cómo realiza el algoritmo "súmese 1", y con un poco de imaginación puede verse cómo podría realizar el algoritmo de multiplicación descrito antes. Dije también que podría usted escribir un algoritmo para una receta sencilla, tal como cocer un huevo, o para imaginar qué llave abre determinada cerradura. ¿Cómo puede hacer eso una máquina de Turing? ¿De seguro una máquina de Turing sólo puede calcular con números, y eso es todo lo que puede ser escrito en su cinta?

Por supuesto, una máquina de Turing no puede cocer un huevo o abrir una puerta. El algoritmo que mencioné es una *descripción* de cómo cocer un huevo. Y estas descripciones pueden ser codificadas en una máquina de Turing dada la notación adecuada.

¿Cómo? Hay una manera sencilla de lograrlo. Nuestros algoritmos fueron escritos en español, de manera que necesitamos primero un modo de codificar instrucciones en texto español como números. Podemos hacer esto sencillamente asociando cada letra del alfabeto español y cada

fragmento significativo de puntuación con un número, como sigue:

A – 1, B – 2, C – 3, D – 4, y así sucesivamente.

De manera que mi nombre se escribiría:

20 9 13
3 18 1 14 5

Evidentemente, la puntuación es decisiva. Necesitamos una manera de decir cuándo una letra se detiene y comienza otra, y otra manera de decir cuándo una palabra concluye y otra empieza, y una manera más de saber cuándo un fragmento completo de texto (por ejemplo una tabla de la máquina) se detiene y comienza otro. Esto no presenta ningún problema de principio. (Piénsese en cómo los telegramas anticuados usaban palabras para la puntuación, por ejemplo separando oraciones con "ALTO".) Una vez que hemos codificado un fragmento de texto en números, podemos volver a escribir estos números en binario.

O sea que podemos entonces convertir cualquier algoritmo escrito en español (o en cualquier otro idioma) a código binario. Y esto podría entonces ser escrito en una cinta de la máquina de Turing, y servir como entrada a la máquina de Turing universal.

Por supuesto, los genuinos programas de computadora no usan este sistema de notación para el texto. Pero no me interesan los detalles reales por el momento: la cuestión con la cual estoy tratando de salir adelante es precisamente que una vez que se da uno cuenta de que cualquier fragmento de texto puede ser codificado en términos de números, en-

tonces es evidente que cualquier algoritmo que pueda ser escrito en español (o en cualquier otro idioma) puede ser realizado en una máquina de Turing.

Esta manera de representar es plenamente *digital* en el sentido de que cada elemento representado (una letra o palabra) se representa de una manera enteramente "positivo-negativo". Cualquier cuadrado en una cinta de la máquina de Turing tiene bien un número 1 en ella o un 0. No hay etapas "intermedias". Lo opuesto a la forma digital de representación es la forma *analógica*. La distinción queda mejor ilustrada por el ejemplo familiar de los relojes analógicos y digitales. Los relojes digitales representan el transcurso del tiempo de un modo paso a paso, con números distintos para cada segundo (digamos), y nada de por medio entre estos números. Los relojes analógicos, en cambio, marcan el paso del tiempo por el movimiento tranquilo de una manecilla en la carátula. Las computadoras analógicas no son directamente pertinentes a las cuestiones planteadas aquí; las computadoras discutidas en el contexto de computadoras y pensamiento son todas computadoras digitales.[7]

Ahora, por último, nos estamos acercando a nuestra caracterización de las computadoras. Recuérdese que dije que una computadora es un dispositivo que procesa representaciones de una manera sistemática. Para entender esto necesitamos dar un sentido claro a dos ideas: *(i)* "procesa de un modo sistemático" y *(ii)* "representación". La primera ha sido explicada en términos de la idea de algoritmo, que a su vez ha sido iluminada por la idea de una máquina de Turing. La segunda está implícita en la idea de la máquina de

[7] Acerca de la distinción, véase John Haugeland, *Mind Design* (Cambridge, MIT Press, 1981), introducción, § 5.

Turing: para que la máquina sea entendida como una función realmente computadora, los números de su cinta tienen que ser tomados como *estando en el lugar* o *representando algo*. Otras representaciones —por ejemplo, las oraciones en español— pueden ser codificadas en estos números.

Algunas veces las computadoras se llaman "procesadores de información". Otras "manipulaciones simbólicas". En mi terminología esto es lo mismo que decir que las computadoras procesan representaciones. Las representaciones llevan información en el sentido de que "dicen" algo o se les interpreta como "si dijeran" algo. Esto es *lo que* los procesos de computadora manipulan. *Cómo* procesan o manipulan es realizando procedimientos eficaces.

Ejemplificación y computación de una función

Este comentario sobre las representaciones nos permite ahora hacer una distinción muy importante que es decisiva para comprender cómo la idea de computación se aplica a la mente.[8]

Recuérdese que la idea de una función puede ser extendida más allá de las matemáticas. En la teorización científica, por ejemplo, los científicos a menudo describen el mundo en términos de funciones. Considérese un ejemplo famosamente sencillo: la segunda ley del movimiento de Newton, que dice que la aceleración de un cuerpo es determinada por su masa y las fuerzas aplicadas a ella. Esto puede repre-

[8] Véase D. H. Mellor, "How Much of the Mind is a Computer?", en D. H. Mellor, *Matters of Metaphysics*.

sentarse como $F = ma$, que se lee "Fuerza = masa × aceleración". Los detalles de esto no tienen importancia: la cuestión es que la fuerza o las fuerzas que actúan sobre cierto cuerpo serán iguales a la masa multiplicada por la aceleración. Una función matemática —la multiplicación— cuyos argumentos y valores son números que representan la relación en la naturaleza entre masas, fuerzas y aceleraciones. Esta relación de la naturaleza también es una función: la aceleración de un cuerpo es una función de su masa y las fuerzas ejercidas sobre ella. Llamemos a esto "función de Newton" para simplificar.

Sin embargo, cuando una fuerza particular se ejerce sobre ella, y acelera a cierta velocidad, no *computa* el valor de la función de Newton. Si lo hiciera, entonces toda relación fuerza-masa-aceleración de la naturaleza sería una computación, y todo objeto físico una computadora. Más bien, como diré, una interacción particular *ejemplifica* la función: es decir, es un *ejemplo* de función de Newton. Igualmente, cuando los planetas del sistema solar siguen sus órbitas en torno al sol, lo hacen de tal manera que es una función de "entrada" gravitacional e inercial. Las leyes de Kepler son una manera de describir esta función. Pero el sistema solar no es una computadora. Los planetas no "computan" sus órbitas a partir de la entrada que reciben: sencillamente se mueven.

Así, la distinción decisiva que necesitamos es entre un sistema que *ejemplifica* una función y un sistema que *computa* una función. Por "ejemplificar" quiero decir "ser un ejemplo de" (si se prefiere puede decirse "ser describible por"). Compárese el sistema solar con una computadora real, digamos una sencilla máquina de sumar. (Quiero decir una máquina física de sumar, no una "máquina" abstracta como la de Turing.) Es natural decir que una máquina sumadora computa

la función suma tomando dos o más números como entrada (argumentos) y dando su suma como salida (valor). Estrictamente hablando, esto no es lo que hace una máquina sumadora. Porque, cualesquiera números que sean, no son el tipo de cosa que puede ser alimentado en máquinas, manipulado o transformado. (Por ejemplo, no se destruye el número 3 destruyendo todos los treses escritos en el mundo; eso no tiene sentido.) Lo que realmente hace la máquina sumadora es tomar *numerales* —esto es, representaciones de números— como entrada, y da numerales como salida. Ésta es la diferencia entre la máquina sumadora y los planetas: aunque ejemplifiquen una función, los planetas no emplean representaciones de su entrada gravitacional y de otro género para formar representaciones de su salida.

Computar una función, entonces, requiere representaciones: representaciones como entrada y representaciones como salida. Una manera perfectamente natural de comprender lo que significa "computar una función" es: cuando computamos con papel y lápiz, por ejemplo, o con un ábaco, usamos representaciones de números. Como ha dicho Jerry Fodor: "¡No hay computación sin representación!"[9]

¿Cómo se relaciona este punto con las máquinas y algoritmos de Turing? Una tabla de la máquina de Turing especifica transiciones entre los estados de la máquina. Según la tesis de Church, cualquier procedimiento que es algorítmico paso a paso puede ser modelado en una máquina de Turing. Así, cualquier procedimiento de la naturaleza que puede ser representado de una manera paso a paso puede representarse por una máquina de Turing. La máquina nada más especifi-

[9] Jerry Fodor, *The Language of Thought* (Hassocks, Harvester, 1975); véase también Gallistel, *The Organisation of Learning*, p. 30.

ca las transiciones entre los estados implicados en el proceso. Esto no significa que estos procesos naturales sean *computaciones,* ni más ni menos que el hecho de que las magnitudes físicas tales como mi temperatura corporal puedan ser representadas por números signifique que la temperatura de mi cuerpo realmente *sea* representada por números. Si una teoría de algún fenómeno natural puede ser representada algorítmicamente, entonces la teoría es *computable,* pero éste es un hecho acerca de teorías, no acerca de los fenómenos mismos. La idea de que las teorías puedan o no ser computables no nos ocupará más en este libro.[10]

Sin pretender elaborar el punto, permítaseme recalcar que es por esto que necesitamos distinguir al principio de este capítulo entre la idea de que algunos sistemas pueden ser *modelados* en una computadora y la idea de que algunos sistemas de veras realizan computaciones. Un sistema puede ser modelado en una computadora cuando una *teoría* de ese sistema es computable. Un sistema realiza computaciones, sin embargo, cuando procesa representaciones usando un procedimiento efectivo.

Algoritmos automáticos

Si ha seguido usted la discusión hasta aquí, entonces se le habrá ocurrido una cuestión muy natural. Las máquinas de Turing describen la estructura abstracta de la computación. No obstante, en la descripción de máquinas de Turing hemos recurrido a ideas como "mover la cinta", "leer la cinta", "escribir un símbolo" y cosas así. Hemos tomado estas ideas

[10] Penrose, sin embargo, piensa que la física "última" no será computable, y que este hecho es pertinente para el estudio de la mente; véase *The Emperor's New Mind,* p. 558.

por descontadas, pero ¿cómo se supone que funcionan?, ¿cómo es que cualquier procedimiento eficaz despega del suelo, en absoluto, sin la intervención de un ser humano en cada etapa del procedimiento?

La respuesta es que las computadoras con las que estamos familiarizados usan algoritmos *automatizados.* Usan algoritmos y representaciones de entrada y salida que están, de cierto modo, "incorporadas" a la estructura física de la computadora. La última parte de nuestra exposición sobre las computadoras será una descripción muy breve de cómo puede hacerse esto. Esta breve discusión no puede, por supuesto, encargarse de todos los rasgos principales de cómo funcionan las computadoras reales, pero espero que será suficiente para darle a usted la idea general.

Considérese una máquina muy sencilla (no una computadora) que se usa para atrapar ratones. Podemos pensar de esta ratonera en términos de entrada y salida: la ratonera toma ratones vivos como entrada, y da ratones muertos (o quizá simplemente atrapados) como salida. Una manera sencilla de representar la ratonera se muestra en la figura III.4.

Desde el punto de vista de la descripción sencilla de la ratonera, no importa realmente qué hay en la "caja" RATONERA: lo que hay "en la caja" es cualquier cosa que atrapa a los ratones. Cajas como ésta son conocidas por los ingenieros como "cajas negras": podemos tratar algo como una caja negra cuando no estamos realmente interesados en cómo funciona internamente, sino que nos interesan nada más las tareas de entrada-salida que ejecuta. Por supuesto podemos "colarnos" en la caja negra de nuestra ratonera y representar su interior como en la figura III.5.

Los dos componentes internos de la caja negra son el cebo y el dispositivo que realmente atrapa a los ratones (la

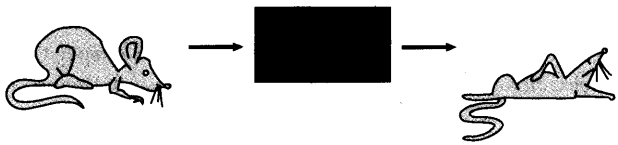


FIGURA III.4. *"Caja negra" ratonera*

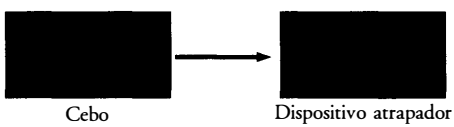


FIGURA III.5. *El interior de la ratonera*

flecha pretende indicar que el ratón se moverá del cebo al dispositivo de atrapar, no a la inversa). En la figura III.4 estamos, en efecto, tratando el CEBO y el DISPOSITIVO ATRAPADOR como cajas negras. Todo nuestro interés está en su función: el CEBO es lo que atrae a los ratones y el DISPOSITIVO ATRAPADOR lo que atrapa al ratón.

Podemos por supuesto irrumpir en *estas* cajas negras y encontrar cómo funcionan. Supóngase que nuestra ratonera es de estilo antiguo, como el de las revistas de tiras cómicas, con una barra de metal mantenida en su lugar por un resorte que se suelta cuando el cebo se quita.

Podemos entonces describir el dispositivo atrapador en términos de sus partes componentes. Y sus partes componentes también —RESORTE, BARRA, etc.— pueden imaginarse como cajas negras. No importa exactamente lo que sean; lo que importa es lo que están *haciendo* en la ratonera.

Estas cajas también pueden ser penetradas, y podemos especificar con mayor detalle cómo funcionan. Lo que es tratado como una caja negra en un nivel puede ser roto en otras cajas negras en otros niveles, hasta que acabamos comprendiendo el funcionamiento de la ratonera.

Este tipo de análisis de máquinas es conocido a veces como "análisis funcional": el análisis de cómo actúa la máquina en las funciones de sus partes componentes. (También a veces se llama "cajología funcional".) Nótese, sin embargo, que la palabra "función" se está usando con un sentido diferente que en nuestra discusión anterior: aquí la función de parte de un sistema es el papel causal que desempeña en el sistema. El uso de "función" corresponde más de cerca al uso cotidiano de la expresión, como en "¿cuál es la función de este trozo?"

Volvamos a las computadoras. Recordemos nuestro sencillo algoritmo de multiplicación. Éste implicaba una serie de tareas, tales como escribir símbolos en los trozos X y Y de papel, y sumar y restar. Piénsese ahora en una máquina que realiza este algoritmo, y pensemos cómo analizarlo funcionalmente. En el nivel más general, claro está, es un multiplicador. Toma números como entrada y le da sus productos como salida. En este nivel puede considerarse como una caja negra (véase figura III.6).

Esto no nos dice gran cosa. Cuando "miramos" dentro de la caja negra, lo que está ocurriendo es lo que se representa por el diagrama de flujo (figura III.7). Cada caja del diagrama de flujo representa una etapa realizada por la máquina. Algunas de estas etapas pueden ser partidas en etapas más sencillas. Por ejemplo, la etapa *(iv)* implica *sumar* el número escrito en Y a la RESPUESTA. Sumar es también un procedimiento paso-a-paso, y así podemos escribir un diagrama de flujo para éste también. Lo mismo que las otras

etapas: restar, "leer" y demás. Cuando analizamos funcionalmente el multiplicador, encontramos que sus tareas se vuelven más sencillas cada vez, hasta que alcanzamos las tareas más simples que puede realizar.

Daniel Dennett ha sugerido un modo vívido de pensar en la arquitectura de las computadoras. Imagínese cada tarea de las cajas del diagrama de flujo ejecutada por un hombrecito, el "homúnculo". La caja más grande (designada como multiplicador en la figura III.6) contiene pues un homúnculo inteligente que, supongamos, multiplica números expresados en notación denaria. Dentro de este homúnculo hay otros homúnculos, menos inteligentes, que pueden sólo sumar y restar y escribir símbolos denarios en el papel. Dentro de estos otros homúnculos hay homúnculos todavía más estúpidos que pueden traducir la notación denaria a binaria. Y dentro de éstos hay homúnculos de veras estúpidos que pueden sólo leer, escribir o borrar números binarios. Así, el comportamiento del multiplicador inteligente es funcionalmente explicado postulando homúnculos progresivamente más y más estúpidos.[11]

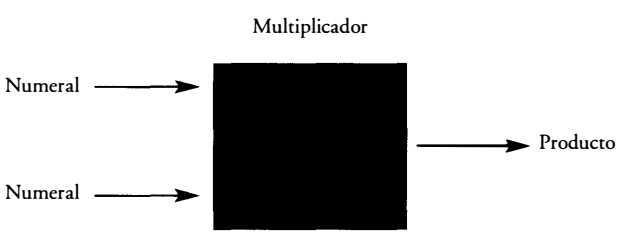


FIGURA III.6. *Caja negra multiplicadora*

[11] Véase Dennett, *Brainstorms* (Hassocks, Harvester Press, 1978).

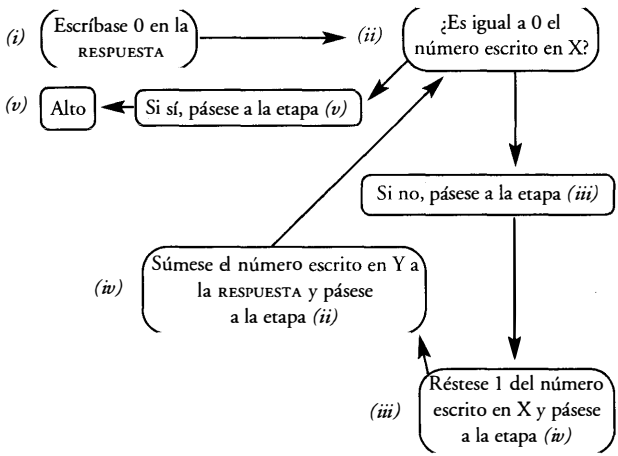


FIGURA III.7. *Diagrama de flujo para el algoritmo de multiplicación, otra vez*

Si tenemos una manera de hacer un dispositivo físico real que funcione como un dispositivo sencillo —un homúnculo estúpido— podemos construir combinaciones de estos dispositivos sencillos dando dispositivos complejos que pueden realizar la tarea del multiplicador. Después de todo, el multiplicador no es más que estos sencillos dispositivos dispuestos de la manera especificada por el diagrama de flujo. Ahora, recuérdese que la gran visión de Turing fue mostrar que cualquier algoritmo podía ser deshecho en tareas lo suficientemente sencillas como para ser ejecutadas por una máquina de Turing. Así, piénsese en los dispositivos más simples como los dispositivos que pueden realizar estas sencillas operaciones de la máquina de Turing: moverse desde la izquierda o la derecha, leer, escribir, etc. Todo

lo que necesitamos hacer es realizar ahora algunos dispositivos que puedan ejecutar estas sencillas operaciones.

Y, por supuesto, tenemos muchas maneras de hacerlos. Para animar la cosa, piénsese en la cinta de alguna máquina de Turing representada por una serie de conmutadores: el conmutador puesto representa 1 y el conmutador apagado representa 0. Entonces cualquier computación puede ser efectuada por una máquina que puede moverse a lo largo de los conmutadores, uno por uno, registrar en qué posición están ("leer") y encenderlos o apagarlos ("escribir"). Mientras tengamos alguna manera de *programar* la máquina (decir por ejemplo qué está imitando la máquina de Turing), entonces hemos construido una computadora de conmutadores.

Las verdaderas computadoras son, en cierto sentido, hechas de "interruptores", aunque no en el sentido sencillo que acabamos de describir. Una de las primeras computadoras (construida en 1944) usaba conexiones telefónicas, en tanto que el famoso esfuerzo de guerra estadunidense llamado ENIAC (usado para calcular trayectorias de proyectiles) estaba construido con válvulas; y las válvulas y las conexiones son, en efecto, simplemente conmutadores. Los auténticos progresos llegaron cuando los procesadores más sencillos (los "conmutadores") pudieron ser construidos a partir de semiconductores, y las computaciones pudieron realizarse más rápido de lo que jamás soñó Turing. Otros progresos importantes llegaron con los "lenguajes de programa" de alto nivel: sistemas de codificación que podían hacer que las operaciones básicas de la máquina realizasen toda suerte de operaciones complejas. Para los fines de este libro, el principio básico que subraya incluso estas máquinas tan complejas puede comprenderse

del modo como he esbozado. (Para mayor información acerca de la historia de la computadora, véase la cronología, al final de este libro.)

Una consecuencia importante de esto es que no importa realmente de qué es la computadora. Lo que importa respecto a una computadora es *qué hace,* esto es, qué tareas computacionales realiza, o qué *programa* pone en marcha. Las computadoras que usamos hoy realizan estas tareas utilizando circuitos electrónicos microscópicos grabados en trocitos de silicio. Sin embargo, aunque esta tecnología es increíblemente eficiente, las tareas que realiza son, en principio, capaces de ser realizadas por ajustes de interruptores, cuentecillas, palillos de cerillos y latas de estaño, y hasta tal vez por la neuroquímica del cerebro. Esta idea se conoce como "realización variable" (o "realización múltiple") del programa (o partes blandas) mediante un mecanismo físico (parte dura), es decir, el mismo programa puede ser "realizado" variablemente o de modo múltiple por diferentes piezas duras.

Agregaré un detalle final acerca de algunas computadoras reales. Es una simplificación decir que todas las computadoras funcionan enteramente de modo algorítmico. Cuando se construyen programas para jugar al ajedrez, por ejemplo, las reglas del ajedrez informan a la máquina, de manera enteramente no ambigua, qué cuenta como un movimiento legal. En cualquier punto del juego sólo las reglas permiten ciertos movimientos. ¿Cómo sabe la máquina *cuál* movimiento realizar, entre todos los posibles? Como una partida de ajedrez concluirá en un número finito —aunque posiblemente muy grande— de movimientos, es posible en principio, para la máquina, recorrerlos por anticipado, representando toda consecuencia de todo movimiento permiti-

do. Sin embargo, esto necesitaría incluso para la computadora más poderosa una cantidad de tiempo enorme. (John Haugeland estima que la computadora tendría que recorrer 10^{120} movimientos, que es un número mayor que el número de estados cuánticos en toda la historia del universo.)[12] Así, los diseñadores de programas para jugar al ajedrez añaden a sus máquinas algunas reglas generales (llamadas *heurísticas*) que sugieren buenos cursos de acción, aunque, a diferencia de los algoritmos, no garanticen una conclusión particular. Una heurística para una máquina jugadora de ajedrez podría ser algo como "trátese y enróquese en el juego tan pronto como sea posible". Las heurísticas han tenido gran influencia en la investigación de la inteligencia artificial. Ahora es tiempo de introducir la idea directriz que hay tras la inteligencia artificial: la idea de una computadora pensante.

¿Computadoras pensantes?

Equipados con un entendimiento básico de lo que son las computadoras, la cuestión que ahora necesitamos preguntar es ¿por qué alguien habría de pensar que una computadora —que procesa representaciones sistemáticamente— puede estructurar pensamientos?

Al principio de este capítulo dije que responder la pregunta de si una computadora piensa requiere que sepamos tres cosas: qué es una computadora, qué es pensar y qué son computadoras y pensamientos que apoyen la idea de que las computadoras podrían pensar. Ahora tenemos cierta idea de lo que es una computadora, y en los capítulos I y II

[12] Véase *Artificial Intelligence* (Cambridge, MIT Press, 1985), p. 178.

discutimos algunos aspectos del concepto de sentido común del pensamiento. ¿Podemos reunir estas cosas?

Hay cierto número de relaciones evidentes entre lo que hemos aprendido acerca de la mente y lo que hemos averiguado acerca de las computadoras. Una es que la noción de *representación* parece figurar en ambas áreas. Uno de los rasgos esenciales de ciertos estados mentales es que representan. Y en este capítulo hemos visto que uno de los rasgos esenciales de las computadoras es que procesan representaciones. Asimismo, los pensamientos de usted le permiten hacer lo que hace a causa de cómo representan el mundo. Y puede discutirse que a las computadoras se les hace producir la salida que producen porque representan: a mi máquina sumadora se le hace producir como salida 5 en respuesta a las entradas 2, +, 3 y =, en parte porque esos símbolos de entrada representan lo que hacen.

Sin embargo, no debemos dejarnos llevar por estas semejanzas. El hecho de que la noción de representación pueda usarse para definir tanto el pensamiento como las computadoras no implica nada acerca de si las computadoras pueden pensar. Considérese esta analogía: la noción de representación puede ser usada para definir tanto el pensamiento como los libros. Uno de los rasgos esenciales de los libros es que contengan representaciones. ¡Empero los libros no pueden pensar! Análogamente, sería una chifladura sostener que las computadoras pueden pensar simplemente porque la noción de representación puede emplearse para definir el pensamiento y las computadoras.

Otra manera de perderse es tomar la noción de "procesamiento de la información" demasiado flojamente. En un sentido, el pensamiento evidentemente implica procesar información: tomamos información de los alrededores, ha-

cemos cosas con ella y la usamos para actuar en el mundo. Sin embargo, sería errado pasar de esto, más el hecho de que las computadoras se conocen como "procesadores de información", a la conclusión de que lo que acontece en las computadoras debe ser un género de pensamiento. Esto descansa en "procesamiento de la información" de una manera muy floja cuando se aplica al pensamiento humano, en tanto que la teoría del procesamiento de información de la computación tiene una definición precisa. La cuestión acerca de las computadoras pensantes es (en parte) si el procesamiento de información que realizan las *computadoras* puede tener algo que ver con el "procesamiento de información" implicado en el *pensamiento*. Esta cuestión no puede ser respondida señalando que las palabras "procesamiento de información" pueden aplicarse tanto a las computadoras como al pensamiento: esto se conoce como "falacia de confusión".

Otra mala manera de discutir, según hemos visto, es decir que las computadoras pueden pensar porque debe haber una tabla de la máquina de Turing para pensar. Decir que hay una tabla de la máquina de Turing para pensar es decir que la *teoría* del pensamiento es computable. Esto puede ser cierto o no serlo. Aun si fuera verdad, evidentemente no implicaría que los pensadores son computadoras. Supóngase que la astronomía fuera computable: esto no implicaría que al universo se le considera una computadora. Una vez más, es decisivo subrayar la distinción entre computar una función y ejemplificar una función.

Por otra parte, tampoco debemos apresurarnos a dejar de lado la idea de las computadoras pensantes. Una crítica destructora familiar es que la gente siempre ha pensado en la mente o el cerebro según las líneas de la última tecnolo-

gía; y el presente entusiasmo por las computadoras pensantes no constituye una excepción. Así es como John Searle plantea las cosas:

> Porque no comprendemos el cerebro muy bien, estamos constantemente tentados a usar la más reciente tecnología como modelo para tratar de comprenderlo. En mi infancia siempre se aseguraba que el cerebro era un tablero telefónico... Sherrington, el gran neurocientífico británico, pensaba que el cerebro funcionaba como un sistema telegráfico. Freud a menudo comparó el cerebro con sistemas hidráulicos y electromagnéticos. Leibniz lo comparó con un molino, y se me ha dicho que algunos de los antiguos griegos pensaban que el cerebro funcionaba como una catapulta. Hoy por hoy, evidentemente, la metáfora es la computadora digital.[13]

Visto de este modo, parece raro que alguien pudiera pensar que el cerebro humano (o la mente), que lleva evolucionando millones de años, explique sus misterios en términos de ideas que surgieron hace 60 o 70 años en especulaciones rebuscadas acerca de los fundamentos de las matemáticas.

Sin embargo, en sí mismo, este punto no prueba nada. El hecho de que una idea sea desarrollada en un contexto histórico específico —¿y qué idea no lo ha sido?— no nos dice nada acerca de la *corrección* de la idea. Pero, también hay una respuesta específica más interesante a la crítica de Searle. Puede ser cierto que la gente haya siempre pensado acerca de la mente por analogía con la tecnología más fre-

[13] Searle, *Minds, Brains and Science* (Harmondsworth, Penguin, 1984), p. 44.

cuente. El caso de las computadoras es muy diferente de los demás casos que Searle menciona. Históricamente, las diferentes etapas en la invención de la computadora siempre han corrido parejas con intentos de sistematizar aspectos del conocimiento y de las habilidades intelectuales humanas, así que nada tiene de sorprendente que los primeros se usaran para modelar (o incluso explicar) las últimas. No ocurre así con la hidráulica, ni con molinos o intercambios telefónicos. Vale la pena insistir en unos cuantos ejemplos.

Junto con muchos de sus contemporáneos, el gran filósofo y matemático G. W. Leibniz (1646-1716) propuso la idea de un "carácter universal" *(characteristica universalis):* un lenguaje matemáticamente preciso, no ambiguo, al cual pudieran traducirse ideas, y mediante el cual las soluciones a las disputas intelectuales podrían ser resueltas por "cálculo". En un pasaje famoso, Leibniz considera las ventajas que semejante lenguaje acarrearía:

> Una vez que los números característicos fueran establecidos para la mayoría de los conceptos, la humanidad poseería un nuevo instrumento que aumentaría las capacidades de la mente en un grado mayor que los instrumentos ópticos que fortifican los ojos, y superarían el microscopio y el telescopio en el mismo grado en que la razón es superior a la visión.[14]

Leibniz no llegó a tanto como a diseñar realmente el carácter universal (aunque es interesante que inventara la no-

[14] G. W. Leibniz, *Selections* (P. Wiener [ed.], Nueva York, Scribner, 1951), p. 23; véase también L. J. Cohen, "On the Project of a Universal Character", *Mind,* 53 (1954).

tación binaria). Sin embargo, con la llamativa imagen de este dispositivo calculador de conceptos, vemos la combinación de intereses que ha preocupado a muchos precursores de las computadoras: por un lado hay un deseo de despojar al pensamiento humano de cualquier ambigüedad y falta de claridad; mientras que, por otro lado, hay la idea de un cálculo o máquina que pudieran procesar estos pensamientos esquemáticos. Estos dos intereses coinciden en las cuestiones que rodean a otra figura muy importante en la historia de las computadoras: el lógico y matemático irlandés George Boole (1815-1864). En su libro *The Laws of Thought* (1854), Boole formuló un álgebra para expresar relaciones lógicas entre enunciados (o proposiciones). Lo mismo que el álgebra ordinaria representa relaciones matemáticas entre números, Boole propuso que pensáramos en las relaciones lógicas elementales entre enunciados o proposiciones —expresadas con palabras como "y", "o", etc.— como expresables en términos algebraicos. La idea de Boole era usar una notación binaria (1 y 0) para representar los argumentos y valores de las funciones expresadas por "y", "o", etc. Por ejemplo, tómense las operaciones binarias $1 \times 0 = 0$ y $1 + 0 = 1$. Ahora, supóngase que 1 y 0 representan *verdadero* y *falso* respectivamente. Entonces podemos pensar en $1 \times 0 = 0$ como si dijera algo así: "Si se tiene una verdad y una falsedad, entonces se tiene una falsedad" y $1 + 0 = 1$ diciendo: "Si se tiene una verdad o una falsedad, se tiene una verdad". Esto es, podemos pensar en $\times$ como representando la "función de verdad" *y*, y pensar + como representando la función de verdad *o*. (Las ideas de Boole serán familiares a los estudiosos de lógica elemental. Una oración "P y Q" es cierta precisamente en caso de que P y Q sean ambos verdaderos, y "P" o "Q," es verdad en caso de que lo sea P o Q.)

Boole sostuvo que construyendo pautas de razonamiento a partir de estas sencillas formas algebraicas podemos descubrir las "leyes fundamentales de esas operaciones de la mente por las cuales es ejecutada la razón".[15] Esto es, aspiró a sistematizar o codificar los principios del pensamiento humano. El hecho interesante es que el álgebra de Boole pasó a desempeñar un papel central en el plano de las computadoras digitales modernas. El comportamiento de la función × en el sistema de Boole puede ser codificado por un sencillo dispositivo conocido como "portal-y" (véase la figura III.8). Un portal-y es un mecanismo que toma corrientes eléctricas de dos fuentes (X y Y) como entradas, y da una corriente eléctrica como salida (Z). El dispositivo está diseñado de tal manera que dé Z cuando, sólo cuando, reciba una corriente de X *tanto* como de Y. En efecto, este dispositivo representa la función de verdad "y". Portales parecidos se construyen para las otras operaciones booleanas: en general estos dispositivos se llaman "portales lógicos" y son decisivos para la planeación de las computadoras digitales de hoy.

A fin de cuentas, las ideas de Boole y Leibniz y otros grandes innovadores, como el matemático inglés Charles Babbage (1792-1871), hicieron nacer la idea de la computadora digital programable de propósito general. La idea

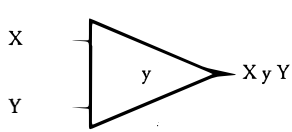


FIGURA III.8. *Un "portal-y"*

[15] George Boole, *The Laws of Thought* (Chicago, Open Court, 1940), vol. II, p. 1.

entonces se convirtió en realidad en los descubrimientos teóricos de Turing y Church, y en los progresos tecnológicos de la electrónica de los años de posguerra (para algunos detalles más véase la "Cronología' al final del libro). Como los casos de Boole y Leibniz ilustran, las ideas que hay detrás de la computadora, por vagas que fueran, estaban vinculadas a menudo con el proyecto general de comprender el pensamiento humano sistematizándolo o codificándolo. Fue simplemente natural que, cuando el público cobró conciencia de las computadoras, fueran saludadas como "cerebros electrónicos".[16]

Estas cuestiones, por supuesto, no *justifican* la pretensión de que las computadoras pueden pensar. Sin embargo, ayudan a ver qué anda mal con algunas reacciones apresuradas ante esta pretensión. En un momento veremos algunos de los argumentos detallados en pro y en contra. Primero necesitamos conceder una breve ojeada a la idea de la inteligencia artificial misma.

Inteligencia artificial

¿Qué es la inteligencia artificial? Es difícil a veces obtener una respuesta rectilínea a esta cuestión, ya que la expresión se aplica a diversos proyectos intelectuales. Algunas personas llaman inteligencia artificial (o IA) a las "máquinas de ciencia de pensar", en tanto que otros, por ejemplo Margaret Boden, son más ambiciosos, y hablan de "la ciencia de la inteligencia en general".[17] Para el recién llegado, la palabra

[16] Véase Haugeland, *Artificial Intelligence,* p. 168 nota 2.

[17] Margaret Boden (ed.), *The Philosophy of Artificial Intelligence* (Ox-

"inteligencia" puede ser una pizca descontrolante aquí, porque sugiere que IA se interesa nada más en tareas que ordinariamente se clasificarían como si requirieran inteligencia (por ejemplo leer libros difíciles o demostrar teoremas de matemáticas). De hecho, abundante investigación de IA se concentra en temas que no consideraríamos, de ordinario, que requirieran inteligencia, tales como ver objetos tridimensionales o entender un texto sencillo.

Algunos de los proyectos que circulan bajo el nombre de IA tienen poco que ver con el pensamiento o las computadoras pensantes. Por ejemplo, hay los llamados "sistemas para expertos", planeados para aconsejar en áreas especializadas del conocimiento (por ejemplo diagnóstico con medicamentos). Refinados como son, los sistemas para expertos no son (ni pretenden ser) computadoras pensantes. Desde el punto de vista filosófico, son sencillamente enciclopedias mejoradas.

La idea filosóficamente interesante detrás de la IA es la de construir una computadora pensante (o alguna otra máquina, puestas así las cosas). Evidentemente es una cuestión interesante en sí misma; pero si Boden y otros están en lo correcto, entonces el proyecto de construir una computadora pensante debe ayudarnos a comprender lo que la inteligencia (o el pensamiento) es en general. Esto es, construyendo una computadora pensante podemos aprender acerca del pensamiento.

Puede no ser evidente cómo se supone que funciona esto. ¿Cómo puede la construcción de una computadora pensante informarnos acerca de lo que pensamos? Considé-

ford, Oxford University Press, 1990), introd., p. 3; la anterior cita es de Alan Garnham, *Artificial Intelligence, an Introduction* (Londres, Routledge, 1988), p. xiii.

rese una analogía: construir una máquina voladora. Los pájaros vuelan y así lo hacen los aeroplanos; pero construir aeroplanos no nos informa gran cosa acerca de lo que los pájaros hacen, de modo que una computadora pensante podría pensar de una manera diferente de como nosotros lo hacemos. De manera que ¿cómo el construir una computadora pensante puede informarnos acerca del pensamiento humano?

Por otra parte, este argumento podría llamar la atención por extraño. Después de todo, pensar es lo que *nosotros* hacemos: la esencia del pensamiento es el pensamiento humano. ¿De manera que podría algo pensar sin pensar del modo como nosotros lo hacemos? Ésta es una buena pregunta. Lo que sugiere es que en lugar de empezar construyendo una computadora pensante y preguntar *entonces* qué nos enseña esto acerca del pensamiento, debemos primero imaginar qué es pensar, y entonces ver si podemos construir una máquina que lo haga. Sin embargo, una vez imaginado lo que es el pensamiento, ¡construir la máquina no nos enseñaría nada que no supiéramos de antemano!

Si la única clase de pensamiento fuera el humano (signifique esto lo que sea), entonces sólo sería posible construir una computadora pensante si el pensamiento humano de hecho *fuera* computacional. Para establecer esto tenemos evidentemente que investigar en detalle qué son el pensamiento y otros procesos mentales. Así, este enfoque requerirá una teoría psicológica que lo sustente, pues se necesitará imaginar qué son los procesos antes de averiguar qué clase de mecanismo computacional realizan dichos procesos. El enfoque entonces implicará una colaboración entre psicología e IA para proporcionar la plena teoría del procesamiento mental humano. Sigo la terminología reciente al

llamar a esta colaboración "ciencia cognitiva"; éste será tema del capítulo IV.[18]

Por otra parte, si algo pudiera pensar, pero *no* del modo como lo hacemos, entonces la IA no se forzaría encontrando cómo funciona la psicología humana. Antes bien, debiera precisamente seguir adelante y hacer una máquina que ejecutara una tarea con pensamiento o inteligencia sin importar el modo como lo hagamos. Ésta fue, precisamente, la manera como procedió la primera investigación sobre IA después de surgir en los años cincuenta. La meta era producir una máquina que hiciera cosas que *requirieran* pensamientos hechos por personas. Se creyó que hacer esto no requeriría un conocimiento detallado de la psicología o la fisiología humana.[19]

Una reacción natural a esto es que tal enfoque puede sólo llegar a producir una *simulación* del pensamiento, no éste en realidad. Para algunos, ello no es un problema: si la máquina pudiera hacer la faena de una manera que simule la inteligencia, entonces ¿por qué preocuparnos de si es "la cosa real" o no? Sin embargo, esta respuesta no es muy provechosa si la IA se supone realmente que es la "ciencia de la inteligencia en general", ya que, confundiendo la distinción entre pensamiento real y simulación, no sería capaz de enseñarnos gran cosa acerca de cómo funciona el pensamiento (presumiblemente real). Así, ¿cómo podría alguien pensar que es aceptable confundir la distinción entre el pensamiento real y su simulación?

[18] Véase David Marr, "Artificial Intelligence: A Personal View", en Margaret Boden (ed.), *The Philosophy of Artificial Intelligence,* y en John Haugeland (ed.), *Mind Design.*

[19] Véase Jack Copeland, *Artificial Intelligence: A Philosophical Introduction* (Oxford, Blackwell, 1993), pp. 26 y 207-208.

La respuesta, creo yo, reside en la historia inicial de la IA. En 1950, Turing publicó un artículo influyente titulado "Maquinaria computadora e inteligencia', que proporcionó algo del fundamento filosófico de la IA. En este artículo, Turing formuló la cuestión: "¿Puede una máquina pensar?" Encontrando esta cuestión demasiado vaga, propuso remplazarla por otra: "¿En qué circunstancias se confundiría una máquina con una persona real?" Turing ideó una prueba en la cual una persona se comunica a distancia con una máquina y otra persona. A grandes rasgos, esta "prueba de Turing" equivale a esto: si la primera persona no puede decir en qué difiere la conversación con la otra persona y la conversación con la máquina, entonces podemos decir que la máquina está pensando.

Hay muchas ramificaciones de esta prueba, y poner en claro detalladamente lo que implica es cosa bastante complicada.[20] Mi propio modo de ver es que los supuestos que hay tras la prueba son behaviorísticos (véase el capítulo II, "Cómo entender otras mentes", p. 88) y que la prueba es por lo tanto inadecuada. El único punto que quiero recalcar aquí es que aceptar la prueba de Turing como decisiva de inteligencia hace posible separar la idea de algo *pensante* de la idea de algo *pensante del modo humano.* Si la propuesta de Turing es una prueba adecuada de pensamiento, entonces todo lo pertinente es cómo la máquina realiza la prueba. No es pertinente si la máquina pasa la prueba del modo humano. La redefinición por Turing de la pregunta "¿Puede una máquina pensar?" permitió a la IA confundir la distinción entre el pensamiento real y su mera simulación.

[20] El artículo de Turing está reimpreso en Boden (ed.), *The Philosophy of Artificial Intelligence.* Para más acerca de la prueba de Turing, véase Ned

Esto nos pone en la posición de distinguir entre las dos preguntas que planteé al principio de este capítulo:

1. ¿Puede pensar una computadora? Esto es, ¿puede pensar algo sencillamente por ser una computadora?
2. ¿Es la mente humana una computadora? Esto es, ¿pensamos (del todo o en parte) computando?

