Filosofía

Paul M. Churchland

# Materia y conciencia

*Introducción contemporánea a la filosofía de la mente*

gedisa editorial

Título original en inglés:
*Matter and Consciousness*
Revised Edition © 1988, The MIT Press, Cambridge, Massachusetts.
Original edition © 1984, Massachusetts Instituteof Technology

*Traducción:* Margarita N. Mizraji
*Diseño de cubierta:* Marc Valls

Segunda edición, abril de 1999, Barcelona



© *by* Editorial Gedisa, S.A.
Muntaner, 460, entlo., 1.ª
Tel. 93 201 60 00
08006 - Barcelona, España
*e-mail:* gedisa@gedisa.com
http://www.gedisa.com

ISBN: 84-7432-323-1
Depósito legal: B-16.568/1999

Impreso en Limpergraf
C/ Mogoda, 29-31. 08210 Barberà del Vallès

Impreso en España
*Printed in Spain*

# Indice general

*A mi padre, que me enseñó a volar,*
*y a mi madre, que me enseñó a ver*

# Prefacio a la edición revisada

Me ha dado mucha satisfacción la generosa acogida que se le ha dispensado a la primera edición de este librito, en especial a los capítulos sobre neurociencia, ciencia cognitiva e inteligencia artificial. Como suele ocurrir, en estas secciones se han concentrado la mayoría de los cambios y agregados para la edición revisada. El motivo de las modificaciones es el progreso espectacular que continúa registrándose en esas disciplinas y su conexión cada vez mayor con los problemas que plantea la filosofía de la mente. Los resultados de esas investigaciones conducen directamente a preguntas como: ¿cuáles son los elementos básicos de la actividad cognitiva?, ¿cómo se los implementa en los sistemas físicos reales? y ¿cómo es posible que el hombre ejecute algunas tareas cognitivas tan rápida y fácilmente mientras que los ordenadores las realizan con mucha dificultad o no pueden llevarlas a cabo en absoluto?

Una de las ideas centrales en la primera edición fue mi convencimiento de que los problemas de la filosofía de la mente no son independientes de los resultados teóricos y experimentales de las ciencias naturales. Esa opinión no ha cambiado. Pero la ciencia ha progresado mucho. En esta nueva edición se hace el intento de lograr que algunos de esos hallazgos resulten accesibles y comprensibles para un público

más amplio. A mi modo de ver, la significación filosófica de esos resultados reside en el apoyo que permite otorgar a las versiones reduccionista y eliminativa del materialismo. Pero mi opinión es sólo una entre muchas otras. Los invito a formarse su propio juicio.

# Prefacio

Por lo común los filósofos escriben sus libros para otros filósofos, y entre paréntesis expresan su esperanza de que la obra también les resulte de utilidad a los estudiantes y a los legos. Esas esperanzas suelen ser vanas. Con la expectativa inversa, he escrito esta obra esencial y explícitamente para quienes no son profesionales de la filosofía ni de la inteligencia artificial ni de las neurociencias. Mi propósito es apelar a la imaginación del lector común y del estudiante. Por supuesto, también abrigo esperanzas de que este conciso volumen les resulte de utilidad, como un resumen completo y como obra de consulta, a mis colegas profesionales y a los estudiantes de posgrado. Pero no escribí el libro para ellos sino para los neófitos en el campo de la filosofía de la mente.

Este libro fue concebido durante un reciente curso de grado sobre filosofía de la mente, que se dictó con la ayuda de textos muy conocidos y ampliamente consagrados. Pero, puesto que en este campo han sucedido muchas cosas en los últimos quince años, esos textos y antologías carecen hoy de actualidad. Y si bien disponemos ahora de algunas buenas compilaciones con trabajos muy recientes, se trata de obras demasiado avanzadas y de muy alto costo como para que puedan utilizarlas fácilmente los estudiantes universitarios. Al finalizar aquel curso, decidí escribir un manual más adecuado y accesible, libre de problemas fosilizados, aligerado de resúmenes históricos y dedicado especialmente a los temas nuevos. Este volumen es el resultado.

Lo escribí durante el verano de 1982, principalmente en una cabaña retirada que poseemos en Moose Lake, en el territorio virgen de Manitoba, donde los extraños somorgujos por la noche ponían una entretenida nota de sonido a mi trabajo. Y lo terminé a mediados de otoño en el Instituto de Estudios Avanzados de Princeton, donde una bandada de gansos canadienses residentes en la zona cumplía una función parecida.

Pero también he utilizado con provecho ideas y enseñanzas más sólidas. En primer lugar debo agradecer a mi amigo y colega, Larry Jordan, por haberme llevado a su laboratorio de neurofisiología durante 1981-82, en el que me permitió participar en el maratón de los "experimentos de los miércoles" organizado por él, y por haber sido el agente de muchísima diversión y enseñanzas invalorables. También debo agradecer a mis compañeros filósofos Daniel Dennett y Stephen Stich por haber gestionado mi participación en una cantidad de reuniones profesionales tanto en los Estados Unidos como en Inglaterra, y por todo lo que me han enseñado durante nuestros múltiples encuentros placenteros y provechosos. Estoy en deuda con mi amigo y colega Michael Stack, por lo que ya se ha convertido en una década de fructíferos debates sobre la mente y su lugar en la naturaleza. Y por sobre todo debo agradecer a mi esposa y colega profesional, Patricia Smith Churchland, quien me ha enseñado más sobre la mente y el cerebro que cualquier otro filósofo en el mundo.

Por último les doy las gracias a Ken Warmbrod, Ned Block, Bob Richardson, Amelie Rorty, Cliff Hooker y David Woodruff Smith por las múltiples formas en que me han alentado y por las valiosas críticas que hicieron a la primera versión de la obra. Y nunca terminaré de agradecer la ayuda del Instituto de Estudios Avanzados que puso a mi disposición los medios necesarios para terminar la obra y me dio la oportunidad de emprender varias otras actividades teóricas.

Paul M. Churchland
Princeton, NJ, 1983

# 1

# De qué trata este libro

La curiosidad del Hombre, y la astucia de su Razón, han revelado muchas cosas que la Naturaleza mantenía ocultas. La estructura del espaciotiempo, la constitución de la materia, las múltiples formas de la energía, la naturaleza de la vida misma: todos estos misterios se han convertido en un libro abierto para nosotros. Sin ninguna duda, preguntas muy profundas quedan por contestar y las revoluciones todavía nos aguardan, pero es inconmensurable la explosión que hemos producido los humanos en el ámbito de la comprensión científica en los últimos 500 años.

Pero, a pesar de este progreso general, hay un misterio central que continúa siéndolo en gran medida: la naturaleza de la *inteligencia consciente*. Ese es el tema de este libro.

Si la inteligencia consciente fuera todavía un misterio absoluto, yo no podría escribir ningún libro que tuviera utilidad. Pero en verdad se han hecho progresos alentadores. Los fenómenos que se han de indagar constituyen hoy el centro de atención normal de una gran variedad de campos de estudio conexos. A la filosofía se han unido la psicología, la inteligencia artificial, la neurociencia, la etología y la teoría de la evolución, para nombrar a las principales. Todas estas ciencias han hecho su aporte a lo que solía ser un debate puramente filosófico, y todas ellas contienen elementos sumamente promisorios.

Este libro es una introducción a los principales elementos del actual debate filosófico-científico: sus problemas funda-

mentales, las teorías alternativas, y sus argumentos y pruebas más importantes. En los últimos treinta años, la propia filosofía ha hecho progresos significativos en lo que respecta a la naturaleza de la mente: en especial al poner de manifiesto el carácter del autoconocimiento de la mente, pero también al proporcionar una concepción más clara de la naturaleza de las posibles teorías alternativas entre las cuales debemos elegir en última instancia, y al poner en claro qué tipos de pruebas son necesarias para poder efectuar entre ellas una elección inteligente.