Estas preguntas son distintas, porque alguien que adoptara el enfoque de IA podría responder "Sí" a 1, permaneciendo agnóstico sobre 2 ("No sé cómo *nosotros* logramos pensar, ¡pero hay una computadora que puede!") Igualmente, alguien pudo responder "Sí" a la cuestión 2 aunque negando que una mera computadora pudiera pensar ("Nada podría pensar *sencillamente* computando; pero computar es parte de la historia acerca de cómo pensamos").

El capítulo IV se ocupará de la cuestión 2, en tanto que el resto de este capítulo se ocupará de algunas de las razones filosóficas más interesantes para responder "No" a la pregunta 1. En aras de la claridad usaré los términos "IA" e "inteligencia artificial" para la opinión de que las computadoras pueden pensar, pero debe tenerse presente que estos términos se usan también de otros modos.

¿Cómo ha respondido la filosofía a las pretensiones así definidas de la IA? Sobresalen dos objeciones filosóficas:

1. Las computadoras no pueden pensar porque pensar requiere capacidades que las computadoras, por su naturaleza misma, no pueden tener nunca. Las com-

Block, "The Computer Model of the Mind", y "Psychologism and Behaviourism".

putadoras tienen que obedecer reglas (ya sean algorítmicas o heurísticas), pero pensar nunca puede ser capturado en un sistema de reglas, sin importar cuán complejas sean. Lo que requiere el pensamiento es un compromiso activo con la vida, la participación en una cultura y un "saber cómo" que nunca puede ser formalizado mediante reglas. Éste es el enfoque asumido por Hubert Dreyfus en su crítica corrosiva de la IA: *What Computers Can't Do.*

2. Las computadoras no pueden pensar porque sólo manipulan símbolos de acuerdo con sus rasgos *formales;* no son sensibles a los *significados* de estos símbolos. Éste es el tema de un argumento bien conocido de John Searle: el "cuarto chino".

En las dos secciones finales de este capítulo valoraré estas objeciones.[21]

[21] Desdeño otra pretensión controvertida: que las computadoras no pueden pensar porque un famoso teorema matemático, el teorema de Gödel, muestra que el pensamiento puede implicar el reconocimiento de verdades que no son demostrables, y así no computables. El argumento fue inicialmente propuesto por J. R. Lucas —véase, por ejemplo, *The Freedom of the Will* (Oxford, Oxford University Press, 1970)— y ha sido revivido por Roger Penrose en *The Emperor's New Mind,* Algunos autores piensan que la tesis de Penrose-Lucas es muy importante; otros la dejan de lado en unos cuantos párrafos. Esto es cierto tanto para los amigos del cuadro computacional de la mente —véase, por ejemplo, Glymour, *Thinking Things Through,* pp. 342-343— como para sus enemigos; véase Dreyfus, *What Computers* Still *Can't Do* (Cambridge, MIT Press, ed. rev., 1992), p. 345. En este libro dejaré de lado esta tesis, ya que los asuntos que cubre no pueden ser propiamente valorados sin abundancia de conocimiento técnico.

¿Puede el pensamiento ser capturado por reglas y representaciones?

El *Arizona Daily Star* del 31 de mayo de 1986 comunicó esta desdichada historia:

> Un conductor de autobús novato, suspendido por no hacer lo indicado cuando una muchacha sufrió un ataque cardiaco en su autobús, seguía reglas perfectamente estrictas que prohíben a los conductores dejar su ruta sin permiso, dijo ayer un funcionario de la unión. "Si hay que acusar a alguien, póngase la acusación en las reglas que estas personas tienen que seguir" (dijo el funcionario). (Un vocero de la compañía de autobuses defendió las reglas:) "Da usted una mínima libertad, y ¿dónde acaba la cosa?"[22]

El comportamiento del desdichado conductor puede usarse para ilustrar un problema perenne para la IA. Ateniéndose a la regla estricta —"Sólo abandonar la ruta si se tiene permiso"—, el conductor no tuvo que vérselas con la emergencia de una manera inteligente, pensante. Sin embargo, las computadoras deben, por su naturaleza misma, adherirse a (al menos algunas) reglas estrictas y, por lo tanto, nunca tendrían que enfrentarse con la clase de respuestas flexibles, espontáneas, que tienen los verdaderos pensantes. La objeción concluye que el pensamiento no puede ser cuestión de usar reglas estrictas; así que las computadoras no pueden pensar.

[22] Esta historia viene de Harry Collins, "Will Machines Ever Think?", *New Scientist* (20 de junio de 1992), p. 36.

Esta objeción es un poco apresurada. ¿Por qué el problema no reside en las reglas *particulares* elegidas antes bien que en la idea de seguir una regla como tal? El problema de la regla en el ejemplo —"Apártese de su ruta sólo si tiene permiso"— es precisamente que resulta demasiado sencillo, no que sea una *regla*. La compañía de autobuses debió haber dado al conductor una regla más bien como: "Sólo abandone su ruta si tiene usted permiso, a menos que en el vehículo haya una emergencia médica, en cuyo caso deberá dirigirse al hospital más cercano". Esta regla se ocuparía del caso de un ataque cardiaco, pero ¿y si el conductor sabe que el hospital más cercano está siendo sitiado por terroristas?, ¿o si sabe que hay un médico en el autobús?, ¿debe obedecer la regla que le manda ir a un hospital? Probablemente no, pero si no lo hace, ¿deberá entonces obedecer otra regla? ¿Pero cuál?

Es absurdo suponer que la compañía de autobuses le entregue al conductor una regla como: "Sólo abandone la ruta si tiene permiso, a menos que ocurra en el vehículo una emergencia médica, en cuyo caso deberá usted dirigirse al hospital más cercano, a menos que el hospital esté siendo sitiado por terroristas internacionales, o a menos que haya un médico en el vehículo, o... en cuyo caso deberá usted..." (ni siquiera sabemos cómo llenar los puntos). ¿Cómo podemos obtener una regla que sea suficientemente *específica* para dar a la persona que la sigue instrucciones precisas sobre lo que debe hacer (por ejemplo "diríjase al hospital más cercano" y "haga algo razonable") pero suficientemente *generales* para que se apliquen a todas las eventualidades (por ejemplo no sólo los ataques cardiacos, sino las emergencias en general)?

En su ensayo "Politics and the English Language", Geor-

ge Orwell presenta cierto número de reglas para escribir bien (por ejemplo "Nunca use una palabra larga donde una corta bastaría"), hasta concluir con la regla: "Rompa cualquiera de estas reglas tan pronto se diga algo francamente bárbaro".[23] Podríamos añadir una regla análoga al manojo de reglas dadas al conductor del autobús: "Rompa cualquiera de estas reglas antes de hacer algo estúpido". O, más cortésmente: "¡Use su sentido común!"

Con los seres humanos generalmente podemos contar con que usen su sentido común, y es difícil saber cómo podríamos comprender problemas como el del conductor del autobús sin recurrir (en alguna etapa) a algo como el sentido común, o "lo que es razonable hacer". Si una computadora fuera a vérselas con un problema sencillo como éste, tendría que usar su sentido común también. Sin embargo, las computadoras trabajan manipulando representaciones de acuerdo con reglas (algorítmicas o heurísticas). Así, para que una computadora se enfrente al problema debería almacenar sentido común dentro de ella en términos de reglas y representaciones. Lo que IA necesita, pues, es una manera de programar computadoras con representaciones explícitas de conocimiento de sentido común.

Esto es lo que Dreyfus dice que no puede hacerse. Arguye que la inteligencia humana requiere "el trasfondo de sentido común que los seres humanos adultos poseen en virtud de tener cuerpos, interactuar hábilmente con el mundo material y haber sido adiestrados en una cultura".[24] Y, de acuerdo con Dreyfus, este conocimiento de sentido común no puede ser representado como "una vasta base de conoci-

[23] George Orwell, "Politics and the English Language", *Inside the Whale and other Essays* (Harmondworth, Penguin, 1957), p. 156.

[24] Dreyfus, *What Computers* Still *Can't Do*, p. 3.

miento proposicional", es decir, como un manojo de reglas y representaciones de hechos.[25]

La razón principal por la cual el conocimiento de sentido común no puede ser representado como un manojo de reglas y representaciones es que el conocimiento de sentido común es, o cuenta con serlo, una especie de *saber-cómo.* Los filósofos distinguen entre saber *que* algo es el caso y saber *cómo* hacer algo. La primera clase de conocimiento es cuestión de saber hechos (el tipo de cosas que pueden ser escritas en libros: por ejemplo, que Sofía es la capital de Bulgaria), en tanto que la segunda es cuestión de tener capacidades o habilidades (por ejemplo, ser capaz de andar en bicicleta).[26] Muchos filósofos creen que una capacidad como la de saber montar en bicicleta no es algo que pueda reducirse por entero al conocimiento de ciertas reglas o principios. Lo que hace falta tener cuando se sabe montar en bicicleta no es "aprendizaje-libresco": no se emplean reglas como: "Al dar vuelta a la derecha en una esquina, inclínese ligeramente a la derecha con la bicicleta". Sencillamente se *agarra el modo* mediante un método de prueba y error.

Y, según Dreyfus, agarrarle el modo es lo que uno hace cuando también tiene inteligencia general. Saber *qué es una silla* no es nada más una cuestión de saber la definición de la palabra "silla'. También implica esencialmente saber qué hacer con las sillas, cómo sentarse en ellas, levantarse de ellas, ser capaz de decir qué objetos de la habitación son sillas, o qué clases de cosas pueden usarse como sillas si no las hay a la mano; esto es, el conocimiento presupone un

[25] *Ibid.*, p. xvii.

[26] Véase Gilbert Ryle, *The Concept of Mind* (Londres, Hutchinson, 1949), cap. 2.

"repertorio de capacidades corporales que bien puede ser indefinidamente grande, porque parece haber una variedad indefinida de sillas y de maneras de sentarse en ellas (con gracia, comodidad, seguridad, equilibrio, etc.)".[27] La clase de conocimiento que fundamenta nuestra manera cotidiana de vivir en el mundo es —o descansa en— un saber-cómo práctico de esta clase.

Una computadora es un dispositivo que procesa representaciones de acuerdo con reglas. Y las representaciones y reglas evidentemente no son habilidades. Un libro contiene representaciones, y puede contener representaciones de reglas también, pero un libro no tiene habilidades. Si la computadora tiene conocimiento, debe ser "conocimiento de que esto o lo otro es el caso", más bien que "conocimiento de cómo hacer esto o lo otro". Así, si Dreyfus está en lo correcto, y la inteligencia general requiere sentido común, y el sentido común es una clase de saber-cómo, entonces las computadoras no pueden tener sentido común, y la IA no puede conseguir crear una computadora que tenga inteligencia general. Las dos maneras obvias en que pueden responder los defensores de la IA son *o bien* rechazar la idea de que la inteligencia general requiere sentido común *o* rechazar la idea de que el sentido común es saber-cómo.

La primera opción no es prometedora —¿cómo podría haber inteligencia general que no empleara el sentido común?— ni es popular entre los investigadores de la IA.[28] La segunda es una respuesta más habitual. Los defensores de esta opción pueden decir que requiere trabajo duro hacer explícitos los supuestos implícitos en el punto de vista de

[27] Dreyfus, *What Computers* Still *Can't Do,* p. 37.
[28] *Ibid.,* p. 27.

sentido común del mundo; pero esto no significa que no pueda hacerse. De hecho, se ha ensayado. En 1984, la Corporación de Microelectrónica y Tecnología de la Computación, de Texas, estableció el proyecto CYC, cuya intención era construir una base de conocimiento de una gran cantidad de conocimiento de sentido común. (El nombre "CYC" deriva de la palabra en inglés *encyclopaedia.*) Quienes trabajan en el proyecto de CYC intentan entrar en los supuestos de sentido común acerca de la realidad, tan fundamentales y evidentes que normalmente se descuidan (por ejemplo que los objetos sólidos no suelen ser penetrables por otros objetos sólidos, etc.). La meta es expresar un gran porcentaje del conocimiento de sentido común en términos de aproximadamente 100 millones de proposiciones, codificadas en una computadora. En los primeros seis años del proyecto, un millón de proposiciones estaban en su sitio. El director del proyecto CYC, Doug Lenat, pretendió una vez que, para 1994, habrían almacenado entre 30 y 50% del conocimiento de sentido común (o, como dicen, "realidad de consenso").[29]

Las ambiciones que sustentan esquemas como CYC han sido muy criticadas por Dreyfus y otros. Sin embargo, aun si todo el conocimiento de sentido común pudiera almacenarse como un manojo de reglas y representaciones, esto sólo sería el principio de los problemas de la IA. Pues no es suficiente para la computadora tener sencillamente almacenada la información; debe estar en condiciones de recuperarla y usarla de una manera inteligente. No es suficiente

[29] Para una discusión sobre el proyecto CYC, véase Jack Copeland, *Artificial Intelligence: A Philosophical Introduction* (Oxford, Blackwell, 1993), cap. 5, § 6, de donde tomé estos detalles. Dreyfus discute CYC en detalle en la introducción a *What Computers* Still *Can't Do.*

tener una enciclopedia, hay que estar en condiciones de saber buscar cosas en ella.

Aquí es decisiva la idea de *pertinencia.* Si la computadora no puede saber qué hechos son pertinentes para cuáles otros hechos, no funcionará bien usando el sentido común que ha almacenado para resolver problemas. Pero que una cosa sea pertinente para otra varía conforme varían las concepciones del mundo. El sexo de una persona ya no se juzga pertinente para saber si tiene derecho a votar; pero así fue hace 200 años.

La pertinencia va de la mano con un sentido de lo que está fuera de lugar o lo que es excepcional o desacostumbrado. Aquí está lo que dice Dreyfus acerca de un programa destinado a comprender historias de restaurantes:

> El programa no ha entendido una historia de restaurantes del modo como procede la gente en nuestra cultura, mientras no pueda responder cosas tan sencillas como: cuando el camarero llegó a la mesa, ¿llevaba ropa?, ¿caminaba hacia adelante o hacia atrás?, ¿comió el cliente la comida con la boca o con la oreja? Si el programa contesta "no lo sé", sentimos que todas las respuestas acertadas eran trucos o conjeturas felices y que no ha entendido nada de nuestro comportamiento cotidiano en el restaurante.[30]

Dreyfus arguye que es sólo porque tenemos una manera de vivir en el mundo fundada en habilidades e interacción con cosas (más bien que la representación del conocimiento proposicional o "conocimiento de que esto y lo otro"), que somos capaces de saber qué clases de cosas están fuera de lugar, y qué es pertinente para qué.

[30] *What Computers* Still *Can't Do,* p. 43.

Hay mucho más en la crítica de Dreyfus a la IA de lo que sugiere este breve resumen, pero cuento con que esto dé una idea de la línea general de ataque. Los problemas suscitados por Dreyfus se agrupan a menudo bajo el encabezado del "problema del marco",[31] y suscitan algunos de los puntos más difíciles para el enfoque tradicional a la IA, la clase de IA descrita en este capítulo. Hay múltiples maneras de responder a Dreyfus. Una respuesta es la del proyecto CYC: intentarlo y enfrentarnos al desafío de Dreyfus introduciendo la "realidad de consenso". Otra respuesta es conceder que la IA "clásica", basada en reglas y representaciones, ha fracasado al captar las habilidades fundamentales para el pensamiento: la IA necesita un enfoque radicalmente diferente. En el capítulo IV esbozaré un ejemplo de esta actitud, conocida como "conexionismo". Otra respuesta, ni qué decir tiene, es levantar las manos con desesperación, y dejar todo el proyecto de hacer una máquina pensante. Muy por lo menos, los argumentos de Dreyfus presentan un reto para el programa de investigación de la IA: el reto es representar el conocimiento de sentido común en términos de reglas y representaciones. Y, cuando mucho, los argumentos señalan la caída final de la idea de que la esencia del pensamiento es manipular símbolos según reglas. Cualquier modo de ver que se escoja, creo que la aspiración de Dreyfus autoriza cierto grado de escepticismo acerca de la idea de construir una computadora pensante.

[31] Para el problema del marco, véase Daniel Dennett, "Cognitive Wheels: The Frame Problem of AI", en Margaret Boden (ed.), *The Philosophy of Artificial Intelligence;* Jack Copeland, *Artificial Intelligence,* cap. 5.

El cuarto chino

Dreyfus sostiene que los programas convencionales de IA no tienen probabilidades de producir cualquier cosa que logre pasar por inteligencia general, o sea, pasar plausiblemente la prueba de Turing. John Searle adopta una actitud diferente. Admite, por mor del argumento, que un programa de IA podría pasar la prueba de Turing. Entonces sostiene que, aunque la pasara, sólo sería una *simulación* del pensamiento, y no el auténtico.[32]

Para establecer su conclusión, Searle usa un experimento mental que llama "el cuarto chino". Se imagina a sí mismo dentro de un cuarto con dos ventanas, llamémoslas 1 y 0, respectivamente. A través de la ventana entran trozos de papel con marcas complejas. En el cuarto hay un gran libro escrito en español, en donde hay escritas instrucciones de la forma "Siempre que reciba usted un trozo de papel por la ventana con *esta* clase de marcas encima, haga ciertas cosas con él, y tire un trozo de papel con *esa* clase de marcas, a través de la ventana". Hay también un montón de pedazos de papel con marcas dentro del cuarto.

Ahora supóngase que las marcas son en realidad caracteres chinos, los que vienen por la ventana 1 son preguntas, y los que pasan por la ventana 0 son respuestas razonables a las preguntas. La situación se asemeja ahora al establecimiento de una computadora: un manojo de reglas (el programa) opera sobre símbolos, dando algunos símbolos a

[32] Véase "Minds, Brains and Programs", *Behavioral and Brain Sciences,* 1980, y *Minds, Brains and Science* (Harmondsworth, Penguin, 1984), capítulo 2.

través de la ventana de salida en respuesta a otros símbolos llegados por la ventana de entrada.

Searle acepta, para llevar adelante la argumentación, que, con un programa adecuado, el sistema podría pasar la prueba de Turing. Desde fuera del cuarto, hablantes de chino podrían pensar que estaban conversando con la persona del cuarto. En realidad, la persona del cuarto (Searle) no entiende el chino. Searle no hace sino manipular los símbolos de acuerdo con su forma (a grandes rasgos, su figura), no tiene idea de lo que quieren decir los símbolos. El cuarto chino, por lo tanto, se supone que muestra que hacer funcionar un programa de computadora nunca puede constituir un genuino entendimiento o pensamiento, ya que todo lo que las computadoras pueden hacer es manipular símbolos de acuerdo con su forma.

La estructura general de la argumentación de Searle es como sigue:

1. Los programas de computadora son puramente formales o "sintácticos": a grandes rasgos, son sensibles únicamente a las "figuras" de los símbolos que procesan.
2. La comprensión genuina (y, por extensión, todo pensamiento) es sensible al significado (o "semántica") de los símbolos.
3. La forma (o sintaxis) nunca puede constituir, ni ser suficiente para ello, un significado (o semántica).
4. Por lo tanto, hacer funcionar un programa de computadoras nunca puede ser suficiente para la comprensión o el pensamiento.

El meollo de la argumentación de Searle es la premisa 3. Las premisas 1 y 2 se supone que no despiertan controver-

sia, y la defensa de la premisa 3 es proporcionada por el experimento mental del cuarto chino. (Los términos "sintaxis" y "semántica" se explicarán con mayor detalle en el capítulo IV. Por el momento, tómense como si significaran "forma" y "significado", respectivamente.)

La respuesta evidente a la argumentación de Searle es que la analogía no funciona. Searle sostiene que la computadora no entiende chino porque en el cuarto chino *él* no entiende chino. Pero sus críticos responden que esto no es lo que debiera decir la IA. Searle-en-el-cuarto es análogo a sólo una *parte* de la computadora, no la computadora misma. La computadora misma es análoga a Searle + el cuarto + las reglas + los otros pedazos de papel (los datos). Así, afirman los críticos, Searle está proponiendo que la IA pretende que una computadora comprende porque una *parte* de ella comprende; pero nadie trabajando sobre la IA diría eso. Antes bien, dirían que todo el cuarto (o sea toda la computadora) entiende chino.

Searle no puede resistir divertirse con la idea de que un cuarto pueda comprender, pero, por supuesto, esto es filosóficamente impertinente. Su respuesta seria a esta crítica es ésta: supóngase que *memorizo* todas las reglas y los datos. Puedo entonces hacer todas las cosas que hice dentro del cuarto, excepto que, por haber memorizado las reglas y los datos, lo puedo hacer fuera del cuarto. No obstante, sigo sin entender chino. Así el recurso al entendimiento del cuarto no contesta la cuestión.

Algunos críticos objetan a esto diciendo que memorizar las reglas y los datos no es un asunto trivial, que es como decir que una vez que lo ha hecho usted ¿no entendería? Arguyen que es una falla de la imaginación, por parte de Searle, lo que lo hace descartar esta posibilidad. (Regresaré a esto luego.)

Otra manera de objetar a Searle aquí es decir que si Searle no hubiese sólo memorizado las reglas y los datos, sino también empezado a *actuar* en el mundo de los chinos, entonces es plausible que antes de mucho tiempo hubiese advertido lo que significan estos símbolos. Supóngase que los datos se referían a una conversación de restaurante (en el estilo de algunos programas de la IA reales), y Searle fuese de hecho un camarero en un restaurante chino. Llegaría a la conclusión, por ejemplo, de que determinado símbolo iba siempre asociado a solicitudes de arroz frito, otro a demandas de albóndigas de aleta de tiburón, y así sucesivamente. Y esto sería el principio (en cierto sentido) para ver lo que significan.

La objeción de Searle a esto es que el defensor de la IA ahora ha concedido lo que él quería: no es bastante para comprender que un programa esté en funciones; necesitamos interacción con el mundo para el entendimiento genuino. La idea original de la IA, sostiene, era que "correr" un programa era suficiente *de por sí* para comprender. Así, esta respuesta admite efectivamente que la idea principal que hay detrás de la IA está equivocada.

Estrictamente hablando, Searle tiene razón aquí. Si dice usted que, a fin de pensar, necesita interactuar con el mundo, entonces habrá abandonado la idea de que una computadora puede pensar *sencillamente porque* es una computadora. Pero adviértase que esto no significa que la computación no esté implicada con el pensamiento en algún nivel. Alguien que ha realizado la tarea (tal vez prácticamente imposible) de memorizar las reglas y los datos sigue manipulando símbolos de una manera gobernada por reglas o algoritmos. Es precisamente que él o ella necesita interactuar con el mundo para dar sentido a estos símbolos. ("Interactuar con el mundo" es, por supuesto, algo muy vago. Algo más se dirá

al respecto en el capítulo v.) Así, la argumentación de Searle no toca la idea general de la ciencia cognitiva: la idea de que pensar podría ser realizar computaciones, aun cuando no sea todo lo que está en juego. Searle tiene plena conciencia de esto, y también ha suministrado una argumentación separada contra la ciencia cognitiva, de la cual veremos aspectos en el capítulo IV.

¿Qué conclusión debemos sacar de la argumentación de Searle? Un punto acerca del cual creo que está muy en lo cierto es su premisa 3 en la argumentación anterior: la sintaxis no es suficiente para la semántica. Esto es, los símbolos no "se interpretan solos". Esto, en efecto, es un enunciado seco del problema mismo de la representación. Si fuera falso, entonces en cierto sentido no habría problema de representación. ¿Significa esto que no puede haber explicación de cómo los símbolos significan lo que significan? No por fuerza, algunas explicaciones serán examinadas en el capítulo v. Debemos siempre tener presente que cuando estamos dando una explicación semejante, no introducimos subrepticiamente lo que estamos tratando de explicar (comprensión, significado, semántica, etc.). Considero que ésta es la lección principal de la argumentación de Searle contra la IA.

Sin embargo, algunos filósofos han puesto en tela de juicio si Searle tiene derecho incluso a esta premisa. Los materialistas eliminativos Paul y Patricia Churchland usan una analogía física para ilustrar este punto. Supóngase que alguien aceptara *i)* que la electricidad y el magnetismo fueran fuerzas y *ii)* que la propiedad esencial de la luz es la luminosidad. Entonces podrían sostener *iii)* que las fuerzas no pueden ser suficientes para la luminosidad ni pueden constituirla. Pueden apoyar esto mediante el siguiente experimento mental (el "cuarto luminoso"). Imagínese a alguien en un

cuarto oscuro, moviendo un imán. Esto generará ondas electromagnéticas, pero sin importar cuán rápido mueva el imán el cuarto seguirá oscuro. Se extrae la conclusión de que la luz no puede ser radiación electromagnética.

Si la luz es radiación electromagnética, entonces ¿cuál fue el error? Los Churchland dicen que el error está en la tercera premisa: las fuerzas no pueden ser suficientes, o no constituyen la luminosidad. Esta premisa es falsa, y el experimento mental del cuarto luminoso no puede establecer su verdad. Asimismo, sostienen que el yerro de la argumentación de Searle está en la tercera premisa, en sostener que la sintaxis no es suficiente para la semántica, y que el cuarto chino no puede establecer su verdad. Para los Churchland, que la sintaxis sea suficiente para la semántica es una cuestión empírica, científica, y no algo que pueda establecerse sobre la base de experimentos mentales imaginativos como el cuarto chino:

> Goethe encontraba inconcebible que partículas diminutas por sí mismas pudieran constituir el fenómeno de la luz, o ser suficientes para él. Incluso en este siglo ha habido gente que encontraba inimaginable que la materia inanimada por sí misma, y organizada de cualquier manera, pudiera constituir la vida o ser suficiente para ésta. Claramente, lo que la gente puede o no imaginar no tiene nada que ver con lo que es o no el caso, aun cuando la gente en cuestión sea sumamente inteligente.[33]

Esto es una versión de la objeción de que Searle está atascado por los límites de lo que puede imaginar. Como respues-

[33] Paul M. Churchland y Patricia Smith Churchland, "Could a Machine Think?", *Scientific American* (enero de 1990), p. 29.

ta, Searle ha negado que sea o pueda ser una cuestión empírica si la sintaxis es suficiente para la semántica, así que el cuarto luminoso no es una buena analogía. Para comprender esta respuesta necesitamos saber un poquito más acerca de las nociones de sintaxis y semántica, y cómo se aplicarían a la mente. Ésta será una de las metas del capítulo IV.

Conclusión: ¿puede pensar una computadora?

¿Qué debiéramos entonces hacer con la IA y la idea de las computadoras pensantes? En 1965 uno de los precursores de la IA, Herbert Simon, predijo que "habrá máquinas capaces, dentro de 20 años, de hacer cualquier labor que un hombre pueda hacer".[34] Casi 40 años después todavía no parece haber oportunidad de que esta predicción se cumpla. ¿Es éste un problema-de-principio para la IA, o es sencillamente un asunto de más tiempo y más dinero?

Dreyfus y Searle creen que es un problema-de-principio. El resultado de la argumentación de Dreyfus era, cuando menos, éste: si una computadora ha de tener inteligencia *general*—es decir, ser capaz de razonar acerca de cualquier tema— entonces tiene que poseer conocimiento de sentido común. La cosa ahora, para la IA, es si el conocimiento de sentido común podría representarse en términos de reglas y representaciones. Hasta la fecha, todos los intentos para lograrlo han fracasado.[35]

La lección de la argumentación de Searle, a mi parecer,

[34] Citado por Dreyfus, *What Computers* Still *Can't Do,* p. 129.

[35] Véase Copeland, *Artificial Intelligence,* caps. 5 y 9, para una apreciación ecuánime de los fracasos de la IA.

es harto diferente. La argumentación misma de Searle es un círculo vicioso contra la IA (en efecto) con sólo negar su tesis central: que el pensamiento es manipulación de símbolos formales. El supuesto de Searle, sin embargo, me parece correcto. Sostuve que la respuesta apropiada a lo que sostiene Searle es: sí, claro, Searle-en-el-cuarto, si no es que el cuarto sólo, no puede entender chino. Pero si se deja que el mundo exterior tenga algún efecto sobre el cuarto, el significado o "semántica" podría empezar a tener un sustento. Por supuesto, esta respuesta concede que el pensamiento no puede ser sencillamente manipulación de símbolos. Nada puede pensar con sólo ser una computadora.

Sin embargo, esto no significa que la idea de computación no se aplique de ningún modo a la mente. Pues pudiera ser cierto que nada pensara *sencillamente* siendo una computadora, y también que la manera como *nosotros* pensamos sea *parcialmente* por computación. Esta idea será discutida en el siguiente capítulo.

LECTURAS ADICIONALES

Una introducción muy buena (aunque técnica) a la inteligencia artificial es la de S. J. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach* (Englewood Cliffs, Prentice Hall, 1995). Los mejores libros filosóficos sobre el tema de este capítulo son de John Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge, MIT Press, 1985) y Jack Copeland, *Artificial Intelligence: A Philosophical Introduction* (Oxford, Blackwell, 1993). Hay cierto número de buenos libros generales que exponen los conceptos medulares de la computación de un

modo claro y no técnico. Uno de los mejores es de Joseph Weizenbaum, *Computer Power and Human Reason* (Harmondsworth, Penguin, 1984), capítulos 2 y 3. El capítulo 2 de Roger Penrose, *The Emperor's New Mind* (Oxford, Oxford University Press, 1989) da una exposición muy clara de las ideas de algoritmo y de la máquina de Turing, con ejemplos útiles [ed. en el FCE, *La mente nueva del emperador,* México, 1996]. Una introducción llana al fundamento lógico y matemático de la computación se debe a Clark Glymour, en *Thinking Things Through* (Cambridge, MIT Press, 1992), capítulos 12 y 13. El libro de Hubert Dreyfus ha sido reimpreso, con una nueva introducción, como *What Computers* Still *Can't Do* (Cambridge, MIT Press, 1992). La famosa crítica de Searle sobre la IA puede encontrarse en su libro *Minds, Brains and Science* (Harmondsworth, Penguin, 1984), y también en un artículo que antecedió al libro, "Minds, Brains and Programs", que está reimpreso en la útil antología de Margaret Boden, *The Philosophy of Artificial Intelligence* (Oxford, Oxford University Press, 1990). Ésta contiene también el famoso artículo de Turing: "Computing Machinery and Intelligence", y un importante trabajo de Dennett acerca del problema del marco. El artículo de Searle, junto con varios interesantes ensayos debidos a algunos de los fundadores de la IA, está reimpreso también en la antología de John Haugeland, *Mind Design* (Cambridge, MIT Press, 1981; 2ª ed., sustancialmente revisada, 1997), que incluye una buena introducción de Haugeland.

IV. LOS MECANISMOS DEL PENSAMIENTO

La idea central del punto de vista mecánico acerca de la mente es que esta forma es parte de la naturaleza, algo que tiene una estructura regular, gobernada por ley. Es otra cosa decir que la estructura causal de la mente es también una estructura *computacional,* que el pensamiento es computador. Sin embargo, muchos creyentes en la mente mecánica creen en la mente computacional también. De hecho, la asociación entre pensamiento y computación es tan vieja como la misma imagen mecánica del mundo:

> Cuando un hombre razona, no hace otra cosa sino concebir una suma, de la *Adición* de partículas; o concebir un residuo, de la *Sustracción* de una suma de otra: lo cual (si se ha de hacer con Palabras) es concebir la consecuencia de los nombres de todas estas partes para nombrar el conjunto; o de los nombres del todo y de una parte para nombrar la otra parte... A partir de lo cual podemos definir (esto es, determinar) lo que es, qué significa esta palabra *Razón,* cuando la contamos entre las Facultades de la mente. Pues la razón, en este sentido, no es sino *Calcular* (esto es, Sumar y Restar) las consecuencias de los nombres generales en los que se convino para *marcar* y *significar* nuestros pensamientos; digo *marcar* éstos, cuando calculamos por nosotros mismos; y *significar,* cuando

> demostramos o aprobamos nuestros cálculos en otros hombres.[1]

Éste es un extracto del *Leviathan* (1651) de Thomas Hobbes. La idea de Hobbes de que el razonamiento es "calcular" ha llamado la atención de algunos autores como prefiguración del punto de vista computacional del pensamiento.[2] El fin de este capítulo es considerar dicho punto de vista.

Según recalqué en el capítulo III, el punto de vista computacional del pensamiento es distinto de la pretensión de que algo puede pensar sencillamente volviéndose una computadora de determinado género. Aun si negáramos que cualquier cosa podría pensar sin más que computar, podríamos sostener que nuestros pensamientos tienen una base computacional. Esto es, podríamos pensar que *algo* de *nuestros* estados y procesos mentales es, de alguna manera, computacional, sin pensar que la idea de computación agote la naturaleza del pensamiento.

La idea de que algunos estados y procesos mentales son computacionales es dominante en la filosofía de la mente en este momento y en la psicología cognitiva, y, por esta razón cuando menos, es una idea digna de ser explorada en detalle. Antes de discutir estas teorías necesitamos saber qué fenómenos mentales podrían ser considerados plausiblemente computacionales. Sólo entonces sabremos de qué fenómenos pudieran ser ciertas estas teorías.

[1] Hobbes, *Leviathan,* parte I, "Of Man", capítulo v, "Of Reason and Science". [Ed. en el FCE: *Leviatán,* parte I, "Del hombre", cap. v, "De la razón y la ciencia", México, 2003.]

[2] Véase John Haugeland, *Mind Design,* introducción.

Cognición, computación y funcionalismo

He hablado acerca de la idea de que la *mente* es una computadora; pero ahora necesitamos ser un poco más precisos. En nuestra discusión de los fenómenos mentales en el capítulo 1 ("La tesis de Brentano", véase p. 73) sacamos a luz una disputa acerca de si todos los estados mentales son representacionales (o exhiben intencionalidad). Algunos filósofos piensan que ciertos estados mentales —como las sensaciones corporales, por ejemplo— tienen propiedades no representacionales conocidas como "qualia". Desde este punto de vista, pues, no todos los estados mentales son representacionales. Si este punto de vista es correcto, no será posible para la mente completa ser una computadora, porque la computación se define en términos de representación; recuérdese que una computadora es un dispositivo que procesa representaciones de una manera sistemática. Así, sólo aquellos estados mentales que son puramente representacionales podrían ser candidatos a ser estados computacionales. El otro punto de vista posible (conocido como "representacionalismo" o "intencionalismo") dice que todos los estados mentales, en todos sus aspectos, son de naturaleza representacional. Basándose unos en este modo de ver, no hay obstáculo en principio para que todos los estados mentales sean de naturaleza computacional.

No decidiré esta disputa aquí, sino que volveré a ella brevemente en el capítulo VI.[3] Mi estrategia en este capítulo será apoyar lo mejor posible la teoría computacional de la

[3] Para más discusión acerca del intencionalismo, véase mis *Elements of Mind*, especialmente los caps. 3 y 5.

mente, es decir, considerar los ejemplos más firmes de estados y procesos mentales que tenemos la pretensión más plausible de que son computacionales por naturaleza, y los argumentos de que hay semejantes estados y procesos computacionales. Podemos entonces ver hasta dónde estos argumentos se aplican a todos los demás estados mentales. En un sentido, esto es simplemente el método filosófico correcto: debe siempre apreciarse una teoría en su versión más plausible. A nadie le interesa una crítica de una caricatura. Sin embargo, en este caso, el argumento para la naturaleza computacional de los estados mentales representacionales tiene interés independiente, sin importar lo que uno piense del punto de vista que afirma que *todos* los estados mentales son computacionales. Así, por el momento, dejaremos de lado la cuestión de si puede haber una teoría computacional del dolor.[4]

Ahora se requiere una breve digresión acerca de un asunto de historia filosófica. Aquellos lectores que estén familiarizados con la filosofía funcionalista de la mente de los años sesenta acaso hallen confuso esto. Pues ¿no fue la meta de esta teoría mostrar que los estados mentales podían ser clasificados por sus tablas de máquina de Turing, y no fue el *dolor* el ejemplo paradigmático empleado (entrada = daño de tejidos; salida = quejas/comportamiento lamentoso)? Estos filósofos pueden haber estado equivocados en lo de si la mente será una máquina de Turing, pero ¿de fijo no podrían haber sido tan *confusos* como yo diciendo que lo eran?

[4] Para brillantes discusiones de estos asuntos, sin embargo, véase Dennett, "Towards a Cognitive Theory of Consciousness" y "Why You Can't Make a Computer that Feels Pain", *Brainstorms*. Véase también John Haugeland, "The Nature and Plausibility of Cognitivism", en Haugeland J. (ed.), *Mind Design*.

Sin embargo, no estoy diciendo que fueran confusos. Tal como veo las cosas, la idea de que los estados mentales tienen tablas de máquina fue una reacción contra la teoría materialista que ligaba los estados mentales demasiado estrechamente con clases particulares de estados cerebrales ("dolor = disparo de fibras C", etc.). Así, una tabla de la máquina de Turing era una manera de dar una especificación relativamente *abstracta* de tipos de estado mental que no se ligaban a estructuras neurales particulares. Muchas clases de entidades físicas podían estar en el mismo estado mental; la intención de la analogía de la tabla de la máquina era mostrar cómo podría ser esto.[5] Según vimos en el capítulo III —"Ejemplificación y computación de una función" p. 169)— necesitamos distinguir entre la idea de que una transición entre estados puede ser descrita por una tabla de la máquina de Turing y la idea de que una transición entre estados *implica* realmente computación. Para distinguir entre estas ideas, necesitamos recurrir a la idea de representación: las computadoras procesan representaciones, en tanto que (por ejemplo) el sistema solar no lo hace. Se sigue que debemos diferenciar entre la teoría funcionalista de la mente, que dice que ésta es definida por su estructura causal, y la teoría computacional de la mente, que dice que esta estructura causal es computacional, o sea una serie disciplinada de transiciones entre representaciones. Esta distinción es fácil de ver, por supuesto, ya que no todas las estructuras causales son computaciones.

[5] Tal fue de seguro la meta de Hilary Putnam en "The Nature of Mental States" y "Philosophy and Our Mental Life", *Mind, Language and Reality* (Cambridge, Cambridge University Press, 1975) y otros artículos que proponen la teoría funcionalista. Creo que no fue su intención adelantar una teoría computacional de la mente.

Regresemos a la cuestión del alcance de la teoría computacional de la mente. Dije que es discutible si los dolores son puramente representacionales, y por lo tanto igualmente discutible si puede haber una teoría puramente computacional de los dolores. Así, ¿qué estados y procesos mentales podrían ser ejemplos más plausibles de estados y procesos computacionales? La respuesta es ahora evidente: aquellos estados que son esencialmente representacionales, sin más, en la naturaleza. En el capítulo I sostuve que las creencias y los deseos (las actitudes proposicionales) son así. Su esencia es representar el mundo, y aunque a menudo aparezcan en la conciencia no es esencial para ellos que sean conscientes. No hay razón para creer, al menos desde la perspectiva de la psicología del sentido común, que tengan ninguna propiedad diferente de las representacionales. La naturaleza de una creencia se agota por lo que representa acerca de cómo es el mundo, y las propiedades que tiene como consecuencia de ello. Así, las creencias tienen el aire de ser los mejores candidatos, si es que hay alguno, de ser estados computacionales de la mente.

La pretensión principal de lo que es a veces llamado *teoría computacional de la cognición* es que estos estados representacionales se relacionan uno con otro de una manera computacional. Esto es, están relacionados entre sí de modo algo parecido a como los estados representacionales de una computadora lo están: son procesados por medio de reglas algorítmicas (y quizás heurísticas). El término "cognición" indica que lo que le importa a la teoría son los procesos *cognitivos,* tales como el razonamiento y la inferencia, procesos que asocian estados cognitivos como la creencia. La teoría computacional del conocimiento es, por lo tanto, la base filosófica de la ciencia cognitiva (véase el capítulo III,

"¿Computadoras pensantes?", p. 180, acerca de la idea de la ciencia cognitiva).

Otro término para esta hipótesis es la de *teoría representacional de la mente*. Este término es menos apropiado que el de "teoría computacional de la cognición", por dos razones al menos. La primera es que aspira a describir la mente completa, lo cual, como hemos visto, es problemático. La segunda es que la idea de que los estados de la mente representan el mundo es, en sí misma, una idea muy inocua: casi todas las teorías de la mente pueden aceptar que ésta "representa' el mundo en algún sentido. Lo que no todas las teorías aceptarán es que la mente *contenga representaciones*. Jean-Paul Sartre, por ejemplo, dijo que las "representaciones... son ídolos inventados por los psicólogos".[6] Una teoría de la mente podría aceptar la tautología sencilla de que la mente "representa el mundo" sin sostener que la mente "contiene representaciones".

¿Qué significa decir que la mente "contiene" representaciones? A grandes rasgos significa que: en la mente de los pensadores hay distintos estados que tienen el lugar de cosas del mundo. Por ejemplo, ahora estoy pensando sobre mi inminente viaje a Budapest. De acuerdo con la teoría computacional de la mente, hay en mí —en mi cabeza— un estado que representa mi visita a Budapest. (Análogamente: hay, en el disco duro de mi computadora, un archivo —un estado complejo de la computadora— que representa este capítulo.)

Esto podría recordarle a usted la teoría controvertida de las ideas como "imágenes en la cabeza' que descartamos en

[6] Citado por Gregory McCulloch, *Using Sartre* (Londres, Routledge, 1994), p. 7.

el capítulo 1. La teoría computacional no se entrega a imágenes en la cabeza: hay muchas clases de representación distintas de las imágenes. Esto alza el problema: ¿qué dice la teoría computacional de la cognición acerca de cómo son estas representaciones mentales?

Hay múltiples respuestas a esta cuestión; el resto del capítulo esbozará las más influyentes. Comenzaré con el punto de vista que ha provocado más debate durante los últimos 20 años: la idea de que las representaciones mentales son, muy literalmente, *palabras* y *oraciones* en un lenguaje: el "lenguaje del pensamiento".

El lenguaje del pensamiento

A menudo expresamos nuestro pensamiento con palabras, y a menudo también pensamos en palabras, silenciosamente, para nosotros mismos. Aunque es implausible decir que todo pensamiento es imposible sin lenguaje, es indiscutible que el lenguaje que hablamos nos da la capacidad de formular pensamientos extremadamente complejos. (Es difícil imaginar cómo alguien podría pensar en, digamos, el *posmodernismo* sin estar en condiciones de hablar un lenguaje.) Esto no es lo que las personas piensan cuando dicen que pensamos en un lenguaje de pensamiento.

Lo que quieren decir es que cuando se tiene un pensamiento —una creencia, supongamos, en que *el precio de la propiedad está volviendo a subir*— hay (literalmente) escrita en la cabeza de usted una oración que significa lo mismo que la oración en español: "El precio de la propiedad está volviendo a subir". Esta oración en la cabeza de usted no es en sí misma (normalmente) considerada una oración en es-

pañol, o una oración en cualquier lenguaje público. Antes bien, es una oración de un lenguaje mental postulado: el lenguaje del pensamiento, a veces abreviado como LP, y a menudo llamado mentalés. La idea es que es una hipótesis científica o empírica plausible suponer que hay semejante lenguaje mental, y que la ciencia cognitiva debe funcionar suponiendo esto y tratar de descubrir el mentalés.

A quienes encuentran esta teoría por primera vez bien puede parecerles sumamente rara: ¿por qué debiera alguien querer creerla? Antes de responder a esta cuestión hay otra, anterior: ¿qué significa exactamente la hipótesis del mentalés?

Podemos dividir esta cuestión en otras dos: ¿qué quiere decir que un símbolo, cualquier símbolo, está escrito en la *cabeza* de alguien? Y ¿qué significa decir que una *oración* está escrita en la cabeza de alguien?

Podemos dirigir estas preguntas volviendo a la naturaleza de los símbolos en general. Tal vez, cuando pensamos primeramente en las palabras y otros símbolos (por ejemplo imágenes), pensamos al respecto como visualmente identificables: vemos palabras en la página, signos de tráfico y así por el estilo. Por supuesto, en el caso de las palabras, es igualmente común escuchar oraciones cuando oímos hablar a otras personas. Y muchos de nosotros estamos familiarizados con otros modos de almacenar y transmitir oraciones: a través de ondas de radio, pautas en cinta magnética, y en los discos magnéticos y circuitos electrónicos de una computadora.

Hay muchas maneras, pues, como los símbolos pueden ser almacenados y transmitidos. De hecho, hay muchas maneras como los *mismísimos* símbolos pueden ser almacenados, transmitidos o (diré) *realizados*. La oración en español: "El hombre que hizo quebrar la banca en Monte Carlo

murió en la miseria" puede escribirse, decirse, o almacenarse en cinta magnética o en un disco de computadora. En algún sentido sigue siendo la misma oración. Podemos hacer las cosas absolutamente precisas aquí si distinguimos entre *tipos* y *prendas* de palabras y oraciones. En la lista de palabras "Est! Est! Est!", el mismo tipo de palabra aparece tres veces: son, como dicen los filósofos y los lingüistas, tres *prendas* del mismo *tipo*. En nuestro ejemplo de una oración, el mismo *tipo* de oración tiene muchas *prendas* físicas, y las prendas pueden ser realizadas de maneras muy diferentes.

Llamaré a estos diferentes modos de almacenar distintas prendas del mismo tipo de oración *medios* diferentes en que son realizadas. Las palabras españolas escritas son un medio, las palabras españolas dichas son otro y las palabras en cinta magnética otro más. La misma oración puede ser realizada según muy diferentes medios. Sin embargo, para la discusión que sigue necesitamos otra distinción. Necesitamos distinguir no sólo entre los medios diferentes de almacenar los mismos *símbolos,* sino también las diferentes maneras como el mismo *mensaje* o el mismo *contenido* puede ser almacenado.

Considérese una señal de tránsito con un dibujo esquemático en un triángulo rojo, de dos niños agarrados de la mano. El mensaje que esta señal porta es: "¡Cuidado! ¡Cruce de niños!" Compárese esto con una señal verbal que dice, en español: "¡Cuidado! ¡Cruce de niños!" Estas dos señales expresan el mismo mensaje, pero de modos muy diferentes. Esta diferencia no es captada por la idea de un medio, ya que ese término fue destinado a expresar la diferencia entre los distintos modos como la misma *oración* española (por ejemplo) puede ser realizada por diferentes materiales físicos. En el caso de la señal de tránsito, no tenemos frase en absoluto.

Llamaré a esta clase de diferencia por el modo como puede ser almacenado un mensaje, diferencia en el *vehículo* de la representación. El mismo mensaje puede ser almacenado en diferentes vehículos, y estos vehículos pueden ser "realizados" mediante diferentes medios. La distinción más obvia entre vehículos de representación es que puede hacerse entre oraciones e imágenes, aunque hay otros tipos. Por ejemplo, algunos filósofos han pretendido que hay una clase de representación natural, que llaman "indicación". Éste es el tipo de representación en el cual los anillos de un árbol, por ejemplo, representan o indican la edad del árbol.[7] Claramente, no hay representación lingüística ni pictórica: está implicada otra clase de vehículo. (Véase en el capítulo v, "Teorías causales de la representación mental", p. 277.) Más adelante encontraremos otro tipo de vehículo en el apartado "Computadoras 'sesudas'" (p. 254).

Ahora tenemos la distinción entre el medio y el vehículo de representación, y podemos empezar a formular la hipótesis del mentalés. La hipótesis dice que las oraciones son escritas en la cabeza. Esto significa que cuando alguien cree, digamos, que *los precios aumentan,* el vehículo de este pensamiento es una oración. Y el medio en el cual es realizada esta oración es la estructura neural del cerebro. La idea tosca que está tras esta segunda afirmación es: piénsese el cerebro como una computadora, con sus neuronas y sinapsis constituyendo sus "procesadores primitivos". Para hacer vívido esto, piénsese en neuronas, las células constituyentes del cerebro, como, más bien, las puertas lógicas del capítulo III: emiten una señal de salida ("disparo") cuando

[7] El ejemplo procede de Dennis Stampe, "Toward a Causal Theory of Linguistic Representation", *Midwest Studies in Philosophy.*

capítulo III: emiten una señal de salida ("disparo") cuando sus entradas son de la clase apropiada. Entonces podemos suponer que las combinaciones de estos procesadores primitivos (de algún modo) constituyen la oración del mentalés cuya traducción al español es "Suben los precios".

Hasta aquí, la primera cuestión. La segunda era: supóngase que hay representaciones en la cabeza; ¿qué significa pensar estas representaciones como *oraciones*? Esto es, ¿por qué habría un *lenguaje* del pensamiento, mejor que algún otro sistema de representación (por ejemplo imágenes en la cabeza)?

Sintaxis y semántica

Decir que un sistema de representación es un lenguaje es decir que sus elementos (oraciones y palabras) tienen una estructura sintáctica y semántica. Encontramos los términos "sintaxis" y "semántica' en nuestra discusión del argumento del cuarto chino de Searle, y ahora es tiempo de decir más al respecto. (Se notará que lo que sigue es sólo un bosquejo y, como tantos términos en esta área, "sintaxis" y "semántica" son términos harto controvertidos, usados de maneras sutilmente distintas por diferentes autores. Aquí sólo aspiro a capturar los bosquejos no controvertidos.)

Esencialmente, los rasgos sintácticos de las palabras y las oraciones en un lenguaje son las que conciernen a su *forma* más bien que a su *significado.* Una teoría de la sintaxis para un lenguaje nos enseñará cuáles son las maneras fundamentales de expresión en el lenguaje, y qué combinaciones de expresiones son legítimas en el lenguaje, esto es, qué combinaciones de expresiones son gramaticales o "bien formadas". Por ejemplo, es un rasgo sintáctico de la expresión com-

sólo legítimamente darse en oraciones en determinadas posiciones: "el papa lleva una vida dichosa" es gramatical, pero "la vida lleva una dichosa al papa" no lo es. La tarea de una teoría sintáctica es decir cuáles son las categorías sintácticas fundamentales, y qué reglas gobiernan la producción de expresiones gramaticalmente complejas a partir de combinaciones de las expresiones sencillas.

¿En qué sentido pueden los símbolos que hay en la cabeza tener sintaxis? Pues bien, algunos símbolos se clasificarán como símbolos sencillos y operarán reglas sobre estos símbolos para producir símbolos complejos. La tarea a que se enfrenta el teórico mentalista es encontrar estos sencillos símbolos, y las reglas que operan sobre ellos. Esta idea no es claramente absurda —una vez que de algún modo hemos aceptado la idea de símbolos en la cabeza—; dejemos la sintaxis por el momento y pasemos a la semántica.

Los rasgos semánticos de las palabras y las oraciones son aquellos que se relacionan con su significado. Aunque es un rasgo sintáctico de la palabra "pusilánime" el ser un adjetivo, y así sólo poder aparecer en ciertos lugares de las oraciones, es un rasgo semántico de "pusilánime" que signifique... *pusilánime,* es decir, fofo, débil, incauto. Una teoría del significado para un lenguaje se llama "teoría semántica", y "semántica" es esa parte de la lingüística que se ocupa del estudio sistemático del significado.

En efecto, es porque los símbolos tienen rasgos semánticos por lo que, ni más ni menos, son símbolos. Está en la naturaleza misma de los símbolos no poder ocupar el lugar de cosas o representarlas; *ocupar el lugar* y *representar* son relaciones semánticas. La semántica no trata nada más del modo como las palabras se relacionan con el mundo, sino que también trata del modo como las palabras se relacionan

una con otra. Una oración como "Cleopatra ama a Antonio" tiene tres constituyentes, "Cleopatra", "ama a" y "Antonio", todos los cuales pueden presentarse en otras oraciones, digamos "Cleopatra se suicidó", "Desdémona ama a Casio" y "Antonio abandonó su deber". Desconociendo, por conveniencia, complejidades introducidas por la metáfora, los dichos, la ambigüedad y el hecho de que más de una persona pueda compartir un nombre —omisiones no insignificantes, pero que podemos hacer en esta etapa—, se reconoce generalmente que cuando estas palabras se presentan en estas otras oraciones tienen el mismo sentido que cuando se presentaron en la oración original.

Este hecho, aunque podría parecer trivial y evidente al principio, es en realidad muy importante. El significado de las oraciones es determinado por los significados de sus partes y su modo de combinación, o sea su sintaxis. Así, el significado de la frase "Cleopatra ama a Antonio" está enteramente determinado por los significados de los constituyentes "Cleopatra", "ama a" y "Antonio", el orden en el cual se presentan y el papel sintáctico de estas palabras (el hecho de que la primera y la tercera palabra sean nombres y la segunda un verbo). Esto significa que, cuando entendemos el significado de una palabra, podemos comprender su contribución a *cualquier otra* oración en la cual se presente. Y mucha gente cree que es este hecho el que explica cómo es que conseguimos comprender oraciones que no hemos encontrado previamente. Por ejemplo, dudo que usted haya encontrado alguna vez esta oración antes: "Hay catorce cuartos en el puente".

Por rara que pueda parecer, ciertamente se sabe lo que significa porque se sabe lo que significan las palabras constituyentes y cuál es su lugar sintáctico en la oración. (Por

ejemplo, puede usted contestar la siguiente pregunta acerca de la oración: "¿Qué hay en el puente?" "¿Dónde están los cuartos?" "¿Cuántos cuartos hay ahí?") Este hecho acerca de los lenguajes se llama "composicionalidad semántica". De acuerdo con muchos filósofos y lingüistas, es este rasgo de los lenguajes lo que nos permite aprenderlos por completo.[8]

Para captar este punto, puede ayudar el contraste de un lenguaje con un sistema representacional que no es composicional de esta manera: el sistema de banderines con colores y dibujos de los barcos. Supóngase que hay un banderín que advierte: "Hay fiebre amarilla a bordo"; otro que pide: "Que venga el inspector aduanal". Sin embargo, dados sólo estos recursos, no puede combinarse su conocimiento de los significados de estos símbolos para producir otro símbolo, por ejemplo, uno que dice: "Que vengan los inspectores de fiebre amarilla". Lo que es más, cuando se encuentra un banderín que nunca ha sido visto antes, no hay cantidad de conocimiento de los demás banderines que pueda ayudar a comprenderlo. Hay que aprender el significado de cada banderín individualmente. La diferencia con un lenguaje es que, aun cuando pueda aprenderse el significado de palabras individuales una por una, este entendimiento proporciona la capacidad de formar y entender *cualquier número* de nuevas oraciones. De hecho, el número de oraciones en un lenguaje es potencialmente infinito. No obstante, por las razones dadas, es claro que si un lenguaje ha de ser aprendible, el número de elementos significantes básicos (palabras) ha de ser finito. De otra manera, encontrar una nueva oración sería siempre como encontrar

[8] Véase Donald Davidson, "Theories of Meaning and Learnable Languages", *Inquiries into Truth and Interpretation* (Oxford, Oxford University Press, 1984).

un nuevo banderín sobre el barco, lo cual evidentemente no es el caso.

¿En qué sentido pueden los símbolos en la mente tener rasgos semánticos? La respuesta debe ahora resultar bien llana. Pueden tener rasgos semánticos porque representan las cosas del mundo, o están en lugar de éstas. Si hay oraciones en la mente, estas oraciones tendrán partes semánticamente significativas (palabras) y dichas partes se referirán a cosas del mundo o se aplicarán así. Lo que es más, los significados de las oraciones serán determinados por los significados de sus partes más su modo de combinación. A manera de sencilla exposición, supongamos de modo chauvinista que el mentalés es español. Entonces, para decir que creo que los precios están subiendo, pienso que hay una oración escrita en mi mente, "Los precios están subiendo", cuyo significado es determinado por los significados de las palabras constituyentes "precios", "están" y "subiendo" y por su modo de combinación.

El argumento del lenguaje del pensamiento

Así, ahora que tenemos una captación elemental de las ideas de sintaxis y semántica, podemos decir con precisión qué es la hipótesis mentalesa. Ésta se da cuando un pensante tiene una creencia o deseo con el contenido *P;* hay una oración (o sea una representación con estructura semántica y sintáctica) que significa que *P* está escrita en su mente. Los vehículos de representación son lingüísticos, en tanto que el medio de representación es la estructura neural del cerebro.

El lector atento habrá advertido que algo falta en esta

descripción. Pues, según vimos en el capítulo I, diferentes pensamientos pueden tener el mismo contenido: puedo creer que los precios bajarán, puedo desear que los precios bajen, puedo esperar que los precios bajarán, y así sucesivamente. La hipótesis del mentalés dice que estos estados implican todos tener una oración con el significado *los precios bajarán* escrita en la mente de los que piensan. De seguro, creer que los precios bajarán es una clase muy diferente de estado mental de esperar que los precios bajen; ¿cómo explica la hipótesis mentalesa esa diferencia?

La respuesta breve es que no lo hace. Una respuesta más larga es que la hipótesis mentalesa no tiene por objeto explicar la diferencia entre creencia y deseo, o entre creencia y esperanza. Lo que quiere explicar no es la diferencia entre *creer* algo y *desearlo,* sino entre creer (o desear) una cosa y algo más. Es la terminología de las actitudes y los contenidos, introducida en el capítulo I, cuyo propósito es explicar qué es tener una actitud con cierto contenido, no qué es tener esta actitud antes que la otra. Por supuesto, los creyentes en el mentalés piensan que habrá una teoría científica de lo que es tener una creencia más bien que un deseo, pero esta teoría será independiente de la hipótesis mentalesa misma.

Podemos ahora volver a nuestra pregunta inicial: ¿por qué debiéramos creer que el vehículo de la representación mental es un lenguaje? El inventor de la hipótesis mentalesa, Jerry Fodor, ha adelantado dos argumentos influyentes para contestar a esta pregunta, y los esbozaré brevemente. El segundo requerirá un poco más de exposición que el primero.

La primera argumentación descansa en una comparación entre la "composicionalidad" de la semántica, discuti-

da en la oración anterior, y un fenómeno en apariencia similar en el pensamiento mismo. Recuérdese que si alguien comprende la frase en español: "Cleopatra ama a Antonio", está *ipso facto* en la posición de comprender otras oraciones que contienen estas palabras, con tal que comprenda las demás palabras de las oraciones. O que, cuando menos, pueda comprender la oración: "Antonio ama a Cleopatra". Similarmente, Fodor sostiene que si alguien es capaz de pensar que *Cleopatra ama a Antonio,* entonces también es capaz de pensar que *Antonio ama a Cleopatra.* Cualquier cosa que se requiera para pensar el primer pensamiento, nada más es preciso para conseguir pensar el segundo. Por supuesto, puede que no *crea* que Antonio ama a Cleopatra meramente porque cree que Cleopatra ama a Antonio; pero puede cuando menos considerar la idea de que Antonio ama a Cleopatra.

Fodor sostiene que la mejor explicación del fenómeno es que el pensamiento mismo tiene una estructura composicional, y que tener una estructura composicional equivale a tener un lenguaje de pensamiento. Nótese que no está diciendo que el fenómeno *acarrea lógicamente* que el pensamiento tenga una sintaxis y una semántica composicionales. Es *posible* que el pensamiento pudiera exhibir el fenómeno sin ser un lenguaje de pensamiento, pero Fodor y sus seguidores creen que la hipótesis del lenguaje de pensamiento es la mejor explicación científica de este aspecto del pensamiento.

La segunda argumentación de Fodor se apoya en ciertos supuestos acerca de los procesos o cursos de pensamiento mentales. Este argumento ayudará a ver en qué sentido exactamente la hipótesis mentalesa es una teoría *computacional* de la cognición o pensamiento. Para captar esta argu-

mentación, considérese la diferencia entre los siguientes dos procesos de pensamiento:

1. Supóngase que quiero ir a Ljubljana, y puedo llegar allí por tren o por autobús. El autobús es más barato, pero el tren será más agradable, y parte a una hora más conveniente. Sin embargo, el tren tarda más, porque la ruta del autobús es más directa. Pero el tren implica una parada en Viena, que me gustaría visitar. Sopeso los factores de cada lado y decido sacrificar tiempo y dinero a causa del medio más sano del tren y las atracciones de una visita a Viena.
2. Supóngase que quiero ir a Ljubljana, y puedo llegar allí por tren o por autobús. Despierto por la mañana y miro por la ventana. Veo dos palomas en la azotea de enfrente. Las palomas siempre me hacen pensar en Venecia, que una vez visité en tren. De modo que decido ir allá por tren.

Mi conclusión es la misma en los dos casos, pero los métodos son muy diferentes. En el primer caso, uso la información que tengo, ponderando la deseabilidad relativa de los resultados diferentes. En una palabra, *razono:* tomo una decisión razonada a partir de la información disponible. En el segundo caso, simplemente asocio ideas. No hay conexión racional particular entre palomas, Venecia y trenes; sencillamente se me ocurrieron las ideas. Fodor arguye que, a fin de que funcionen las explicaciones psicológicas de sentido común (del tipo que examinamos en el capítulo II) mucho más de nuestro pensamiento debe ser como en el primer caso que como en el segundo. En el capítulo II defendí la idea de que, si hemos de encontrar sentido en el com-

portamiento de la gente, debemos verla como en busca de proyectos, razonando, sacando conclusiones razonables a partir de lo que cree y quiere. Si todo pensamiento fuese del estilo de "libre asociación", sería muy difícil hacer esto: desde la salida, sería muy difícil ver la conexión entre los pensamientos de la gente y su comportamiento. El hecho de que no sea muy difícil sugiere fuertemente que la mayoría del pensamiento no es la asociación libre.

Fodor no está negando que la asociación libre proceda. Lo que aspira a recalcar es la naturaleza sistemática, racional, de muchos procesos mentales.[9] Un modo como puede ser sistemático el pensar está en el anterior ejemplo 1, cuando razoné acerca de qué hacer. Otra manera más es al razonar acerca de qué pensar. Por tomar un ejemplo sencillo: creo que el obispo Berkeley, filósofo irlandés, pensó que la materia es una noción contradictoria. También creo que nada contradictorio puede existir, y creo que el obispo Berkeley creía esto también. Concluyo que el obispo Berkeley creía que la materia no existe y que si la materia existe, se equivoca. Porque creo que la materia existe, concluyo que el obispo Berkeley estaba equivocado. Éste es un ejemplo de razonamiento acerca de qué pensar.

Inferencias como ésta son el tema de la lógica. La lógica estudia esos rasgos de inferencia que no dependen del contenido específico de las inferencias; esto es, la lógica estudia la *forma* de las inferencias. Por ejemplo, desde el punto de vista de la lógica, las siguientes simples inferencias pueden verse como de la misma forma o estructura:

[9] Fodor a veces utiliza una bonita comparación entre el pensamiento y la clase de deducciones que realiza Sherlock Holmes para resolver sus casos. Véase "Fodor's Guide to Mental Representation", en *A Theory of Content and Other Essays* (Cambridge, MIT Press, 1990), p. 21.

Si quiero visitar Ljubljana, iré en tren.
Visitaré Ljubljana.
Por lo tanto: iré en tren.

y

Si la materia existe, el obispo Berkeley estaba equivocado.
La materia existe.
Por lo tanto: el obispo Berkeley estaba equivocado.

Lo que los lógicos hacen es representar la forma de inferencias como éstas, sin importar qué podría significar cualquier caso particular, esto es, sin importar su contenido específico. Por ejemplo: usando las letras P y Q para representar las oraciones constituyentes anteriores, y la flecha "→" para representar "si... entonces...", podemos representar la forma de las inferencias anteriores como sigue:

$P \rightarrow Q$
P
Por lo tanto: Q

Los lógicos llaman a esta forma particular de inferencia *modus ponens*. La argumentación con esta forma se mantiene firme precisamente porque tiene esta forma. ¿Qué significa "mantiene firme"? No que sus premisas y conclusiones siempre sean verdad: la lógica sola no puede dar verdades acerca de la naturaleza del mundo. Lo que pasa es que el sentido en el cual se sostiene firme es que *preserva la verdad:* si se comienza con verdades como premisas, se preservará la verdad en la conclusión. Una forma de argumentación que preserva la verdad es lo que los lógicos llaman una argu-

mentación *válida:* si las premisas son ciertas, entonces las conclusiones deben ser ciertas.

Los defensores de la hipótesis del mentalés piensan que muchas transiciones entre estados mentales —muchos procesos mentales, o cursos de pensamiento, o inferencias— son como ésta: son *preservadoras de la verdad a causa de su forma.* Cuando la gente razona lógicamente a partir de premisas a conclusiones, las conclusiones alcanzadas serán ciertas si las premisas con que se iniciaron son ciertas *y* usan un método o regla preservadora de la verdad. Así, si esto es verdad, los ítems que realizan los procesos mentales tienen más bien *forma.* Y esto, por supuesto, es lo que la hipótesis del mentalés sostiene: las oraciones en nuestra mente tienen una forma sintáctica, y es por tener esta forma sintáctica como pueden interactuar en procesos mentales sistemáticos.

Para entender esta idea necesitamos comprender el nexo entre tres conceptos: semántica, sintaxis/forma y causalidad. El nexo puede plantearse usando la comparación con computadoras. Los símbolos de una computadora tienen semántica y propiedades "formales", pero los procesadores de la computadora son sensibles únicamente a las propiedades formales. ¿Cómo? Recuérdese el sencillo ejemplo del "portal-y" (capítulo III: "¿Computadoras pensantes?", p. 180). Las propiedades *causales* del portal-y son aquellas propiedades a las cuales es causalmente sensible la máquina: la máquina producirá una corriente eléctrica cuando, y sólo cuando, tome corrientes eléctricas de ambas entradas. Este proceso causal codifica la estructura formal de "y": una oración si "P entonces Q" será verdad cuando y sólo cuando P sea cierta y Q sea cierta. Y esta estructura formal refleja el significado de "y": cualquier palabra con esa estructura for-

mal tendrá el significado que tiene "y". Así, las propiedades *causales* del dispositivo reflejan sus propiedades *formales,* y éstas a la vez reflejan las propiedades *semánticas* de "y". Esto es lo que permite a la computadora realizar computaciones ejecutando operaciones puramente causales.

Lo mismo pasa con el lenguaje del pensamiento. Cuando alguien razona, de su creencia de que P → Q (es decir, *si P entonces Q*) y su creencia de que existe P, tiene la conclusión Q; hay dentro un proceso causal, el cual refleja la relación puramente formal de *modus ponens*. Así, los elementos del proceso causal deben tener componentes que reflejen las partes componentes de la inferencia, es decir, la *forma debe tener una base causal*.

Todo lo que necesitamos hacer ahora es el enlace entre sintaxis y semántica. El punto esencial, aquí, es mucho más complicado, pero puede ser ilustrado con la simple forma del argumento lógico discutido antes. *Modus ponens* es válido en virtud de su forma: pero este rasgo puramente formal de la argumentación garantiza algo acerca de sus propiedades semánticas. Lo que garantiza es que la propiedad semántica de la *verdad* es preservada: si comienza usted su razonamiento con verdades, y sólo usa un argumento de la forma *modus ponens,* entonces tendrá usted la garantía de obtener sólo verdades al final de su razonamiento. Así, el razonamiento con su regla puramente formal asegurará que sus propiedades semánticas serán "reflejadas" por las propiedades formales. La sintaxis no crea semántica, pero la lleva a remolque. Como ha dicho John Haugeland: "Si se cuida usted de la sintaxis, *la semántica se cuidará a sí misma*".[10]

[10] Haugeland, "Semantic Engines: An Introduction to Mind Design", en Haugeland (ed.), *Mind Design,* p. 23.

Ahora tenemos el enlace que queríamos entre tres cosas: los rasgos semánticos de las representaciones mentales, sus rasgos sintácticos y sus rasgos causales. La pretensión de Fodor es que, pensando en procesos mentales como en computaciones, podemos ligar estas tres clases de rasgos:

> Las computadoras nos muestran cómo conectar la semántica con las propiedades causales *para símbolos* [...] Se conectan las propiedades causales de un símbolo con sus propiedades semánticas a través de la sintaxis [...] Podemos pensar en su estructura sintáctica como un rasgo abstracto de su... *forma*. En vista de que, para cualquier fin y propósito, la sintaxis se reduce a forma, y porque la forma de un símbolo es un determinante potencial de su papel causal, es sumamente fácil [...] imaginar prendas simbólicas interactuando causalmente *en virtud de sus* estructuras sintácticas. La sintaxis de un símbolo podría determinar sus causas y efectos [...] de manera muy análoga a como la geometría de una llave determina qué cerraduras abrirá.[11]

Lo que la hipótesis nos da, entonces, es una manera de conectar las propiedades representacionales del pensamiento (su contenido) con su naturaleza causal. El enlace es suministrado por la idea de una sintaxis mental que se realiza en la estructura causal del cerebro, y cómo las propiedades formales de los símbolos de computadora se realizan en la estructura causal de la computadora. Las propiedades sintácticas o formales de las representaciones en una computadora son interpretables como cálculos, o inferencias, o trozos de razonamiento —son semánticamente interpretables—, y

[11] "Fodor's Guide to Mental Representation", p. 22.

esto nos proporciona un vínculo entre propiedades causales y propiedades semánticas. De forma parecida, se espera, sucederá con el enlace entre el contenido y la causación del pensamiento.

La hipótesis del mentalés es una hipótesis computacional porque invoca representaciones que son manipuladas o procesadas de acuerdo con reglas formales. No dice qué son estas reglas: ésta es una materia que incumbe descubrir a la ciencia cognitiva. Usé el ejemplo de una regla lógica sencilla, para simplificar la exposición, pero no es parte de la hipótesis del mentalés que sólo las reglas descubiertas serán las leyes de la lógica.

¿Qué podrían ser estas otras reglas? Los defensores de la hipótesis a menudo recurren a teorías computacionales de la visión como ilustración de la clase de explicación que tienen en mente. La teoría computacional de la visión ve la tarea, para la psicología de la visión, como la de explicar cómo nuestro sistema visual produce una representación del medio visual 3D a partir de la distribución de la luz en la retina. La teoría sostiene que el sistema visual hace esto creando una representación de la pauta de luz sobre la retina y haciendo inferencias computacionales en varias etapas, para llegar finalmente a la representación 3D. A fin de lograrlo, el sistema debe tener construido en sí mismo el "conocimiento" de ciertas reglas o principios, para hacer la inferencia de una etapa a la siguiente. (En este breve libro no puedo dar una descripción detallada de esta clase de teoría, pero hay muchas buenas introducciones disponibles: véase el apartado "Lecturas adicionales", p. 266.)

Por supuesto, no podemos afirmar estos principios nosotros mismos sin conocimiento de la teoría. Los principios no son accesibles a la introspección. De acuerdo con la teo-

ría, "conocemos" estos principios en el sentido de que son representados de alguna manera en nuestra mente, podamos o no llegar a ellos por medio de introspección. Esta idea se origina en la teoría lingüística de Noam Chomsky,[12] quien ha sostenido durante muchos años que el mejor modo de explicar nuestra realización lingüística es postular que tenemos conocimiento de las reglas gramaticales fundamentales de nuestro lenguaje. Sin embargo, el hecho de que tengamos este conocimiento *no* implica que podamos llevarlo a nuestra mente consciente. La hipótesis del mentalés propone que así es como son las cosas con las reglas que gobiernan los procesos mentales. Según mencioné en el capítulo II, los defensores de esta clase de conocimiento lo llaman a veces "conocimiento tácito".[13]

Adviértase, finalmente, que la hipótesis del mentalés no está forzosamente ligada a la idea de que toda vida mental implique procesar representaciones lingüísticas. Es coherente con la hipótesis sostener, por ejemplo, que las sensaciones no son del todo representacionales. Es también coherente con la hipótesis sostener que podría haber procesos que "manipularan" representaciones *no lingüísticas.* Un área particularmente activa de investigación en la ciencia cognitiva, por ejemplo, es el estudio de la imaginería mental. Si le pregunto a usted si "tienen labios las ranas", hay buenas probabilidades de que considerará esta cuestión formando una imagen mental y "revisándola" mentalmente. De acuerdo con algunos científicos cognitivos, hay un sentido en el cual de hecho hay representaciones en la mente

[12] Para una introducción bien accesible a las ideas filosóficas de Chomsky, véase sus *Rules and Representations* (Oxford, Blackwell, 1980).

[13] Para una discusión crítica de esta noción, véase Stephen P. Stich, "What every Speaker Knows", *Philosophical Review,* 80 (1971).

que tienen una estructura pictórica, que pueden ser "giradas", "barridas" e "inspeccionadas". ¡Tal vez hay imágenes en la mente, después de todo! Así, un científico cognitivo *podría* sostener coherentemente que hay tales representaciones pictóricas sin dejar de sostener que los vehículos del *razonamiento* son lingüísticos. (Para sugerencias de cómo ir más allá en este fascinante tema, véase el apartado "Lecturas adicionales", p. 266.)

La modularidad de la mente

La argumentación en pro de la hipótesis del mentalés, tal como la he presentado, es un ejemplo de lo que se llama una inferencia a la mejor explicación. Se alude a determinado hecho innegable o evidente, y entonces se muestra que este hecho obvio tiene sentido, dada la verdad de nuestra hipótesis. En vista de que no hay mejor hipótesis rival, esto nos da una razón para creer en nuestra hipótesis. Ésta es la forma general de una inferencia a la mejor explicación, y es un método medular y variable de explicación que se usa en la ciencia.[14] En nuestro caso, el hecho evidente es la naturaleza sistemática de las propiedades semánticas del pensamiento: el hecho general que es revelado antes por fenómenos descritos en el ejemplo de Antonio y Cleopatra. La argumentación de Fodor se funda en el hecho de que los procesos mentales explotan esto sistemáticamente en las transiciones racionales de pensamiento en pensamiento. Los cursos de pensamiento tienen una estructura racional, y

[14] Véase Peter Lipton, *Inference to the Best Explanation* (Londres, Routledge, 1991).

tienen consecuencias causales que dependen de esta estructura. La mejor explicación de esto, sostiene Fodor, es que hay un medio interno de representación —el mentalés, el lenguaje del pensamiento (LP)— con las propiedades semánticas y sintácticas descritas antes.

En muchas zonas de la mente, aunque hay buena razón para suponer que hay representación mental, no parece haber nada como un proceso racional en curso. ¿Qué diría un defensor del mentalés acerca de esto? Tómese el caso de la percepción visual, por ejemplo. Como vimos en el apartado anterior, los psicólogos que estudian la visión tienden a tratar el sistema visual como procesamiento de representaciones, a partir de la representación de la distribución de luz reflejada en la retina, hasta la construcción eventual de una representación de la escena objetiva que rodea a quien percibe. Hay un sentido en el cual la percepción visual no es un proceso racional del modo como lo es el pensamiento, y esto omitiría la motivación inmediata de postular un lenguaje del pensamiento para la percepción visual. Este punto es una manera de introducir otra propuesta importante de Fodor acerca de la estructura de la mente: la propuesta de que la mente es *modular*.

Estamos familiarizados con el fenómeno de la ilusión visual, en que algo parece visualmente lo que no es. Considérense las bandas de Mach (nombradas según el gran físico Ernst Mach, quien descubrió la ilusión) representadas en la figura IV.1. Al verlas por primera vez, la reacción inicial será que cada franja no es uniformemente gris, sino que el matiz se vuelve ligeramente más claro del lado de la franja más próximo a la franja más oscura. Así se ven las cosas. Sin embargo, al revisar más de cerca puede notarse que cada franja es en realidad uniformemente gris. Aíslese una de las

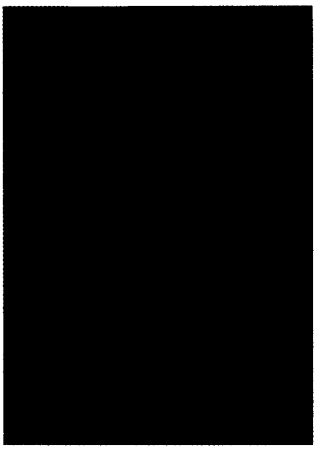


FIGURA IV.1. *Las bandas de Mach. Las franjas son realmente de un matiz uniforme de gris, pero parecen más claras en los bordes más cercanos a las franjas más oscuras*

franjas entre dos trozos de papel, y esto se vuelve evidente. Así, ahora sabe usted, y por lo tanto cree, que cada franja es uniformemente gris. ¡No obstante, sigue pareciendo como si no lo fuera, pese a lo que sabe usted! Para nuestros propósitos presentes, lo interesante no es tanto que su sistema visual es víctima de esta ilusión, sino que la ilusión *persista* aun cuando sepa usted que es una ilusión.

Una cosa que esto muestra claramente es que percibir no es lo mismo que juzgar o creer. Pues si percibir no fuese sino una forma de creer, entonces su estado psicológico del momento sería un conflicto entre creer que *las franjas están uniformemente coloreadas* y creer que *las franjas no están uniformemente coloreadas.* Éste sería un caso de creencia explí-

citamente contradictoria: cree usted que algo es el caso y que no lo es, simultánea y conscientemente. Ninguna persona racional puede vivir con semejantes contradicciones explícitas en sus creencias. Es imposible saber qué conclusión puede extraerse razonablemente de la creencia de que P y no-P; y es imposible saber cómo actuar sobre la base de esta creencia. Por lo tanto, la persona racional intenta eliminar contradicciones explícitas en su creencia, so pena de irracionalidad. Enfrentado a una situación donde uno se inclina a creer una cosa y la opuesta, uno tiene que ponerse de acuerdo y decidir en un sentido o el otro. Uno está obligado, como pensador racional, a tratar de eliminar la incoherencia en el pensamiento propio.

En el caso de la ilusión de las bandas de Mach, no hay manera de eliminar la incoherencia. Nada hay que pueda uno hacer para que las líneas dejen de parecer desigualmente sombreadas, sin importar cuánto se afane uno. Si la percepción fuera nada más una forma de creencia, como han sostenido algunos, entonces éste sería un caso de irracionalidad.[15] Es claro que no es así: uno no tiene dificultad, una vez informado de los hechos, para saber qué conclusión extraer de esta combinación de creencia y percepción, y actuar al respecto. La racionalidad de uno no queda para nada socavada por esta experiencia ilusoria. Por lo tanto, percepción no es creencia.

¿Qué clase de imagen total de la mente sugieren por fenómenos como éste? Jerry Fodor ha sostenido que atestiguan el punto de vista de que el sistema visual es un "módulo mental" relativamente aislado, un sistema proce-

[15] D. M. Armstrong sostiene que la percepción es creencia en *A Materialist Theory of the Mind,* cap. 7.

sador de información que, en importantes aspectos, es independiente del "sistema central" responsable de la creencia y el razonamiento.[16] Fodor sostiene también que otros "sistemas de entrada" —por ejemplo los sistemas que procesan entradas lingüísticas— son modulares de esta manera. La tesis de que la mente tiene esta estructura total —sistema central más módulos— se llama tesis de la modularidad de la mente, y ha tenido gran influencia en la psicología y la ciencia cognitiva. Muchos psicólogos creen en alguna versión de la tesis, aunque sea discutible cuánto hay de modular en la mente. Aquí trataré brevemente de dar algún sentido de la naturaleza y el alcance de la tesis.

¿Qué es exactamente un módulo? Según la introducción original de Fodor sobre esta noción, un módulo funcionalmente definido es una parte de la mente cuyo rasgo más importante es lo que llama *encapsulación informacional*.[17] ("Funcionalmente definido" significa en términos de lo que hace, y no aquello de que está hecho.) Un mecanismo cognitivo es informacionalmente encapsulado cuando sistemáticamente no tiene acceso a toda la información en la mente de quien piensa, al realizar sus operaciones características. Un mecanismo computacional informacionalmente encapsulado puede dar como salida la conclusión P, incluso si en algún otro lugar de la mente del sujeto está el

[16] Jerry A. Fodor, *The Modularity of Mind* (Cambridge, MIT Press, 1983).

[17] Sobre la encapsulación informacional como el rasgo más importante de los módulos, véase *The Modularity of Mind,* pp. 37 y 71; a pesar de algunos cambios en sus modos de ver con los años, este punto ha permanecido constante. Véase *The Mind Doesn't Work That Way* (Cambridge, MIT Press, 2000), p. 63.

conocimiento de que no-P: pero, lo que es más, el conocimiento de que no-P *no puede cambiar la salida del mecanismo computacional*. Por usar una frase de Zenon Pylyshyn, la salida del mecanismo no es "penetrable cognitivamente": no puede ser penetrado por otras áreas del sistema cognitivo, específicamente por creencias y conocimientos.

El punto es fácil de comprender cuando se aplica a un ejemplo concreto. Por mucho que se empeñe uno, no puede ver las franjas de las bandas de Mach uniformemente matizadas de gris, aun cuando se sepa que lo están. El conocimiento que se tiene acerca del modo como están realmente coloreadas no puede penetrar la salida de nuestro sistema visual. La explicación de Fodor de esto es que el sistema visual (y otros "sistemas de entrada") están encapsulados informacionalmente, y que tal es la esencia de aquello que será un módulo. Por supuesto, las ilusiones como las bandas de Mach necesitan explicación detallada en términos de la elaboración detallada de los sistemas visuales; para Fodor, la cuestión es que esta explicación debe realizarse dentro del contexto de un punto de vista modular de la percepción, y no de acuerdo con un modo de ver de la percepción que la trata como una clase de cognición o creencia.

Fodor contrasta los módulos, como el sistema visual, con los "sistemas centrales" o "mente central". La mente central es la morada de las actitudes proposicionales normales, los estados que participan en el razonamiento y la inferencia, y la solución de problemas intelectuales y prácticos. Donde es cuestión de creencia, la estructura del sistema de creencias permite usar la información, al razonar, procedente de cualquier parte de la dotación personal de creencias y conocimientos. Por supuesto, la gente es irracional, tiene puntos ciegos, y se engaña. La cuestión es que estas limitaciones

son idiosincrasias personales; no están construidas en el sistema de creencias mismo. La situación es diferente con el procesamiento visual y los otros módulos.