Hay algo más importante aún, y es que las ciencias empíricas mencionadas suministran un caudal constante de informaciones pertinentes para poder efectuar esa elección racional. La psicología nos ha enseñado cosas sorprendentes acerca de la profundidad y fiabilidad de nuestro conocimiento introspectivo. (Esta es una cuestión importante, puesto que algunas teorías de la mente se basan en gran medida en aquello que presuntamente revela la introspección autoconsciente.) La psicología cognitiva y la inteligencia artificial han elaborado modelos cognitivos muy interesantes que, cuando se les "da vida" dentro de un ordenador adecuadamente programado, imitan con mucha fidelidad algunas de las complejas actividades de la inteligencia impulsada por un propósito. Las neurociencias han comenzado a dilucidar el inmenso microsistema de células cerebrales interconectadas que, en los seres vivos, supuestamente realizan esas actividades. La etología nos ha permitido comprender mejor la continuidad, y la discontinuidad, entre la inteligencia humana y la de otras criaturas. Y la teoría de la evolución ha dilucidado los extensos e intrincados procesos de selección a partir de los cuales ha surgido poco a poco la inteligencia consciente. Sin embargo, los datos son todavía ambiguos y aún no se ha hecho una elección entre todas las teorías pertinentes, de modo que el lector de esta obra podrá experimentar el placer y el entusiasmo de participar en una aventura intelectual en plena marcha.

Comenzamos nuestro estudio con la pregunta más obvia dentro de este campo. ¿Cuál es la verdadera naturaleza de los

estados y procesos mentales? ¿En qué medio se llevan a cabo y cómo se relacionan con el mundo físico? Si nos referimos a la mente, el objeto de estas preguntas es lo que los filósofos denominan el *problema ontológico*. (En el lenguaje filosófico, se trata de una pregunta acerca de qué cosas *existen* realmente y cuál es su esencia.) Esta cuestión es más conocida como el *problema mente-cuerpo*, y es muy probable que el lector ya esté familiarizado con las dos concepciones básicas que expongo a continuación. Por una parte están las teorías *materialistas* de la mente, que afirman que lo que denominamos estados y procesos mentales son simplemente estados y procesos muy sofisticados de un sistema físico complejo: el *cerebro*. Por otra parte están las teorías *dualistas*, que afirman que los estados y procesos mentales no son simplemente estados y procesos de un sistema puramente físico, sino que constituyen un tipo específico de fenómeno de naturaleza esencialmente no física.

Muchos de nosotros tenemos opiniones muy firmes sobre este tipo de cuestiones, y habrá quienes piensen que es muy fácil o muy evidente la elección entre estas alternativas, pero es prudente estar alerta, sean cuales fueren nuestras propias convicciones, por lo menos hasta saber bien cómo son las cosas. Por ejemplo, existen al menos cinco versiones del dualismo radicalmente diferentes, y una cantidad comparable de teorías materialistas, también muy diferentes entre sí. No son sólo *dos* teorías entre las que debemos elegir, sino que son ¡cerca de *diez*! y algunas de ellas se han formulado hace muy poco tiempo. En el capítulo 2 me propongo exponer todas esas teorías, una por una, y tratar de evaluar sus cualidades y defectos.

Sin embargo, cualquier decisión que se tome fundándose solamente en el material presentado en el capítulo 2 será prematura, puesto que existen muchas otras cuestiones urgentes absolutamente interrelacionadas con el problema mente-cuerpo.

Una de ellas es el *problema semántico*. ¿En dónde adquieren su *significado* los términos que utilizamos corrientemente para referirnos a los estados mentales? ¿En qué consis-

tiría una definición o un análisis adecuados de esos conceptos específicos que nos aplicamos a nosotros mismos y a otras criaturas dotadas de inteligencia consciente ? Una sugerencia —tal vez la que parece más admisible en un comienzo— es que el significado de términos como "dolor" o "sensación de calor" se aprende simplemente conectando la palabra correspondiente con un tipo determinado de estado mental, según la propia experiencia personal. Pero esta idea suscita una cantidad de problemas, uno de los cuales tal vez ya se le habrá planteado al lector en algún momento. ¿Cómo puede estar seguro de que la sensación interna a la que un amigo suyo (digamos) le atribuyó el término "dolor" es cualitativamente la misma que *usted* denomina así? Tal vez el estado interno de su amigo es radicalmente diferente del suyo, a pesar de estar conectado con conductas, lenguaje y circunstancias causales muy parecidas a las que se dan en su caso. Por ejemplo, su amigo podría actuar exactamente igual que usted, aun cuando hubiese una diferencia interna oculta. El problema es que, una vez que aparece esta inquietud escéptica, parece imposible de resolver porque no hay modo posible de que alguien llegue a tener una experiencia *directa* de los estados mentales de *otra* persona, y nada podría zanjar la cuestión salvo ese tipo de experiencia.

Si esto es así, entonces parece que ninguno de nosotros sabe, ni puede saber, cuál es el significado que tienen para otras personas los términos que aluden a los estados mentales, si es que en verdad significan algo para ellos. Sólo podemos saber el significado que tienen para cada uno de nosotros. Esta es una conclusión bastante extraña que sacamos acerca de una parte muy importante de nuestro lenguaje, ya que, después de todo, el propósito del lenguaje es la comunicación pública en el marco de una red de conocimientos compartidos.

Otra teoría del significado, que rivaliza con la anterior, postula que la significación de nuestro vocabulario psicológico corriente tiene un origen distinto. Aprender el significado del vocablo "dolor", explica, es saber que el dolor es un estado, ocasionado a menudo por un daño en el cuerpo, que a su vez

produce otros estados internos como un leve malestar o directamente pánico y da lugar a tipos característicos de conductas como sobresaltarse, protegerse y lamentarse. En suma, se dice que el rasgo esencial del dolor es una *red de relaciones causales* que conectan esa sensación con una multiplicidad de otras cosas, especialmente cosas que todo el mundo puede observar.

Los materialistas de todas las tendencias suelen preferir este último enfoque, en parte porque deja abierta la posibilidad de que los estados mentales sean realmente estados físicos. No hay ningún problema en suponer que en un estado puramente físico se establezcan las adecuadas conexiones *causales* esenciales para constituir un dolor. Y este enfoque no nos lleva directamente al escepticismo. Por otra parte, se desembaraza sin más trámite del aspecto interno, susceptible a la introspección, de nuestros estados mentales, que era la base del primer enfoque. Es comprensible que los dualistas hayan preferido en general la primera explicación del significado, a pesar de que evidentemente lleva al escepticismo. Las cualidades introspectibles o "subjetivamente evidentes" de nuestros estados mentales constituyen en cierto modo para ellos la esencia misma de lo mental, que trasciende una explicación meramente física.

Ya se habrá dado cuenta el lector de que ninguna solución del problema mente-cuerpo se sostendría fácilmente sin una solución simultánea del problema semántico. En el capítulo 3 se examinarán en detalle las principales soluciones alternativas, que son varias. En uno de los casos será necesario hacer una breve descripción de algunos conceptos elementales de la filosofía de la ciencia contemporánea, de modo que es posible anticipar la aparición de nuevos e inesperados elementos teóricos.

Estas cuestiones desembocan naturalmente en el *problema epistemológico.* (Epistemología es el estudio del conocimiento, qué es y de dónde proviene.) Este problema se divide en dos partes, ambas bastante intrincadas. La primera surge rápidamente a partir de una preocupación ya analizada: ¿sobre qué bases tenemos el derecho de suponer que otros seres

res adecuadamente programados? Una respuesta preliminar es: "En un grado inimaginable", aunque los investigadores de la IA serían los primeros en admitir que algunos de los problemas básicos siguen refractariamente sin resolverse.