Como resultado de esta encapsulación informacional, otras propiedades "se apiñan" en torno a un módulo. Los módulos son de *dominio específico:* usan información únicamente de un dominio cognitivo restringido; esto es, no pueden representar nada más cualquier proposición acerca del mundo, a diferencia del pensamiento. El sistema visual representa sólo propiedades visualmente perceptibles del medio, por ejemplo. Asimismo, los módulos tienden a ser obligatorios: no puede uno impedirse ver cosas de determinado modo, oír una oración como gramatical o no, etc. Son innatos, no adquiridos; nacemos con ellos. Pueden bien estar conectados con fuerza, es decir, realizados en una parte del cerebro dedicada a ello; si se dañan no pueden ser remplazados por actividad en otro lugar del cerebro. Y son rápidos, mucho más rápidos que los procesos de la mente central. Estas características proceden todas como resultado de encapsulación informacional: "Lo que la encapsulación compra es velocidad, y compra velocidad al costo de inteligencia".[18] Precisamente como contrasta los módulos con la mente central, Fodor gusta de compararlos con reflejos. Un reflejo, tal como el parpadeo, es rápido y no constreñido por lo que uno pudiera creer o saber; esto tiene perfecto sentido, dada la función del reflejo del parpadeo para proteger los ojos. No se quiere dejar de pensar acerca de si la avispa realmente va a posársele a uno en el ojo; el ojo hace cortocircuito con el pensamiento. Los módulos no son reflejos, ya que contienen estados con contenido representacional; pero la

[18] *The Modularity of Mind,* p. 80.

comparación aclara por qué todas (o algunas, o la mayoría) de las propiedades anteriores tienden a estar asociadas con lo que Fodor llama módulos. (Vale la pena mencionar que Chomsky ha usado el término "módulo" de una manera diferente: para él un módulo es simplemente un cuerpo de conocimiento innato. La idea de Chomsky de un módulo no compromete la encapsulación informacional.)[19]

Desde que Fodor propuso esta tesis en 1983, ha habido un debate activo entre los psicólogos y filósofos acerca del grado de modularidad. ¿Cuántos módulos más hay? Fodor fue inicialmente muy cauto: sugirió que cada sistema perceptual es modular, y que había un módulo para el procesamiento del lenguaje. Otros han sido más aventurados: algunos han argüido, por ejemplo, que el conocimiento tácito de la teoría de las otras mentes es un módulo innato, de acuerdo con la hipótesis de que puede ser dañado —y así dañar las interacciones interpersonales— mientras deja intacta gran parte de la inteligencia general. (A menudo se pretende que ésta es la fuente del autismo: es típico que los niños autistas tengan alta inteligencia en general pero carezcan de "teoría de la mente".)[20] Otros van incluso más lejos y sostienen que la mente es "modular en masa': hay un mecanismo distinto, más o menos encapsulado, para cada clase de tarea cognitiva. Podría haber un módulo para reconocer pájaros, un módulo para creencias acerca de la cocina y acaso inclusive un módulo para la filosofía. Y así sucesivamente.

Si la modularidad masiva es cierta, entonces no hay distinción entre mente central y módulos, simplemente porque no hay tal cosa como una mente central: nada como un

[19] Véase Chomsky, *Rules and Representations.*

[20] Véase S. Baron-Cohen, *Mindblindness: An Essay on Autism and Theory of Mind* (Cambridge, MIT Press, 1995).

mecanismo cognitivo no encapsulado y sin especificidad de dominio. Nuestras facultades mentales estarían mucho más fragmentadas de lo que parecen desde el punto de vista de la psicología del sentido común. Supóngase que tengo un módulo para pensar acerca de comida (no estoy diciendo que nadie haya propuesto semejante módulo, sino que podemos usarlo como ejemplo para ilustrar la tesis). ¿Podría realmente ser cierto que mi razonamiento acerca de qué cocinar para la comida se limite a información disponible a este módulo alimentario solo?, ¿tiene sentido suponer que debe también ser sensible a la información acerca de si quiero salir después, si quiero perder peso, si quiero impresionar y complacer a mis amigos, y así por el estilo? Tal vez éstos podrían considerarse como fragmentos de información pertenecientes al mismo módulo; pero ¿cómo, entonces, distinguimos un módulo de otro?

Asimismo, como ha mostrado Fodor, la tesis está sometida a un problema muy general: si no hay mecanismo de propósito general, ni específico para el dominio, entonces ¿cómo decide la mente?; para cualquier entrada dada, *¿cuál* módulo debe encargarse de dicha entrada? El procedimiento de decisión para asignar entrada a los módulos no puede ser modular, ya que debe seleccionar entre información que va a ser tratada por muchos módulos diferentes. Me gustaría saber si la tesis de la modularidad masiva no terminará socavándose a sí misma.[21]

[21] Véase Fodor, *The Mind Doesn't Work That Way,* cap. 4, especialmente pp. 71-78.

Problemas para el lenguaje del pensamiento

La discusión de la modularidad fue en cierto grado una digresión. Espero que nos haya dado un sentido de la relación entre la tesis de la modularidad y la teoría computacional de la cognición. Ahora volvamos a la hipótesis del mentalés. A mucha gente la hipótesis le parece —tanto en filosofía como fuera de ella— un estrafalario fragmento de información, fácil de refutar por argumentación filosófica o por evidencia empírica. De hecho, me parece que las cosas no son tan sencillas, y que la hipótesis puede defenderse contra los más vigorosos de estos ataques. Discutiré aquí dos de las críticas más interesantes a la hipótesis del mentalés, ya que tienen interés filosófico general, y nos ayudarán a refinar nuestra comprensión de la hipótesis. A pesar del poder de estas argumentaciones, creo que Fodor puede defenderse solo contra sus críticos.

1. ¿Homúnculos de nuevo?

Hemos hablado muy libremente acerca de oraciones en la mente, y sus interpretaciones. Usando la comparación con las computadoras dije que los estados electrónicos de la computadora son "interpretables" como cálculo o como el procesamiento de oraciones. Tenemos una idea bien clara de cómo estos estados pueden tener contenido semántico o significado: son diseñados por ingenieros y programadores en computación, de tal modo que sean interpretables por sus usuarios. Los rasgos semánticos de los estados de una

computadora derivan por lo tanto de las intenciones de los diseñadores y usuarios de la computadora.[22]

O consideremos oraciones en un lenguaje natural como el español. Según vimos en el capítulo II, sigue siendo un problema establecer cómo las oraciones obtienen su significado. Sin embargo, una idea influyente es que lo obtienen a causa del modo como son *usadas* por los hablantes en la conversación, la escritura, el soliloquio, etc. Qué signifique exactamente esto, no importa aquí; lo que importa es la idea plausible de que las oraciones pasan a significar lo que hacen, a causa de los usos que les dan los hablantes.

¿Qué pasa con el mentalés? ¿Cómo llegan a significar algo sus oraciones? Claramente, no alcanzan su significación siendo usadas conscientemente por quienes piensan, pues de otra manera podríamos saber por introspección si la hipótesis del mentalés era cierta. Decir que alcanzan su significado siendo usadas por *alguien más,* parece generar lo que en ocasiones se denomina "falacia del homúnculo". Esta argumentación podría expresarse como sigue.

Supóngase que explicamos el significado de frases en mentalés diciendo que hay un subsistema u homúnculo en el cerebro que usa estas oraciones. ¿Cómo consigue el homúnculo usar estas oraciones? Aquí hay un dilema. Por un lado, si decimos que el homúnculo usa las oraciones teniendo su propio lenguaje interior, entonces tenemos que explicar cómo las oraciones, en este lenguaje, obtienen su significado: pero recurrir a otro homúnculo, menor, claramente sólo suscita el mismo problema otra vez. Por otra parte, si decimos que el homúnculo consigue usar estas oraciones

[22] Véase Fred Dretske, "Machines and the Mental", *Proceedings and Addresses of the American Philosophical Association,* 59 (septiembre de 1985).

sin tener un lenguaje interior, entonces ¿por qué no podemos decir lo mismo acerca de la gente?

El problema es éste. O bien las oraciones en mentalés obtienen su significado de la misma manera que las oraciones del lenguaje público, o bien obtienen su sentido de alguna otra manera. Si obtienen su significado del mismo modo, entonces parecemos presas de un encadenamiento de homúnculos. Pero si pueden obtener su sentido de una manera diferente, entonces necesitamos decir qué camino es éste. O lo uno, o lo otro, no tenemos explicación de cómo significan algo las oraciones en mentalés.

Algunos autores creen que esta clase de objeciones traba la hipótesis del mentalés.[23] Empero, bajo una luz más positiva, podría verse no como una objeción sino como un reto: explicar los rasgos semánticos del lenguaje del pensamiento, sin recurrir a las ideas que se está tratando de explicar. Hay dos maneras posibles de responder al reto. La primera sería aceptar la metáfora del homúnculo pero negar que los homúnculos generen necesariamente un círculo vicioso. La idea se origina a partir de un concepto de Daniel Dennett (mencionada en "Algoritmos automáticos", p. 172). Lo que necesitamos garantizar es que, cuando postulamos un homúnculo para explicar las capacidades de otro, no le atribuimos las capacidades que estamos tratando de explicar. Cualquier homúnculo que postulemos debe ser más estúpido que aquel cuyo comportamiento estamos tratando de explicar, pues de otro modo no hemos explicado nada.[24]

[23] Searle, por ejemplo, piensa que "la falacia del homúnculo es endémica en los modelos computacionales de cognición", en *The Rediscovery of the Mind* (Cambridge, MIT Press, 1992), p. 226.

[24] La actitud tomada en este párrafo es más cercana a la de William G. Lycan, *Consciousness* (Cambridge, MIT Press, 1987).

Sin embargo, como lo ha señalado Searle, si en el nivel computacional profundo el homúnculo sigue manipulando *símbolos,* éstos deben tener un significado, aun si son unos y ceros. Y, si hay un homúnculo realmente estúpido bajo este nivel —piénsese en él como uno que sólo mueve la cinta de una máquina de Turing de lado a lado—, entonces sigue siendo difícil ver cómo la mera existencia de este homúnculo movedor de cinta sólo puede explicar el hecho de que los unos y ceros tienen significado. El problema de obtener a partir de una actividad sin sentido una actividad significativa simplemente parece surgir otra vez en este nivel más bajo.

El segundo enfoque, más popular, al reto es que las oraciones en mentalés tienen su significado de un modo muy distinto al modo como lo tienen las oraciones del lenguaje público. Las oraciones de este tipo pueden adquirir su significado siendo intencionalmente usadas por hablantes, pero esto no puede ser como es con el mentalés. Las oraciones del mentalés, como ha dicho Fodor, tienen sus efectos sobre el comportamiento de quien piensa, "sin tener que ser comprendidas".[25] No son comprendidas no porque no sean usadas conscientemente: el uso consciente de oraciones se interrumpe en el mundo exterior. No hay homúnculos que usen oraciones del modo como lo hacemos nosotros.

Esto no evita la objeción. Ahora, por supuesto, la cuestión es: ¿cómo obtienen *su* sentido las oraciones del mentalés? Éste es un problema difícil, que ha sido objeto de intenso debate. Será considerado en el capítulo v.

[25] "A Situated Grandmother?", *Mind and Language,* 2 (1987), p. 67.

2. Seguir una regla y conformarse a una regla

Searle también patrocina la segunda objeción que mencionaré aquí, que deriva de algunas objeciones bien sabidas provocadas por W. V. Quine ante la tesis de Chomsky de que tenemos conocimiento tácito de la gramática.[26] Recuérdese que la hipótesis del mentalés dice que pensar es regla gobernada, e incluso que en algún sentido "tácito" conocemos estas reglas. ¿Cómo se distingue esta pretensión de la pretensión de que nuestro pensamiento *conforma* una regla, que simplemente actuamos y pensamos *de acuerdo con una regla*? Como vimos en el capítulo III, los planetas se ajustan a las leyes de Kepler, pero no "siguen" o "saben" estas leyes en ningún sentido literal. La objeción es que si la hipótesis del mentalés no pudiera explicar la diferencia entre seguir una regla y conformarse nada más a una regla, entonces se pierde mucho de su sustancia.

Nótese que no ayudará decir que la mente contiene una representación explícita de la regla (es decir, una oración que enuncie la regla). Pues una representación de una regla no es sino otra representación: necesitaríamos *otra* regla más para conectar esta representación de regla con las otras representaciones a que se aplica. Y decir que esta regla "superior" debe ser representada explícitamente, no hace sino suscitar el mismo problema otra vez.

La cuestión no es "¿qué hace que la hipótesis del mentalés sea computacional?"; es computacional porque las oraciones del mentalés son representaciones gobernadas por reglas computacionales. La cuestión es "¿qué sentido puede

[26] Véase Quine, "Methodological Reflections on Current Linguistic Theory", en Donald Davidson y Gilbert Harman (eds.), *Semantics of Natural Language* (Dordrecht, Reidel, 1972).

darse a la idea de 'gobernado por reglas computacionales'?" Pienso que el defensor del mentalés debiera responder explicando qué es para una regla ser representada *implícitamente* en la estructura causal de los procesos mentales. Decir que las reglas son representadas implícitamente es decir que el comportamiento de un pensante puede ser *mejor explicado* suponiendo que el que piensa tácitamente sabe una regla, en lugar de suponer que no la conoce. Lo que ahora necesita ser explicado es la idea del conocimiento tácito, pero esto debo dejarlo a las investigaciones posteriores del lector, ya que hay un punto más a propósito de las reglas que necesita ser expuesto.[27]

A algunas personas pudiera importarles el uso de un ejemplo lógico en mi exposición de la hipótesis del mentalés. Pues es claro que los seres humanos no siempre razonan de acuerdo con las leyes de la lógica. No obstante, si reglas como *modus ponens* se supone que gobiernan causalmente el pensamiento real, ¿cómo es posible esto? Una posibilidad es decir que las reglas de la lógica no *describen* el pensamiento humano, sino que más bien *prescriben* maneras como los humanos debieran pensar. (Esto se plantea a veces diciendo que las reglas de la lógica son "normativas" y no "descriptivas".) Una manera de presentar la diferencia es decir que, si fuéramos a encontrar muchas excepciones a las leyes físicas, pensaríamos que habíamos errado al plantear las leyes, de alguna manera. Si encontramos una persona que se comporta ilógicamente, no pensamos que hemos obtenido mal las leyes de la lógica; preferimos designar a la persona como irracional o ilógica.

[27] Para una discusión útil del conocimiento tácito, véase Martin Davies, "Tacit Knowledge and Subdoxastic States", en Alexander George (ed.), *Reflections of Chomsky* (Oxford, Blackwell, 1989).

No surge el punto sólo porque el ejemplo fue tomado de la lógica. Pudimos igualmente tomar un ejemplo de la teoría del razonamiento práctico. Supóngase que la regla es "actuar racionalmente". Cuando encontramos a alguien que actúa coherentemente de una manera que está en conflicto con esta regla, podríamos hacer una de dos cosas: rechazar la regla como una verdadera descripción del comportamiento de la persona, o mantener la regla y decir que la persona es irracional. El reto que estoy considerando dice que deberíamos hacer esto último.

La hipótesis del mentalés no puede permitir que las reglas que gobiernan el pensamiento sean normativas de esta manera. Así, ¿qué es lo que habrían de decir? Pienso que habrían de decir dos cosas, una defensiva y otra más agresiva. La pretensión defensiva es que la hipótesis no está, en esta etapa, entregada a la idea de que las leyes normativas de la lógica y la racionalidad son las reglas que operan en las oraciones del mentalés. Es una cuestión científica/empírica si las reglas gobiernan la mente, y las reglas que hemos mencionado pueden no figurar entre éstas. La pretensión agresiva es que, aun si algo como estas reglas gobernase la mente, serían *idealizaciones* del comportamiento complejo, revuelto y real de las mentes. Para enunciar las reglas propiamente, habríamos de añadir una cláusula diciendo que "todas las demás cosas son iguales" (lo que se llama una cláusula *ceteris paribus*). Esto no carcome la naturaleza científica del mentalés, ya que las cláusulas *ceteris paribus* se usan también en otras teorías científicas.[28]

Estos cuidados acerca de reglas son fundamentales para

[28] Véase Fodor, *Psychosemantics* (Cambridge, MIT Press, 1987), capítulo 1.

la hipótesis del mentalés. La clave plena de la hipótesis es que pensar es la manipulación, gobernada por regla, de oraciones mentales. Como uno de los principales argumentos en pro de la estructura sintáctica fue la idea de que los procesos mentales son sistemáticos, resulta que la cuestión crucial es: ¿está la regla humana de pensar gobernada en el sentido en que dice la hipótesis? ¿Hay leyes del pensamiento para que las descubra la ciencia cognitiva? De hecho, ¿puede la naturaleza del pensamiento humano ser capturada en términos de reglas o leyes, a fin de cuentas?

Hemos encontrado esta cuestión antes, al discutir las objeciones de Dreyfus a la inteligencia artificial. Dreyfus se opone a la idea del pensamiento humano que inspira a la ciencia cognitiva y la hipótesis del mentalés: la idea de que el pensamiento humano puede ser capturado exhaustivamente mediante un conjunto de reglas y representaciones. Opuestamente a esto, sostiene que una actividad práctica, una red de capacidades corporales que no pueden ser reducidas a reglas, subyace a la inteligencia humana. En el capítulo anterior examinamos múltiples modos como la IA podría responder a estas críticas. Sin embargo, algunas personas juzgan posible aceptar algunas de las críticas de Dreyfus sin dar una visión ampliamente computacional de la mente.[29] Esta posibilidad podría parecer muy difícil de captar; el propósito del apartado siguiente es explicarlo.

[29] Véase H. Dreyfus y S. Dreyfus, "Making a Mind versus Modelling the Brain", en Boden (ed.), *The Philosophy of Artificial Intelligence.*

Computadoras "sesudas"

Piénsese en las cosas para las que sirven las computadoras. Se han construido computadoras que sobresalen en los cálculos rápidos, en el almacenamiento eficiente de información y su rápida recuperación. Los programas de inteligencia artificial han sido proyectados para poder jugar un excelente ajedrez, y pueden demostrar teoremas de lógica. A menudo se señala que, comparados con las computadoras, la mayoría de los seres humanos no son muy buenos calculando, jugando ajedrez, demostrando teoremas o recuperando información rápida del género logrado por las modernas bases de datos (la mayoría de nosotros seríamos fatales memorizando algo como nuestras listas de direcciones: por eso usamos computadoras que lo hagan). Lo que es más, las clases de tareas que acuden muy naturalmente a los humanos —como reconocer caras, percibir habilidades lingüísticas y habilidades corporales prácticas— han sido precisamente tareas que ha encontrado la IA tradicional y la ciencia cognitiva más difíciles de simular y/o explicar.

La ciencia cognitiva y la IA tradicional han considerado estos problemas como desafíos, que requieren más tiempo de investigación y algoritmos y heurísticas afinados más delicadamente. Desde más o menos mediados de los años ochenta, estos problemas han empezado a verse como sintomáticos de una debilidad más general en el enfoque ortodoxo de la ciencia cognitiva, mientras otro enfoque computacional ha comenzado a ganar influencia. Mucha gente piensa que este nuevo enfoque —conocido como "conexionismo"— representa una posibilidad seria ante explicaciones como la hipótesis del mentalés de Fodor. Que esto sea

cierto es una cuestión muy controvertida, pero lo que parece ser cierto es que la existencia del conexionismo amenaza la defensa "pragmática" del mentalés, de que es "el único juego de la ciudad". (En *The Language of Thought* Fodor cita la famosa observación de Lyndon B. Johnson: "Soy el único presidente que tienen ustedes".) La existencia del conexionismo también pone en tela de juicio el argumento en favor del mentalés esbozado antes, basado en una inferencia de ser la mejor explicación; ya que, si hay otras buenas explicaciones a la vista, entonces el mentalés tiene que combatir más para mostrar que es la mejor.

Las cuestiones que rodean al conexionismo son sumamente técnicas, y caería más allá del alcance de este libro dar una descripción detallada del debate. Así, el propósito de este apartado final es simplemente dar una impresión de estas cuestiones, a fin de mostrar que podría haber un tipo de teoría computacional de la mente que sirva en lugar de la hipótesis del mentalés y su parentela. Quienes no estén interesados en esta cuestión, bastante más técnica, pueden saltarse este apartado y pasar de una vez al capítulo siguiente. Quienes aspiran a seguir adelante pueden consultar las sugerencias de las "Lecturas adicionales" al final del capítulo. Empezaré diciendo qué define a los enfoques "ortodoxos" y en qué difieren los modelos conexionistas.

La hipótesis del mentalés construye la computación del modo que ahora se llama ortodoxo o "clásico". Las máquinas con una "arquitectura" computacional clásica (que se llama a veces arquitectura de Von Neumann) suelen de ordinario incluir una distinción entre *estructuras de datos* (esencialmente representaciones explícitas de fragmentos de información) y reglas o *programas* que operan sobre estas estructuras. Las representaciones en las arquitecturas clásicas

tienen estructura sintáctica, y las reglas se aplican a las representaciones en virtud de la estructura, según ilustré ya. Así también, las representaciones son típicamente procesadas en serie en lugar de hacerlo en paralelo; todo lo que esto significa es que el programa opera sobre los datos paso a paso (representado, por ejemplo, por el diagrama de flujo del programa), en oposición a realizar multitud de operaciones al mismo tiempo. (Esta clase de arquitectura computacional se llama a veces imagen de "reglas y representaciones"; aplicada a IA, John Haugeland lo ha bautizado como "GOFAI", acrónimo de "good old-fashioned artificial intelligence".)[30]

La arquitectura conexionista es muy diferente. Una máquina conexionista es una red que consiste en gran número de unidades o nodos: simples dispositivos de entrada-salida capaces de ser excitados o inhibidos por corrientes eléctricas. Cada unidad está conectada con otras unidades (de ahí lo de "conexionismo"), y las conexiones entre las unidades pueden ser de varias fuerzas o "pesos". Que una unidad dé cierta salida —de manera estándar, una corriente eléctrica— depende de su umbral de disparo (la entrada mínima requerida para que responda) y las fuerzas de sus conexiones con otras unidades. Esto es, una unidad es encendida cuando la fuerza de sus conexiones con las otras unidades rebasa su umbral. Esto a su vez afectará la fuerza de todas sus conexiones con otras unidades, y con mayor razón si esas unidades están encendidas.

Las unidades están dispuestas en "capas"; normalmente hay una capa de entrada de unidades, una capa de salida y una o más capas de unidades "ocultas", que median entre entrada y salida. (Véase la figura IV.2 para un diagrama idealizado.)

[30] Véase Haugeland, *Artificial Intelligence,* pp. 112 y ss.

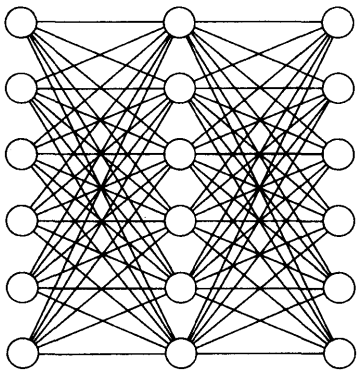


FIGURA IV.2. *Diagrama de una red conexionista*

La computación en las redes conexionistas incluye primero fijar las unidades de entrada en alguna combinación de "encendidos" y "apagados". Como las unidades de entrada están conectadas con las otras unidades, fijar su estado inicial causa una pauta de activación que se extiende a través de la red. Esta pauta de activación es determinada por las fuerzas de las conexiones entre las unidades y el modo como estén fijadas las unidades de entrada. A fin de cuentas, la red "se asienta" en un estado estable —las unidades se han colocado en equilibrio con los estados fijos de las unidades de entrada— y la salida puede leerse en la capa de unidades de salida. Un carácter notable es que este proceso acontece en paralelo, o sea que los cambios en los estados de la red están ocurriendo a través de la red al mismo tiempo, no paso a paso.

Para que esto sea computación, por supuesto, necesitamos interpretar las capas de unidades de entrada y salida como *representantes* de algo. Precisamente como en una máquina clásica, se asignan representaciones a redes conexionistas por las personas que las construyen, pero las maneras como son asignadas son muy diferentes. La representación conexionista puede ser de dos tipos: interpretaciones *localistas,* en que a cada unidad se asigna un rasgo que representa, o interpretaciones *distribuidas,* en las que es el estado de la red en conjunto lo que representa. La representación distribuida es, se pretende a menudo, uno de los rasgos distintivos del conexionismo; el enfoque mismo a menudo es conocido como procesamiento paralelo distribuido o PPD. Diré unas palabras más acerca de la representación distribuida, dentro de un momento.

Un rasgo distintivo de las redes conexionistas es que parece que pueden "adiestrarse para aprender". Supóngase que quisiera usted obtener que la máquina produjera cierta salida en respuesta a la entrada (por ejemplo, hay una red que convierte el tiempo presente de los verbos españoles en sus formas de tiempo pasado).[31] Empezará alimentando la entrada y dejará que una pauta bastante casual de activación se difunda a través de la máquina. Verifíquese la salida, y véase hasta dónde diverge de la salida deseada. Entonces alterará usted repetidamente la fuerza de las conexiones entre las unidades hasta que la unidad de salida sea la deseada. Esta clase de método de prueba y error es conocido como "entrenar la red". Lo interesante es que, una vez que ha

[31] W. Bechtel y A. Abrahamsen, *Connectionism and the Mind* (Oxford, Blackwell, 1991), cap. 6; Andy Clark, *Microcognition* (Cambridge, MIT Press, 1989), cap. 9.

sido entrenada una red, puede aplicar el proceso de prueba y error, *por sí sola,* a nuevas muestras con algún éxito. Así es como los sistemas conexionistas "aprenden" cosas.

Las máquinas conexionistas se llaman a veces "redes neurales", y este nombre da una clave aparte de su atractivo para algunos científicos cognitivos. Con su vasto número de unidades interconectadas (aunque sencillas) y las fuerzas variables de la conexión entre las unidades, se asemejan a la estructura del cerebro mucho más de cerca que ninguna máquina clásica. Los conexionistas, por lo tanto, tienden a sostener que sus modelos son más plausibles biológicamente que los de arquitectura clásica. Sin embargo, estas pretensiones pueden ser exageradas: hay muchas propiedades de las neuronas que estas unidades no tienen.[32]

Muchos conexionistas pretenden también que sus modelos son más psicológicamente plausibles, es decir, que las redes conexionistas se comportan de un modo más próximo a como funciona la mente humana, que el de las máquinas clásicas. Según mencioné antes, las computadoras clásicas son muy malas realizando innumerables tareas que encontramos muy naturales, reconocimiento de caras y pautas, por ejemplo. Los conexionistas entusiastas arguyen a menudo que éstos son precisamente los tipos de tareas en que sus máquinas sobresalen.

Espero que esta imagen muy rudimentaria le haya dado a usted alguna idea de la diferencia entre la ciencia conexionista y la cognitiva clásica. Puede usted preguntarse, a pesar de todo, cómo son en algún sentido computadoras las máquinas conexionistas. Ciertamente, la idea de una pauta de activación que se difunde a través de una red no tiene un

[32] Véase Jack Copeland, *Artificial Intelligence,* cap. 10, § 5.

aire muy parecido a la clase de computación que vimos en el capítulo III. Algunos autores insisten en una definición estricta de "computadora" en términos de manipulación de símbolos y descartan las máquinas conexionistas con este fundamento.[33] Otros son dichosos al ver las redes conexionistas como casos de la noción muy general de una computadora, como algo que transforma una representación de entrada en una representación de salida, de una manera disciplinada.[34]

En parte, esto es una cuestión de terminología: todo el mundo convendrá en que hay algo en común entre lo que hace una máquina conexionista y lo que hace una computadora clásica, y todo el mundo convendrá en que también hay diferencias. Si no concuerdan acerca de si llamar a las similaridades "computación", esto no puede ser cuestión de gran importancia. Sin embargo, me pongo de parte de quienes dicen que las máquinas conexionistas son computadoras. Después de todo, las redes conexionistas procesan funciones de entrada-salida de una manera sistemática, usando representaciones (localizadas o distribuidas). Y, cuando aprenden, lo hacen empleando "algoritmos o reglas de aprendizaje". Así que hay suficiente en común para llamar computadoras a ambas, aunque esto puede ser resultado de la definición suficientemente general que di de una computadora en el capítulo III.

Sin embargo, ésta no es la cuestión interesante, sino la de que las diferencias fundamentales son entre máquinas conexionistas y máquinas clásicas, y cómo estas diferencias

[33] Véase, por ejemplo, Jack Copeland, *Artificial Intelligence*, cap. 9, § 8, y cap. 10, § 4.

[34] Véase, por ejemplo, Robert Cummins, *Meaning and Mental Representation* (Cambridge, MIT Press, 1989), pp. 147-156.

repercuten en la teoría de la mente. Al igual que muchas cuestiones en esta área, no hay consenso general acerca de cómo hay que contestar a esta pregunta. Empero, trataré de esbozar lo que veo como los puntos más importantes.

La diferencia no es nada más que una red conexionista pueda ser descrita en el nivel computacional más sencillo en términos que no tienen interpretación en lenguaje psicológico (o científico, por ejemplo como la creencia de que "pasado" es el tiempo pasado de "pasar"). Pues, en una máquina clásica, hay un nivel de procesamiento —el nivel de los "bits" o dígitos binarios de información— en el cual los símbolos procesados carecen de interpretación psicológica natural.[35] Como vimos en el capítulo III, una computadora funciona descomponiendo las tareas que realiza en tareas simples más sencillas: en el nivel más simple, no hay interpretación de los símbolos procesados como, digamos, oraciones o contenidos de creencias y deseos.

Sin embargo, lo atractivo de las máquinas clásicas era que estas operaciones básicas podían realizarse de una manera sistemática para construir símbolos complejos —como pueden ser palabras y oraciones en el lenguaje del pensamiento— según los cuales operan los procesos computacionales. De acuerdo con la hipótesis del mentalés, el proceso opera sobre los símbolos en virtud de la forma o sintaxis de éstos. La hipótesis es que las oraciones en mentalés son *a)* procesadas "formalmente" por la máquina y *b)* representaciones: son interpretables como dotadas de significado. Esto es, una y la misma cosa —la oración en mentalés— es el vehículo de la computación y el vehículo del contenido mental.

[35] Véase la discusión de Cummins en *Meaning and Mental Representation,* pp. 150-152.

Esto no tiene que ser así con las redes conexionistas. Según lo plantea Robert Cummins, "los conexionistas no suponen que los objetos de computación sean los objetos de interpretación semántica".[36] Esto es, las computaciones son realizadas por la activación (o la inhibición) de unidades que acrecientan (o disminuyen) la fuerza de las conexiones entre ellas. El "aprendizaje" acontece cuando las relaciones entre las unidades están alteradas sistemáticamente de tal modo que producen una salida próxima al blanco. Así se realiza la computación en el nivel de unidades sencillas. Pero no tiene por qué haber representación en este sencillo nivel: cuando está implicada la representación distribuida, los estados de la red *en conjunto* son lo que es interpretado como lo que se está representando. Los vehículos de la computación —las unidades— no necesitan ser los vehículos de la representación o interpretación psicológica. Los vehículos de la representación pueden ser estados de la red entera.

Este asunto puede ser planteado en términos de sintaxis. Supóngase, para simplificar, que hay una palabra del mentalés, "perro", que tiene los mismos rasgos sintácticos y semánticos que la palabra española "perro". Entonces el defensor del mentalés dirá que, siempre que se tiene un pensamiento acerca de perros, el mismo tipo de estructura sintáctica se halla en la mente de uno. Así, si piensa usted: "algunos perros son mayores que otros" y piensa también: "hay demasiados perros por este rumbo", la palabra "perros" aparece ambas veces en su mente. Los conexionistas niegan que esto tenga que ser así: dicen que cuando uno tiene estos dos pensamientos, los mecanismos de la mente no necesitan tener *nada no semántico* en común. Como dicen

[36] *Meaning and Mental Representation,* p. 157 nota 6.

dos de los precursores del conexionismo: "La moneda de nuestros sistemas no son símbolos, sino excitación e inhibición".[37] En otras palabras: los pensamientos no tienen sintaxis.

Una analogía de Scott Sturgeon podría ayudar a ilustrar esta diferencia entre los vehículos de la computación y los vehículos de la representación.[38] Imagínese una amplia dotación rectangular de luces eléctricas tan grande como un estadio de futbol. Cada luz individual puede o no brillar en mayor o menor grado. Cambiando la iluminación de cada luz, todo el estadio puede exhibir pautas que cuando se ven a distancia son oraciones en español. Una pauta podría decir: "sabemos su secreto", otra: "compre su boleto temprano para evitar quedar desilusionado". Estas palabras son creadas nada más alterando la iluminación de las luces individuales (no hay nada en este nivel de "procesamiento" que corresponda a la sintaxis o semántica de las palabras). La palabra "su" es exhibida por un grupo de luces del primer conjunto y por otro grupo de luces del segundo: pero al nivel de "procesamiento" estos conjuntos de luces no necesitan tener nada más en común (no necesitan ni siquiera tener la misma forma: considérese SU y su). Los objetos de "procesamiento" (las luces individuales) no son los objetos de representación (las pautas de todo el estadio).

Esta analogía podría ayudar a darle a usted una impresión de cómo el procesamiento básico puede producir

[37] D. E. Rumelhart y J. L. McClelland, "PDP Models and General Issues in Cognitive Science", en D. E. Rumelhart y J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* vol. 1 (Cambridge, MIT Press, 1986), p. 132.

[38] Véase Scott Sturgeon, "Good Reasoning and Cognitive Architecture", *Mind & Language,* 9 (1994).

representación sin ser "sensible" a la sintaxis de los símbolos. Sin embargo, hay quien pensará que la analogía es muy engañosa, pues sugiere que el procesamiento en el nivel de las unidades está más cerca del *medio* de representación que del *vehículo* (por usar la terminología introducida antes en este capítulo). Una teoría clásica convendrá en que sus palabras y oraciones son implementadas o realizadas en la estructura del cerebro; y no puede tener objeciones a la idea de que podría haber un nivel "intermedio" de realización en una estructura de tipo conexionista. No obstante, puede insistir de todos modos en que si la cognición es sistemática, entonces su vehículo necesita ser sistemático también; y, como las redes conexionistas no son sistemáticas, no pueden servir como vehículo de cognición, sino sólo como medio.

Ésta es, en efecto, una de las líneas principales de la crítica que Fodor y Zenon Pylyshyn sostienen contra el conexionismo como teoría del procesamiento mental.[39] Como vimos antes, es fundamental para la teoría de Fodor que la cognición sea sistemática: si alguien puede pensar que *Antonio ama a Cleopatra,* entonces debe ser capaz de, cuando menos, considerar el pensamiento de que *Cleopatra ama a Antonio.* Fodor juzga que éste es un hecho fundamental acerca del pensamiento o cognición que toda teoría tiene que explicar, y piensa que un mecanismo análogo al lenguaje *puede* explicarlo; pues está construido según la idea misma de la sintaxis y semántica composicionales. Él y Pylyshyn sostienen entonces que no hay garantía de que las redes conexionistas produzcan representaciones sistemáticas pero, si lo hacen, simplemente estarán "implementando" un meca-

[39] J. Fodor y Z. Pylyshyn, "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition,* 28 (1988).

nismo de estilo mentalés. En la terminología de este capítulo: o la red conexionista será el simple medio de una representación cuyo vehículo es lingüístico o la red no puede conducirse con sistematicidad.

¿Cómo responderían los conexionistas a esto? En líneas generales, podrían aceptar uno de dos enfoques. Podrían sostener que la cognición no es sistemática en el sentido de Fodor, o podrían argüir que mientras que la cognición *es* sistemática, las redes conexionistas pueden ser sistemáticas también. Si asumen la primera actitud, tienen que hacer un gran trabajo para mostrar cómo la cognición puede no lograr ser sistemática. Si toman por el segundo camino, entonces será arduo para ellos evitar el cargo de Fodor y Pylyshyn, de que sus máquinas acabarán simplemente "implementando" mecanismos mentaleses.

Conclusión: ¿explica la computación la representación?

¿Qué conclusiones debemos extraer acerca del debate entre el conexionismo y la hipótesis del mentalés? Es importante recalcar que ambas teorías son altamente especulativas: sugieren imágenes en gran escala de cómo los mecanismos del pensamiento podrían funcionar, pero las teorías detalladas del razonamiento humano están muy lejos en el porvenir. Además, igual que lo correcto de la teoría computacional de la cognición en general, la cuestión no puede a fin de cuentas ser resuelta filosóficamente. Es una cuestión empírica o científica si nuestras mentes tienen una arquitectura clásica de estilo mentalés, una arquitectura conexionista o alguna mezcla de ambas, o, a decir verdad, si nuestras men-

tes tienen, como sea, alguna clase de estructura computacional. Ahora, al fin, tenemos alguna idea de lo que habría que establecer en la disputa entre la teoría computacional y sus rivales.

Volvamos al problema de la representación. ¿Dónde dejó a este problema esta discusión sobre mentes y computadoras? En un sentido, el problema queda intacto con la teoría computacional de la cognición. Como la computación ha de ser definida de acuerdo con la idea de representación, la teoría computacional de la cognición da por descontada la representación. De modo que, si aún queremos explicar la representación, tenemos que mirar en otro sentido. Éste será el tema del capítulo final.

LECTURAS ADICIONALES

The MIT Encyclopedia of the Cognitive Sciences, redactada por Robert A. Wilson y Frank A. Keil (Cambridge, MIT Press, 1999), es la mejor obra de referencia en un volumen acerca de todos los aspectos de la psicología de la ciencia cognitiva, la lingüística, la neurociencia y la filosofía. Una introducción más adelantada a las cuestiones discutidas en este capítulo es la obra de Kim Sterelny, *The Representational Theory of Mind: An Introduction* (Oxford, Blackwell, 1990). Fodor fue el primero en introducir su teoría en *The Language of Thought* (Hassocks, Harvester, 1975), pero la mejor exposición es probablemente *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MIT Press, 1987; especialmente el capítulo 1 y el apéndice), que, como todo lo de Fodor, está escrito con un estilo animado, legible y humorístico. Véase también el

ensayo "Fodor's Guide to Mental Representation", en su colección *A Theory of Content and Other Essays* (Cambridge, MIT Press, 1990). La tesis influyente de la modularidad fue introducida en *The Modularity of Mind* (Cambridge, MIT Press, 1983), y los últimos puntos de vista de Fodor sobre esta tesis y acerca de la teoría computacional de la mente en general se pueden hallar en *The Mind Doesn't Work That Way* (Cambridge, MIT Press, 2000). Uno de los críticos persistentes de Fodor ha sido Daniel Dennett; su temprano ensayo "A Cure for the Common Code?", en *Brainstorms* (Hassocks, Harvester, 1978; reimpreso por Penguin Books en 1997), sigue siendo una importante fuente de ideas para quienes se oponen a la hipótesis del mentalés. Una colección de artículos, muchos de los cuales se ocupan de cuestiones estudiadas en este capítulo, es la de William G. Lycan (ed.), *Mind and Cognition* (Oxford, Blackwell, 2ª ed., 1998). *Vision,* de David Marr (San Francisco, Freeman, 1982), es un texto clásico sobre la teoría computacional de la visión; el capítulo 4 del libro de Sterelny (véase antes) da una buena exposición desde el punto de vista de un filósofo. *The Language Instinct* (Harmondsworth, Penguin, 1994), de Steven Pinker, es una exposición brillante y legible del modo de ver chomskiano del lenguaje, y muchas cosas más. Para la imaginería mental, véase Stephen Kosslyn, *Image and Brain* (Cambridge, MIT Press, 1994). Una introducción sencilla al conexionismo puede encontrarse en el capítulo sobre conexionismo de la segunda edición de *Matter and Consciousness* (Cambridge, MIT Press, 1988), de Paul Churchland, y hay también un capítulo acerca del conexionismo en el libro de Sterelny. Un resumen excelente, inteligible para el no especialista, es el de Brian McLaughlin, "Computationalism, Connectionism and the Philosophy of Mind", en *The Blackwell Guide to Computation and Information* (Oxford, Blackwell 2002).

V. EXPLICANDO LA REPRESENTACIÓN MENTAL

Los últimos dos capítulos han dado, en cierto sentido, un giro a través de algunas de las controversias filosóficas que rodean la teoría computacional de la mente y la inteligencia artificial. Es tiempo ahora de volver al problema de la representación, introducido en el capítulo I. ¿Cómo la discusión de la teoría computacional de la mente nos ha ayudado a entender estos problemas?

Por un lado, ha ayudado a sugerir respuestas, pues vimos que la idea de una computadora ilustra cómo las representaciones pueden también ser cosas que tienen causas y efectos. Asimismo, la idea ordinaria de un proceso computacional —esto es, un proceso causal gobernado por reglas implica representaciones estructuradas— nos permite ver cómo un dispositivo puramente mecánico puede digerir, almacenar y procesar representaciones. Y aunque puede no ser plausible suponer que la mente entera sea como esto, en el capítulo IV examinamos algunos modos como los procesos mentales por lo menos podían ser computacionales.

Por otra parte, la teoría computacional de la mente nos enseña en sí misma qué hace de algo una representación. La razón de esto es sencilla: la noción de computación da por descontada la representación. Un proceso computacional es, por definición, una relación gobernada por regla, o sistemática, entre representaciones. Decir que algún proceso o estado es computacional no explica su naturaleza represen-

tacional; la presupone. O, para plantearlo de otro modo, decir sencillamente que hay un lenguaje del pensamiento no es decir qué hace que las palabras y oraciones que lo componen *signifiquen* algo.

Esto nos conduce, pues, al tema de este capítulo final: ¿cómo habría de explicar la representación el punto de vista mecánico de la mente?

Reducción y definición

El punto de vista mecánico de la mente es uno de carácter *naturalista:* trata la mente como parte de la naturaleza, donde "naturaleza" se entiende como el tema de la ciencia natural. Vistas así las cosas, una explicación de la mente necesita una explicación de cómo la mente se ajusta al resto de la naturaleza, así entendida. En este libro he ido considerando una cuestión más específica: ¿cómo puede la representación mental ajustarse al resto de la naturaleza? Una manera de contestar a esta pregunta es sencillamente aceptar la representación como un rasgo natural básico del mundo. Hay muchas clases de objetos naturales y rasgos naturales del mundo —organismos, hormonas, carga eléctrica, elementos químicos, etc.— y algunos de ellos son fundamentales en tanto que otros no. Por "fundamental" quiero decir que no necesitan o no pueden ser más explicados en términos de otros hechos o conceptos. En física, por ejemplo, el concepto de *energía* se acepta como fundamental; no hay explicación de energía en términos de otros conceptos. ¿Por qué no tomar, entonces, la *representación* como uno de los rasgos fundamentales del mundo?

Este modo de ver puede defenderse recurriendo a la idea

de que la representación es una noción *teórica:* una noción cuya naturaleza es explicada por las teorías a las que pertenece (algo como la noción de *electrón*). Recuérdese la discusión de las teorías en el capítulo II. Allí vimos que, de acuerdo con un punto de vista muy aceptado, la naturaleza de una *entidad teórica* se agota por las cosas que la teoría dice acerca de ello. Lo mismo puede decirse acerca de la representación: la representación es precisamente lo que la teoría de la representación nos dice que es. No hay necesidad de preguntar más acerca de su naturaleza.

Regresaré a esta clase de teoría al final del capítulo. Para la mayoría de los filósofos naturalistas, es un enfoque insatisfactorio del problema. Dirían que la representación sigue siendo un concepto filosóficamente problemático, y no obtenemos auténtica comprensión de él aceptándolo (o su teoría) como primitivo. Dirían: considérese lo que sabemos a propósito del resto de la naturaleza. Sabemos, por ejemplo, que la luz es radiación electromagnética. Al aprender cómo la luz se relaciona con otros fenómenos electromagnéticos, hallamos algo "más profundo" sobre la naturaleza de la luz. Encontramos lo que la luz fundamentalmente *es*. Éste es el tipo de entendimiento que necesitamos de la noción de representación. Jerry Fodor plantea la cuestión de esta manera:

> Supongo que antes o después el físico completará el catálogo que se está compilando de las propiedades últimas e irreducibles de las cosas. Cuando se haga, las (propiedades microfísicas) *espín, encantamiento* y *carga* tal vez aparezcan en la lista. Sin embargo, con seguridad no aparecerá la *acerquidad:* la intencionalidad simplemente no llega tan hondo.[1]

[1] Fodor, *Psychosemantics,* p. 97.

Piénsese lo que se quiera acerca de semejantes puntos de vista; es claro que lo que Fodor y muchos otros filósofos desean es una explicación de la intencionalidad en *otros términos,* esto es, en términos de conceptos distintos de los conceptos de representación. Hay múltiples maneras de lograr hacer esto. Una de ellas, muy evidente, sería dar *condiciones necesarias y suficientes* para pretensiones de la forma "X representa Y". (Los conceptos de condiciones necesarias y suficientes se explicaron en el capítulo 1.) Necesarias y suficientes para "X representa Y" serán aquellas condiciones que rigen cuando, y sólo cuando, X representa Y, descrito en términos que no mencionan el concepto de representación en absoluto. Para plantear esto precisa y limpiamente, necesitamos el término técnico "si y sólo si". (Recuérdese que, así como "A *si* B" expresa la idea de que B es una condición suficiente para A y "A *sólo si* B" expresa la idea de que B es una condición necesaria para A, podemos expresar la idea de que B es una condición necesaria *y* suficiente para A, diciendo "A *si y sólo si* B".)

La presente presunción acerca de la representación puede entonces describirse por el principio de la siguiente forma, que designaré como (R):[2]

(R) X representa Y si y sólo si _____.

Así, por ejemplo, en el capítulo 1 consideré la idea de que la base de la representación pictórica podría ser el parecido. Podríamos expresar esto como sigue:

X representa (pictóricamente) Y si y sólo si X se parece a Y.

[2] Véase Fodor, "Semantics Wisconsin Style", en *A Theory of Content*

Aquí la "_____" es llenada por la idea del parecido. (Por supuesto, encontramos esta idea inadecuada, pero aquí está precisamente usándose como ejemplo.)

El principio (R) define el concepto de representación *reduciéndolo* a otros conceptos. Por esta razón, puede llamarse una *definición reductiva* del concepto de representación. Muchos filósofos han considerado que las definiciones reductivas dan la naturaleza o esencia de un concepto. Es importante tener conciencia de que no todas las definiciones son reductivas. Para ilustrar esto, tomemos el ejemplo del color. Muchos filósofos naturalistas han querido dar una exposición reductiva del lugar de los colores en el mundo natural. Otros han tratado de formular una definición reductiva de lo que es para un objeto tener determinado color en términos de (digamos) la longitud de onda de la luz que refleja. Así, podrían expresar semejante definición como sigue:

1. X es rojo si y sólo si X refleja luz de longitud de onda *N,* donde *N* es algún número.

Hay un debate fascinante entre si los colores pueden ser definidos reductivamente de este modo (algo así).[3] Mi interés del momento no es la teoría del color, sólo quiero usarla para ilustrar un punto acerca de la definición. Pues algunos filósofos piensan que es un error perseguir una definición reductiva del color. Piensan que lo más que realmente podemos esperar es una definición del color en términos de cómo ven las cosas los que lo perciben, siempre que sean normales. Por ejemplo:

and Other Essays, p. 32. Nótese que Fodor después ("A Theory of Content") debilita el requisito a sólo una condición suficiente.

[3] Véase C. L. Hardin, *Color for Philosophers* (Indianapolis, Hackett, 1988).

2. X es rojo si y sólo si X se ve rojo por los percibidores normales en circunstancias normales.

Ésta no es una definición plenamente reductiva, porque el ser rojo no se define en otros términos: el lado derecho de la definición menciona *verse rojo.* Algunos filósofos piensan algo parecido acerca de la noción de representación o contenido; no debemos esperar conseguir definir el concepto de representación en otros términos. Volveré a esto al final del capítulo.

Definiciones conceptuales y naturalistas

El ejemplo del color sirve para ilustrar otro punto acerca de las definiciones en términos de condiciones necesarias y suficientes. Una razón por la cual uno podría preferir 2 (la definición no reductiva de ser rojo) a 1 es que 2 no va más allá de lo que *sabemos* cuando comprendemos el concepto del color rojo. En cuanto entendemos el concepto de rojo, podemos entender que las cosas rojas se ven rojas por los percibidores normales en circunstancias normales, y que las cosas que se ven rojas por los percibidores normales en circunstancias normales son rojas. A fin de comprender el concepto de rojo, no necesitamos saber nada acerca de longitudes de onda de la luz o de reflexión. De modo que 1 nos dice más de lo que sabemos cuando sabemos el concepto.

Podemos plantear esto diciendo que 2, a diferencia de 1, intenta dar condiciones *conceptualmente* necesarias y suficientes para ser rojo. Da estas condiciones que en algún sentido "definen el concepto" de rojo. Por otra parte, 1 no define el concepto de rojo. De fijo hay personas que tienen

el concepto de rojo, que pueden usar el concepto *rojo* y nunca, sin embargo, han oído de longitudes de onda, por no decir que la luz es una radiación electromagnética. En lugar de esto, 1 da lo que podríamos llamar condiciones necesarias y suficientes *naturalistas* de ser rojo: nos dice en términos científicos qué es para algo ser rojo. (Las condiciones naturalistas necesarias y suficientes para ser rojo se llaman a veces condiciones "nomológicas", ya que caracterizan el concepto en términos de leyes naturales, y *nomos* significa "ley" en griego.)

La idea de una condición naturalista necesaria (o suficiente) no debe ser difícil de captar en general. Cuando decimos que usted necesita oxígeno para mantenerse con vida, estamos diciendo que el oxígeno es una condición necesaria para la vida: si usted está vivo, entonces está recibiendo oxígeno. Éste no es, puede sostenerse, parte del *concepto* de vida, porque nada tiene de malo decir que algo *podría* estar vivo de una manera que no requiera oxígeno. Puede tener sentido la idea de que hay vida en Marte sin suponer que haya oxígeno en Marte. Así que la presencia de oxígeno es una condición naturalista necesaria para la vida, más bien que una condición conceptual necesaria.

Algunos filósofos dudan de si habrá ciertas condiciones reductivas conceptualmente necesarias y suficientes, esto es, condiciones que den definiciones conceptuales reductivas de conceptos.[4] Arguyen, inspirados por Quine o Wittgenstein, que aun la clase de ejemplos que se ha usado tradicionalmente para ilustrar la idea de las condiciones conceptuales necesarias y suficientes es problemática. Tómese el famoso ejemplo de Quine del concepto *soltero*. Parece sumamente

[4] Fodor es uno: véase, por ejemplo, *A Theory of Content*, p. x.

plausible, al principio, que el concepto de un soltero sea el de un hombre que no se ha casado, por ponerlo en términos de condiciones necesarias y suficientes:

X es un soltero si y sólo si X es un hombre no casado.

Esto parece razonable, hasta que consideramos algunos casos raros. ¿Tiene un soltero que ser un hombre que nunca se ha casado o puede aplicarse la expresión a alguien que está divorciado o ha enviudado? ¿Qué pasa con un joven varón de 15 años, es un soltero, o hay que haber rebasado cierta edad?, de ser así ¿qué edad? ¿Es el papa un soltero o una vocación religiosa evita su inclusión? ¿Era Jesús un soltero? ¿O el concepto sólo se aplica a hombres en ciertos tiempos y ciertas culturas?

Por supuesto, podemos siempre legislar que los solteros son todos aquellos hombres de más de 25 años de edad que nunca han estado casados y que no pertenecen a ninguna orden religiosa... y así sucesivamente, como queramos. No obstante, el punto es que *estamos* legislando, estamos tomando una nueva decisión, y así yendo más allá de lo que sabemos cuando conocemos el concepto. La sorprendente verdad es que el concepto no nos informa, por sí mismo, dónde trazar la frontera alrededor de todos los solteros. La argumentación dice que, como muchos conceptos (tal vez la mayoría) son así, empieza a parecer imposible dar condiciones conceptuales necesarias y suficientes, informativas, acerca de estos conceptos.[5]

[5] Para esta clase de escepticismo, véase Stephen Stich, "What is a Theory of Mental Representation?", *Mind,* 101 (1992), y Michael Tye, "Naturalism and the mental", *Mind,* 101 (1992).

Ahora no quiero entrar en este debate acerca de la naturaleza de los conceptos. Menciono la cuestión sólo para ilustrar una manera como puede uno sentir sospechas ante la idea de condiciones conceptualmente necesarias y suficientes, que sean también reductivas. La idea es que es bastante difícil obtener tales condiciones para un concepto bien sencillo, como el de *soltero,* ¿cuánto más difícil será para los conceptos como *representación mental*?

Muchos filósofos han sacado la conclusión de que si queremos definiciones reductivas debemos buscar más bien las condiciones naturalistas necesarias y suficientes para el concepto de representación mental. La "_____" en nuestro principio (R) se llenaría con una descripción de los hechos naturales (por ejemplo, hechos físicos, químicos o biológicos) que sustentan la representación. Éstas serían condiciones naturalistas reductivas necesarias y suficientes para la representación.

¿Cuáles podrían ser estas condiciones? Jerry Fodor ha dicho que sólo se han propuesto dos opciones, nunca seriamente: parecido y causalidad.[6] Esto es, o bien la "_____" se llena con alguna pretensión acerca de que X se parece a Y de alguna manera, o se llena con alguna pretensión acerca de la relación causal entre X y Y. Por supuesto, puede haber otras posibilidades para teorías reductivas de la representación, pero Fodor tiene indudablemente razón en que el parecido y la causalidad han sido las ideas principales a las que en verdad han recurrido los filósofos naturalistas. En el capítulo I discutí, e hice a un lado, las teorías de semejanza de la representación pictórica. Una teoría del parecido para otras clases de representación (por ejemplo palabras) parece

[6] "Semantics, Wisconsin Style", en *A Theory of Content,* p. 33.

aún menos plausible, y la idea de que toda representación puede ser explicada en términos de representación pictórica es, según vimos, cosa sin esperanza. Así, la mayor parte del resto de este capítulo esbozará los elementos de la alternativa principal: teorías causales de la representación.

Teorías causales de la representación mental

En cierto modo, es obvio que los filósofos naturalistas tratarían de explicar la representación mental en términos de causalidad. Pues parte del naturalismo es lo que estoy llamando imagen causal de estados de la mente: la mente se ajusta al orden causal del mundo y su comportamiento es cubierto por la misma clase de leyes causales que otras cosas de la naturaleza (véase el capítulo II). La cuestión que hemos estado encarando en favor de los naturalistas es: ¿cómo se ajusta a todo esto la representación mental? Es casi evidente que deben contestar que la representación es a fin de cuentas una relación causal o, más precisamente, *se basa en* ciertas relaciones causales.

De hecho, parece que el sentido común ya reconoce un sentido en el cual la representación o significado puede ser un concepto causal. H. P. Grice advirtió que el concepto de significado se usa de modos muy diferentes en las siguientes oraciones:[7]

a) Una luz roja significa *alto.*
b) Esas manchas significan sarampión.

[7] Véase H. P. Grice, "Meaning", *Philosophical Review,* 66 (1957).

Es una tautología que el hecho de que una luz roja signifique *alto* es materia de convención. No hay nada acerca del color rojo que lo conecte con detenerse. El color ámbar hubiera servido igual. Por otra parte, el hecho de que las manchas "signifiquen" sarampión no es materia de convención. A diferencia de la luz roja, *hay* algo en las manchas que las conecta con el sarampión. Las manchas son síntomas del sarampión, y a causa de esto se pueden usar para identificar la presencia del sarampión. Las luces rojas, en cambio, no son síntomas de detenerse. Las manchas son, si quiere usted, signos naturales o representaciones naturales del sarampión: *representan* la presencia de sarampión. Análogamente, decimos que el "humo significa fuego", "esas nubes significan tormenta", y lo que queremos decir es que el humo y las nubes son signos naturales (o representaciones) del fuego y la tormenta. Grice llamó a este tipo de representación "significado natural".

El significado natural no es sino una clase de correlación causal. Precisamente como las manchas son los efectos del sarampión, el humo es un efecto del fuego y las nubes son efectos de una causa que es también la de la tormenta. Las nubes, el humo y las manchas están todos *correlacionados* causalmente con las cosas que decimos que "significan": la tormenta, el fuego y el sarampión. Algunas teorías causales de la representación mental opinan que las correlaciones causales entre los pensamientos y las cosas que representan pueden formar la base natural de la representación. ¿Cómo, exactamente?

Sería por supuesto demasiado sencillo decir que X representa Y cuando, y sólo cuando, Y causa X. (Esto es lo que Fodor llama la "teoría causal cruda".)[8] Puedo tener pensa-

[8] *Psychosemantics,* cap. 4.

mientos acerca de borregos, pero ciertamente no es verdad que cada uno de estos pensamientos sea causado por borregos. Cuando un niño se duerme por la noche contando borregos, estos pensamientos acerca de los borregos no necesitan ser causados por éstos. A la inversa, no tiene que ser verdad que, cuando un estado mental es causado por un borrego, representará un borrego. En una noche oscura un borrego asustado podría espantarme, pero lo haría porque represento el borrego como un perro, o un fantasma.

En ambos casos, lo que falta es la idea de que hay algún enlace *natural* y/o *regular* causal entre los borregos y los pensamientos en cuestión. Es mera convención la que asocia *borregos* con el deseo de dormirse uno, y es un mero accidente que un borrego me cause espanto. Si la representación mental va a basarse en correlación causal, deberá hacerlo en regularidades naturales —como con el humo y el fuego—, no simplemente en una conexión causal sola.[9]

Introduzcamos una expresión técnica estándar para esta clase de regularidad natural: denominemos *indicación confiable* la relación entre X y Y, cuando X es un signo natural de Y. En general, X indica de manera confiable Y cuando hay un enlace causal confiable entre X y Y. Así, el humo confiablemente indica fuego, las nubes confiablemente indican tormenta, y las manchas confiablemente indican sarampión. Nuestro siguiente intento en pos de una teoría de la representación puede entonces ponerse así:

[9] Para este punto, véase Fred Dretske, *Knowledge and the Flow of Information* (Cambridge, MIT Press, 1981), p. 76, y "Misrepresentation", en R. Bogdan (ed.), *Belief* (Oxford, Oxford University Press, 1985), p. 19.

X representa Y si y sólo si X confiablemente indica Y.

Aplicado a estados mentales, podemos decir que un estado mental representa Y si y sólo si hay una correlación causal entre este tipo de estado mental y Y.

Una dificultad inicial evidente es que podemos tener muchas clases de pensamiento que no están *causalmente* correlacionadas con nada en absoluto. Puedo pensar acerca de unicornios, acerca de Santa Claus y acerca de otras cosas no existentes, pero estas "cosas" no pueden hacer nada, ya que no existen. También puedo pensar en números y en otras entidades matemáticas tales como conjuntos y funciones, pero, aun si estas cosas existen, no pueden causar nada porque ciertamente no existen en el espacio y el tiempo. (Una causa y sus efectos deben existir en el tiempo si uno va a preceder a la otra.) Y, finalmente, puedo pensar que los sucesos en el porvenir no pueden causar nada en el presente porque las causas deben preceder a sus efectos. ¿Cómo pueden las teorías causales de la representación vérselas con estos casos?

Los teóricos de la causalidad normalmente tratan esta clase de casos como de algún modo especial, y el resultado de los mecanismos productores de pensamiento, muy complicados, que poseemos. Tomemos las cosas con calma, dirán: empiécese con los casos sencillos, los pensamientos básicos acerca del medio percibido, los impulsos básicos (hacia la comida, la bebida, el sexo, el calor, etc.). Si podemos explicar los poderes representacionales de estos estados en términos de una noción como la indicación, entonces también podemos tratar los casos complejos y ocuparnos de ellos después. Al fin y al cabo, si *no podemos* explicar los casos sencillos en términos de nociones como la indicación,

no tendremos gran suerte con los casos complejos. Por lo tanto, no tiene objeto comenzar con este tipo de casos.

Las ventajas de una teoría causal de la representación mental para los filósofos naturalistas son patentes. La indicación confiable está por doquier: siempre que hay esta clase de correlación causal, hay indicación. Así, como la indicación no es un fenómeno misterioso ni único para la mente, sería un claro progreso si pudiéramos explicar la representación mental en términos de ella. Si la sugerencia funciona, entonces estaríamos en camino de explicar cómo la representación mental es constituida por relaciones causales naturales y, a fin de cuentas, cómo la representación mental se ajusta al mundo natural.

El problema del error

Sin embargo, la ubicuidad de la indicación también presenta algunos de los problemas máximos para el enfoque causal. Por un lado, *a)* como las representaciones siempre indicarán *algo,* es difícil ver cómo pueden incluso representarse mal. Por otro, *b)* hay muchos fenómenos que están correlacionados causal y confiablemente con representaciones mentales, y sin embargo no son en ningún sentido los puntos representados por ellos. Estos dos problemas están vinculados, ambos son rasgos del hecho de que las teorías causales de la representación pasen un mal rato dando razón de *errores* en el pensamiento. Esto requerirá una pequeña explicación.

Tómese el primer problema *a).* Considérese de nuevo el ejemplo de Grice del sarampión. Dijimos que las manchas representan sarampión porque son indicadores confiables

de sarampión. En general, si no hay manchas, entonces no hay sarampión. ¿Es verdad lo contrario, puede haber manchas sin sarampión? Esto es como decir, ¿pudieron las manchas representar *mal* el sarampión? Pues bien, alguien puede tener manchas análogas, porque tiene alguna otra clase de enfermedad; varicela, por ejemplo. Sin embargo, *estas* manchas serían entonces indicadores de varicela. Así, la teoría tendría que decir que no representan mal al sarampión; representan lo que indican, a saber, varicela.

Por supuesto, *nosotros* podríamos cometer un error y, mirando las manchas de la varicela, concluir: ¡sarampión! Esto no viene al caso. La teoría pretende explicar las capacidades representacionales de nuestras mentes en términos de indicación confiable; con esta teoría no podemos recurrir a la interpretación que *nosotros* damos de un fenómeno al explicar lo que representa. Esto pondría de cabeza las cosas, para mal.

El problema es que, porque lo que X representa es explicado en términos de indicación confiable, X no puede representar algo que no indica. Grice subrayó el asunto observando que, donde se trata de significado natural, *X significa que p* acarrea *p* (el humo que significa fuego hace que haya fuego). En general, parece que, cuando X significa naturalmente Y, esto garantiza la existencia de Y, pero pocas representaciones mentales garantizan la existencia de aquello que representan. Es innegable que nuestros pensamientos pueden representar algo como este caso, incluso cuando no sea así: el error en la representación mental es posible. Así, una teoría de la representación que no permita el error nunca podrá formar la base de representación mental. A falta de una expresión mejor, llamemos a esto el "problema de la representación equivocada".

Este problema está estrechamente ligado al otro problema para la teoría de la indicación, que se conoce (por razones que explicaré) como el "problema de la disyunción". Supóngase que soy capaz de reconocer borregos, soy capaz de percibir borregos cuando hay borregos a la vista. Mis percepciones de los borregos son representaciones de algún género —llamémoslo "representación S" para abreviar— y son indicadores confiables de borregos, y por lo tanto la teoría dice que representan borregos. Hasta aquí, todo va bien.

Supóngase también que, en ciertas circunstancias —digamos, a distancia, con mala luz—, soy incapaz de distinguir borregos y cabras. Y supóngase que esta conexión es plenamente sistemática: hay una conexión confiable entre cabras-en-ciertas-circunstancias y percepciones de borregos. Tengo una representación S cuando veo una cabra. Éste parece ser un caso claro de representación errónea: mi representación S malrepresenta una cabra como un borrego. No obstante, si mis representaciones S son indicadores confiables de cabras-en-ciertas-circunstancias, entonces ¿por qué no habríamos de decir en lugar de esto que representan cabras-en-ciertas-circunstancias así como borregos? De hecho, seguramente la teoría de la indicación *tendrá* que decir algo como esto, ya que se supone que la indicación confiable sola es la fuente de la representación.

El problema, pues, es que tanto borregos como cabras-en-ciertas-circunstancias son indicados confiablemente por representaciones S. Así que parece, diríamos, que una representación S representa que hay presente un borrego *o* una cabra-en-ciertas-circunstancias. El contenido de la representación, entonces, serían *borregos o cabras-en-ciertas-circunstancias.* Esto se llama el "problema de la disyunción"

porque los lógicos llaman *disyunción* al enlace entre dos o más términos con un "o".[10]

En caso de que usted piense que esta clase de ejemplo es nada más una fantasía filosófica, considérese este otro de etología cognitiva en la vida real. Los etólogos D. L. Cheney y R. M. Seyfarth han estudiado los llamados de alarma de los monos, y han conjeturado que diferentes tipos de llamado tienen distintos significados, dependiendo de por qué sea provocado el llamado particular. Un tipo particular de llamado, por ejemplo, es producido en presencia de leopardos, y, así, lo etiquetan como una "alarma por leopardo". Pero:

> El significado del llamado por leopardo es, desde el punto de vista del mono, sólo tan preciso como hace falta que sea. En Amboseli, donde los leopardos cazan monos pero los leones y guepardos no lo hacen, la alarma por leopardo pudiera significar "gran gato moteado que no es un guepardo" o "gran gato con patas más cortas" [...] En otras áreas de África, donde los guepardos cazan monos, el llamado por leopardo pudiera significar "leopardo o guepardo".[11]

Estos etólogos están muy contentos atribuyendo contenido disyuntivo a los llamados por leopardo de los monos. El problema de la disyunción surge cuando preguntamos qué sería representar mal un guepardo como leopardo.

[10] Para el problema de la disyunción, véase Fodor, *A Theory of Content*, cap. 3, especialmente pp. 59 y ss.; Papineau, *Philosophical Naturalism* (Oxford, Blackwell, 1993), cap. 3, pp. 58-59.

[11] D. L. Cheney y R. M. Seyfarth, *How Monkeys See the World: Inside the Mind of Another Species* (Chicago, University of Chicago Press, 1990), p. 169. Agradezco a Pascal Ernst por este ejemplo.

Decir que el significado del llamado es "sólo tan preciso como hace falta que sea" no responde a esta cuestión sino que la evita.

Permítaseme resumir la estructura de los dos problemas. El problema de la malrepresentación es que, si se supone que la indicación confiable es una condición necesaria de la representación, entonces X no puede representar Y en ausencia de Y. Si es una condición necesaria que algunas manchas representen sarampión e indiquen sarampión, entonces las manchas no pueden representar sarampión en ausencia de sarampión.

El problema de la disyunción es que, si se supone que la indicación confiable es una condición suficiente de la representación, entonces lo que X indique será representado por X. Si es una condición suficiente para una representación S el representar un borrego que confiablemente indique borrego, entonces será también una condición suficiente para una representación S representar una cabra-en-ciertas-circunstancias que indica una cabra-en-ciertas-circunstancias. Lo que sea indicado por una representación es representado por ella: así, el contenido de la representación S será *borregos o cabras-en-ciertas-circunstancias.*

Evidentemente, los dos problemas están relacionados. Ambos son aspectos del problema de que, según la teoría de la indicación, el error no es realmente posible.[12] El problema de la mala representación hace imposible el error *descartando* alguna situación (sarampión) en que la situación no existe. El problema de la disyunción, sin embargo, hace imposible el error *regulando* la representación de demasiadas situaciones (borregos-o-cabras). En ambos casos la teoría de

[12] Fodor, *A Theory of Content,* p. 90, adopta un punto de vista diferente.

la indicación da una respuesta equivocada a la pregunta "¿qué representa esta representación?"

¿Cómo puede la teoría de la indicación responder a estos problemas? La manera ordinaria de responder es sostener que, cuando algo se representa mal, eso significa que las condiciones de representación (ya sea dentro o fuera del organismo) no son perfectas: como Robert Cummins lo plantea, la mala representación es mal funcionamiento.[13] Cuando las condiciones son ideales, entonces no habrá ninguna falla que representar: las manchas representarán sarampión en condiciones ideales, y mis representaciones S representarán un borrego (y no una cabra) en condiciones ideales.

La idea, pues, es que la representación es definible como indicación confiable en condiciones ideales:

X representa Y si y sólo si X es un indicador confiable de Y en condiciones ideales.

El error resulta de que las condiciones no llegan a ser ideales de alguna manera: mala luz, distancia, defecto de los órganos sensoriales, etc. (Las condiciones ideales se llaman a veces condiciones "normales".) ¿Cómo caracterizaríamos, en general, cuáles son las condiciones ideales? Evidentemente no podemos decir que las condiciones ideales son aquellas en que ocurre la representación, pues de otro modo nuestra exposición será circular y no informativa:

X representa Y si y sólo si X indica confiablemente Y en esas condiciones en las cuales X representa Y.

[13] Para uno de los enunciados originales de esta idea, véase Dennis Stampe, "Toward a Causal Theory of Linguistic Representation". Para

Lo que necesitamos es una manera de especificar condiciones ideales sin mencionar la representación.

Fred Dretske, uno de los precursores del enfoque de la indicación, trató de resolver este problema recurriendo a la idea de la *función teleológica* de una representación.[14] Éste es un sentido diferente de "función", ante la noción matemática descrita en el capítulo III: "teleológica" significa "dirigida a la meta". Las funciones teleológicas son normalmente atribuidas a mecanismos biológicos, y las explicaciones teleológicas son explicaciones en términos de funciones teleológicas. Un ejemplo de una función teleológica es la función de bombeo de la sangre por el cuerpo debido al corazón. La idea de función es útil aquí porque *a)* es una noción que entiende bien la biología y *b)* es generalmente aceptado que algo puede tener una función teleológica aun si no está ejerciéndola: es la función del corazón bombear sangre por el cuerpo aun cuando no esté haciéndolo realmente. Así la idea es que X puede representar Y, aun cuando Y no esté a mano, precisamente en este caso es *función* de X indicar Y. Las condiciones ideales son por lo tanto de "buen funcionamiento":[15] cuando todo está funcionando como debiera.

Esto sugiere cómo el recurso a las funciones teleológicas puede ocuparse en lo que vengo llamando problema de la

una excelente discusión crítica, véase Cummins, *Meaning and Mental Representation,* pp. 40 y ss.

[14] Véase "Misrepresentation". Para la idea general de una función teleológica, véase Karen Neander, "The Teleological Notion of 'Function'", *Australasian Journal of Philosophy,* 69 (1991), y David Papineau, *Philosophical Naturalism,* cap. 2.

[15] La expresión es de Stampe; véase "Toward a Causal Theory of Linguistic Representation", especialmente pp. 51-52.

mala representación. X puede representar Y si tiene la función de indicar Y; y puede tener la función de indicar Y aun si no hay Y a mano. Hasta en la oscuridad mis ojos tienen la función de indicar la presencia de objetos visibles. Hasta aquí, muy bien, pero ¿puede esta teoría vérselas con el problema de la disyunción?

Algunos filósofos, incluyendo a Fodor (quien inicialmente favoreció esta clase de enfoque), han sostenido que no es posible. El problema es que algo muy parecido al problema de la disyunción se aplica a las funciones teleológicas también. El problema está bien ilustrado por un hermoso ejemplo de Dretske:

> Algunas bacterias marinas tienen magnetos internos (llamados magnetosomas) que funcionan como agujas de brújula, alineándose (y como resultado las bacterias) paralelamente al campo magnético terrestre. En vista de que estas líneas magnéticas se inclinan hacia abajo (hacia el norte geográfico) en el hemisferio norte (hacia arriba en el hemisferio sur) las bacterias del hemisferio norte... se mueven hacia el norte geomagnético. El valor de supervivencia de la magnetotaxia (según se llama este mecanismo sensorial) no es evidente, pero es razonable suponer que funciona a fin de permitir a las bacterias que eviten el agua superficial. Como estos organismos son capaces de vivir sólo en ausencia de oxígeno, el movimiento hacia el norte geomagnético apartará a las bacterias del agua superficial rica en oxígeno y hacia el sedimento del fondo, comparativamente libre de oxígeno.[16]

[16] Dretske, "Misrepresentation", p. 26.

Convengamos en que el mecanismo del organismo tiene una función teleológica. ¿Qué función tiene? ¿Es su función *impulsar a la bacteria hacia el norte geomagnético* o es *impulsar a la bacteria a la ausencia de oxígeno*? Por un lado, el mecanismo es él mismo un *magneto;* por otra parte, el objeto de tener dentro el magneto es que el organismo llegue a áreas libres de oxígeno.

Tal vez tiene ambas funciones. Sin embargo, como no necesita tener ambas juntas, debemos realmente decir que tiene la función compleja que pudiéramos describir como "impulsar a la bacteria hacia el norte geomagnético *o* impulsar a la bacteria hacia la ausencia de oxígeno". Y he aquí por qué podemos ver que las funciones teleológicas tienen el mismo género de "problemas disyuntivos", igual que la indicación. Como dicen algunos, las funciones teleológicas están sometidas a cierta "indeterminación": es literalmente indeterminado qué función tiene algo. Si esto es cierto, entonces no podemos usar la idea de la función teleológica para resolver el problema de la disyunción, ya que la representación es ella misma determinada.

Por esta razón, algunos teóricos causales se han apartado de las funciones teleológicas. Notable entre ellos es Fodor, que ha defendido una teoría causal no teleológica de la representación mental, que él llama la teoría de la "dependencia asimétrica".[17] Examinémosla brevemente. (Los principiantes pueden desear saltarse la sección siguiente.)

[17] La teoría fue propuesta primero en *Psychosemantics,* cap. 4, y luego refinada en *A Theory of Content,* cap. 4. Para una discusión, véase Cummins, *Meaning and Mental Representation,* cap. 5, y los ensayos en George Rey y Barry Loewer (eds.), *Meaning in Mind* (Oxford, Blackwell, 1991).

Supóngase que hay algunas circunstancias en las cuales (por volver a nuestro ejemplo) los borregos causan que tengamos representaciones S. Fodor observa que, si hay condiciones en las cuales las cabras-en-ciertas-circunstancias también causan que tengamos representaciones S, tiene sentido suponer que las cabras hacen esto porque los *borregos* ya causan representaciones S. Si bien tiene sentido suponer que sólo los borregos podrían causar representaciones de borregos, Fodor piensa que no lo tiene suponer que sólo las cabras podrían causar representaciones de borregos. Puede sostenerse, en tal caso, que las representaciones S serían representaciones de *cabras,* no de borregos en absoluto. Decir que el enlace causal cabra-a-representación S es un error, pues es decir que las cabras no causarían representaciones S a menos que lo hicieran los borregos. Sin embargo, los borregos causarían todavía representaciones S aun si las cabras no lo hicieran.

Tal vez resulta más fácil captar el asunto en el contexto de la percepción. Supóngase que algunas de mis percepciones-de-borregos fueran causadas por borregos. Pero algunas cabras tienen el aire de borregos, esto es, algunas de mis percepciones de cabras (es decir aquellas *causadas* por cabras) me parecen percepciones de borregos. Las percepciones causadas por cabras no parecerían percepciones de borregos *a menos* que las percepciones causadas por borregos también parecieran como percepciones de borregos. Y lo contrario no es el caso; es decir, las percepciones causadas por borregos seguirían pareciendo percepciones de borregos aun si no hubiera percepciones de borregos causadas por cabras.

Fodor expresa esto diciendo que la relación causal entre cabras y representaciones de borregos es *asimétricamente*

dependiente de la relación causal entre borregos y representaciones de borregos. ¿Qué significa este término técnico? Abreviemos "causa" con una flecha, →, y abreviemos "representación de borregos" con mayúsculas, BORREGOS. También ayudará si subrayamos las pretensiones causales que se hacen. Fodor dice que la relación causal cabra→BORREGO es *dependiente* de la relación causal borregos→BORREGOS en el siguiente sentido:

> Si no hubiese habido una conexión borregos→BORREGOS, entonces no hubiera habido una conexión cabra→BORREGOS.

La conexión cabra→BORREGOS es *asimétricamente* dependiente de la conexión borregos→BORREGOS porque:

> Si no hubiera habido una conexión cabra→BORREGOS, seguiría habiendo una conexión borregos→BORREGOS.

Por lo tanto, hay una dependencia entre la conexión cabra→BORREGOS y la conexión borregos→BORREGOS, pero no es simétrica.

Hay dos puntos que merecen señalarse acerca de la teoría de Fodor. Primero, el papel que la idea de la dependencia asimétrica desempeña es sencillamente contestar el problema de la disyunción. Fodor queda esencialmente feliz con teorías de la representación por indicación: piensa nada más que usted necesita algo como la dependencia asimétrica para manejar el problema de la disyunción. Así, evidentemente, si tiene usted alguna otra manera de vérselas con tal problema —o si tiene una teoría en la cual tal problema no se presente—, entonces no tiene usted que encarar la

cuestión de si la dependencia asimétrica ofrece un caso de representación mental.

Segundo, Fodor propone la dependencia asimétrica sólo como una condición *suficiente* de representación mental. Esto es, sostiene solamente que *si* estas condiciones (indicación y dependencia asimétrica) se dan entre X y Y, entonces X representa Y. No afirma que *cualquier* posible clase de representación mental deba exhibir la estructura de dependencia asimétrica, sino que si algo exhibe de hecho esta estructura, entonces es una representación mental.

Por lo que a mí toca, soy incapaz de ver cómo la dependencia asimétrica sea algún avance hacia la *explicación* de la representación mental. Creo que las condiciones que Fodor describe probablemente sean verdad sobre las representaciones mentales. Pero no veo cómo esto nos proporcione una comprensión más profunda de cómo funciona realmente la representación mental. En efecto, Fodor dice: el error es parásito en la creencia cierta. Es difícil no objetar que esto es, ni más ni menos, lo que sabíamos ya. La pregunta es más bien: *¿qué es el error?* Hasta que podamos dar alguna razón del error, no nos ayuda realmente decir que es parásito de la creencia cierta. Fodor, por supuesto, ha respondido a quejas como ésta, pero tal vez valga la pena buscar un enfoque diferente.

Representación mental y éxito en la acción

En los términos más generales, las teorías causales de la representación mental que he esbozado hasta aquí tratan de identificar el contenido de una creencia —lo que representa— con su causa. Y, vistas así las cosas, es evidente por qué

esta teoría habría de encontrar el problema del error: si cada creencia tiene una causa, y el contenido de toda causa es aquello que la cause, entonces toda creencia representará correctamente su causa, más bien que (en algunos casos) representar incorrectamente algo distinto.

Sin embargo, hay otra manera de abordar el asunto. En vez de concentrarnos en las *causas* de las creencias, como hacen las teorías de la indicación, podemos concentrarnos en los *efectos* que tienen sobre el comportamiento. Según vimos en el capítulo II, lo que hace usted es causado por aquello en que usted cree (es decir, cómo juzga usted que es el mundo) y por lo que usted quiere. Quizá la base causal de la representación no pueda hallarse sencillamente entre las causas de estados mentales, sino entre sus efectos. La reducción de la representación debiera ser vista no nada más como las *entradas* en los estados mentales, sino como sus *salidas.*

Aquí hay una idea siguiendo esta línea, cuyos elementos ya encontramos en el capítulo II. Cuando actuamos, intentamos alcanzar alguna meta o satisfacer algún deseo. Y *qué* deseamos depende en parte de cómo pensemos que son las cosas; si piensa usted que no ha tomado un poco de vino, puede usted desear *vino;* pero si piensa usted que ha tomado algo de vino, puede usted desear *más vino.* Esto es, desear *vino* y desear *más vino* son evidentemente diferentes tipos de deseo: puede usted desear más vino sólo si piensa que ya ha tomado algo de vino. Ahora, el que *logre* sus intentos de obtener lo que desea, dependerá de si el modo como considera usted que son las cosas —creencia suya— es la misma que la manera como son las cosas. Si quiero algo de vino, y creo que hay un poco de vino en el refrigerador, entonces el que consiga vino yendo al refrigerador dependerá de si esta

creencia es correcta: esto es, dependerá de si hay vino en el refrigerador.

(El éxito de la acción —ir al refrigerador— dependerá también de otras cosas, tales como si existe el refrigerador, y si puedo mover mis miembros. Podemos ignorar estos factores en el momento, ya que podemos suponer que mi creencia de que hay vino en el refrigerador implica la creencia de que el refrigerador existe, y que normalmente no probaría mover mis miembros a menos que pensara que podía. Así, el fracaso en estas cosas implicaría fracaso en estas otras creencias.)

Hasta aquí, la idea general debe de ser bastante clara: si nuestras acciones tienen éxito al satisfacer nuestros deseos es cosa que depende de si nuestras creencias representan el mundo correctamente. Es difícil objetar esta idea, excepto tal vez a causa de su vaguedad. Sin embargo, es posible convertir la idea en parte de la definición del contenido representacional de la creencia. La idea es ésta. Una creencia dice que el mundo es de cierta manera: que hay vino en el refrigerador, por ejemplo. Esta creencia puede o no ser correcta. Desconociendo por el momento las complicaciones mencionadas en el párrafo anterior, podemos decir que, si la creencia es correcta, entonces las acciones causadas por ellas más algún deseo (por ejemplo el deseo de vino) *lograrán* satisfacer dicho deseo. Así, las condiciones en que la acción tiene éxito son precisamente aquellas especificadas por el contenido de la creencia: el modo como la creencia dice que es el mundo. Por ejemplo, las condiciones bajo las cuales mi intento de conseguir vino se logran son precisamente aquellas especificadas por el contenido de mi creencia: hay vino en el refrigerador. En forma de eslogan: el contenido de una creencia es idéntico a las "condiciones de éxito" de las accio-

nes que causa. Llamemos a esto la "teoría del éxito" del contenido de creencia.[18]

La teoría del éxito nos ofrece así una manera de reducir el contenido representacional de las creencias. Recuérdese la forma de una explicación reductiva de la representación:

(R) X representa Y si y sólo si _____.

La idea era llenar el "_____" sin mencionar la idea de representación. La teoría del éxito hará esto de una manera aproximadamente así:

Una creencia B representa la condición C si y sólo si las acciones causadas por B tienen éxito cuando se da C.