El segundo programa de investigación es el campo en rápido desarrollo de las diversas *neurociencias,* que son las que se ocupan del estudio empírico del cerebro y el sistema nervioso. ¿En qué medida (cabe preguntarse) la neurofisiología, la neuroquímica y la neuroanatomía comparada pueden ayudar a esclarecer temas tales como la enfermedad mental, el aprendizaje, la visión tridimensional y la vida mental de los delfines? La respuesta es: "en gran medida", aunque los neurocientíficos serían los primeros en admitir que sólo han rozado la superficie de estos problemas.

He incluido estos capítulos para ofrecer al menos un muestreo instructivo de las investigaciones que actualmente se llevan a cabo en esas disciplinas. Por cierto no sirven para introducir en esos campos a quien aspire a convertirse en un especialista en ordenadores o en un neurocientífico. Pero sí permitirán comprender bien el modo en que la investigación empírica se conecta con los problemas analizados en este texto. (Esto es algo importante porque, como espero dejar bien en claro, la mayor parte de estas cuestiones filosóficas en última instancia son de naturaleza empírica, y se irán resolviendo a medida que progresen los distintos programas de investigación científica y vayan logrando relativo éxito.) En estos capítulos se ofrecerá también un marco de referencia conceptual que pueda servir en todo momento para abordar los nuevos problemas que se irán presentando en este campo del estudio de la mente. Y tal vez al leerlos se despierte el interés del lector por obtener más datos empíricos. Si logran sólo eso, habrán cumplido su propósito.

El último capítulo es abiertamente especulativo, como corresponde, y comienza con el intento de calcular cuál es la distribución de la inteligencia consciente en el universo en su conjunto. Es muy probable que la inteligencia constituya un fenómeno muy difundido en el universo, y en sus instancias más avanzadas inevitablemente se le presentará el problema

de tener que elaborar una concepción útil sobre lo que es la inteligencia. Ese proceso de autodescubrimiento, de medir con la propia vara, no necesariamente es algo fácil. Ni se podrá completar en un período breve, si es que en verdad puede llegar a estar *acabado* alguna vez. Pero también aquí es posible el progreso, como en cualquier tarea humana, y debemos estar dispuestos a esperar revoluciones en las ideas acerca de lo que *nosotros* somos, del mismo modo que una y otra vez hemos capeado revoluciones en nuestra concepción del universo que nos incluye. En la última parte del libro se investigan las consecuencias que tendría una revolución conceptual de ese tipo para la capacidad humana de la autoconciencia.

Con esto concluye la exposición de mis buenos deseos. Pasemos ahora a los problemas mismos.

# 2

# El problema ontológico (el problema mente-cuerpo)

¿Cuál es la verdadera naturaleza de los estados y procesos mentales? ¿En qué medio se producen y cómo se relacionan con el mundo físico? ¿Sobrevivirá mi conciencia a la desintegración de mi cuerpo físico? ¿O desaparecerá para siempre cuando mi cerebro deje de funcionar? ¿Es posible construir un sistema puramente físico como un ordenador y que posea verdadera inteligencia consciente? ¿De dónde proviene la mente? ¿Qué es?

Estas son algunas de las preguntas que abordaremos en este capítulo. Las respuestas que les demos dependen de cuál de las teorías sobre la mente resulte más racional en función de los datos, tenga mayor poder explicativo, capacidad predictiva, coherencia y simplicidad. Examinemos las teorías existentes y sometamos a consideración los argumentos a favor y en contra de cada una.

## 1. Dualismo

El enfoque dualista de la mente abarca varias teorías muy diferentes, pero todas coinciden en que la esencia de la inteligencia consciente reside en algo *no físico,* que jamás podrá entrar en la esfera de acción de ciencias como la física, la neurofisiología y la informática. El dualismo no es la concepción más difundida dentro de la comunidad filosófica y científica actual, pero es la teoría más popularizada sobre la

mente, tiene un profundo arraigo en la mayor parte de las religiones universales y ha sido la hipótesis dominante a lo largo de la historia de Occidente. De modo que resulta muy adecuado comenzar nuestro análisis a partir de aquí.

## Dualismo sustancial

La afirmación característica de este enfoque es que cada mente es una cosa no física distinta, un "paquete" individual de sustancia no física, algo que tiene una identidad independiente de cualquier cuerpo físico al que pudiera estar temporariamente "unida". Los estados y actividades mentales adquirirían su carácter específico por el hecho de ser estados y actividades de esta sustancia no física única en su género.

Esta formulación nos deja con muy pocos elementos para elaborar una caracterización *positiva* de cuál es la materia de la mente que se propone. Con frecuencia se ha acusado a este enfoque porque hasta ahora sólo había sido definido su objeto de un modo casi absolutamente negativo. Sin embargo, éste no tiene por qué ser un defecto irremediable, puesto que sin ninguna duda tenemos mucho que aprender acerca de la naturaleza básica de la mente, y tal vez con el tiempo se pueda remediar esta deficiencia. En este sentido, el filósofo René Descartes (1596-1650) es quien más se ha esforzado por dar una definición positiva de cuál es la materia de la mente, por lo cual vale la pena examinar sus ideas al respecto.

Según la teoría de Descartes, la realidad se divide en dos tipos básicos de sustancias. La primera es la materia común, cuya naturaleza consiste solamente en que es una cosa extensa: en cualquier caso tiene longitud, ancho, altura y ocupa una determinada posición en el espacio. Descartes no intentó restarle importancia a este tipo de materia. Por el contrario, fue uno de los físicos más talentosos de su época y defendía con gran entusiasmo lo que se denominaba entonces "la filosofía mecanicista". Pero existía un único aspecto de la realidad que en su opinión no podía ser explicado en términos de la mecánica de la materia: la razón consciente del Hombre. Por

este motivo postuló la existencia de un segundo tipo de sustancia, radicalmente diferente, que no tiene ninguna extensión ni posición espacial y cuya característica esencial es la actividad de *pensar*. Esta concepción se conoce con el nombre de *dualismo cartesiano.*

Según Descartes, el verdadero *sí mismo* no es el cuerpo material sino una sustancia pensante no espacial, una unidad individual de elementos mentales muy diferente de nuestro cuerpo material. Esta mente no física interactúa con el cuerpo en forma causal y sistemática. El estado físico de los órganos sensoriales de nuestro cuerpo, por ejemplo, genera en la mente experiencias visuales, auditivas o táctiles. Y los deseos y decisiones de la mente no física hacen que nuestro cuerpo ejecute conductas intencionales. Las conexiones causales que tiene con la mente hacen que nuestro cuerpo sea nuestro y no de otra persona.

Las razones principales ofrecidas en apoyo de esta concepción fueron bastante directas. En primer lugar, Descartes creía que podía determinar, solamente por medio de la introspección directa, que él era esencialmente una sustancia pensante y nada más que eso. Y, en segundo lugar, no concebía que un sistema puramente físico pudiera utilizar el *lenguaje* en forma apropiada ni efectuar *razonamientos* matemáticos, como cualquier ser humano normal. Enseguida hemos de analizar si se trata de buenas razones o no, pero antes reparemos en una dificultad que hasta el propio Descartes consideraba un problema.

Si el "material de la mente" es de naturaleza tan absolutamente diferente al del "material de la materia" —diferente hasta el punto de no tener ni masa ni forma alguna ni ninguna posición en el espacio—, entonces ¿cómo es posible que mi mente tenga siquiera algún tipo de influencia causal sobre mi cuerpo? Como sabía el propio Descartes (quien fue uno de los primeros en formular la ley de conservación del momentum), la materia común funciona en el espacio de acuerdo con leyes muy rígidas, y el movimiento ( = momentum) de los cuerpos no surge de la nada. ¿Cómo es que esta "sustancia pensante" absolutamente insustancial puede tener alguna influencia so-

bre la materia mensurable? ¿Cómo pueden tener algún tipo de contacto causal dos cosas tan diferentes? Descartes postuló la existencia de un principio material muy sutil —los "espíritus animales"— que transmiten la influencia de la mente al cuerpo en general. Pero esto no constituye ninguna solución, puesto que volvemos al mismo problema inicial: cómo algo mensurable y espacial (aun los "espíritus animales") puede interactuar con algo totalmente no espacial.