Aquí el "_____" es llenado de una manera que, vista así, no menciona la representación: sólo menciona acciones causadas por creencias, el éxito de estas acciones y las condiciones que se dan en el mundo.[19]

Una objeción inicial evidente es que muchas creencias no causan acciones de ningún género. Creo que el actual primer ministro del Reino Unido no lleva bigote. Esta creencia nunca ha provocado que yo haga nada hasta ahora, pues ¿qué acciones sería posible que causara?

La pregunta es fácil de contestar, si nos concedemos su-

[18] La teoría ha sido defendida por J. T. Whyte, "Success Semantics", *Analysis,* 50 (1990), y David Papineau, *Philosophical Naturalism.* Las simientes de la idea están en F. P. Ramsey, "Facts and Propositions", *Philosophical Papers,* y fueron desarrolladas por R. B. Braithwaite, "Belief and Action", *Proceedings of the Aristotelian Society,* volumen suplementario 20 (1946).

[19] Compárese Robert Stalnaker, *Inquiry,* cap. 1.

ficiente imaginación. Imagínese, por ejemplo, estar en una exhibición donde le han preguntado a usted por gobernantes actuales del mundo, sin bigotes. Su acción (dar el nombre del primer ministro actual) tendría éxito si la condición representada por la creencia de usted —que el actual primer ministro no usa bigote— se da. Lo que importa es que siempre es *posible* pensar en alguna situación en que una creencia podría desembocar en acción. Sin embargo, esto significa que tenemos que revisar nuestra definición de la teoría del éxito, para incluir situaciones posibles. Un sencillo cambio del indicativo al subjuntivo puede lograrlo:

> Una creencia B representa la condición C si y sólo si acciones que *serían* causadas por B *tendrían* éxito en caso de que se diera C.

Esta formulación daría la idea general de lo que afirma la teoría del éxito.

Hay una dificultad general concerniente a la definición de la idea clave de *éxito*. ¿A qué equivale de hecho el éxito de una acción? Según presenté la teoría antes, es el hecho de que la acción satisfaga el deseo que en parte lo causa. Mi deseo es vino: creo que hay vino en el refrigerador; esta creencia y deseo conspiran para hacerme ir al refrigerador. Mi acción tiene éxito si encuentro vino, esto es, si mi deseo es satisfecho. De este modo habríamos de llenar la definición de la teoría como sigue:

> Una creencia B representa condiciones C si y sólo si acciones que serían causadas por B y un deseo D satisficieran D si se diera C.

Aunque un poco más complicada, ésta sigue siendo una definición reductiva: la idea de representación no aparece en la parte de la definición después de "si y sólo si".

Sin embargo, podríamos todavía preguntarnos cuál es la satisfacción del deseo.[20] No puede sencillamente ser la *cesación* de un deseo, porque hay demasiadas maneras en las que un deseo puede cesar y que no sean modos de satisfacer el deseo. Mi deseo de vino puede cesar si repentinamente deseo alguna otra cosa, o si se cae el techo, o si muero. Éstas no son maneras de satisfacer el deseo. Ni puede la satisfacción de mi deseo ser cuestión de mi *creencia* de que el deseo se ha satisfecho. Si usted me hipnotiza para que crea que he bebido algo de vino, realmente no ha satisfecho mi deseo. Pues no conseguí lo que quería, a saber, vino.

No: la satisfacción de mi deseo de vino es un asunto de armar un estado de cosas en el mundo. ¿Qué estado de cosas? La respuesta es obvia: el estado de cosas representado por el deseo. Así, para llenar nuestra definición de la teoría del éxito debemos decir:

> Una creencia B representa la condición C si y sólo si acciones que serían causadas por B y un deseo D provocasen el estado de cosas representado por D si se diera C.

Ahora el problema es evidente: la definición de la representación para las creencias contiene la idea del *estado de cosas representado por un deseo.* La naturaleza representacional de las creencias es explicada en términos de la naturaleza

[20] Véanse los artículos de Whyte, "Success Semantics" y "The Normal Rewards of Success", *Analysis,* 51 (1991).

representacional de los deseos. Hemos vuelto adonde empezamos.[21]

Así, si la teoría del éxito va a ir en pos de su meta de una teoría reductiva de la representación mental, tiene que explicar la naturaleza representacional de los deseos sin emplear la idea de representación. Hay múltiples maneras de hacerlo. Aquí enfocaré la idea de que los estados mentales tienen funciones teleológicas, funciones biológicas específicamente. Llamo a esto la teoría biológica de la representación mental; versiones de la teoría han sido defendidas por Ruth Millikan y David Papineau.[22]

Representación mental y función biológica

La teoría biológica supone que los deseos tienen algún propósito o función evolutivos, esto es, que desempeñan algún papel en el aumento de la supervivencia del organismo, y con ello de la especie. En algunos casos, parece haber una conexión evidente entre ciertos deseos y el valor para la supervivencia de los organismos de la especie. Tómese el deseo de agua. Si organismos como nosotros no consiguen agua, entonces no sobreviven mucho. Así, desde el punto de vista de la selección natural, es claramente buena cosa tener esta-

[21] Este punto fue anticipado por Chisholm, *Perceiving*, cap. 11, nota 13, contra la versión de Braitwaite de la teoría del éxito en su artículo "Belief and Action".

[22] Véase Papineau, *Philosophical Naturalism*, cap. 3, y Ruth Garrett Millikan, *Language, Thought and other Biological Categories* (Cambridge, MIT Press, 1986). En esta sección sigo la versión de Papineau de la teoría, que no es exactamente la misma que la de Millikan, por razones que no tienen que entretenernos.

dos que nos motivan o nos mueven a obtener agua: y esto de seguro forma parte de lo que es el deseo de agua.

Sin embargo, una cosa es decir que los deseos deben de haber tenido algún origen evolutivo, o incluso propósito evolutivo, y otra decir que estos contenidos —lo que representan— pueden explicarse en términos de tales propósitos. La teoría biológica se enfrenta a esto según una línea más radical. Sostiene que la selección natural ha garantizado que estemos en estados cuya función es causar una situación que incrementa nuestra supervivencia. Estos estados son deseos, y las situaciones son entonces contenidos. Así, por ejemplo, conseguir agua aumenta nuestra supervivencia, de modo que la selección natural ha garantizado que estemos en estados que nos inducen (siendo iguales las demás cosas) a conseguir agua. El contenido de estos estados es (algo como) *tengo agua,* porque nuestra supervivencia ha sido incrementada cuando estos estados provocan un estado de cosas en que tengo agua.

El éxito de una acción, pues, tiene que ver con provocar un estado incrementador de la supervivencia en las cosas. Al reducir el contenido representacional de creencias y deseos, la teoría trabaja de "fuera adentro": primero establece qué estados de cosas incrementarán la supervivencia del organismo; entonces encontrará estados cuya función es causar dichos estados de cosas. Éstos son deseos, y representan tales estados de cosas. Así es como se explican las capacidades representacionales de los deseos.

Una vez que tenemos una explicación del poder representacional de los deseos, podemos empalmarla con nuestra explicación de las capacidades representacionales de las creencias. (Así es como funcionan todas las versiones de la teoría biológica: pero es una sugerencia natural.) Recuérdese que

la teoría del éxito las explicaba en términos de la satisfacción de deseos por acciones. No obstante, descubrimos que la satisfacción de deseos implica un recurso tácito a lo que los deseos representan. Esto puede ser explicado ahora en términos de función biológica de los deseos incrementando la supervivencia del organismo. Si funciona esta ingeniosa teoría, entonces claramente nos da una explicación reductiva de la representación mental.

Sin embargo, ¿funciona? La teoría explica el contenido representacional de una creencia dada en términos de las condiciones en las cuales las acciones causadas por la creencia y un deseo consiguen satisfacer el deseo. La satisfacción del deseo se explica en términos del deseo que suscita condiciones que incrementan la supervivencia del organismo. Dejemos aparte un momento la cuestión evidente de que la gente puede tener muchos deseos —por ejemplo, el deseo de *ser famosos saltando desde el puente Golden Gate*—, que claramente tienen poco que ver con aumentar la supervivencia. Recuérdese que la teoría trata de vérselas con nuestros pensamientos y motivaciones más básicos —creencias y deseos acerca de comida, sexo, calor, etc.— y no todavía con estados mentales más rebuscados. Más adelante en este capítulo examinaremos esto un poco más ("Contra la reducción y la definición", p. 315).

Lo que trato de enfocar aquí es una consecuencia evidente de la teoría biológica: si un ser tiene deseos, entonces ha evolucionado. Esto es, la teoría torna condición de algo que tiene deseos que sea el producto de la evolución por selección natural. Pues la teoría dice que un deseo es precisamente un estado en el cual la selección natural ha otorgado cierta función biológica: causar el comportamiento que incrementa la supervivencia del organismo. Si un organismo

está en uno de estos estados, entonces la selección natural ha garantizado que está en él. Si el estado no hubiera sido seleccionado, entonces el organismo no estaría en aquel estado.

El problema con esto es que no parece imposible que hubiera un ser que tuviese pensamientos pero no hubiese evolucionado. Supóngase, por mor de la argumentación, que los pensadores están hechos de materia, que si quita usted toda la materia del pensador no quedaría nada. De seguro es precisamente lo contrario de esto que sí es posible en principio *reconstruir* al pensador, juntar toda su materia otra vez y seguiría siendo un pensador. Y si puede usted reconstruir un pensador, entonces ¿por qué no puede construir otro pensador según el mismo plan? A primera vista parece que la teoría biológica de la representación mental descartaría esta posibilidad. Empero, aunque sea muy improbable, no parece ser absolutamente imposible; de hecho, la coherencia de "teletransportación" del tipo descrito en *Star Trek* parece depender de él.

La teoría biológica no necesita admitir que esto sea imposible. Lo que es medular para la teoría es que los estados del ser tengan una función. Las funciones pueden ser adquiridas de varias maneras. En el caso de un pensador artificialmente creado, la teoría puede decir que sus estados obtienen su función porque son funciones atribuidas por su creador. Así, tal como un corazón artificial puede adquirir una función por ser designado y usado como corazón, así los estados internos de una persona artificial podrían adquirir funcionamiento siendo diseñados y usados como deseos. Estos estados sólo tienen intencionalidad *derivada,* más bien que intencionalidad *original* (véase el capítulo 1, "Intencionalidad"). No obstante, la intencionalidad derivada sigue siendo intencionalidad de cierta clase.

Sin embargo ¿por qué no habría de haber un pensador que no sea designado en absoluto? ¿No podría haber un pensador que viniese a la existencia por accidente? Donald Davidson ha descrito una situación imaginaria en la cual cae un rayo en un pantano y por una pasmosa coincidencia sintetiza las sustancias del pantano para crear una réplica de un ser humano.[23] Esta persona —llamada "pantanombre"— tiene todos los estados físicos y químicos de un ser humano normal; imaginemos que es una copia física de mí. El pantanombre (o panta-yo) no tiene historia evolutiva, es sólo un accidente extraño. Es igual que yo, camina como yo, hace sonidos como yo: pero no ha evolucionado.

¿Tendría el pantanombre algunos estados mentales? Los fisicalistas que creen que los estados mentales están completamente determinados por los estados físicos locales del cuerpo deben responder que sí. De hecho deben decir que en el momento de su creación accidental, el pantanombre tendrá casi *todos* los mismos estados mentales que yo —pensamientos y estados conscientes—, excepto aquellos que, por supuesto, dependen de nuestros diferentes contextos y localizaciones espaciotemporales. La teoría biológica de la representación mental niega que los pantanombres tengan cualquier estado mental representacional, en absoluto, pues para tener estados mentales representacionales un ser debe haber sido producto de la evolución por selección natural. Así, si el pantanombre es un pensador, entonces la teoría

[23] El ejemplo del "swampman" de Davidson está en "Knowing One's Own Mind", reimpreso en Q. Cassam (ed.), *Self-Knowledge* (Oxford, Oxford University Press, 1994). Cummins usa esta objeción contra Millikan y Papineau en *Meaning and Mental Representation,* cap. 7. Véase Millikan, *Language, Thought and other Biological Categories,* p. 94, para la respuesta de ella.

biológica de la representación mental es falsa. Así, la teoría biológica tiene que negar la posibilidad de pantanombres. ¿Cómo pueden negar esta mera posibilidad? He aquí cómo responde David Papineau:

> La teoría aspira a ser una *reducción teórica* de la noción cotidiana de contenido representacional, no como fragmento de *análisis conceptual*. Y como tal puede esperarse que eche por tierra algunos de los juicios intuitivos que estamos inclinados a hacer sobre la base de la noción cotidiana. Considérese, por ejemplo, la reducción teórica de la noción cotidiana de líquido a la noción de estado de la materia en el cual las moléculas se juntan pero no constituyen un orden de largo alcance. Esto claramente no es un análisis conceptual del concepto cotidiano, ya que el concepto cotidiano no presupone nada acerca de estructura molecular. En consecuencia, esta reducción *corrige* algunos de los juicios que se desprenden del concepto cotidiano, tales como el juicio de que el vidrio no es un líquido.[24]

Ya hemos hecho la distinción, en este capítulo, entre definiciones conceptuales y naturalistas, y, según deja en claro esta cita, la teoría biológica está ofreciendo la última. La defensa contra el ejemplo del pantanombre es que nuestros juicios intuitivos acerca de lo que es o no es posible nos desconcierta. Si la teoría de Papineau es correcta, entonces lo que pensamos que era permitido por el concepto ordinario no lo es de hecho. Análogamente, el concepto ordinario de un líquido parece descartar que el vidrio sea un líquido: pero a pesar de todo, lo es.

[24] Papineau, *Philosophical Naturalism*, p. 93.

Esta respuesta puede parecer que negara que el pantanombre sea un pensador y no es sino un efecto lateral contraintuitivo e infortunado de la teoría biológica, que debemos aceptar a causa de otras ventajas explicativas de la teoría. De hecho, la situación es mucho más extrema que eso. De la denegación de que el pantanombre tenga algún pensamiento procede la denegación de que sus mecanismos formadores de creencia tengan alguna función biológica, donde el hecho de que un mecanismo tenga función se entiende en términos de su historia causal real de provocar ciertos efectos que en realidad incrementaban la supervivencia del ser que es su huésped. (Ésta es la llamada lectura "etiológica" de la noción de función biológica.)[25] De modo que sin historia evolutiva real no hay función.

Por supuesto, esta manera de comprender la función biológica no está restringida a lo mental. La noción de función también se aplica a todos los demás órganos biológicos que pasan por tener una función. Así, si el pantanombre no tiene pensamientos; tampoco tiene cerebro, porque un cerebro es definido en términos de sus muchas funciones y, por la concepción etiológica, el cerebro del pantanombre carece de función. Por el mismo razonamiento, el pantanombre no tiene corazón. Y como la sangre es sin duda definida por su función, tampoco tiene sangre. Nada más tiene algo que parece un corazón, que bombea algo que parece sangre por algo que parece un cuerpo humano, manteniendo la actividad de algo que tiene el aire de un cerebro, y generando algo que "parece" pensamiento. De hecho, ¿por qué estoy llamando al pantanombre "él" en algún sentido? Vistas así las cosas, no es un hombre, sino apenas algo que parece un hombre.

[25] Véase L. Wright, "Functions", *Philosophical Review,* 82 (1973).

Así, si la teoría biológica de la representación mental sostiene que el pantanombre no tiene pensamientos, al parecer se compromete a que el pantanombre no es un organismo, por la misma razón. Lo que hace el trabajo aquí es la concepción de función biológica que la teoría emplea. Si encontramos implausible la consecuencia de la teoría, entonces podemos rechazar ese concepto de función, o podemos rechazar de plano la teoría.[26] Dado lo que acabamos de decir, y las dificultades que esbozaré en breve ("Contra la reducción y la definición"), preferiría rechazar la teoría. Pero la idea de que la representación tiene una base en hechos biológicos acerca de los organismos no carece de plausibilidad para un creyente en la mente mecánica. Por supuesto, un creyente en la mente mecánica sostiene que los seres humanos son fundamentalmente entidades biológicas. La cuestión es, sin embargo, ¿dé *qué* manera pueden las explicaciones biológicas ayudarnos a comprender la naturaleza de las capacidades mentales, y la representación mental en particular? ¿hay una respuesta general a esta pregunta? Algunos filósofos influidos por la psicología evolutiva piensan que la hay. Será útil, por lo tanto, hacer una breve digresión acerca de la psicología evolutiva, antes de regresar a nuestro tema principal de la representación mental.

La evolución y la mente

Una manera de comprender la teoría biológica de la representación mental es verla como parte del proyecto más am-

[26] Para una crítica de la naturaleza de la función, véase Fodor, *The Mind Doesn't Work That Way*, p. 85.

plio de entender las capacidades mentales en términos de explicación biológica evolutiva, lo que se llama psicología evolutiva.[27] Esta psicología no es nada más la pretensión (aceptada por toda la gente científicamente informada) de que los seres humanos, seres con capacidades mentales, evolucionaron a partir de especies anteriores de monos en un proceso largo y complejo que se inició hace unos siete millones de años. Ésta es una verdad tan sólida como la que más en la ciencia, y (dando o tomando algunos detalles y fechas) no es cosa que se discuta. La psicología evolutiva es la pretensión más específica y controvertida de que muchas capacidades y facultades mentales pueden explicarse considerándolas como *adaptaciones* en el sentido del biólogo evolutivo. Una adaptación es un rasgo o capacidad cuya naturaleza puede ser explicada como producto de la selección natural. El plumaje monótono de algunos pájaros, por ejemplo, puede ser explicado por el hecho de que aquellos de sus antepasados remotos con plumaje monótono eran más capaces de camuflarse entre plantas y así sobrevivir a los depredadores, y por lo tanto reproducirse, transmitiendo con ello su plumaje a su descendencia... y así sucesivamente. El plumaje de los pájaros, se concluye, es una adaptación.[28]

Hay un debate entre los biólogos evolutivos acerca de cuáles serán las unidades de la "moneda" de la selección natural. ¿Qué es lo que selecciona la selección natural? Algu-

[27] Véase J. L. Barkow, L. Cosmides y J. Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (Nueva York, Oxford University Press, 1992).

[28] Para una excelente introducción a estas cuestiones, véase Paul Griffiths y Kim Sterelny, *Sex and Death: An Introduction to the Philosophy of Biology* (Chicago, University of Chicago Press, 1999).

nos dicen que selecciona de entre los organismos los más adecuados para la supervivencia. Otros, como Richard Dawkins, piensan que esto no llega hasta el meollo del asunto, y sostienen que la unidad básica de selección es el gen: los organismos son "vehículos" para llevar los genes y transportar ese material genético replicándolo en generaciones futuras (esto es lo que Dawkins llamó la hipótesis del "gen egoísta").[29] Nótese que la creencia de que algunos, o muchos, rasgos humanos son adaptaciones no es lo mismo que creer que la unidad básica de selección es el gen. Ni creer en adaptaciones es lo mismo que ser un *adaptacionista.* El adaptacionismo es definido de varias maneras: algunos dicen que es considerar que todos los rasgos son adaptaciones (punto de vista tonto, según veremos); otros lo definen como una manera de examinar que la adaptación es *óptima:* tal como un comentarista lo plantea, el punto de vista es que "un modelo censurado de todos los mecanismos evolutivos excepto la selección natural podría predecir con exactitud la evolución".[30]

Dos rasgos del concepto de adaptación merecen ser señalados. Primero, la inferencia de que algo es una adaptación es deducir a la mejor explicación (véase el capítulo IV, "La modularidad de la mente"). La explicación adaptativa

[29] Véase Richard Dawkins, *The Selfish Gene* (Oxford, Oxford University Press, 1976). No hay espacio para discutir las ideas de Dawkins en este libro. Algunas de ellas son defendidas por Daniel C. Dennett en *Darwin's Dangerous Idea* (Londres, Allen Lane, 1995). Mis propias simpatías están con las críticas de Fodor en su reseña de Dawkins, *Climbing Mount Improbable* en su colección *In Critical Condition* (Cambridge, MIT Press, 1998), especialmente pp. 167-169.

[30] Paul Griffiths, "Adaptation and Adaptationism", en R. Wilson y F. Keil (eds.), *The MIT Encyclopedia of Cognitive Science* (Cambridge, MIT Press, 1999), p. 3.

del plumaje del pájaro es mejor que las otras posibilidades, cualesquiera que sean, lo que nos da razón para apoyar la pretensión de que el plumaje es una adaptación. En segundo lugar, y de manera relacionada, la explicación es una forma de "ingeniería inversa": del rasgo observable del pájaro, el biólogo infiere el tipo de orígenes del medio en el cual tal rasgo sería adaptativo, es decir, ayudaría a la supervivencia de seres con dicho rasgo. Por lo tanto, la evidencia de la explicación adaptativa propuesta contendría al menos dos cosas: primero, que la explicación adaptativa es mejor que las otras posibilidades, cualesquiera que puedan ser; y, segundo, que tenemos alguna clase de conocimiento independiente de la clase de medio en el cual la presencia de tal rasgo ayuda a la supervivencia.

¿Cómo podrían capacidades y rasgos psicológicos ser interpretados como productos de la selección natural? Debemos tener claro, primero que nada, qué estamos tratando de explicar. Si enfocamos las pautas de comportamiento de los individuos, entonces no encontraremos ejemplos remotamente plausibles de adaptaciones. Sólo encontraremos el género de seudociencia que llena los suplementos dominicales. Es absurdo explicar el comportamiento de un viejo rico que compra una comida costosa en un restaurante para una mujer joven, diciendo que el hombre quería propagar sus genes y era atraído por la mujer ya que la juventud es una buena indicación de fertilidad; e igualmente absurdo explicar el comportamiento de la mujer al aceptar la compra de comida diciendo que quería propagar sus genes y era atraída por el hombre porque su evidente opulencia era una buena indicación de que podía sostener la descendencia de ella. Esta clase de explicaciones es absurda, en parte porque la disposición a comprar comidas en restaurantes sencilla-

mente no podría ser una adaptación, y no sólo porque los restaurantes fueron inventados en el París del siglo XVIII y no en la era pleistocénica.[31] Comprar comidas en restaurantes es una actividad social compleja que tiene implicaciones para otras muchas instituciones y prácticas sociales (dinero, estructuras sociales y de clase, gastronomía, viticultura, etc.). Comparar casos como éste con cosas como la colorida cola del pavorreal macho es sencillamente negarse a reconocer las diferencias reales y vastas entre estos fenómenos. Y, sin reconocer estas diferencias, nunca pasaremos de la comprensión más superficial de lo que ocurre en los restaurantes (y, por tanto, la psicología humana).

Más aún, según señalé antes, los argumentos en favor de las adaptaciones deben confiar fundamentalmente en la inferencia sobre la mejor explicación (de la cual los argumentos de "ingeniería inversa" son un caso especial). Tal vez la explicación del comportamiento del hombre en términos adaptacionistas podría tener algo en su favor, en caso de no haber explicaciones disponibles. Sin embargo, donde se trata de la explicación del comportamiento humano, no estamos en esta situación. Situaciones como la que acabo de describir no las encontramos misteriosas o desconcertantes con la perspectiva de la psicología del sentido común. Podemos imaginar cualquier número de explicaciones de psicología del sentido común que tengan mucho mayor sentido de esta situación que cualquier hipótesis acerca de los deseos de la pareja de propagar sus genes. A menos que agreguemos algunos supuestos más —por ejemplo, el mate-

[31] Para la historia del restaurante, véase el excelente libro de Rebecca Spang, *The Invention of the Restaurant* (Cambridge, Harvard University Press, 2000).

rialismo eliminativo— la explicación de este comportamiento en términos de genes es probablemente una de las peores explicaciones que hay. En cualquier caso, tiene poca probabilidad de ser la mejor.

Alguien podría responder concebiblemente que es cierto que la gente en esta clase de situación no tiene *creencias y deseos* conscientes acerca de la propagación de sus genes. No obstante, podría decirse que hay mecanismos inconscientes profundos que conducen a hacer cosas como ésta, y estos mecanismos son adaptaciones. ¿Qué razón hay para creer esta explicación, aun en su forma modificada? La razón no puede ser que todos los rasgos sean adaptaciones; poca razón hay para creer esto. En algunos casos, los rasgos que plausiblemente evolucionaron con un propósito han sido usados para otros (esto se llama "exaptaciones"). Un ejemplo clásico son las plumas de los pájaros, las cuales, según se cree, originalmente evolucionaron para el aislamiento y sólo más tarde se usaron para el vuelo. Más aún, hay casos en los cuales carecemos de cualquier razón para suponer que un rasgo realmente surgió como resultado de la selección natural, en absoluto. Por tomar un ejemplo controvertido: algunos pensadores, incluyendo a Chomsky, sostienen que éste es el caso con el lenguaje. Dicen que no hay razón para creer que el lenguaje humano sea un producto de la selección natural. Como no conocemos las circunstancias en las cuales poseer un lenguaje realmente ayudó a la supervivencia de nuestros antepasados, no podemos de hecho suponer que fue una adaptación. Por supuesto, podemos imaginar casos en los cuales el lenguaje *podría* haber ayudado a la supervivencia. Mas no hay argumento válido que nos lleve desde "X podría haber ayudado a la supervivencia en la circunstancia Y" hasta "X es una adaptación".

Sólo porque algo pudo haber surgido por dar a un organismo cierta ventaja en la supervivencia, esto en nada contribuye a mostrar que realmente lo hiciera.[32]

Tampoco deberíamos suponer (pocos lo hacen) que todo lo que hacemos esté determinado por nuestros genes. Organismos con material genético idéntico pueden desarrollarse de maneras muy diferentes en diferentes medios. El desarrollo y comportamiento de los organismos es determinado por muchos factores, incluyendo sus disposiciones genéticas internas y sus condiciones ambientales generales, aparte de sucesos estrafalarios y desastres ambientales tales como inundaciones y épocas glaciales. La evolución, el desarrollo de formas de vida a lo largo del tiempo, no se basa nada más en la selección natural.

En una famosa discusión, Stephen J. Gould y Richard Lewontin describieron una analogía entre las explicaciones adaptacionistas de rasgos y las explicaciones espurias de por qué algunos artefactos tienen la forma que tienen.[33] Al contemplar los fabulosos mosaicos de los arcos o la entrada de la basílica de San Marcos, en Venecia, uno podría ser llevado a pensar que los espacios entre los arcos (llamados "tímpanos") fueron planeados a fin de que los mosaicos pudieran ser colocados ahí. Esto no es así: los tímpanos son un simple efecto lateral de la construcción de los arcos, y el artista o los artistas inspirados aprovecharon el espacio para

[32] Para un enunciado particularmente claro de esta cuestión, véase Fodor, *In Critical Condition,* pp. 163-166.

[33] S. J. Gould y R. Lewontin, "The spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme", *Proceedings of the Royal Society of London,* 205 (1979), pp. 581-598. Véase también la colección de ensayos de Lewontin, *It Ain't Necessarily So* (Londres, Granta Books, 1999). Crítica de Gould y Lewontin puede encontrarse en Dennett, *Darwin's Dangerous Idea* (Londres, Allen Lane, 1995).

crear algo bello. Los tímpanos no fueron construidos a fin de hacer los mosaicos. Sostener que lo fueron es cometer un error análogo a ver adaptaciones por doquier. Los rasgos de un organismo pueden surgir a través de muchos procesos históricos, y necesitamos buena evidencia empírica antes de pretender que la selección natural es uno de ellos. En ausencia de tal testimonio, no habría que hacer historias adaptacionistas de las circunstancias en las cuales determinados rasgos ayudarían a la supervivencia.

Así, parece que no tenemos razón para creer que todo rasgo de un organismo es una adaptación. Tal vez esto no sería realmente objeto de mucha controversia, y el adaptacionismo extremo mencionado anteriormente es realmente un espantapájaros. Paul Bloom resume la actitud presente de los biólogos evolutivos como sigue:

> Los biólogos modernos han elaborado la visión darwiniana de que aunque la selección natural es el más importante de todos los mecanismos evolutivos, no es el único. Muchos caracteres que los animales poseen no son adaptaciones, pero aparecen como subproductos de adaptaciones, o a través de procesos enteramente no seleccionistas, tales como la deriva genética casual. La selección natural es necesaria sólo a fin de explicar la evolución de lo que Darwin llamó "órganos de extrema perfección y complejidad", tales como el corazón, la mano y el ojo [...] Aunque hay controversia acerca del alcance auténtico de las teorías seleccionistas, en lo anterior hay cuando menos acuerdo al respecto, incluso por parte de aquellos que son más cautos en la aplicación de explicaciones adaptativas.[34]

[34] Paul Bloom, "Evolution of language", en *The MIT Encyclopedia of Cognitive Science*, p. 292.

Suponiendo que, a grandes rasgos, ésta es una exposición correcta del estado actual del conocimiento, la consecuencia es que necesitamos razones positivas para creer que cualesquiera rasgos son adaptaciones. Nuestro ejemplo del rico y la joven bien puede haber sido una caricatura de cierto tipo de explicación adaptativa. ¿Qué clase de ejemplo sería más plausible?

Partiendo de la observación de Darwin citada antes, tal vez deberíamos buscar "órganos de extrema perfección y complejidad" en la mente. O por lo menos deberíamos buscar *órganos mentales* de alguna clase, independientemente identificados como tales. Entonces estaríamos en condiciones de plantear la cuestión de "ingeniería inversa": ¿en qué medio la posesión de semejante órgano habría ayudado a la supervivencia de los seres de los cuales es el órgano? El psicólogo necesitaría entonces buscar testimonios de que el organismo en cuestión vivió en tal o cual ambiente, y testimonios de que los organismos se desarrollaron siguiendo las líneas propuestas.

Los mejores candidatos para tales órganos mentales serían mecanismos relativamente aislados, resilentes, probablemente innatos en la mente, dedicados a tareas específicas de procesamiento de la información. En otras palabras, serían módulos mentales de algún modo, como el sentido descrito en el capítulo IV ("La modularidad de la mente"). El sistema visual es un ejemplo de primera de semejante módulo. Para establecer que el sistema visual es una adaptación —supuesto que quizá sería plausible aun para el más escéptico de los antiadaptacionistas— habría que dar una especificación de su tarea, y del medio en el cual la realización de dicha tarea ayudaría a la supervivencia. En posesión de un módulo mental bastante bien entendido, podemos

plantear cuestiones acerca de su función y su historia evolutiva, con la esperanza de encontrar si es una adaptación, ni más ni menos que podemos hacerlo acerca de otros órganos. (Una dificultad, por supuesto, es encontrar los testimonios reales de la existencia pasada de capacidades cognitivas: como dice Fodor, "la cognición es demasiado blanda para dejar un registro paleontológico".)[35] No es sorprendente, entonces, que los psicólogos evolutivos hayan tendido a adoptar la tesis de la modularidad compacta descrita en el capítulo IV, la tesis de que todos los aspectos de la cognición pueden ser descompuestos en módulos. Y tampoco sorprenderá que los críticos de la psicología evolutiva, como Fodor, sean también los que rechazan la modularidad compacta. No habrá explicación adaptacionista de la cognición que subraye, por ejemplo, el comportamiento "apareador" humano, sencillamente porque es imposible aislar estas actividades cognitivas de todas las demás actividades entrelazadas dentro de las cuales tienen sentido.

La única conclusión que podemos extraer de esta clase de discusión es que las cuestiones que rodean a la psicología evolutiva están enredadas con cuestiones controvertidas en la propia teoría evolutiva —tales como el alcance de la explicación adaptacionista, y a qué equivale esa clase de explicación—, pero esa psicología evolutiva es más fuerte cuando lo que ha de explicar son módulos mentales. Que debamos creer que cualesquiera módulos son adaptaciones depende, de manera no sorprendente, de la evidencia, no de la teorización filosófica, ni de la disponibilidad de explicaciones posibles. En todo caso, parece claro que la imagen mecánica de la mente no necesita una exposición evolutiva de la mente.

[35] Fodor, *In Critical Condition*, p. 166.

Ésta puede integrarse en el mundo de las causas y efectos aun si la mayoría de las capacidades mentales carecen de explicación evolutiva.[36]

Contra la reducción y la definición

Volvamos ahora al proyecto de explicar la representación mental dando una definición reductiva de ella. Incluso si este enfoque reductivo consigue resolver el problema de la disyunción, uno de los problemas que pospusimos antes sigue en pie: cómo explicamos las capacidades representacionales de conceptos diferentes de otros, muy sencillos, como *agua, comida, depredador* y otras. Las teorías reductivas de la representación tienden a tratarla en gran medida como cuestión de detalle, su enfoque es: que nuestros conceptos sencillos sean correctos antes de pasar a los conceptos complejos. Aun si se obtienen correctamente los conceptos sencillos, ¿cómo exactamente supusimos que se mueven hacia los conceptos complejos? ¿Cómo se supone que explicamos un concepto como (por ejemplo) *arquitectura barroca* en términos causales o biológicos?

Esta cuestión se le plantea también a Fodor. Tal vez Fodor diría que las representaciones mentales de la arquitectura barroca son asimétricamente dependientes de fragmentos de arquitectura barroca, por ejemplo, un fragmento de arquitectura barroca causa la representación mental *arquitectura barroca,* y, aun cuando un fragmento de arquitectura renacentista puede causar esta representación mental, no lo

[36] Fodor da una argumentación decisiva en pro de esta opinión en *The Mind Doesn't Work That Way,* pp. 80-84.

haría si tampoco lo hiciera la arquitectura barroca. Esto es muy implausible. Sin ir más lejos, mucha gente ha estado en contacto con la arquitectura barroca sin formar ninguna representación de ella como barroca; y algunas personas habrán pasado por el concepto en libros sin jamás haber tenido contacto causal con la arquitectura barroca. De este modo ¿qué debiera decir Fodor?

Las teorías reductivas de la representación aspiran a proporcionar algún modo de llenar el esquema

(R) X representa Y si y sólo si _____.

en términos que no mencionan la representación. Como ha dicho Fodor, "si la acerquidad es real, debe realmente ser algo más".[37] El problema que estoy planteando es que, si una teoría reductiva va a ser una teoría de todas las clases de contenido mental, entonces *o bien* tiene que explicarnos cómo podemos llenar plausiblemente el "_____" de modo directo para todos los conceptos y contenidos, *o bien* tiene que darnos un método sistemático de pasar desde los conceptos con los que puede directamente tratar (los conceptos "sencillos"), a aquellos con los cuales no puede tratar (los conceptos "complejos"). He sugerido que ni la teoría de Fodor ni la teoría biológica pueden tomar el camino directo. Así, estas teorías deben proporcionarnos alguna idea de cómo pasar de conceptos "sencillos" a conceptos "complejos". Y mientras no exista tal idea tenemos el derecho de suspender la creencia acerca de si puede haber algo como una teoría reductiva de la representación en absoluto.

[37] Fodor, *Psychosemantics,* p. 97.

(La teoría del éxito, por otra parte, no tiene ninguna dificultad para vérselas con todos los contenidos directamente. Pues puede ser nada más que una creencia tenga el contenido P sólo en caso de que las acciones causadas por esa creencia y un deseo D consiguieran satisfacer D sencillamente cuando P es verdad, y P puede ser una situación concerniente a cualquier cosa. Como vimos, la teoría del éxito no puede suministrar una reducción genuina de la representación a menos que pueda dar una reducción del contenido de los deseos. Tal como se plantea, la teoría del éxito es incompleta.)

Esta línea de pensamiento puede conducir a tomar con muchas precauciones la idea de explicar la representación mental reduciéndola por medio de una definición como (R). Pues, después de todo, definir algo (ya sea de modo naturalista o no) no es el único modo de explicarlo. Si quisiera yo explicarle a usted la arquitectura barroca, por ejemplo, podría llevarlo a ver algunas construcciones barrocas, señalando los rasgos distintivos —los frontones rotos, los cartuchos, el uso extravagante de la línea y el color— y contrastar el estilo con estilos anteriores y posteriores de arquitectura hasta que usted fuese adquiriendo gradualmente una noción del concepto. Lo que no haría es decir: "Una construcción es barroca si y sólo si _____", con el blanco lleno con términos que no mencionan el concepto *barroco*. Para este caso, captar el concepto no es captar una definición, por usar la frase de Wittgenstein, "la luz se alza gradualmente sobre el conjunto".[38]

Esto no es decir que una definición reductiva no pueda ser una explicación, sólo que no es la única clase de explica-

[38] Wittgenstein, *On Certainty* (Oxford, Blackwell 1979), § 141.

ción. Hasta aquí en este capítulo me he concentrado en intentos filosóficos por explicar la representación reduciéndola por definición. En lo que queda deseo retornar a la posibilidad no reductiva que mencioné al iniciar este capítulo.

Tal como introduje la idea en el capítulo III, la noción de computación depende de la noción de representación. Así, de acuerdo con los reduccionistas como Fodor, por ejemplo, la línea de investigación es la que sigue. Lo que distingue a los sistemas que son simplemente *describibles* como funciones de computación (tales como el sistema solar) de sistemas que genuinamente computan funciones (tales como una máquina sumadora) es que los últimos contienen y procesan representaciones; no hay computación sin representación. La meta, pues, es explicar la representación: necesitamos una teoría reductiva de la representación para vindicar nuestra teoría de la cognición de acuerdo con los supuestos naturalistas mencionados antes ("Reducción y definición").

Este movimiento final podría ser rechazado con el fundamento de que los supuestos naturalistas mismos deberían ser rechazados. O podría ser rechazado sobre el fundamento de que la teoría computacional de la cognición no requiere una exposición reductiva de la representación a fin de emplear la noción de representación. Me concentraré en esta segunda línea de pensamiento. Quiero considerar, de un modo muy abstracto, una teoría de la representación mental que adopta la siguiente estrategia.[39] Lo que la teoría se

[39] Mi descripción de esta estrategia ha sido tomada de Cummins, *Meaning and Mental Representation,* y Frances Egan, "Individualism, Computation and Perceptual Content", *Mind,* 101 (1992), especialmente pp. 444-449. No quiero dar a entender que todos estos filósofos convengan con todos los aspectos de la estrategia tal como yo la defino.

encarga de explicar es el comportamiento de organismos en sus medios. Este comportamiento es visto plausiblemente como representacional, como dirigido a metas, como intentando satisfacer los deseos y metas del organismo (por ejemplo buscar comida). La teoría sostiene que la mejor explicación de cómo se produce este comportamiento es considerarlo como producto de procesos computacionales, verlo, pues, como computando una "función cognitiva": una función cuyos argumentos y valores son representaciones que tienen alguna relación cognitiva con otro (del modo descrito en el capítulo IV: "El argumento del lenguaje del pensamiento"). Como las computaciones están (por su propia naturaleza) definidas en términos de representaciones, algunos estados interiores del organismo, así como las salidas y entradas, deben tratarse como representaciones. Estos estados son los implicados en la computación, así que deben tener una especificación que no es dada en términos de lo que representan: una especificación en términos puramente formales o "sintácticos". Y tratar un estado como una representación es especificar un mapa del estado mismo —descrito en términos puramente formales— dando su contenido abstracto representacional. Este mapeo es conocido como "función de interpretación". El cuadro que resulta es lo que Cummins llama la imagen del "Puente de la Torre" (véase la figura V.1).[40]

Basados en esta perspectiva, no es como si tuviéramos que encontrar los estados del organismo que podamos decir que son representaciones *con fundamentos diferentes,* esto es, fundamentos independientes de las computaciones que

[40] Véase Egan, "Individualism, Computation and Perceptual Content", pp. 450-454; y Cummins, *Meaning and Mental Representation,* cap. 8.

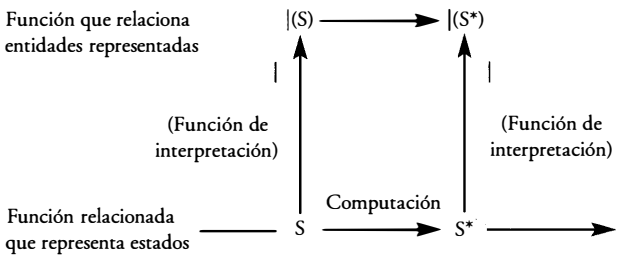


FIGURA V.1. *La imagen de Cummins, "Puente de la Torre", de la computación. El tramo superior representa la función cuyos argumentos y valores son los encabezados representados. El "tramo" inferior representa la función cuyos argumentos y valores son estados del mecanismo,* S, *y* S*. |, *la función de interpretación mapea los estados del mecanismo en las entidades representadas. "|(S)" puede leerse: "la entidad representada por el estado* S *bajo la interpretación |". Por ejemplo, tratar las entidades representadas como números y el mecanismo como una máquina sumadora. La función del tramo superior es la adición. La función | mapea estados de la máquina (apretar botones, exhibir, etc.) como números. Una computación de la función de suma es una transición causal entre los estados de la máquina que refleja la "transición" entre números adicionados.*

atribuimos al organismo. Lo que hacemos es tratar cierto sistema como ejecutante de computaciones, donde la computación es la transición disciplinada entre estados internos, formalmente especificados. Definimos entonces una función de interpretación que "mapea" los estados internos en contenidos. Este enfoque concuerda con la pretensión de Fodor de que no hay computación sin representación. Esto no significa que necesitemos hacer una exposición reductiva de lo que es una representación. La representación es sola-

mente otro concepto de la teoría; no necesita defensa y reducción filosóficas externas. Por eso llamo a este enfoque "no reductivo".