De todas maneras, el principio básico de la división que utilizó Descartes ya no parece tan admisible como lo era en su época. Ya no es útil ni exacto definir la materia común como aquello que tiene extensión en el espacio. Los electrones, por ejemplo, son trocitos de materia, pero las mejores teorías actuales los describen como partículas puntuales sin ningún tipo de extensión (inclusive carecen de una determinada posición espacial). Y, de acuerdo con la teoría de la gravitación de Einstein, un astro completo se puede encontrar en esta misma situación si se ve sometido a un colapso gravitacional total. Si verdaderamente existe una división entre la mente y el cuerpo, parece que Descartes no pudo localizar la línea divisoria.

Estas dificultades que planteaba el dualismo cartesiano llevaron a que se pensara en una forma menos extrema de dualismo sustancial, que es la que encontramos en una concepción a la que denominaré *dualismo popular*. Según esta teoría, una persona es literalmente un "fantasma dentro de una máquina": la máquina es el cuerpo humano y el fantasma es una sustancia espiritual, cuya constitución interna es totalmente diferente de la materia física pero aun así posee plenamente las propiedades espaciales. En particular, la creencia generalizada es que la mente está *dentro* del cuerpo que controla: dentro de la cabeza, es lo más común, en estrecho contacto con el cerebro.

Esta concepción no tiene por qué plantear las mismas dificultades que la de Descartes. La mente está allí mismo, en contacto con el cerebro, y la interacción entre ambos tal vez se puede entender en términos del intercambio de energía de una forma que la ciencia actual todavía no ha podido identificar ni comprender. La materia común, cabe recordarlo, es simplemente una forma o manifestación de la energía. (Se

puede pensar que un grano de arena es una gran cantidad de energía condensada o inmovilizada en un paquetito, de acuerdo con la ecuación de Einstein, $E = mc^2$.) Tal vez la materia de la mente sea una forma o manifestación bien adiestrada también de la energía, pero una forma diferente. De modo que es *posible* que este otro tipo de dualismo sea compatible con las conocidas leyes de conservación de la cantidad de movimiento y de la energía. Es una suerte para el dualismo, ya que esas leyes en particular verdaderamente están muy bien establecidas.

Este enfoque ha de resultar sumamente atractivo para muchos por una razón más, y es que por lo menos contempla la posibilidad (aunque por cierto no da ninguna garantía) de que la mente sobreviva a la muerte del cuerpo. No garantiza la supervivencia de la mente porque deja abierta la posibilidad de que esa forma peculiar de energía de la que presuntamente estaría constituida la mente sea producida y mantenida sólo en conjunción con esa forma sumamente intrincada de la materia que denominamos cerebro y de que necesariamente se desintegre cuando se desintegre el cerebro. De modo que las perspectivas de sobrevivir después de la muerte son bastante confusas, aun suponiendo que el dualismo popular esté en lo cierto. Pero aun cuando la supervivencia se presentara como una clara consecuencia de la teoría, hay aquí una trampa en la que no debemos caer. La esperanza de la supervivencia podría ser una razón para *desear* que el dualismo estuviera en lo cierto, pero no constituye una razón para *creer* que lo *está*. Para eso, sería necesario contar con pruebas empíricas independientes de que la mente en realidad sobrevive a la muerte permanente del cuerpo. Lamentablemente, y a pesar del palabrerío oportunista de la prensa sensacionalista (¡MEDICOS EMINENTES COMPRUEBAN VIDA DESPUES DE LA MUERTE!), no poseemos tales pruebas.

Como veremos más adelante en este capítulo, cuando pasemos a la evaluación, las pruebas positivas acerca de la existencia de esta nueva *sustancia* pensante no material son bastante endebles en términos generales. Esto ha impulsado a muchos dualistas a postular formas menos extremas de

dualismo, con la esperanza de reducir más la distancia entre la teoría y los datos existentes.

## Dualismo de las propiedades

La idea básica de las teorías que se agrupan bajo este nombre es que, si bien aquí no hay que considerar ninguna *sustancia* fuera del cerebro, que es algo físico, éste tiene un conjunto específico de *propiedades* que no posee ningún otro tipo de objeto físico. Estas propiedades especiales son lo no físico; de ahí la expresión *dualismo de las propiedades*. Se trata de las propiedades esperables: la de sentir dolor, la de percibir el color rojo, la de pensar que P, la de desear Q, etc. Estas son las propiedades características de la inteligencia consciente. Se sostiene que son no físicas en el sentido de que jamás podrían reducirse a los conceptos de las ciencias físicas conocidas ni podrían ser explicadas en esos términos. Se requeriría una ciencia autónoma completamente nueva —la "ciencia de los fenómenos mentales"— para poderlas comprender adecuadamente.

A partir de aquí surgen importantes diferencias entre las posiciones sostenidas. Comencemos con la que tal vez sea la versión más antigua del dualismo de las propiedades: el *epifenomenismo*. Este término es bastante ampuloso, pero el significado es simple. El prefijo griego "epi" significa "encima", y esta posición sostiene que los fenómenos mentales no forman parte de los fenómenos físicos del cerebro, que es el que determina en última instancia nuestras acciones y conducta, sino que más bien "están por encima de la refriega". Por eso son epifenómenos. Se dice que simplemente aparecen o surgen cuando el desarrollo del cerebro supera un determinado nivel de complejidad.

Pero hay algo más. El epifenomenista sostiene que si bien la causa de que se produzcan los fenómenos mentales son las diversas actividades del cerebro, estos fenómenos *a su vez no tienen efectos causales*. Son absolutamente incapaces de producir efectos causales en el mundo físico. Son *meros*

epifenómenos. (Para reforzar la idea, nos puede servir una metáfora un tanto imprecisa. Pensemos en nuestros estados mentales conscientes como si fueran chispitas de luz tenue que se producen sobre la superficie rugosa del cerebro, que la causa de que aparezcan es la actividad física del cerebro pero que a su vez no tienen ningún tipo de efecto causal sobre él.) Esto significa que la convicción universal de que nuestras acciones están determinadas por nuestros deseos, decisiones y por nuestra voluntad ¡es falsa! Nuestras acciones están escrupulosamente determinadas por hechos físicos del cerebro, que *también* son la causa de los epifenómenos que denominamos deseos, decisiones y actos voluntarios. Por lo tanto existe una constante conjunción entre volición y acciones. Pero, según el epifenomenismo, es una mera ilusión que la primera sea la causa de las últimas.

¿Qué podía motivar una concepción tan extraña? En realidad, no es muy difícil entender por qué alguien podría tomarla en serio. Pongámonos en el lugar de un neurocientífico que se ocupa de rastrear los orígenes de la conducta desde los nervios motores hasta las células activas de la corteza motora del cerebro, y de rastrear a su vez la actividad de estas células, que consiste en entradas de estímulos provenientes de otras partes del cerebro y de los diversos nervios sensoriales. Vemos que se encuentra con un sistema cabalmente físico, con una estructura y sensibilidad impresionantes, que desarrolla una actividad sumamente intrincada. Todos los elementos son inequívocamente químicos o eléctricos y el científico no encuentra absolutamente ningún indicio de esas entradas de estímulos no físicos del tipo que propone el dualismo sustancial. ¿Qué es lo que va a pensar? Desde el punto de vista de las investigaciones, la conducta humana es sin duda una función de la actividad del cerebro. Y esa opinión recibe un apoyo adicional, porque el científico sabe que el cerebro tiene esos rasgos de control de la conducta precisamente porque esos rasgos han sido despiadadamente seleccionados a lo largo de la prolongada historia evolutiva del cerebro. En suma, la sede de la conducta humana se presenta como algo enteramente físico en su constitución, en sus orígenes y en sus actividades internas.

Por otra parte, nuestro neurocientífico también cuenta con el testimonio de su propia introspección para explicar estos hechos. No podría negar que tiene experiencias, creencias y deseos ni que éstos están conectados de alguna manera con su conducta. Aquí podemos cerrar trato y admitir la *realidad* de las propiedades mentales, como propiedades no físicas, pero bajándolas a la categoría de epifenómenos impotentes que no tienen nada que ver con la explicación científica de la conducta humana y animal. Esta es la posición que adopta el epifenomenismo y ahora el lector puede darse cuenta de cuál es su fundamento. Se trata de un compromiso entre el deseo de respetar un enfoque rigurosamente científico para la explicación de la conducta y el deseo de respetar el testimonio de la introspección.

Esta "degradación" de las propiedades mentales que hace el epifenomenismo —reduciéndolas a subproductos causalmente impotentes de la actividad cerebral— les ha parecido demasiado extremada a la mayor parte de los dualistas de las propiedades y por eso ha gozado de mayor popularidad una teoría que se acerca más a las certezas del sentido común. Esta concepción, a la que podemos denominar *dualismo interaccionista de las propiedades,* difiere de la anterior sólo en un aspecto esencial: el interaccionista afirma que las propiedades mentales efectivamente tienen efectos causales sobre el cerebro y, debido a eso, también sobre la conducta. Las propiedades mentales del cerebro constituyen una parte integrante de la refriega causal general, en interacción sistemática con las propiedades físicas del cerebro. Por lo tanto, se sostiene que después de todo nuestros deseos y actos voluntarios son la causa de nuestras acciones.

Como antes, en este caso se dice que las propiedades mentales son propiedades *emergentes*, es decir, que no aparecen de ninguna manera hasta que la materia física común haya podido organizarse, a través del proceso evolutivo, para llegar a constituir un sistema lo suficientemente complejo. Ejemplos de propiedades emergentes de este tipo serían la propiedad de ser *sólido,* la de tener *color* y la de estar *vivo.* Todas requieren una organización adecuada de la materia

antes de poder aparecer. Con todo esto estaría de acuerdo cualquier materialista. Pero un dualista de las propiedades afirma además que los estados y propiedades mentales son *irreductibles*, es decir que no son simplemente rasgos organizativos de la materia física, como en los ejemplos citados, sino que son propiedades nuevas que las ciencias físicas no pueden explicar ni predecir.

Esta última condición —la irreductibilidad de las propiedades mentales— tiene mucha importancia, puesto que es lo que hace que esta posición sea dualista. Pero no casa bien con la afirmación compartida de que las propiedades mentales no aparecen hasta que la materia física haya logrado organizarse. Si así es como se producen las propiedades mentales, entonces sería posible esperar que tuviesen una versión física. Postular simultáneamente la aparición evolutiva *y* la irreductibilidad física es prima facie algo abstruso.

No es absolutamente inevitable que un dualista de las propiedades insista en ambas afirmaciones. Podría abandonar la tesis de la aparición evolutiva y sostener que las propiedades mentales son propiedades *fundamentales* de la realidad, que han estado presentes desde los comienzos del universo y que son equiparables a la longitud, la masa, la carga eléctrica y otras propiedades básicas. Inclusive tenemos un precedente histórico para una posición como esta. A fines del siglo pasado todavía existía la creencia muy generalizada de que los fenómenos electromagnéticos (como la carga eléctrica y la atracción magnética) eran simplemente una sutil manifestación inusual de fenómenos puramente *mecánicos*. Había científicos que pensaban que era bastante posible hacer una reducción del electromagnetismo a la mecánica. Creían que las ondas electromagnéticas, por ejemplo, resultarían ser simplemente oscilaciones que se propagan por un éter muy sutil pero solidificado que llena el espacio. Pero resultó que el éter no existía. De modo que las propiedades electromagnéticas pasaron a ser propiedades fundamentales por derecho propio y nos vimos obligados a agregar la carga eléctrica a la lista existente de propiedades básicas (masa, longitud y duración).

Tal vez las propiedades mentales se encuentren en las mismas condiciones que las electromagnéticas: irreductibles

pero no emergentes. Esta concepción se puede denominar *dualismo de las propiedades elementales* y tiene la ventaja de que es más clara que la anterior. Lamentablemente, la comparación con los fenómenos electromagnéticos tiene una falla muy evidente. A diferencia de las propiedades electromagnéticas, que aparecen en todos los niveles de la realidad, desde el nivel subatómico para arriba, las propiedades mentales aparecen solamente en sistemas físicos grandes que evolutivamente han llegado a tener una organización interna muy compleja. La postulación de la aparición evolutiva de las propiedades mentales a través de la organización de la materia es extremadamente fuerte. No parece que se trate de propiedades básicas ni elementales. Por lo tanto, volvemos a la cuestión de su carácter irreductible. ¿Por qué debemos aceptar esta afirmación fundamental del dualismo? ¿Por qué ser dualista?

## Argumentos en favor del dualismo

Examinaremos aquí algunas de las principales consideraciones que suelen hacerse comúnmente en apoyo del dualismo. Pospondremos un poco las críticas para poder apreciar mejor la fuerza colectiva que tienen estos argumentos.

Una de las fuentes principales de los argumentos que sostiene el dualismo es la creencia religiosa que muchos de nosotros ponemos en juego en estas cuestiones. Cada una de las religiones importantes a su manera es una teoría sobre la causa o la finalidad del universo y sobre el lugar que ocupa en él el Hombre, y muchas de ellas han adoptado la idea de que existe un alma inmortal, es decir, una forma de dualismo sustancial. Suponiendo que uno sea coherente, considerar la posibilidad de no creer en el dualismo significa pensar en descreer de la propia herencia religiosa, y para algunos esto resulta muy difícil. A éste lo denominaremos el *argumento religioso.*

Una consideración más universal es el *argumento de la introspección,* que es el siguiente. Cuando fijamos la atención

en el contenido de nuestra conciencia, no percibimos con claridad la pulsación de una red nerviosa con actividad electroquímica sino que percibimos un flujo de pensamientos, sensaciones, deseos y emociones. Parece que los estados y propiedades mentales, según se revelan en la introspección, son algo absolutamente diferente de las propiedades y los estados físicos. Por lo tanto, el veredicto de la introspección parece apoyar firmemente alguna forma de dualismo... como mínimo el dualismo de las propiedades.

Otro conjunto de importantes consideraciones se pueden agrupar en el *argumento de la irreductibilidad.* Aquí nos encontramos con una multiplicidad de fenómenos mentales a propósito de los cuales parece claro que no existe ningún tipo posible de explicación puramente física. Descartes ya había reparado en nuestra capacidad para usar el lenguaje en forma adecuada en circunstancias muy distintas y también le había causado una gran impresión la facultad humana de la Razón, en especial del modo en que se manifiesta en nuestra capacidad para el razonamiento matemático. Estas aptitudes, reflexionó, seguramente deben de superar la capacidad de cualquier sistema físico. Más recientemente, también se ha hecho referencia a las cualidades introspectibles de nuestras sensaciones (los "qualia" sensoriales) y al contenido significativo de nuestros pensamientos y creencias, como fenómenos que jamás podrían ser reducidos a lo físico. Considérese, por ejemplo, el hecho de ver un color o de oler el perfume de una rosa. Un físico o un químico podrán saberlo todo acerca de la estructura molecular de la rosa y de la del cerebro humano, argumenta el dualista, pero ese conocimiento no les permitirá predecir ni anticipar la cualidad de estas experiencias inefables.