Una analogía puede ayudar a mostrar cómo la representación figura en la teoría computacional en virtud de esto.[41] Cuando medimos un peso, por ejemplo, usamos números para captar los pesos de los objetos, de acuerdo con cierta unidad de medida. Usamos el número 2.2 para captar el peso (en kilogramos) de una bolsa estándar de azúcar. Habiendo captado un peso "mapeándolo" en un número, podemos ver que las operaciones aritméticas con números "reflejan" relaciones físicas entre pesos determinados. Así, por ejemplo, si sabemos que una bolsa de azúcar pesa 2.2 kilos, sólo necesitamos conocer aritmética elemental para saber que dos bolsas de azúcar así pesarán 4.4 kilos, y así sucesivamente.

Análogamente, cuando "medimos" los pensamientos de una persona, usamos oraciones para captar estos pensamientos (sus creencias, deseos y demás). Usamos la oración "El hombre que hizo quebrar la banca en Monte Carlo murió en la miseria" para captar la creencia de alguien de que el hombre que hizo quebrar la banca en Monte Carlo murió en la miseria. Habiendo captado la creencia "mapeándola' en una oración, podemos ver que las relaciones lógicas entre oraciones "reflejan" relaciones psicológicas entre creencias específicas. Así, por ejemplo, si sabemos que Vladimir

[41] Para esta analogía véase Hartry Field, suplemento de "Mental Representation", en Ned Block (ed.), *Readings in the Philosophy of Psychology,* vol. II (Londres, Methuen, 1980). Field atribuye la analogía a David Lewis. Para un uso de la analogía en un sentido algo semejante al usado aquí, véase Robert Matthews, "The Measure of Mind", *Mind,* 103 (1994).

cree que el hombre que hizo quebrar la banca en Monte Carlo murió en la miseria, necesitamos sólo lógica elemental para saber que Vladimir cree que alguien murió en la miseria, y así sucesivamente.

O, más bien, así dice la historia, la analogía suscita muchas cuestiones complicadas. (Recuérdese, por ejemplo, la cuestión discutida en el capítulo IV de si la lógica puede realmente proporcionar una *descripción* de los procesos del pensamiento humano.) El objeto de emplear la analogía aquí es nada más para ilustrar cómo estados concretos de la mente podrían ser mapeados en entidades evidentemente "abstractas", tales como números u oraciones, y cómo el comportamiento de estas entidades abstractas refleja ciertas interesantes relaciones entre los estados. La analogía también ilustra cómo la teoría puede permitirse ser no reductiva: así como no surge la cuestión de cómo "reducimos" la relación de un objeto con un número que capta su peso, así tampoco surge ninguna cuestión acerca de cómo reducimos la relación de una persona con las oraciones que expresan los contenidos de sus pensamientos.

Dos rasgos del caso de los pesos mostrado antes son dignos de nota. En primer lugar, debe haber un modo independiente de caracterizar los pesos de objetos, aparte de hacerlo en términos numéricos. Piénsese en balanzas arcaicas de cocina, donde el peso de algo se mide comparándolo sencillamente con otros pesos. No hay por qué usar números. En segundo lugar, tenemos que aceptar que no hay número único que mida el peso de un objeto. Pues qué número es usado para medir el peso es relativo a la unidad de medida que se elija. El peso de nuestra bolsa de azúcar es de 2.2 kilos, pero es también cierto número de libras. No hay límite, en principio, a los números que pueden usarse para

medir nuestra bolsa de azúcar, de modo que no podemos hablar de "el" número que expresa su peso.

¿Conducen estos caracteres al caso análogo de la representación mental? El primer caso convencería sin controversia a aquellos que aceptan una teoría computacional de la cognición. Pues aceptarán que los estados mentales que participan en las computaciones tienen una descripción formal que no es dada en términos de las oraciones que expresan sus contenidos.

El segundo rasgo es un poco más problemático. Pues, en el caso de una creencia, por ejemplo, tenemos una fuerte convicción de que hay una oración única que expresa su contenido. El contenido de una creencia es lo que hace que sea la creencia que es: a tal grado es esencial el contenido de una creencia para ella. Si la creencia de que la nieve es blanca tuviera un contenido diferente (digamos, que la *hierba es verde*), de seguro sería una creencia diferente. No obstante, si la analogía con los números ha de funcionar, entonces debe de haber muchas oraciones diferentes que capten el mismo estado de creencia. ¿Qué oración, entonces, expresa el contenido de la creencia?

El modo evidente de esquivar esto es decir que el contenido de la creencia es expresado por todas las oraciones con el *mismo sentido.* La creencia de que la nieve es blanca, por ejemplo, puede ser captada usando la frase en inglés "Snow is white", la oración italiana "La neve è bianca", la oración alemana "Schnee ist weiss", o el húngaro "A hó fehér", y así sucesivamente.[42] Estas oraciones son intertraducibles; todas significan la misma cosa. Es el significado, más bien que las

[42] Véase Davidson, "Reality without Reference", en *Inquiries into Truth and Interpretation,* especialmente pp. 224-225.

oraciones que tienen el significado, el contenido de la creencia. Así se preserva la idea de que cada creencia tiene un contenido único que es esencial para ella.

Sin embargo, pudiera decirse que, si bien este enfoque puede funcionar tranquilamente para estados como la creencia, no hay necesidad de aplicarlo a toda clase de estados postulados por una teoría computacional de la mente (por ejemplo una teoría computacional de la visión).[43] Pues, desde el punto de vista de la computación defendido por el enfoque no reductivo, debiéramos abandonar la idea de que todos los estados mentales tienen contenidos únicos que son esenciales para ellos.[44] La razón, esencialmente, es que una función interpretativa es sólo un mapeo de los estados interiores en una estructura abstracta que "preserva" la estructura de los estados interiores. Y hay muchos mapeos que harán esto. Pues existen funciones interpretativas que asignarán distintas interpretaciones a los símbolos; el que escojamos es determinado no por el escurridizo "contenido único" del estado, sino por cuál interpretación da a la teoría mayor poder explicativo.

Podría objetarse que este enfoque vuelve la naturaleza de la representación y la computación demasiado dependiente de las decisiones de teóricos humanos. Pues justamente acabo de hablar acerca de "tratar" los estados del sistema como representaciones, y de "especificar" mapeos de estados a contenidos, "asignando" interpretaciones a esta-

[43] Véase David Marr, *Vision* (San Francisco, Freeman, 1982). Una exposición accesible, no técnica, de la teoría de Marr se da en Kim Sterelny, *The Representational Theory of Mind* (Oxford, Blackwell, 1990).

[44] Éste es de hecho el punto de vista adoptado por Egan y Cummins; véase "Individualism, Computation and Perceptual Content", p. 452, y *Meaning and Mental Representation,* pp. 102-108.

dos, y así sucesivamente. Podría objetarse que la ejecución o no por un organismo es un tema de hecho objetivo, no de nuestras especificaciones o atribuciones. Esta crítica está mal situada. Pues, en tanto que la aplicación de una teoría a un organismo es claramente asunto de decisión humana, no lo es que esta aplicación caracterice correctamente al organismo. La cuestión es: ¿alguno de los procesos cognitivos del organismo es correctamente caracterizable como computaciones? Para probar una hipótesis acerca del carácter computacional de los procesos de un organismo tenemos que interpretar los elementos del proceso. Pero esto no hace de la existencia del proceso un asunto de decisión humana en mayor grado que el hecho de que podamos captar y etiquetar las fuerzas físicas que actúan individualmente sobre un cuerpo, y así calcular la fuerza neta, torna esta interacción física un asunto de decisión humana.

Resumiendo: la respuesta no reductiva a la pregunta "¿qué es una representación mental?" sería dada estableciendo la lista de modos como el concepto de representación figura en la teoría. Aquellos estados de un organismo que son interpretables como ejemplificadores de las etapas en la computación de una función cognitiva son representaciones. Esta exposición, más la teoría general de la computación, nos enseña todo lo que necesitamos saber acerca de la naturaleza de las representaciones mentales. Las tareas difíciles están ahora delante de nosotros: encontrar cuáles sistemas tratar como computacionales y hallar qué computaciones realizan.

El atractivo de esta teoría no reductiva de la representación es que puede decir muchas de las cosas que la teoría reductiva quiere decir acerca de la estructura computacio-

nal de los estados mentales, sin tener que proporcionar una reducción definitiva de la noción de representación, y sin tener que vérselas con los problemas intratables del error. El precio que se paga por esto es permitir que la idea de los estados mentales computacionales no tenga un contenido único que sea esencial para ellos.

¿Por qué habría de ser éste un problema? En parte porque parece tan obvio para nosotros que nuestros pensamientos tengan contenidos únicos. Es evidente para mí que mi actual creencia de que ahora está lloviendo, por ejemplo, pudo sencillamente no tener otro contenido sin ser una creencia diferente. Sin embargo, puede responderse que esta referencia a cómo nuestras mentes nos parecen a nosotros es, hablando estrictamente, irrelevante para la teoría computacional de la mente. Pues esa teoría se ocupa de los mecanismos inconscientes del pensamiento y los procesos de pensamiento; no contesta directamente a la introspección, de cómo nuestros pensamientos nos afectan. Después de todo, nuestros pensamientos no nos llaman la atención como computacionales —excepto tal vez cuando estamos abriéndonos paso conscientemente a través de un algoritmo explícito—, pero nadie pensaría que ésta es una objeción adecuada a la teoría computacional de la cognición.

Hay una tensión, pues, entre cómo nuestros pensamientos nos parecen a nosotros y ciertas cosas que la teoría computacional de la cognición afirma acerca de ellos. La significación de esta tensión será discutida mejor en el capítulo VI.

Conclusión: ¿puede la representación ser explicada reductivamente?

Los intentos filosóficos por explicar la noción de representación reduciéndola no han tenido éxito notable. Todos se topan con los problemas del error, que los estorban. Esto no tiene nada de sorprendente: la idea de error y la idea de representación van codo con codo. Representar el mundo como si fuera de cierta manera es aceptar implícitamente una brecha entre cómo la representación dice que es el mundo y cómo el mundo realmente es. Esto es precisamente conceder la posibilidad del error. Así, cualquier reducción que capture la esencia de la representación debe capturar cualquier cosa que permita esta posibilidad. He aquí por qué la posibilidad de error nunca puede ser un tema lateral para una teoría reductiva de la representación.

No obstante, hay un problema más. Las teorías reductivas de la representación tienen que estar en condiciones de dar razón de toda clase de contenido mental, no nada más las clases sencillas conectadas con (digamos) la comida y la reproducción. Hasta ahora no han proporcionado explicaciones de cómo hacer esto. Así, parece aconsejable cierto grado de escepticismo.

Mientras que la teoría no reductiva que describí al final del capítulo evita ambos problemas, esta teoría acepta la consecuencia de que a muchos estados mentales no se les asignará un contenido único. La idea de que nuestros estados mentales tienen contenido único parece ser esencial para los estados mentales representacionales tal como de ordinario los entendemos. De este modo, incluso al comprender la teoría computacional de la cognición de esta

manera no reductiva, empezamos a apartarnos de la noción ordinaria de cognición y pensamiento. La cuestión del grado en el cual es aceptable esto tendrá que ser abordada en el siguiente capítulo.

LECTURAS ADICIONALES

Un buen lugar adonde dirigirse desde aquí es el *Meaning and Mental Representation,* de Robert Cummins (Cambridge, MIT Press, 1989), que contiene un excelente examen crítico de las principales teorías naturalistas de la representación mental que eran populares en los años ochenta (y que, cosa interesante, no cambiaron mucho en los noventa). La antología más útil es *Mental Representation,* editada por Stephen Stich y Ted Warfield (Oxford, Blackwell, 1994). Un intento innovador de gran escala por defender una teoría causal de la representación mental es *Knowledge and the Flow of Information,* de Fred Dretske (Cambridge, MIT Press, 1981). Una versión abreviada de algunas de las ideas de Dretske es su artículo "The Intentionality of Cognitive States", en David Rosenthal (ed.), *The Nature of Mind* (Oxford, Oxford University Press, 1991). Dretske responde a los problemas del error en su ensayo "Misrepresentation", en R. Bogdan (ed.), *Belief: Form, Content and Function* (Oxford, Oxford University Press, 1985). La teoría de Jerry Fodor aparece en los capítulos 3 y 4 de *A Theory of Content and Other Essays* (Cambridge, MIT Press, 1990). Una versión menos complicada de la teoría de Fodor es *Psychosemantics* (Cambridge, MIT Press, 1987), capítulo 4. Un enfoque de la representación naturalista que no se discute aquí, pero que necesitaría estar en un tratamiento más amplio, es el papel fun-

cional de la semántica; véase Ned Block, "Advertisement for a Semantics for Psychology", *Midwest Studies in Philosophy,* 10 (1986). David Papineau defiende su teoría biológica-teológica de la representación mental en *Philosophical Naturalism* (Oxford, Blackwell, 1993), y Ruth Millikan defiende una clase un tanto diferente de teoría biológica en *Language, Thought and Other Biological Categories* (Cambridge, MIT Press, 1984). El texto clave de la psicología evolutiva es la obra de J. L. Barkow, L. Cosmides y J. Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (Nueva York, Oxford University Press, 1992); pero para una exposición y una síntesis más accesibles de varias áreas de la ciencia cognitiva, véase Steven Pinker, *How the Mind Works* (Nueva York, Norton, 1997). Todo el enfoque es atacado vigorosamente por Fodor en *The Mind Doesn't Work That Way* (Cambridge, MIT Press, 2000). Entre las teorías antinaturalistas de la representación (no cubiertas en ningún detalle en este libro), sobresale la obra de John McDowell. Véase *Mind and World* (Cambridge, Harvard University Press, 1994) y su artículo "Singular Thought and the Extent of Inner Space", en Philip Pettit y John McDowell (eds.), *Subject, Thought and Context* (Oxford, Oxford University Press, 1986). Esta antología también contiene "Scientism, Mind and Meaning", de Gregory McCulloch, una introducción más accesible a esta clase de enfoque antinaturalista.

VI. LA CONCIENCIA Y LA MENTE MECÁNICA

La historia hasta aquí

¿Qué habremos de hacer con la visión mecánica de la mente?[1] En este libro hemos considerado varios modos en los que este punto de vista ha tratado con el fenómeno de la representación mental, con nuestro conocimiento de los pensamientos de otros y cómo (complementado por supuestos adicionales) forma la base filosófica del punto de vista computacional del pensamiento. Y, en el capítulo anterior, vimos los intentos de explicar la representación mental en otros términos, o de "reducirla".

Hay muchos asuntos no resueltos: ¿cuán adecuada es la "teoría teoría" y su exposición de nuestra comprensión de los pensamientos de otros?, ¿tienen nuestras mentes una "arquitectura" conexionista o clásica, o alguna combinación de ambas?, ¿debe una teoría de la representación mental intentar reducir el contenido de estados mentales a pautas causales de indicación y por el estilo, o es preferible un enfoque no reductivo? Acerca de algunas de estas cuestiones —por ejemplo el conexionismo ante el clasicismo— no se sabe todavía lo suficiente para que la respuesta razonable sea otra que una cautelosa mente abierta. Otras veces —por

[1] Los interesados sólo en el problema de la conciencia pueden saltarse este apartado introductorio, que tiene por propósito enlazar el asunto de la conciencia con el resto del libro.

ejemplo la "teoría teoría" frente a la simulación— me parece que el debate no ha sido todavía formulado con suficiente agudeza para saber exactamente qué está en tela de juicio. Debe ser claro, sin embargo, que la ausencia de respuestas definidas aquí no debe darnos razón para rechazar el punto de vista mecánico de la mente. Por su esencia, el punto de vista mecánico tal como lo he caracterizado es muy difícil de rechazar. Implica esencialmente una entrega al punto de vista abrumadoramente plausible de que la mente es un mecanismo causal que tiene sus efectos sobre el comportamiento. Todo lo demás —computación, "teoría teoría", teorías reductivas del contenido— es cuestión de detalle.

Sin embargo, hay filósofos que rechazan el punto de vista en conjunto, y no a causa de lo inadecuado de los detalles. Creen que el problema real con el punto de vista mecánico de la mente es que deforma —o incluso no ofrece exposición de ello— la manera como nuestras mentes nos parecen. Deja fuera lo que en ocasiones se llama *fenomenología* de la mente, donde "fenomenología" es la teoría ("logía") de cómo nos parecen (los "fenómenos"). Estos críticos objetan que la mente mecánica deja fuera todos los hechos acerca de cómo nos llaman la atención nuestras mentes, lo que se siente al tener un punto de vista del mundo. Por lo que concierne al enfoque mecánico de la mente, dicen que este lado de tener una mente bien podría no existir. El enfoque mecánico trata la mente como "un fenómeno muerto, un agente en blanco marcado con rastros causalmente eficaces de encuentros recuperables con fragmentos del ambiente".[2] O, por tomar una llamativa frase de Francis

[2] Esta observación viene de Gregory McCulloch, "Scientism, Mind and Meaning", en P. Pettit y J. McDowell (eds.), *Subject, Thought and Context* (Oxford, Clarendon Press, 1986), p. 82. Véase su obra *The Mind*

Bacon, la crítica es qué punto de vista mecánico "escudará y plegará la mente a la naturaleza de las cosas".[3]

De hecho, algo como esto es un elemento común en algunas de las críticas a la mente mecánica que hemos encontrado a lo largo de este libro. En el capítulo II, por ejemplo, vimos que la "teoría teoría" era atacada por teóricos de la simulación debido a su representación inadecuada de lo que hacemos cuando interpretamos a otros. Por "lo que hacemos cuando interpretamos a otros", los teóricos de la simulación hablan acerca de cómo la interpretación nos *llama la atención.* La interpretación no nos *parece* como aplicar una teoría; es mucho más parecida a un acto de identificación imaginativa. (No quiero afirmar que los teóricos de la simulación se opongan necesariamente a toda la imagen mecánica; pero pueden hacerlo.) Sin embargo, ¿por qué negaría alguien que la interpretación a veces nos parece esto? En particular, ¿por qué habrían de negarla los teóricos de la teoría? Y, si lo negaran, ¿entonces de qué se supone que trata el debate? La "teoría teoría" puede replicar que la cuestión no es lo que la interpretación *nos parece,* sino lo que hace que tenga *éxito* la interpretación. La mejor explicación para el éxito de la interpretación es postular el conocimiento tácito o implícito de una teoría de la interpretación. Llamar a la teoría "tácita" es en parte para indicar que no está fenomenológicamente disponible, esto es, no podemos necesariamente afirmar por introspección si la teoría es correcta. Pero de acuerdo con la "teoría teoría", esto no viene al caso.

La misma pauta argumentativa surgió cuando exami-

and its World (Londres, Routledge, 1995) para una presentación más cabal. La discusión de este capítulo tiene particularmente una deuda con conversaciones con el difunto Greg McCulloch.

[3] Francis Bacon, *Avance de la ilustración,* libro 2, iv, 2.

namos la crítica de Dreyfus de la IA en el capítulo III. Dreyfus sostuvo que pensar no puede ser un asunto de manipular representaciones de acuerdo con reglas. Esto es porque pensar requiere un "saber cómo" que no puede reducirse a representación o reglas. Parte de la argumentación de Dreyfus para esto es fenomenológica: el pensamiento no nos parece manipulación de símbolos gobernada por reglas. Mucho tendría de caricatura representar a Dreyfus diciendo: "Sólo inténtelo: piense en alguna tarea cotidiana como ir, digamos, a un restaurante, alguna tarea que requiera capacidades cognitivas básicas. Entonces hágase la prueba y represéntese qué reglas está siguiendo, y cuáles 'símbolos' está manipulando. No puede usted decir lo que son, excepto del modo más inconcluso e impreciso".

Y, una vez más, la réplica a esta clase de objeción en favor de la IA y la teoría computacional de la cognición es que Dreyfus yerra la cuestión. Porque el meollo de la hipótesis computacional es explicar la naturaleza sistemática de las transiciones causales que constituyen la cognición. Los procesos computacionales que la teoría postula no se supone que sean accesibles a la introspección. Así, no puede ser una objeción a la teoría computacional decir que no podemos introspectarlas.

En múltiples debates, entonces, parece haber un tipo general de objeción a las hipótesis mecánicas sobre la mente, que dejan fuera, desconocen o no pueden dar razón de hechos acerca de cómo nos parecen nuestras mentes, acerca de la fenomenología de la mente. En respuesta, el punto de vista mecánico arguye que cómo nos parezcan nuestras mentes es irrelevante para la hipótesis mecánica en cuestión.[4]

[4] Para un ejemplo de este tipo de respuesta, véase Michael Tye, *The Imagery Debate* (Cambridge, MIT Press, 1992), caps. 1 y 2.

Debe admitirse que hay algo de insatisfactorio en torno a esta respuesta. Pues el punto de vista mecánico no puede negar el fenómeno de cómo las mentes (nuestras y ajenas) nos parecen a nosotros. Y, lo que es más, muchos aspectos de la idea de la mente mecánica surgen al considerar cómo la mente nos parece, en un sentido muy general de "parece". Considérese, por ejemplo, el camino que seguí en el capítulo II desde la interpretación de otras mentes hasta la hipótesis de que los pensamientos son mecanismos causales internos, los resortes de la acción. Éste es un modo ordinario de motivar la imagen causal de los pensamientos, y sus puntos de partida son observaciones de sentido común acerca de cómo usamos las conjeturas sobre las mentes de la gente para explicar su comportamiento. Otro ejemplo es la invocación de Fodor de la naturaleza sistemática del pensamiento a fin de motivar la hipótesis del mentalés. Los ejemplos que típicamente usa Fodor conciernen a creencias ordinarias, tal como las concibe el sentido común: si alguien cree que Antonio ama a Cleopatra, entonces deben *ipso facto* tener los recursos conceptuales para (cuando menos) considerar el pensamiento de que Cleopatra ama a Antonio. Los puntos de partida en muchas argumentaciones en pro de aspectos de la mente mecánica son observaciones de sentido común acerca de cómo las mentes nos afectan. Así, sería taimado por parte de los defensores de la mente mecánica decir que no tienen interés en absoluto en cómo nos parecen las mentes.

El inconveniente aquí es que, aunque puede comenzar en hechos de sentido común acerca de cómo nos llaman la atención las mentes, el punto de vista mecánico de la mente termina por decir cosas que parecen ignorar cómo las mentes nos afectan, y así apartarse de su punto de partida en el

sentido común. ¿Cuál es la base de este escepticismo acerca de la mente mecánica?, ¿es precisamente que ningún defensor de este modo de ver ha producido hasta ahora una exposición de la fenomenología de la mente? O ¿hay alguna objeción más profunda, más de principio, que derive de la fenomenología y muestre por qué la imagen mecánica debe ser incorrecta? En el capítulo v vimos que muchos suponen que la *normatividad* de lo mental es una razón por la cual una reducción general de la representación mental tiene que fracasar. La idea es que los hechos de que el pensamiento es cierto o falso, correcto o incorrecto, que el razonamiento es atinado o no, se supone que evitan todos una explicación del contenido mental en términos puramente causales. Sin embargo, sostuve que una reducción conceptual del contenido mental puede no ser esencial para la imagen mecánica de la mente. La representación puede tener que ser considerada como un concepto básico o fundamental en la teoría de la mente, sin mayor análisis. Si esto es cierto, entonces la normatividad es un concepto básico o fundamental en la teoría de la mente también, porque la idea de la representación lleva consigo esencialmente la idea de corrección e incorrección. Pero no vimos razón en esto para negar que los mecanismos subyacentes de la representación mental son de naturaleza causal, y por lo tanto ninguna razón para negar en conjunto la imagen mecánica.

Hay otra área de la investigación de la mente donde se han adelantado argumentos generales de que ninguna imagen causal o mecánica puede dar una exposición adecuada de los fenómenos de la mente. Ésta es la investigación de la conciencia, pospuesta desde el capítulo I. Se dice a menudo que la conciencia es lo que presenta el máximo obstáculo a una exposición científica de la mente. Nuestra tarea en

este capítulo es comprender cuál se supone que es este obstáculo.

La conciencia, "lo parecido" y los qualia

La conciencia es a la vez el rasgo más evidente de la vida mental y uno de los más difíciles de definir o caracterizar. En un sentido, por supuesto, no necesitamos definirla. En la vida cotidiana no hallamos dificultad para emplear la noción de conciencia (como cuando el médico pregunta si el paciente ha perdido la conciencia, o cuando nos preguntamos si una langosta está consciente de algún modo cuando es echada viva en una sartén de agua hirviendo). Podemos carecer de cualesquiera pruebas infalibles que establezcan si un ser está consciente o no; pero parece que no tenemos dificultad para decidir lo que está en juego cuando tratamos de establecer esto.

O por lo menos no tenemos dificultad para decidir qué está en juego en tanto que no intentemos reflexionar acerca de lo que está ocurriendo. Considerando la pregunta "¿qué es el tiempo?", san Agustín observó, cosa famosa, que cuando nadie se lo pregunta lo sabe bastante bien, pero si esto ocurre, entonces no sabría cómo contestar. La situación parece la misma con "¿qué es la conciencia?" Estamos perfectamente a gusto con la distinción entre lo consciente y lo no consciente cuando lo aplicamos en la vida ordinaria; pero cuando nos preguntamos a nosotros mismos *"¿qué es la conciencia?"*, nos atascamos buscando una respuesta. ¿Cómo debemos proceder?

Pues bien, ¿cuál es la distinción cotidiana entre lo consciente y lo no consciente? Atribuimos conciencia a los seres,

organismos vivos, y también a estados de la mente. La gente y los animales son conscientes; pero así también son sus sensaciones y (algunos) de sus pensamientos. El primer uso del concepto de conciencia ha sido llamado "conciencia de seres" y el segundo "conciencia de estados".[5] La conciencia de seres y la conciencia de estados son evidentemente interdependientes. Si un ser es consciente, es cuando es consciente de estados de la mente; y los estados mentales conscientes son *ipso facto* los estados de un ser consciente. No hay razón para suponer que debemos definir una idea en términos de la otra. Sin embargo, con todo, es tal vez más fácil empezar nuestra exploración de la conciencia considerando lo que es para un ser el ser consciente. Thomas Nagel dio a los filósofos una manera vívida de hablar acerca de la distinción entre seres conscientes y no conscientes: un ser es consciente, dijo, cuando hay algo que *es parecido* a ser tal ser.[6] No hay nada parecido a ser una bacteria, nada es como tener un pedazo de queso, pero algo es parecido a ser un perro o un ser humano o (por usar el famoso ejemplo de Nagel) un murciélago. Este giro de lo "parecido" puede ser fácilmente transferido a conciencia de estados también: hay algo que es parecido a saborear (estar en estado de saborear) el helado de vainilla o de oler (estar en estado de oler) hule quemado. Esto es, hay algo que es parecido a estar en estos estados mentales. No hay nada que sea como estar en gran medida compuesto de agua, o tener alta presión arterial. Estos no son estados de la mente.

[5] David Rosenthal, "A Theory of Consciousness", en Block, Flanagan y Güzeldere (eds.), *The Nature of Consciousness* (Cambridge, MIT Press, 1995).

[6] Thomas Nagel, "What is it Like to Be a Bat?", en Nagel, *Mortal Questions* (Cambridge, Cambridge University Press, 1979).

La oración "lo que es parecido" no se supone que sea una definición de la conciencia. Como he dicho ya, no estamos buscando aquí una definición. Nadie que carezca de concepto de conciencia (si semejante persona fuera posible) sería capaz de captarlo con sólo decirle que hay algo que se parece a ser consciente, o estar en estados conscientes. Podemos decir un par de cosas acerca del significado de esta oración que nos ayude a aclarar su papel en la discusión de la conciencia. Primero, la oración no estaba pensada de una manera *comparativa.* Podría preguntarse: ¿parecido a qué es la Vegemite? Y la respuesta pudiera darse: es como Marmite. (Para los no iniciados, Vegemite y Marmite son maravillosos condimentos basados en levadura, el primero de Australia, el segundo británico.) Aquí, preguntar a qué es parecido es preguntar qué cosas son *como ello;* esto es, qué cosas parecen ello. Éste no es el sentido de "lo que es parecido" a que aludía Nagel cuando dijo que hay algo que es parecido a ser un murciélago. En segundo lugar, la oración no aspira sencillamente a significar *lo que se siente,* si "sentir" tiene su significado normal. Pues hay algunos estados de la mente donde tiene sentido decir que hay algo que es parecido a lo que es estar en estos estados, aun cuando esto no implique sentir en ningún sentido ordinario. Considérese el proceso de evaluar en algún problema, tratando de comprender alguna cuestión difícil, en la cabeza. Hay, intuitivamente, algo que es parecido a estar evaluando este problema; pero no necesita "sentirse" como ninguna cosa. No tiene por qué haber sentimientos o sensaciones especiales que intervengan. Así, aunque haya algo que es como sentir una sensación, no todos los casos en que hay algo parecido son casos de sentimientos.

"Lo que es parecido", pues, no significa *lo que parece* y

no significa (nada más) *lo que se siente ser*. Lo que se trata de expresar es lo que parecen las cosas para nosotros cuando estamos conscientes, o en estados conscientes, lo que llamé en el apartado anterior la *apariencia* o los *fenómenos* de la mente. Esto se supone que es diferente de ser sólo la clase de ser que tiene una mente: qué es ser un murciélago es una cosa; qué es *como* ser un murciélago es otra. Ahora bien, la expresión de "conciencia fenoménica" se usa algunas veces para esta idea de cómo parecen ser las cosas para un ser consciente; y el término es etimológicamente atinado, dado que la palabra española "fenómeno" deriva de la palabra griega para *apariencia*. Un ser es fenoménicamente consciente cuando hay algo que es parecido a ser ese ser; un estado mental es fenoménicamente consciente cuando hay algo que es como estar en ese estado. El modo especial de ser de un estado mental, que constituye lo que es estar en ese estado, se llama asimismo el *carácter fenoménico* del estado.

A veces la conciencia fenoménica es descrita en términos de *qualia* (encontramos qualia por vez primera en el capítulo 1, "La tesis de Brentano"). Los qualia (plural: el singular es *quale*) se supone que son las propiedades no representacionales, no intencionales, pero fenoménicamente conscientes, de estados de la mente.[7] Los creyentes en qualia dicen que el carácter particular del aroma del café no puede, sencillamente, captarse en términos de la manera como el olor representa el café; esto no conseguiría captar el modo como se *siente* oler café. Aun cuando haya usted descrito todos los modos como su experiencia del olor a café

[7] Ned Block usa la expresión de esta manera: véase "Inverted Earth", en Block, Flanagan y Güzeldere (eds.), *The Nature of Consciousness*.

representa café, habrá usted dejado algo afuera: estos son los qualia de la experiencia de oler café, las *propiedades intrínsecas* de la experiencia, que son independientes de la representación del café. Alguien que crea en qualia niega la tesis de Brentano de que todos los fenómenos mentales son intencionales: algunas propiedades conscientes de estados de la mente no son en absoluto intencionales. Y se supone que son las propiedades que es tan difícil que tengan sentido, desde un punto de vista naturalista. De ahí que el problema de la conciencia se denomine a menudo "problema de los qualia".[8]

Sin embargo, aunque no es discutible que hay una cosa como la conciencia fenoménica, es controvertido que haya qualia. Algunos filósofos niegan que haya ningunos qualia, y con esto no quieren decir que no haya conciencia fenoménica .[9] Lo que significa es que no hay nada en la conciencia fenoménica por encima de las propiedades representacionales de estados de la mente. En el caso de la percepción visual, por ejemplo, estos filósofos —conocidos como *intencionalistas* o *representacionalistas*— afirman que cuando percibo algo azul no tengo conciencia de alguna propiedad *intrínseca* de mi estado mental, aparte de lo azul que percibo. Miro una pared azul y de lo único que tengo conciencia es de la pared y de su color azul. No estoy, además, consciente de algunas propiedades intrínsecas de mi estado mental.[10] Y este modo de ver dice cosas parecidas acerca de la

[8] Así es como David Chalmers lo expresa en *The Conscious Mind* (Oxford, Oxford University Press, 1996).

[9] Véase Daniel Dennett, "Quining Qualia", en Lycan (ed.), *Mind and Cognition.*

[10] Para las teorías intencionalistas de la mente, véase Michael Tye, *Ten Problems of Consciousness* (Cambridge, MIT Press, 1995), y Gilbert Har-

sensación. El creyente en qualia dice que, en semejante caso, uno tiene también noción de lo que Ned Block ha llamado "pintura mental": las propiedades intrínsecas del estado mental de uno.

Las cosas pueden volverse confusas aquí a causa de que otros filósofos usan la palabra "qualia" sencillamente como un sinónimo de "carácter fenoménico", así que tener conciencia fenoménica es, por definición, tener qualia. Esto ayuda muy poco porque hace imposible comprender lo que filósofos como Tye y Dennett bien podrían significar cuando niegan que haya qualia. Para hacer un primer intento de aclarar las cosas aquí, debemos distinguir dos maneras de usar la expresión "qualia": *i)* tener qualia es sencillamente tener experiencia con un carácter fenoménico; o *ii)* los qualia son cualidades no intencionales (no representacionales) de la experiencia.

El debate acerca de la conciencia implica, tal parece, un alto grado de confusión terminológica. Necesitamos establecer una distinción a grandes rasgos entre la conciencia fenoménica —la cosa por explicar— y aquellas propiedades a las que se recurre a fin de explicar la conciencia fenoménica. A menos que hagamos esto no entenderemos qué es lo que los filósofos están haciendo cuando niegan la existencia de qualia. Superficialmente, podría parecer como si rechazasen los fenómenos de la conciencia, en tanto que lo que parecen estar realmente rechazando es un modo determinado de explicar la conciencia fenoménica: en términos de qualia, propiedades no intencionales, no representacionales, de estados mentales.

man, "The Intrinsic Qualities of Experience", en Block, Flanagan y Güzeldere (eds.), *The Nature of Consciousness*. Para un repaso general, véase *Elements of Mind*, cap. 3.

Hechas estas aclaraciones, debemos finalmente pasar a un tema que se ha quedado atrás, el problema mente-cuerpo.

Conciencia y fisicalismo

En el capítulo II ("El problema mente-cuerpo") dije que el problema mente-cuerpo puede ser expresado en términos del desconcierto que sentimos al tratar de entender cómo un simple fragmento de materia como el cerebro puede ser fuente de algo como la conciencia. Por un lado, sentimos que nuestra conciencia debe estar, ni más ni menos, fundada en la materia; pero, por otra parte, encontramos imposible comprender cómo puede ser así. Hay ciertamente algo que hace a mucha gente pensar que la conciencia es misteriosa; pero este, por sí mismo, no es un pensamiento suficientemente preciso para generar un problema filosófico. Supóngase que alguien fuera a mirar una planta, y habiendo averiguado acerca de los procesos de la fotosíntesis y el crecimiento celular en las plantas, siguiera encontrando increíble que las plantas pudieran crecer sólo con ayuda del sol, el agua y la tierra. Muy bien. Ningunas consecuencias filosóficas han de extraerse de la incapacidad de esta persona para comprender los hechos científicos. Por supuesto, la vida y la reproducción pueden parecer fenómenos notables y misteriosos; pero la respuesta adecuada a esto es sencillamente aceptar que algunos fenómenos de la naturaleza son notables y tal vez hasta misteriosos. Ello no significa que no puedan ser explicados por la ciencia. La capacidad de los seres para reproducirse es cosa que ahora entienden bastante bien los científicos; puede ser notable y misteriosa, por añadidura.

Por enfocar la cuestión de otra manera, considérese el argumento de que los puntos de vista fisicalista o materialista dan, típicamente, como su opinión, que los estados mentales (tanto los pensamientos como los estados conscientes) son idénticos a estados del cerebro. Hablando toscamente, arguyen, primero, que los estados mentales conscientes y otros tienen efectos sobre el mundo físico (tal vez usando los tipos de argumentación que empleamos en el capítulo II, "La imagen causal de los pensamientos", p. 100); y, segundo, que todo suceso físico es resultado de causas puramente físicas, de acuerdo con la ley física (esto a veces se denomina "el cierre causal de lo físico").[11] No puedo entrar en las razones de este segundo supuesto con ningún detalle aquí. Digamos nada más que los fisicalistas creen que ésta es la consecuencia de lo que hemos aprendido de la ciencia: la ciencia alcanza sus aspiraciones explicativas buscando los mecanismos subyacentes de cosas que pasan. Y buscando mecanismos subyacentes acaba revelando mecanismos físicos, la clase de mecanismos descubiertos por los físicos, la ciencia del espacio-tiempo, la materia y la energía. Según lo plantea David Lewis:

> Hay algún cuerpo unificado de teorías científicas de la clase que ahora aceptamos, que suministran reunidas una exposición verdadera y exhaustiva de todos los fenómenos físicos. Están unificados en ser acumulativos: la teoría que gobierna cualquier fenómeno físico es explicada por teorías que gobiernan fenómenos a partir de los cuales ese fenó-

[11] Para la discusión de este principio, que él denomina la "compleción de la física", véase David Papineau, *Thinking about Consciousness* (Oxford, Oxford University Press, 2002). Para algo de discusión crítica, véase Tim Crane, *Elements of Mind,* cap. 2.

> meno está compuesto y por la manera como está compuesto a partir de ellos. Lo mismo es cierto de los fenómenos posteriores, y así sucesivamente hasta alcanzar partículas elementales o campos gobernados por unas cuantas leyes sencillas, más o menos como se conciben por la física teórica del día de hoy.[12]

Es esta clase de cosa la que fundamenta la confianza de los fisicalistas en la idea de que, a fin de cuentas, todos los efectos físicos son resultado de causas físicas. Concluyen entonces que, si las causas mentales realmente tienen efectos en el mundo físico, entonces deben ser físicas ellas mismas. Pues si las causas mentales no fuesen físicas, entonces serían efectos físicos ocasionados por causas no físicas, lo cual contradice el segundo supuesto.

Hay una discusión muy general para identificar los estados mentales con los estados físicos (por ejemplo, estados del cerebro). Llamémosla "argumento causal del fisicalismo". Aunque descansa en algún supuesto científico o empírico acerca de la estructura causal del mundo físico, la argumentación causal en pro del fisicalismo no descansa en que los científicos hayan de hecho descubierto la base en el cerebro (lo que tienden a llamar "correlato neural")[13] o cualquier estado mental particular. Aunque la mayoría de los físicos piensan que semejantes correlatos neurales serán encontrados a fin de cuentas, no están *presuponiendo* que se encontrarán; todo lo que están presuponiendo en esta

[12] David Lewis, "An Argument for the Identity Theory", en *Philosophical Papers*, vol. I (Oxford, Oxford University Press, 1985), p. 105.

[13] Véase Ned Block, "How to Find the Neural Correlate of Consciousness", en A. O'Hear (ed.), *Contemporary Issues in the Philosophy of Mind* (Cambridge, Cambridge University Press, 1998).

argumentación es la naturaleza causal de los estados mentales y el cierre causal del mundo físico. Se sigue que uno podría objetar la conclusión de la argumentación ya fuera objetando la naturaleza causal de los estados mentales, o bien objetando cierre causal del mundo físico, o diciendo que hay alguna confusión o falacia al pasar de estos dos supuestos a la conclusión de que los estados mentales son estados del cerebro.

Adviértase que no es una objeción seria a esta conclusión decir sencillamente: "los estados mentales no *parecen* ser los estados del cerebro". Esto es, debe admitirse un pensamiento muy natural. Pues es verdad que cuando uno introspecciona sus estados mentales —en el caso de intentar poner en claro qué está uno pensando, por ejemplo— no *parece* como si estuviéramos obteniendo alguna clase de acceso directo a las neuronas y sinapsis de nuestros cerebros. Si el argumento anterior es correcto, entonces este testimonio de introspección no viene al caso. Pues si es verdad que los estados mentales son estados del cerebro, entonces será verdad que, de hecho, ser cierto estado cerebral le parecerá a usted que es de determinada manera, aunque pudiera no parecer un estado cerebral. Eso está bien; puede parecerle a usted que George Orwell escribió *1984* sin que le parezca que Eric Blair lo hizo, aun cuando, de hecho, Eric Arthur Blair escribiera *1984*. (Los lógicos dirán que "me parece que..." es un *contexto intensional:* véase el capítulo 1, "Intencionalidad", p. 63). La conclusión de la argumentación causal para el fisicalismo es que los estados mentales son estados cerebrales. Objetar diciendo "¡pero de fijo los estados mentales no pueden ser estados cerebrales porque no nos parece que lo sean!" no es plantear una objeción genuina: es nada más rechazar la conclusión de la ar-

gumentación. Es como si alguien dijera, en respuesta a la pretensión de que la materia es energía, "¡la materia no puede ser energía porque no parece energía!" En general, cuando alguien afirma alguna proposición, P, no es una auténtica objeción decir "¡pero P no parece ser cierto; por lo tanto no es cierto!" Y el punto no es que uno pudiera no estar en lo *cierto* al negar P. El punto es más bien que hay una distinción entre plantear una objeción a una tesis y negar la tesis.