Por último, a veces se mencionan los fenómenos parapsicológicos como argumento en favor del dualismo. La telepatía (lectura de la mente), la precognición (ver el futuro), la telequinesis (control mental de objetos materiales) y la clarividencia (conocimiento de objetos distantes) son todos fenómenos muy difíciles de explicar dentro de los límites normales de la psicología y de la física. Si estos fenómenos son

reales, bien podrían ser el reflejo de la naturaleza metafísica que el dualista le atribuye a la mente. En términos superficiales, son fenómenos *mentales*, y si además nunca serán susceptibles de recibir una explicación física, entonces por lo menos algunos fenómenos mentales deben de ser irreductiblemente *no* físicos.

En términos colectivos, estas consideraciones suelen tener mucha fuerza. Pero se pueden hacer serias críticas a cada una y debemos examinarlas también. Tomemos en primer lugar el argumento religioso. Sin duda no hay nada de malo en principio en recurrir a una teoría más general que venga en apoyo de la propia, como creo que es el caso cuando se echa mano a la religión. Pero este recurso sólo puede ser válido si lo son las credenciales de la(s) religión(es) a la(s) que apela, y en este caso no parece que se cumplan de ninguna manera estas condiciones. Por lo general, los intentos de decidir las cuestiones científicas apelando a la ortodoxia religiosa tienen una historia muy lamentable. Que las estrellas son otros soles, que la Tierra no es el centro del universo, que las enfermedades son provocadas por microorganismos, que la Tierra tiene miles de millones de años de antigüedad, que la vida es un fenómeno fisicoquímico: todos estos conocimientos decisivos fueron resistidos con mucha fuerza y a veces malintencionadamente, porque la religión dominante en la época pensaba justamente de otra manera. Giordano Bruno murió en la hoguera por propugnar la primera idea mencionada; Galileo fue obligado bajo amenaza de tortura en el sótano del Vaticano a retractarse de la segunda; la firme creencia de que la enfermedad era un castigo infligido por el Diablo llevó a utilizar prácticas de salud pública que produjeron epidemias crónicas en la mayor parte de las ciudades europeas, y las ideas acerca de la edad de la Tierra y sobre la evolución tuvieron que librar una penosa batalla contra el prejuicio religioso, inclusive en una época presuntamente ilustrada.

Dejando de lado los datos históricos, la opinión casi universal de que las propias convicciones religiosas son el resultado racional de una evaluación desapasionada de todas las principales alternativas se podría demostrar que es falsa para

toda la humanidad. Si ésa fuera verdaderamente la génesis de las convicciones de la mayoría de la gente, entonces se podría esperar que las principales creencias religiosas estuvieran distribuidas en forma más o menos aleatoria o uniforme en toda la Tierra. Pero en realidad manifiestan una tendencia muy fuerte a agruparse: el Cristianismo tiene su centro en Europa y las Américas, el Islam en Africa y el Cercano Oriente, el Hinduismo en la India y el Budismo en Oriente. Lo cual ilustra lo que todos sospechábamos de algún modo: que los principales determinantes de las creencias religiosas de la gente en general son las *fuerzas sociales*. Por lo tanto, decidir cuestiones científicas apelando a la ortodoxia religiosa equivaldría a poner las fuerzas sociales en el lugar de los datos empíricos. Por todas estas razones, los científicos y filósofos profesionales que se ocupan de la naturaleza de la mente por lo general hacen lo posible por mantener completamente fuera de la discusión las instancias religiosas.

El argumento de la introspección es mucho más interesante, puesto que trata de recurrir a la experiencia directa de cada uno. Pero el argumento es profundamente sospechoso porque supone que nuestra facultad de observación interna o introspección revela las cosas tal como son verdaderamente en su naturaleza más íntima. Este supuesto es sospechoso porque ya sabemos que nuestras otras formas de observación —vista, oído, tacto, etc.— de ninguna manera hacen eso. La superficie roja de una manzana no *se ve* como un arreglo de moléculas que reflejan fotones en determinadas longitudes de ondas, pero eso es precisamente lo que es. El sonido de una flauta no *suena* como un tren de ondas de presión sinusoidales en la atmósfera, pero eso es lo que es. El calor del aire estival no *se siente* como la energía cinética media de millones de diminutas moléculas, pero eso es lo que es. Si nuestros dolores, esperanzas y creencias *introspectivamente* no tienen el aspecto de estados electroquímicos en una red nerviosa, tal vez eso sólo se deba a que nuestra facultad de introspección, como nuestros otros sentidos, no es lo suficientemente aguda como para poner de manifiesto esos detalles ocultos. Lo cual de alguna manera es lo que cabría esperar. Por lo tanto, el argumento de la introspección no tiene ninguna fuerza, a

menos que de alguna manera pudiésemos demostrar que la facultad de introspección es totalmente diferente de todas las otras formas de observación.

El argumento de la irreductibilidad nos plantea un problema más serio, pero también en este caso tiene una fuerza menor de lo que parece a primera vista. Consideremos primero nuestra capacidad para el razonamiento matemático que tanta impresión le causó a Descartes. Desde hace unos diez años, cualquiera que disponga de cincuenta dólares puede conseguirse calculadoras electrónicas cuya capacidad para el razonamiento matemático —por lo menos la parte de cálculo— sobrepasa en mucho a la de cualquier ser humano normal. El hecho es que, en los siglos transcurridos desde la época de Descartes, filósofos, lógicos, matemáticos y científicos especializados en informática han logrado identificar los principios generales del razonamiento matemático, y los ingenieros electrónicos han inventado máquinas que calculan de acuerdo con esos principios. El resultado es un objeto manuable que hubiese dejado pasmado a Descartes. Esta derivación es muy notable, no sólo porque se ha probado que las máquinas poseen algunas de las capacidades que ostentaba la razón humana, sino también porque algunos de esos logros invaden zonas de la razón humana que según los filósofos dualistas anteriores jamás serían accesibles a ningún dispositivo meramente físico.

Aunque el debate sobre esta cuestión aún continúa abierto, el argumento de Descartes acerca del uso del lenguaje también suscita dudas. La idea de un *lenguaje de ordenador* es hoy en día un lugar común: piénsese en el BASIC, PASCAL, FORTRAN, APL, LISP y otros. Es cierto que la estructura y el contenido de estos "lenguajes" artificiales son mucho más simples que los del lenguaje natural humano, pero las diferencias son sólo de grado y no de género. Además, el trabajo teórico de Noam Chomsky y el enfoque lingüístico de la gramática generativa se han ocupado en gran medida de explicar la capacidad humana para el uso del lenguaje en términos utilizables para la simulación por ordenadores. No trato de sugerir que los ordenadores que puedan sostener una verda-

dera conversación ya sean casi una realidad. Todavía tenemos mucho que aprender y quedan problemas fundamentales por resolver (la mayor parte de los cuales tienen que ver con nuestra capacidad para el razonamiento inductivo o teórico). Pero los progresos recientes realizados en este campo no permiten dar apoyo a la afirmación de que el uso del lenguaje necesariamente le estará vedado para siempre a un sistema puramente físico. Por el contrario, tal afirmación parece bastante arbitraria y dogmática, como veremos en el capítulo 6.

La próxima cuestión también es un problema candente: ¿Sería posible llegar a explicar o predecir las cualidades intrínsecas de nuestras sensaciones, o el contenido significativo de nuestras creencias y deseos, en términos puramente físicos? Este es un reto mayúsculo para el materialismo. Pero, como veremos en secciones posteriores, ya se están llevando a cabo programas de investigación activos sobre ambos problemas y se están indagando sugerencias concretas. En realidad, no es imposible imaginar cómo serían esas explicaciones, aunque el materialismo no puede pretender aún haber resuelto ninguno de estos problemas. Hasta que lo haga, el dualismo se quedará con una carta en la manga, pero no puede pasar de allí. Lo que necesitan los dualistas para ganar la partida es la conclusión de que una reducción física es absolutamente imposible, conclusión que no han podido llegar a establecer. Las preguntas retóricas, como la que encabeza este párrafo, no constituyen argumentos. Y adviértase que es igualmente difícil imaginar de qué modo los fenómenos pertinentes podrían explicarse o predecirse exclusivamente en términos de la materia mental no física del dualismo sustancial. Aquí el problema de la explicación es un reto mayúsculo para todos, no sólo para el materialista. De modo que en esta cuestión tenemos aproximadamente un empate.

El último argumento en apoyo del dualismo recurre a la existencia de fenómenos parapsicológicos como la telepatía y la telequinesis, y sostiene que tales fenómenos mentales a) son reales, y b) están más allá de la explicación puramente física. En realidad este argumento es otra versión del de la irreductibilidad que analizamos antes y, como en aquel caso,

no resulta totalmente claro que esos fenómenos, aun siendo reales, jamás puedan ser susceptibles de recibir una explicación puramente física. El materialista podría proponer inmediatamente un mecanismo posible para la telepatía, por ejemplo. Según su concepción, pensar es una actividad eléctrica producida dentro del cerebro. Pero, de acuerdo con la teoría electromagnética, tales movimientos alternados de las cargas eléctricas deben producir ondas electromagnéticas que irradian en todas direcciones a la velocidad de la luz, y que contendrán información sobre la actividad eléctrica que las produjo. Posteriormente esas ondas pueden tener efectos sobre la actividad eléctrica de otros cerebros, es decir, sobre su actividad de pensar. Llamemos a esta teoría de la telepatía la teoría del "transmisor y receptor de radio".

De ninguna manera sugiero que esta teoría sea cierta: las ondas electromagnéticas emitidas por el cerebro son increíblemente débiles (mil millones de veces más débiles que la corriente electromagnética ambiental producida por las estaciones comerciales de radiodifusión) y es casi seguro que también van a terminar irremediablemente mezclándose entre sí. Esta es una de las razones por las cuales, en ausencia de datos sistemáticos, precisos y repetibles de la existencia de la telepatía, debemos dudar de que sea posible. Pero resulta significativo que el materialista cuente con los recursos teóricos para proponer una minuciosa explicación posible de la telepatía, si fuera algo real, lo cual es más de lo que cualquier dualista ha hecho hasta el momento. De modo que no hay por qué afirmar rotundamente que el materialista *necesariamente* se encuentra en una posición desventajosa para explicar estas cuestiones. Todo lo contrario.

Dejemos de lado lo expuesto, si les parece, porque la principal dificultad que plantea el argumento basado en los fenómenos parapsicológicos es muchísimo más simple. A pesar de los constantes anuncios y anécdotas que aparecen en la prensa popular, y a pesar de la poca información que brindan las investigaciones serias sobre estas cosas, no existen pruebas significativas ni dignas de confianza de que tales fenómenos hayan existido alguna vez. La gran distancia que media

entre las creencias populares al respecto y los datos reales es algo que en sí mismo requiere investigación, puesto que no hay ni un solo efecto parapsicológico que se pueda producir en forma repetida o fiable en ningún laboratorio adecuadamente equipado para ejecutar y controlar el experimento. Ni uno solo. Investigadores honestos han sido repetidamente embaucados por charlatanes "psíquicos" con habilidades provenientes del arte de la magia, y la historia del tema se compone en gran medida de episodios de credulidad, selección de datos, escasos controles experimentales y también directamente fraude cometido por el investigador ocasional. Si alguien efectivamente descubriera un efecto parapsicológico repetible, entonces tendríamos que volver a evaluar la situación pero, tal como están las cosas, no hay aquí ningún elemento que sirva para apoyar una teoría dualista de la mente.

Una vez que se los somete a un examen crítico, los argumentos en favor del dualismo pierden gran parte de su fuerza. Pero todavía no hemos terminado: existen argumentos en contra del dualismo que también requieren que se los analice.

## Argumentos en contra del dualismo

El primer argumento contra el dualismo que esgrimen los materialistas es la mayor *simplicidad* de su propia concepción. Uno de los principios de la metodología racional es que, en igualdad de condiciones, debe preferirse la más simple de dos hipótesis rivales. Este principio se conoce como "la navaja de Occam" —por Guillermo de Occam, el filósofo medieval que fue el primero en enunciarlo— y también se puede expresar del modo siguiente: "Para explicar los fenómenos, no se deben multiplicar las entidades más allá de lo necesario". El materialista postula un solo tipo de sustancia (la materia física) y una única clase de propiedades (las propiedades físicas) mientras que el dualista postula dos tipos de materias y/o dos clases de propiedades. Y si no hay ninguna ventaja explicativa, gana el materialista.

Sin embargo, éste todavía no es un argumento decisivo

contra el dualismo, puesto que ninguna de las dos teorías puede aún explicar todos los fenómenos estudiados. Pero la objeción tiene bastante fuerza, en especial porque no hay ninguna duda de que la materia física existe, mientras que la materia espiritual no pasa de ser una hipótesis débil.

Si esta última hipótesis nos ofreciera alguna ventaja explicativa definida que no pudiera lograrse de ninguna otra manera, entonces de muy buena gana violaríamos la exigencia de simplicidad, y tendríamos todo el derecho de hacerlo. Pero no es así, afirma el materialista. En realidad, sostiene, es justamente al revés, y esto nos lleva a la segunda objeción planteada al dualismo: la relativa *impotencia explicativa* del dualismo en comparación con el materialismo.

Consideremos, muy brevemente, los recursos explicativos con que ya cuentan las neurociencias. Sabemos que el cerebro existe y de qué está hecho. Conocemos bastante su microestructura: cómo las neuronas están organizadas en sistemas y cómo los distintos sistemas están conectados entre sí, con los nervios motores que salen de los músculos y con los nervios sensoriales que entran en los órganos de los sentidos. Conocemos bastante su microquímica: cómo las células nerviosas emiten diminutos impulsos electroquímicos a lo largo de sus diversas fibras y cómo logran que otras células también los emitan, o dejen de emitirlos. Sabemos cómo por medio de esa actividad se procesa la información sensorial, seleccionando partes importantes o menos importantes para ser enviadas a los sistemas superiores. Y conocemos en parte cómo esa actividad permite iniciar y coordinar la conducta del cuerpo. Principalmente gracias a la neurología (rama de la medicina que se ocupa de la patología cerebral), sabemos mucho acerca de las correlaciones entre lesiones en diversas partes del cerebro humano y diversas deficiencias que padecen sus víctimas. Existen una gran cantidad de deficiencias identificadas —algunas notorias, otras sutiles— que los neurólogos conocen muy bien (incapacidad de hablar, o de leer, o de comprender el lenguaje, o de reconocer rostros, o de sumar/restar, o de mover algún miembro, o de retener información en la memoria por mucho tiempo, etc.) y cuya aparición se rela-

ciona estrechamente con el daño producido en alguna parte específica del cerebro.

No se trata solamente de un catálogo de traumatismos. El crecimiento y desarrollo de la microestructura cerebral es algo de lo que también se ha ocupado la neurociencia y, al parecer, en ese desarrollo se basan diversos tipos de aprendizajes que puede efectuar el organismo. Es decir, el aprendizaje presupone cambios físicos y químicos permanentes en el cerebro. En suma, el neurocientífico puede decirnos muchas cosas sobre el cerebro, sobre su constitución y las leyes físicas que lo rigen; ya está en condiciones de explicar buena parte de nuestra conducta en términos de las propiedades físicas, químicas y eléctricas del cerebro, y cuenta con los recursos teóricos necesarios para explicar mucho más a medida que continúen las investigaciones. (En el capítulo 7 nos ocuparemos con más detalle de la neurofisiología y la neuropsicología.)