Así, los estados mentales podrían ser estados cerebrales, aun si no lo parecen. Podemos ilustrar esto de otra manera, usando una historia famosa acerca de Wittgenstein. "¿Por qué la gente solía pensar que el Sol giraba alrededor de la Tierra?", preguntó una vez Wittgenstein. Cuando uno de sus discípulos replicó: "porque se ve como si el Sol girase alrededor de la Tierra", él contestó: "¿y cómo se vería si la Tierra girase alrededor del Sol?" La respuesta, por supuesto, es: exactamente igual. Así, podemos establecer un punto paralelo en el caso de la mente y el cerebro: ¿por qué algunas personas piensan que los estados mentales no son estados cerebrales? Respuesta: porque los estados mentales no parecen como los estados cerebrales. Respuesta: pero ¿cómo lo parecerían si fueran estados mentales? y la respuesta a esto, por supuesto, es exactamente la misma. Por lo tanto, no hay inferencia sencilla del hecho de que estar en un estado mental hace que las cosas se vean de cierto modo, a cualquier conclusión acerca de si los estados mentales tienen o no una naturaleza física.

Ninguna inferencia *sencilla;* pero tal vez hay una más complicada oculta dentro de ésta (una objeción muy natural, por cierto). Algunos filósofos piensan así; y piensan que es la *conciencia* lo que verdaderamente causa la dificultad para el físico (y, según veremos, para la mente mecánica

también). Hay varias versiones de este problema de la conciencia para el fisicalismo. Aquí intentaré extraer la esencia del problema; la selección de "Lecturas adicionales" (p. 360) indicará al lector cómo puede explorarlo más profundamente.

La esencia del problema de la conciencia deriva del hecho evidente de que cualquier descripción fisicalista de estados conscientes parece ser, en palabras de Nagel, "lógicamente compatible con la ausencia de conciencia". Puede establecerse este punto por comparación con otros casos de identificaciones científicas (identificaciones de fenómenos cotidianos con entidades descritas en lenguaje científico). Considérese, por ejemplo, la identificación del agua con H_2O. La química ha descubierto que la materia que llamamos "agua" está hecha de moléculas que a su vez están hechas de átomos de hidrógeno y oxígeno. No hay nada más en ser agua que estar compuesto de moléculas H_2O; por eso decimos que el agua *es* (o sea que *es idéntica a*) H_2O. Dado esto, pues, no es lógicamente posible para H_2O existir y que el agua no exista; ¡después de todo, son la misma cosa! Preguntar si podía haber agua sin H_2O es como preguntar si pudo haber George Orwell sin Eric Arthur Blair. Claro que no; son la misma cosa.

Si un estado mental consciente —por ejemplo un dolor de cabeza— fuese realmente idéntico a un estado cerebral (llámese "B" para simplificar), entonces en un modo similar sería imposible para B existir y para el dolor de cabeza no existir. Pues, después de todo, se supone que son la misma cosa. Esto parece diferente del caso del agua y de H_2O. Pues en tanto que la existencia de agua sin H_2O parece absolutamente imposible, la existencia de B sin el dolor de cabeza parece ser posible. ¿Por qué? La respuesta breve es: porque podemos coherentemente concebir o imaginar B

sin que exista el dolor de cabeza. Podemos concebir, al parecer, un ser que esté en todos los mismos estados cerebrales en que estoy yo cuando tengo un dolor de cabeza pero que de hecho no padezca el dolor de cabeza. Los seres imaginarios como éste se conocen en la literatura filosófica como "zombis": un zombi es una réplica física de un ser consciente que no es en realidad consciente.[14] La idea básica que hay tras el experimento mental sobre el zombi es que, aunque no parezca posible tener H_2O sin agua, parece posible (en vista de la posibilidad de zombis) tener un estado cerebral sin un estado consciente; así, la conciencia no puede ser idéntica a ningún estado cerebral o constituida por él.

¡Esto parece un modo muy apresurado de refutar el fisicalismo! Sin embargo, aunque es muy controvertida, la argumentación (cuando se plantea claramente) no lleva consigo ninguna falacia evidente. Así que planteémosla más despacio y claramente. La primera premisa es:

1. Si los zombis son posibles, entonces el fisicalismo es falso.

Como vimos en el capítulo 1, el fisicalismo ha sido definido de muchas maneras. Aquí sencillamente consideraremos que el punto de vista significa la conclusión del argumento causal anterior: los estados mentales (incluyendo estados conscientes e inconscientes) son idénticos a los estados del cerebro. La argumentación contra el fisicalismo no cambia sustancialmente, sin embargo, si decimos que, en

[14] Véase Chalmers, *The Conscious Mind,* para una discusión detallada de los zombis; para una versión anterior de la misma idea, véase Ned Block, "Troubles with Functionalism", en Block (ed.), *Readings in the Philosophy of Psychology,* vol. 1.

lugar de ser idénticos a estados del cerebro, los estados mentales están *constituidos* exhaustivamente por estados del cerebro. La identidad y la constitución son relaciones diferentes, ya que la identidad es *simétrica* donde la constitución no lo es (véase el capítulo 1, "Imágenes y parecido", p. 38, para este rasgo de las relaciones). Si Orwell *es idéntico a* Blair, entonces Blair *es idéntico a* Orwell. No obstante, si un parlamento *está constituido por* sus miembros, no se sigue que los miembros *sean constituidos por* un parlamento. Ahora bien, uno podría decir que los estados de conciencia están constituidos por estados del cerebro, o podría uno decir que son idénticos a estados del cerebro. De una o de otra manera, la primera premisa no parece ser cierta. Pues ambas ideas son maneras de expresar la idea de que los estados conscientes *no sean algo por encima y acerca de* estados del cerebro. Expresándolo metafóricamente, la idea básica es que, de acuerdo con el fisicalismo, todo lo que Dios necesita hacer para crear mis estados de conciencia es crear mi cerebro físico. Dios no necesita añadir nada más. Así, si pudiera mostrarse que crear mi cerebro no es suficiente para crear mis estados de conciencia, entonces el fisicalismo sería falso. Mostrar que son posibles los zombis es una manera de mostrar que crear mi cerebro no es suficiente para crear mis estados de conciencia. Por eso es por lo que la premisa 1 es verdadera.

La siguiente premisa es:

2. Los zombis son concebibles (o imaginables).

Lo que esto significa es que podemos imaginar de modo coherente una réplica física de un ser consciente (por ejemplo yo) sin absolutamente ninguna conciencia. Este zombi-

yo tendría todos los mismos estados físicos que yo, la misma apariencia externa, y el mismo cerebro y así sucesivamente. Sin embargo, no sería consciente: no tendría sensaciones ni percepciones ni pensamientos ni imaginación, nada. Quizá podemos permitirle que tenga toda clase de estados mentales inconscientes (la clase descrita en el capítulo 1, "Pensamiento y conciencia", p. 57). Lo que no tiene es conciencia de cualquier clase. Evidentemente, cuando estamos imaginando al zombi, lo imaginamos desde el "exterior"; no podemos imaginarlo desde "adentro", desde el punto de vista propio del zombi. Pues, por supuesto, no hay tal cosa como el punto de vista de un zombi.

Seamos claros acerca de lo que dice la premisa 2. Si alguien afirma la premisa 2, no está diciendo que *realmente haya zombis* o que *hasta donde pudieran todos ustedes ser zombis,* o que sea posible en cualquier sentido *realista* o *científico.* En absoluto. Puede negarse de plano que haya zombis, negar que tenga yo cualquier duda acerca de si usted es consciente, y negar que pudiera haber, coherente con las leyes de la naturaleza tal como las conocemos, nada semejante, y uno puede sostener de todos modos la premisa 3. Ésta afirma la mera posibilidad, desnuda, de réplicas físicas que no son conscientes.

No hay contradicción evidente en enunciar la hipótesis del zombi. Tal vez hay alguna, no evidente, oculta de alguna manera en los supuestos que estamos haciendo, que muestre por qué la premisa 2 es realmente falsa. Tal vez estamos solamente *pensando* que imaginamos al zombi, pero realmente no estamos imaginando nada de manera coherente. Puede ocurrir que alguien trate de imaginar algo y parezca imaginarlo, pero no logre realmente imaginar precisamente *esa cosa* porque no es realmente posible. Podría yo,

por ejemplo, tratar de ser mi hermano e imaginarlo. Pienso que puedo imaginar esto, vivir donde él vive, hacer lo que él está haciendo. Mas por supuesto no puedo literalmente *ser* mi hermano, nadie puede literalmente *ser idéntico* a algo distinto. Esto es imposible. Así, puede ser que falle al imaginar literalmente que soy mi hermano, y realmente imaginar otra cosa. Tal vez lo que estoy realmente imaginando es a mí, a mí mismo, viviendo una vida muy análoga a la vida de mi hermano. Podemos decir algo similar acerca del caso paralelo del agua y H_2O: alguien podría pensar que puede imaginar agua que no sea H_2O, pero que tiene alguna otra estructura química. Puede sostenerse que en realidad no se imagina esto, sino más bien algo que tiene todo el aspecto del agua pero no es agua (ya que el agua es, por hipótesis, H_2O).[15] Así, alguien puede fallar al imaginar algo porque es imposible: la premisa 2 sería falsa.

Hay, sin embargo, otra manera de criticar la argumentación: podríamos convenir en que ser mi hermano es imposible; pero todo lo que muestra esto es que uno puede imaginar cosas imposibles. Con otras palabras, podríamos aceptar las primeras dos premisas de esta argumentación, pero rechazar el traslado a la premisa siguiente:

3. Los zombis son posibles.

Evidentemente, la premisa 3 y la premisa 1 implican la conclusión:

4. El fisicalismo es falso.

[15] Esta idea viene de las influyentes discusiones de Saul Kripke en *Naming and Necessity* (Oxford, Blackwell, 1980), conferencia III.

Así, alguien que desee defender el fisicalismo debe concentrarse en el punto clave de la argumentación, pasar de la premisa 2 a la premisa 3. ¿Cómo se supone que procede este movimiento? La premisa 2 se supone que proporciona la razón de creer en la premisa 3. La argumentación dice que debemos creer la premisa 3 a causa de la verdad de la premisa 2. Nótese que una cosa es decir que si X es concebible entonces X es posible, y muy otra cosa decir que el ser concebible es la *misma cosa* que ser posible. Esto es implausible. Algunas cosas pueden ser imaginables sin ser realmente posibles (por ejemplo alguien podría imaginar un contraejemplo a una ley de la lógica), y algunas cosas son posibles sin ser imaginables (por ejemplo, por lo que a mí toca, encuentro imposible imaginar o visualizar el espacio-tiempo curvado). La imaginabilidad y la posibilidad no son la misma cosa. Sin embargo se relacionan, de acuerdo con este argumento: la imaginabilidad es el mejor testimonio que hay del ser posible algo. En tanto que la percepción está con respecto a lo que es real, así la imaginación está para lo que es posible. Percibir algo es buena evidencia de que es real; imaginar algo es buena evidencia de que algo es posible. Lo real no es nada más lo percibible, precisamente como lo posible no es nada más lo imaginable.

El fisicalista responderá a esto que aunque puede ser verdad, en general, que la imaginación es una buena guía a la posibilidad, no es infalible, y puede despistar (recuérdese el ejemplo de Churchland del cuarto luminoso en el capítulo III, "El cuarto chino", p. 201). Y entonces sostendrán que el debate acerca de la conciencia y los zombis es un área que nos hace despistarnos. Imaginamos algo, y lo creemos posible; pero nos engañamos. Dadas las razones independientes proporcionadas para la verdad del fisicalismo (el argumento

causal, antes), sabemos que no es posible. Así, lo que podemos imaginar es, estrictamente hablando, irrelevante a la verdad del fisicalismo. Esto es lo que el fisicalista diría.

En resumen: hay dos maneras como un fisicalista puede responder al argumento del zombi. La primera es negar la premisa 2 y mostrar que los zombis no son coherentemente concebibles. La segunda es aceptar la premisa 2 y rechazar el tránsito de la premisa 2 a la premisa 3. Así, para el fisicalista, o bien los zombis son inconcebibles e imposibles, o son concebibles pero imposibles. Me parece que la segunda línea de ataque es menos plausible: pues si los fisicalistas convienen en que, en algunos casos, la imaginabilidad es una buena guía a lo posible, entonces ¿qué pasa con este caso particular? Los fisicalistas estarían mejor asumiendo más lo primero, y haciendo un intento por negar que los zombis son realmente, genuinamente, concebibles. Tienen que encontrar alguna confusión o incoherencia oculta en la historia del zombi. Mi propio punto de vista es que no hay tal incoherencia; pero aquí las cosas son muy complicadas.

Los límites del conocimiento científico

Supóngase que el fisicalista puede mostrar que hay una confusión oculta en la historia del zombi, tal vez los zombis son concebibles al fin, pero no realmente posibles. El enlace entre el cerebro y la conciencia es necesario, pese a las apariencias en contra. De todos modos no con esto queda el fisicalismo arreglado como es debido. Pues hay argumentos, relacionados con el del zombi, que aspiran a mostrar que, aun si tal fuese el caso, el fisicalismo seguiría teniendo una limitación epistemológica: habría, a pesar de todo, cosas

que el fisicalismo no podría explicar. Incluso si el fisicalismo fuese metafísicamente correcto —correcto en lo que enuncia en general acerca del mundo— su exposición de nuestro conocimiento del mundo será necesariamente incompleta.

La manera más sencilla de ver esto es bosquejar brevemente un famoso argumento expresado en su forma más rigurosa en años recientes por Frank Jackson: éste lo llamó el "argumento del conocimiento".[16] Planteemos el argumento de este modo. Primero, imagínese que Luis es un científico brillante, experto absoluto en física, fisiología y psicología del gusto, y en todos los hechos científicos acerca de la elaboración de vino, pero nunca en realidad ha probado el vino. Entonces un día Luis prueba algo de vino por vez primera. "¡Increíble!", dice, "¡así es como sabe el Chateau Latour!, ahora ya sé".

Esta pequeña historia puede ahora proporcionar la base de una argumentación con dos premisas:

1. Antes de probar el vino, Luis sabía todos los hechos físicos, fisiológicos, psicológicos y enológicos acerca del vino y de cómo catarlo.
2. Después de probar el vino, aprendió algo nuevo: a qué sabe el vino.

Conclusión: por lo tanto, no todo lo que hay que saber acerca de catar el vino es algo físico. Debe pues haber cosas no físicas que aprender acerca del vino: es decir, a qué sabe.

[16] Véase Frank Jackson, "Epiphenomenal Qualia", en Lycan (ed.), *Mind and Cognition,* y las respuestas a la argumentación reimpresas allí

La argumentación es sorprendente. Pues, si aceptamos la coherencia de la historia imaginaria de Luis, entonces las premisas parecen ser muy plausibles. La conclusión no parece seguirse, directamente, a partir de las premisas. Pues si Luis aprendió algo nuevo, entonces debió de haber algo que aprendiera. No puede aprenderse sin aprender algo. Y, como ya conocía todas las cosas físicas que hay que saber acerca del vino y de cómo catarlo, la nueva cosa que aprende no puede ser nada físico. Si esto es cierto, entonces debe ser que no todo lo que podemos saber cae dentro del dominio de la física. Y no física a secas: cualquier ciencia que uno pudiera aprender sin tener las experiencias descritas por dicha ciencia. Jackson concluyó que el fisicalismo es falso: no todo es físico. ¿Es cierto esto?

La discusión es muy controvertida, y ha inspirado muchas respuestas críticas. Algunas personas no gustan de experimentos de pensamiento como la historia de Luis.[17] Es verdaderamente difícil ver qué podría concebiblemente andar mal con la idea de que, cuando alguien prueba vino por primera vez, aprende algo nuevo: aprende a qué sabe. Así, si vamos a encontrar algo mal en la historia misma, habría de ser la idea de que alguien podía conocer *todos* los hechos físicos acerca del vino sin probarlo. Es muy cierto que resulta difícil imaginar qué sería aprender todos estos hechos. Como dice Dennett, no se imagina a alguien que tuviera todo el dinero del mundo imaginándolo muy rico.[18] Bien, sí; pero si quiere usted verdaderamente imaginar a alguien

por David Lewis, "What Experience Teaches", y Laurence Nemirow, "Physicalism and the Cognitive Role of Acquiantance".

[17] Daniel Dennett es uno; véase *Consciousness Explained* (Londres, Allen Lane, 1991).

[18] *Consciousness Explained,* pp. 380 y ss.

con todo el dinero del mundo, de seguro no erraría gran cosa si empezara a imaginar que alguien fuera muy muy rico y luego más aún, sin tener nunca que imaginar que tuviera más de algo de una *clase diferente,* precisamente más del mismo dinero. Y lo mismo pasa con el conocimiento científico: no tenemos que imaginar a Luis con nada de un *tipo diferente* del tipo de conocimiento científico que la gente tiene hoy: sólo más de lo mismo.

La respuesta ordinaria del fisicalista a esta argumentación es más bien que no muestra que haya *entidades* no físicas en el mundo. Muestra nada más que hay *conocimiento* no físico de dichas entidades. Los *objetos* del conocimiento de Luis, arguye el fisicalista, son todos cosas físicas perfectamente ordinarias: el vino está hecho de alcohol, ácido, azúcar y otros constituyentes físicos. Y no se nos ha mostrado nada que pruebe que el cambio en el estado subjetivo de Luis es algo más que un cambio en la neuroquímica de su cerebro. Nada en la argumentación, sostiene el fisicalista, muestra que hay cualesquiera objetos o propiedades no físicos, en el cerebro de Luis o fuera de él. No obstante, conceden que hay un cambio en el estado de conocimiento de Luis: sabe algo que no sabía antes. Sin embargo, todo lo que esto significa es que los estados de conocimiento son más numerosos que las entidades acerca de las cuales son conocimiento. (Precisamente como podemos conocer al mismo hombre que Orwell y averiguar algo nuevo cuando nos enteramos de que es Blair.)

Éste no es un lugar tan cómodo para que descansen los fisicalistas, según ellos pudieran imaginar. Pues lo que esta respuesta concede es que hay, en principio, límites al tipo de cosa que la ciencia física puede explicarnos. La ciencia puede informarnos acerca de la constitución química del

vino; pero no puede decirnos a qué sabe el vino. Los fisicalistas podrían decir que esto no es gran cosa; pero, si dicen esto, tienen que abandonar la idea de que la física (o la ciencia en general) podría estar en condiciones de afirmar toda la *verdad* acerca del mundo, independientemente de las experiencias y perspectivas de los seres conscientes, pensantes. Pues hay verdades acerca de a qué sabe el vino, y éste es el género de verdades que pueden sólo aprenderse habiendo probado el vino. Son verdades que Luis no habría aprendido antes de probar el vino, creo yo, sin importar cuánta ciencia sepa. Así, hay límites a lo que la ciencia puede enseñarnos, aunque ésta es una conclusión que sólo será sorprendente o perturbadora para quienes creen que la ciencia pudiera decirnos todo, por principio de cuentas.

Así que regresemos finalmente al problema mente-cuerpo. Contrariamente a lo que podríamos haber pensado en el inicio, el problema puede ahora ser formulado clara y precisamente. La forma del problema es la de un dilema. El primer cuerno del dilema concierne a la causalidad mental: si la mente no es una cosa física, entonces ¿cómo se puede tener sentido de sus interacciones causales con el mundo físico? El argumento causal en pro del fisicalismo afirma que debemos por lo tanto concluir que la mente es idéntica a una cosa física. El segundo cuerno del dilema es que, si la mente es una cosa física, ¿cómo podemos explicar la conciencia? Expresado en términos de la argumentación del conocimiento: ¿cómo podemos explicar lo que se *siente* al probar algo, aun si catar es un fenómeno puramente físico? La causalidad nos empuja hacia el fisicalismo, pero la conciencia nos aparta de él.

Conclusión: ¿qué nos enseñan los problemas de la conciencia acerca de la mente mecánica?

¿Qué tiene que ver el problema mente-cuerpo con la mente mecánica? El punto de vista mecánico de la mente es una visión causal de la mente; pero no es necesariamente fisicalista. Así, un ataque al fisicalismo no es necesariamente un ataque a la mente mecánica. El meollo del punto de vista mecánico de la mente es la idea de que la mente es un mecanismo causal que tiene sus efectos sobre el comportamiento. La representación mental indudablemente tiene poderes causales, como vimos en el capítulo II, de modo que esto vincula la mente mecánica directamente con el problema mente-cuerpo. No hemos encontrado buena razón, en nuestra investigación en este libro, para socavar este modo de ver la representación como causalmente potente. El punto de vista mecánico tiene todavía que trabarse con el argumento causal para el fisicalismo esbozado en este capítulo; y, si se recomienda una solución fisicalista, el punto de vista tiene algo que decir acerca de los argumentos de la conciencia que forman la otra mitad del dilema constituido por el problema mente-cuerpo. Dadas las estrechas interrelaciones entre el pensamiento y la conciencia, la cuestión de la conciencia no puede ser ignorada por un defensor de la mente mecánica. (Fodor, característicamente, discrepa: "Nunca trato de pensar acerca de la conciencia. O incluso de escribir acerca de ella".)[19] La conclusión positiva es que no hemos exhumado ningún argumento poderoso con-

[19] *In Critical Condition,* p. 73.

tra la visión de que la mente es un mecanismo causal que tiene sus efectos sobre el comportamiento.

No obstante, nuestras investigaciones acerca de la mente mecánica han llevado a una conclusión amplia y negativa: parece haber un límite a las maneras como podemos dar explicaciones *reductivas* de los rasgos distintivos de la mente. Encontramos en el capítulo III que, aunque hay interesantes vínculos entre las ideas de computación y representación mental, no hay una buena razón para suponer que algo pudiera pensar sencillamente por ser una computadora: razonar no es sólo calcular. En el capítulo IV examinamos la hipótesis del mentalés como una exposición de los mecanismos subyacentes del pensamiento; pero esta hipótesis no explica reductivamente la representación mental, sino que la da por descontada. Los intentos de explicar la representación en términos no mentales, examinada en el capítulo V, se hundió en algunos problemas fundamentales acerca de la mala representación y la complejidad. Y, finalmente, en el presente capítulo hemos visto que, incluso si los ataques al fisicalismo con argumentos de "concebilidad" no tienen éxito, existen variantes que muestran que hay límites fundamentales a nuestro conocimiento científico del mundo. Quizás la lección, propiamente hablando, sea que debemos intentar contentarnos con un entendimiento de los conceptos mentales —representación, intencionalidad, pensamiento y conciencia— que se ocupa de ellos en sus propios términos, y no trata de dar explicaciones reductivas de ellos en términos de otras ciencias. Y tal vez ésta sea una conclusión que, en algún sentido, ya conocíamos. La ciencia, según se supone que observó Einstein, no puede darnos el sabor de la sopa de pollo. No obstante —cuando se piensa en ello— ¿no sería rarísimo si lo hiciese?

LECTURAS ADICIONALES

Una colección excelente de ensayos acerca de la filosofía de la conciencia es la obra *The Nature of Consciousness* editada por Ned Block, Owen Flanagan y Güven Güzeldere (Cambridge, MIT Press, 1997). Contiene el artículo clásico de Thomas Nagel "What is it Like to Be a Bat?"; Colin McGinn, "Can we Solve the Mind-body Problem?", "Epiphenomenal Qualia", de Jackson; "On a Confusion about a Function of Consciousness", de Block, y otros muchos. Véase también *Conscious Experience,* editado por Thomas Metzinger (Paderborn, Schöningh, 1995). Gran parte del proyecto de la filosofía reciente de la conciencia ha sido establecida por David Chalmers en su ambiciosa y rigurosa obra *The Conscious Mind* (Nueva York/Oxford, Oxford University Press, 1996). *Purple Haze,* de Joseph Levine (Nueva York/Oxford, Oxford University Press, 2001), ofrece una exposición clara, aunque a fin de cuentas pesimista, del problema de la conciencia para el materialismo, en términos de lo que Levine ha bautizado "brecha explicatoria". *Thinking About Consciousness,* de David Papineau (Oxford, Oxford University Press, 2002), es una defensa muy buena del punto de vista de que los problemas para el fisicalismo residen en nuestros conceptos más bien que en la sustancia del mundo. Sobre el debate acerca de la intencionalidad y los qualia, *Ten Problems of Consciousness,* de Michael Tye (Cambridge, MIT Press, 1995), es un buen lugar para partir. *Consciousness Explained,* de Daniel Dennett (Londres, Allen Lane, 1991), es una proeza filosófica y literaria, culminación del pensamiento de Dennett acerca de la conciencia; controvertida y muy legible, ningún filósofo de la conciencia puede desconocerla. *The Life of the Mind,* de Gre-

gory McCulloch (Londres/Nueva York, Routledge, 2003), ofrece una perspectiva no ortodoxa, no reductiva, sobre estas cuestiones.

GLOSARIO

actitud proposicional: término inventado por Bertrand Russell para esos estados mentales cuyo contenido es verdadero o falso, o sea proposiciones. Las creencias son las actitudes proposicionales paradigmáticas.

adaptación: rasgo de un organismo cuya naturaleza es explicada por selección natural.

algoritmo: procedimiento paso a paso para computar (hallar el valor de) una *función*. También llamado "procedimiento efectivo" o "procedimiento mecánico".

behaviorismo: en filosofía, el punto de vista de que los conceptos mentales pueden ser analizados exhaustivamente en términos de conceptos relativos al comportamiento. En psicología, la visión de que la psicología puede sólo estudiar el comportamiento, porque los "estados mentales internos" no son científicamente tratables o no existen.

carácter fenoménico: el carácter específico de una experiencia fenoménicamente consciente (véase *conciencia fenoménica*).

composicionalidad: la tesis de que las propiedades semánticas (véase *semántica*) y/o sintácticas (véase *sintaxis*) de expresiones lingüísticas complejas son determinadas por las propiedades semánticas y/o sintácticas de sus partes más sencillas y su modo de combinación.

computación: el uso de un *algoritmo* para calcular el valor de una *función*.

conciencia fenoménica: experiencia consciente en el sentido más amplio. Un ser tiene conciencia fenoménica cuando hay algo

que es parecido a ser ese ser. Un estado mental es fenoménicamente consciente cuando hay algo que es parecido a estar en ese estado mental.

contenido: un estado mental tiene contenido (llamado a veces "contenido intencional" o "contenido representacional") cuando tiene algún carácter o intencionalidad representacional. El contenido es proposicional cuando es determinable como verdadero o falso. Así, la creencia de que los peces nadan tiene contenido proposicional; el amor de Antonio por Cleopatra no lo tiene.

dualismo: en general, una doctrina es dualista cuando postula dos clases fundamentales de entidad o categoría. (Algunas veces el término es reservado para nociones según las cuales estas dos clases de entidad generan una tensión problemática; pero esto no es esencial.) El dualismo de sustancia es el modo de ver según el cual la realidad consiste en dos clases fundamentales de sustancia: sustancia mental y sustancia material (esto se llama también dualismo cartesiano, según la versión latinizada del nombre de Descartes). El dualismo de propiedad es el punto de vista de que hay dos clases fundamentales de propiedad en el mundo: mental y física.

extensión: la entidad en el mundo cuyo lugar ocupa una expresión. Así, la extensión del nombre "Julio César" es el hombre César en persona; la extensión del predicado "es un hombre" es el puesto de todos los hombres.

extensionalidad: un rasgo de los lenguajes lógicos y los contextos lingüísticos (partes de un lenguaje). Un contexto o lenguaje es extensional cuando las propiedades semánticas (verdad y falsedad) (véase *semántica*) de oraciones depende sólo de las extensiones (véase *extensión*) de las palabras constituyentes, o la verdad o falsedad de las oraciones constituyentes.

fenomenología: literalmente, una teoría de los fenómenos o apariencias. Más específicamente, el término ha sido usado por

Edmund Husserl y sus seguidores para un acceso específico al estudio de las apariencias, que implica "poner entre paréntesis" (es decir, desconocer) cuestiones acerca del mundo exterior cuando se estudian fenómenos mentales.

fisicalismo: el punto de vista de que todo es físico o bien todo es determinado por lo físico. "Físico" aquí significa el tema de la física.

función: en matemáticas, una operación que determina una salida para una entrada dada (por ejemplo suma, sustracción); una función computable es una para la cual hay un algoritmo. En biología, el propósito o papel o capacidad de un órgano en la vida del organismo (por ejemplo la función del corazón es bombear sangre por el cuerpo).

funcionalismo: en la filosofía de la mente, el punto de vista de que los estados mentales son caracterizados por sus papeles causales o perfiles causales, esto es, la pauta de entradas y salidas (o causas y efectos típicos) que son característicos de dicho estado. El funcionalismo analítico afirma que el significado del vocabulario de la psicología del sentido común proporciona conocimiento de estos papeles causales; el psicofuncionalismo dice que la psicología empírica proporcionará el conocimiento de los papeles causales.

intencionalidad: la capacidad de la mente para dirigirse a cosas que representan el mundo.

intensionalidad: rasgo de contextos lógicos o lingüísticos. Un contexto es intensional cuando no es extensional (véase *extensionalidad*).

lenguaje del pensamiento (LOT): el sistema de representación mental; hipótesis planteada por Jerry Fodor para explicar el razonamiento y otros procesos mentales. Fodor llama al sistema lenguaje porque tiene sintaxis y semántica, como los lenguajes naturales.

materialismo: algunas veces usado como sinónimo de fisicalismo. O también el punto de vista de que todo es material, esto es, hecho de materia.

mentalés: véase *lenguaje del pensamiento.*

mentalismo: el enfoque general en la filosofía y la psicología, opuesto al *behaviorismo,* que afirma la existencia de estados y procesos mentales internos que son causalmente eficaces para producir comportamiento.

premisa: en una argumentación, una premisa es una pretensión a partir de la cual se extrae una conclusión, de ordinario junto con otras premisas.

programa: conjunto de instrucciones que una computadora usa para computar una *función* dada.

psicología del sentido común: También llamada "psicología popular"; la red de supuestos acerca de los estados mentales que es empleada por los pensadores para explicar y predecir el comportamiento de otros.

psicología popular: véase *psicología del sentido común.*

qualia: el término se usa en dos sentidos: *i)* el uso amplio sostiene que qualia son aquellas propiedades de los estados mentales en virtud de las cuales tienen el *carácter* fenoménico que tienen; *ii)* el uso más estrecho dice que qualia son las propiedades no representacionales (no intencionales) de los estados mentales en virtud de los cuales tienen el *carácter fenoménico* que tienen.

semántica: estrictamente hablando, una teoría que estudia las propiedades semánticas de un sistema de lenguaje o representación. Más generalmente, esas propiedades mismas: las propiedades semánticas son las propiedades de representación que las relacionan con el mundo, o las cosas a las que se refieren. Significado, referencia y verdad son las propiedades semánticas paradigmáticas.

sintaxis: hablando estrictamente, una teoría que estudia las propie-

dades sintácticas de un lenguaje o sistema representacional. Más generalmente, aquellas propiedades mismas: las propiedades sintácticas son las propiedades formales de representaciones, que determinan si una expresión está bien formada.

teleología: la teoría de las metas o propósitos, o del comportamiento dirigido a una meta. Una teoría (por ejemplo la selección natural) puede ser una teoría de teleología aun si concluye explicando propósitos en términos de procesos causales simples.

teoría de la simulación (o simulacionismo): el punto de vista de que la práctica de la psicología del sentido común implica ante todo una técnica de imaginarse uno mismo en la posición de otra persona, y comprender su comportamiento usando este tipo de acto imaginativo.

"teoría teoría": la teoría de que la psicología del sentido común es algo análogo a una teoría científica.

Turing, máquina de: especificación abstracta de una máquina, inventada por Alan Turing, consistente en una cinta infinita con símbolos escritos en ella y un dispositivo que lee la cinta; el dispositivo puede realizar un pequeño número de operaciones sencillas: moverse a través de la cinta, leer un símbolo en la cinta, borrar un símbolo en la cinta. La idea aspira a ilustrar los rasgos más generales de la computación. Véase *tesis de Turing.*

tesis de Turing: la tesis de que cualquier función computable puede ser computada por una máquina de Turing. También llamada tesis de Church-Turing después que Alonzo Church desarrolló algunas ideas semejantes.

zombi: una réplica física imaginaria de un ser humano que carece de conciencia. Algunas veces un zombi se define como una réplica física de un ser humano que carece de qualia; pero este asunto de qualia no es esencial para la hipótesis del zombi.

CRONOLOGÍA

1473 Copérnico pone en tela de juicio la pretensión de que la Tierra fuera el centro del universo.

1616 William Harvey explica la circulación de la sangre.

1632 Galileo publica su *Diálogo de los dos máximos sistemas del mundo.*

1641 Publicación de las *Meditaciones* de René Descartes, donde esboza los principios de su nueva ciencia.

1642 Blaise Pascal inventa la primera máquina sumadora puramente mecánica.

1651 Publicación del *Leviathan* de Thomas Hobbes, donde apoyó una concepción materialista y mecanicista de los seres humanos.

1690 John Locke publica *An Essay Concerning Human Understanding.*

1694 Gottfried Wilhelm Leibniz inventa una máquina calculadora que puede también multiplicar.

1748 David Hume publica *An Inquiry Concerning Human Understanding.*

Julien de la Mettrie publica *L'Homme Machine.*

1786 Luigi Galvani informa de los resultados de estimular músculos de rana por aplicación de una corriente eléctrica.

1810 Franz Josef Gall publica el primer volumen de la *Anatomía y fisiología del sistema nervioso.*

1820 Charles de Colmar inventa una máquina que puede sumar, restar, multiplicar y dividir.

Joseph-Marie Jacquard inventa el "telar de Jacquard" para

hacer telas, que usa tableros perforados que controlan las pautas por tejer.

1822 Charles Babbage propone fabricar una máquina que haga ecuaciones diferenciales, la cual llamó "máquina de diferencia". Babbage trabajó en la máquina de diferencia durante 10 años, después de lo cual empezó a trabajar en su máquina analítica, que era (en concepto al menos) la primera computadora de propósito general.

1854 George Boole publica *The Laws of Thought.*

1856 Hermann von Helmholtz publica el primer volumen de su *Manual de óptica fisiológica.*

1858 Wilhelm Wundt, a menudo considerado uno de los fundadores de la psicología científica, se vuelve asistente de Hermann von Helmholtz.

1859 Charles Darwin publica *El origen de las especies.*

1873 Wundt publica *Principios de psicología fisiológica.*

1874 Franz Brentano publica *La psicología desde un punto de vista empírico.*

1879 Wundt establece el primer laboratorio de psicología en Leipzig.

Gottlob Frege publica su *Begriffsschrift (escrito conceptual),* obra que sentó los fundamentos para la lógica moderna.

1883 El primer laboratorio de psicología en América es establecido en la Universidad Johns Hopkins.

1886 Ernst Mach publica *El análisis de las sensaciones.*

1890 William James publica *Principles of Psychology.*

1895 Sigmund Freud y Josef Breuer publican *Estudios sobre la histeria,* la primera obra de psicoanálisis.

1896 Herman Hollerith (1860-1929) funda la Tabulating Machine Company en 1896 (se volvería la International Business Machines [IBM] en 1924). Usando algo parecido a la idea del telar de Jacquard, usó un lector de tarjetas

perforadas para computar los resultados del censo de los Estados Unidos.

1899 Se usó la aspirina por primera vez para curar dolores de cabeza.

1910 Bertrand Russell y Alfred North Whitehead publican *Principia Mathematica,* que intenta explicar las matemáticas en términos de nociones lógicas sencillas.

1913 El psicólogo behaviorista J. B. Watson publica su artículo "Psychology Behaviorist Views it".

1923 Jean Piaget publica *El lenguaje y pensamiento del niño,* obra fecunda en la psicología del desarrollo.

1931 Vannevar Bush desarrolla un calculador para resolver diferentes ecuaciones.

1932 Kurt Gödel publica sus teoremas de incomplesión en los fundamentos de las matemáticas.

1936 Alan Turing publica su artículo "On Computable Numbers", donde se esboza la idea de una máquina que llevaría su nombre.

1941 El ingeniero alemán Konrad Zuse desarrolla una computadora para diseñar aeroplanos y proyectiles.

1943 La Inteligencia Británica completa una computadora descifradora ("Colossus") para descodificar mensajes militares alemanes.

1944 Howard Aitken, de la Universidad de Harvard, trabajando con IBM, produce su primer calculador electrónico: el calculador automático de secuencia (conocido como Mark I), cuyo propósito fue crear planos balísticos para la armada de los Estados Unidos.

1945 John von Neumann planea la computadora automática variable electrónica (EDVAC). La computadora tenía una memoria que retenía un programa almacenado, así como datos, y una unidad central de procesamiento. Esta "arqui-

tectura de von Neumann" se volvió parte importante en el proyecto de computadoras.

1946 John Presper Eckert y John W. Mauchly, trabajando en la Universidad de Pensilvania, construyen el integrador y calculador numérico electrónico (ENIAC). El calculador fue una computadora para propósitos generales que computó a velocidades mil veces más rápidas que el calculador Mark I de Aitken.

1948 La invención del transistor inicia algunos cambios importantes en el desarrollo de las computadoras. El transistor ya era usado en computadoras hacia 1956.

1949 Se usa litio para tratar la depresión.

1950 Turing publica su artículo "Computing Machinery and Intelligence", que describe la "prueba de Turing" para la inteligencia ("el juego de la imitación").

1953 Francis Crick, James Watson y Maurice Wilkins descubren la estructura del ADN.

1957 Noam Chomsky publica *Syntactic Structures,* donde desarrolla su opinión de que los rasgos superficiales del lenguaje deben ser comprendidos como resultado de operaciones o transformaciones subyacentes.

1958 Jack Kilby, un ingeniero estadunidense, desarrolla el circuito integrado, que combina diferentes componentes electrónicos en un pequeño disco de silicio y permite a las computadoras volverse más pequeñas.

1960 Hilary Putnam publica "Minds and Machines", que constituye su defensa del funcionalismo en la filosofía de la mente.

1963 Donald Davidson publica "Actions, Reasons and Causes".

1971 El científico inglés C. Longuet-Higgins introduce el término "ciencia cognitiva".

1971 Este año marca el desarrollo del chip Intel 4004, que loca-

liza todos los componentes de una computadora (unidad procesadora central, memoria, etc.) en un diminuto chip.

1981 IBM introduce la primera computadora personal (PC).

1982 Publicación póstuma de *Vision* de David Marr.

1984 Apple introduce su primera computadora Macintosh, usando el mouse y la interfaz gráfica del usuario, que había sido desarrollada primeramente por Xerox en los años setenta (e irónicamente fue juzgada no viable comercialmente).

1988 Se da a conocer el Proyecto Genoma Humano, establecido en Washington, D. C.

1997 Gary Kasparov, el gran maestro de ajedrez y campeón mundial, es derrotado por *Deep Blue,* una computadora jugadora de ajedrez.

ÍNDICE DE FIGURAS

I.1. Viejo con un bastón 45
III.1. Diagrama de flujo para el algoritmo de multiplicación 151
III.2. Un diagrama de flujo para cocer un huevo 154
III.3. Una tabla de la máquina para una máquina sencilla de Turing 159
III.4. "Caja negra" ratonera 174
III.5. El interior de la ratonera 174
III.6. Caja negra multiplicadora 176
III.7. Diagrama de flujo para el algoritmo de multiplicación, otra vez 177
III.8. Un "portal-y" 186
IV.1. Las bandas de Mach 239
IV.2. Diagrama de una red conexionista 257
V.1. La imagen de Cummins, "Puente de la Torre", de la computación 320

La mente mecánica, de Tim Crane, se terminó de imprimir y encuadernar en el mes de agosto de 2008 en Impresora y Encuadernadora Progreso, S. A. de C. V. (IEPSA), Calz. San Lorenzo, 244; 09830 México, D. F. En su composición, elaborada en el Departamento de Integración Digital del FCE, por *Juan Margarito Jiménez Piña,* se usaron tipos AGaramond de 12, 10:12, 9.5:12 y 8:10 puntos. La edición estuvo al cuidado de *Nancy Rebeca Márquez Arzate* y consta de 2 000 ejemplares.

¿Cómo puede la mente representar el mundo externo? ¿Qué es el pensamiento? ¿Puede la mente ser explicada por la ciencia o requiere su propio modo no científico de explicación? ¿Puede concebirse la mente como un tipo de máquina? En *La mente mecánica*, Tim Crane ofrece una respuesta a estas preguntas, acercando al lector no especializado a los principales debates contemporáneos en torno a la filosofía de la mente, la inteligencia artificial y la ciencia cognitiva. Escrita en un lenguaje claro y ameno, la obra explica con detalle en qué consiste el problema de la representación mental, qué son y cómo funcionan las computadoras, qué son los pensamientos y cómo podrían producirlos tanto las mentes como las computadoras.

> Es la mejor introducción que conozco a la teoría de la mente computacional.
>
> JOHN SEARLE
> UNIVERSIDAD DE CALIFORNIA, BERKELEY

Tim Crane es profesor de filosofía en el University College de Londres y director del Programa de Filosofía de la Escuela de Estudios Avanzados de la Universidad de Londres. Se ha especializado en filosofía de la mente, filosofía de la percepción y metafísica.

www.fondodeculturaeconomica.com

9 789681 683511