Comparemos ahora lo que puede decirnos el neurocientífico sobre el cerebro, y lo que él puede hacer con ese conocimiento, con lo que puede decirnos el dualista sobre la sustancia espiritual y lo que puede hacer con esos supuestos. ¿El dualista puede decirnos algo sobre la constitución de la materia mental? ¿Sobre los elementos no materiales que la componen? ¿Sobre las leyes que rigen su comportamiento? ¿Sobre las conexiones estructurales entre la mente y el cuerpo? ¿Sobre la modalidad de su funcionamiento? ¿Puede explicar las aptitudes y patologías humanas en términos de sus estructuras y defectos? En realidad el dualista no puede hacer nada de esto, porque nunca se ha formulado una teoría minuciosa sobre la materia mental. Comparado con los abundantes recursos y los logros explicativos del materialismo actual, el dualismo no es tanto una teoría de la mente sino un vacío que aguarda que se lo llene con una auténtica teoría de la mente.

En estos términos discute el materialista. Pero insisto, no se trata de un argumento absolutamente decisivo en contra del dualismo. El dualista puede admitir que el cerebro desempeña un papel muy importante en la administración de la percepción y también de la conducta —dentro de su concepción el cerebro es el *mediador* entre la mente y el cuerpo—

pero tal vez intente argumentar que los éxitos actuales del materialismo y sus perspectivas explicativas futuras sólo tienen que ver con las funciones mediadoras del cerebro, no con las aptitudes *centrales* de la mente no física, tales como la razón, la emoción y la propia conciencia. En lo que se refiere a estos últimos tópicos, diría, ni el dualismo *ni* el materialismo han logrado ningún éxito en la actualidad.

Pero esta respuesta no es muy buena. En lo que respecta a la capacidad de razonamiento, ya existen máquinas que ejecutan en minutos complicadísimos cálculos deductivos y matemáticos que a un ser humano le llevarían toda la vida. Y en lo que respecta a las otras dos aptitudes mentales, estudios realizados sobre la depresión, la motivación, la atención y el sueño han revelado muchos hechos interesantes y enigmáticos acerca de las bases neuroquímicas y neurodinámicas, tanto de la emoción como de la conciencia. Las aptitudes *centrales*, no menos que las periféricas, han sido el objeto de muy provechosos programas de investigación materialistas.

En todo caso, el intento dualista (sustancial) de trazar una distinción muy clara entre las aptitudes "mentales" únicas, propias de la mente no material, y las aptitudes simplemente mediadoras del cerebro, insinúa un argumento que casi llega a ser una abierta refutación del dualismo (sustancial). Si en verdad existe una entidad distinta en la que tienen lugar el razonamiento, la emoción y la conciencia, y si esa entidad sólo depende del cerebro nada más que para la entrada de experiencias sensoriales y la ejecución de actos volitivos como información de salida, *entonces se podría esperar que la razón, la emoción y la conciencia fueran relativamente invulnerables al control inmediato y a los efectos patológicos cuando se produce algún tipo de manipulación o daño cerebrales.* Pero de hecho la verdad es justamente lo opuesto. El alcohol, las drogas o la degeneración senil del tejido nervioso menoscaban, deterioran e inclusive llegan a destruir nuestra capacidad para el pensamiento racional. La psiquiatría conoce cientos de productos químicos que controlan las emociones (litio, clorpromacina, anfetamina, cocaína y otros) que producen sus efectos cuando penetran en el cerebro. Y la vulnerabilidad de

la conciencia frente a los anestésicos, la cafeína, y frente a algo tan simple como un fuerte golpe en la cabeza, demuestra su dependencia muy estrecha de la actividad nerviosa del cerebro. Todo esto tiene sentido si la razón, la emoción y la conciencia son actividades del cerebro mismo. Pero tiene muy poco si son actividades de alguna otra entidad totalmente diferente.

Podemos denominar a éste el argumento de la *dependencia nerviosa* de todos los fenómenos mentales conocidos. Adviértase que el dualismo de las propiedades no se ve amenazado por este argumento puesto que, como el materialismo, este dualismo considera al cerebro como la sede de toda actividad mental. Sin embargo, concluiremos esta sección con un argumento que no favorece a ninguna de las dos variedades del dualismo: el argumento de la *historia evolutiva.*

¿Cuál es el origen de una especie tan compleja y sofisticada como la nuestra? ¿Cuál es, en lo que respecta a esto, el origen del delfín, del ratón o de la mosca doméstica? Gracias a la paleontología, la anatomía comparada y la bioquímica de las proteínas y los ácidos nucleicos, ya no queda ninguna duda importante sobre esta cuestión. Cada especie existente es un tipo sobreviviente de una cantidad de variaciones de un tipo de organismo anterior; cada tipo anterior es a su vez un tipo sobreviviente de una cantidad de variaciones de un tipo de organismo anterior aún, y así sucesivamente descendiendo por las ramas del árbol evolutivo hasta que, hace aproximadamente tres mil millones de años, encontramos un tronco de un solo organismo muy simple o de un puñado de ellos. Esos organismos, al igual que su progenie más compleja, son estructuras moleculares movidas por energía, que se autorrestablecen y se autoduplican. (Este tronco evolutivo tiene sus raíces en una era anterior de evolución puramente química, en la cual los elementos moleculares de la vida se autorreconstruyeron.) El mecanismo evolutivo mediante el cual se ha estructurado este árbol consta de dos elementos principales: 1) la ocasional variación a ciegas en los tipos de reproducción de seres, y 2) la supervivencia selectiva de algunos de esos tipos debida a la relativa ventaja reproductora de algunos individuos de esos tipos. En el trans-

curso de períodos geológicos, este proceso puede producir una enorme variedad de organismos, algunos de ellos verdaderamente muy complejos.

Para los propósitos de nuestro análisis, lo que más nos interesa sobre la historia evolutiva normal es que la especie humana y todos sus rasgos son el resultado enteramente físico de un proceso puramente físico. Como todos los organismos, excepto los más simples, tenemos un sistema nervioso. Y por la misma razón: un sistema nervioso hace posible una orientación discriminada de la conducta. Pero un sistema nervioso no es más que una matriz activa de células, y una célula no es más que una matriz activa de moléculas. Lo único que tenemos de notable es que nuestro sistema nervioso es más complejo y poderoso que el de las otras criaturas del Señor. Entre nuestra naturaleza interior y la de las criaturas más simples hay una diferencia de grado pero no de género.

Si ésta es la versión correcta de nuestros orígenes, entonces parece que no hay ninguna necesidad, ni espacio, para incluir cualquier tipo de sustancias o propiedades no físicas en la explicación teórica sobre nosotros. Somos criaturas hechas de materia. Y debemos aprender a vivir con ese hecho.

Argumentos de este tipo son los que han impulsado a la mayor parte de la comunidad profesional (aunque no a toda) a abrazar alguna forma de materialismo. Sin embargo, no ha habido mucha unanimidad, ya que las diferencias entre las diversas posiciones materialistas son mayores aún que las que dividen al dualismo. En las cuatro secciones siguientes se examinan estas posiciones más recientes.

## Lecturas complementarias

### Sobre el dualismo sustancial

Descartes, René: *Las meditaciones,* meditación II.
Descartes, René: *El discurso del método*, parte 5.
Eccles, John C.: *The Self and its Brain*, con Karl Popper. Nueva York, Springer-Verlag, 1977.

Popper, Karl, *The Self and its Brain*, con John C. Eccles. Nueva York, Springer-Verlag, 1977

Margolis, Joseph, *Persons and Minds: The Prospects Of Nonreductive Materialism*. Dordrecht-Holland, Reidel, 1978.

Jackson, Frank, "Epiphenomenal Qualia", *The Philosophical Quarterly*, vol. 32, nº 127, abril de 1982.

Nagel, Thomas, "What is it like to Be a Bat?", *Philosophical Review*, vol. LXXXIII, 1974. Reproducido en *Readings in Philosophy of Psychology*, vol. 1, N. Block (comp.), Cambridge, MA, Harvard University Press, 1980.

## 2. Conductismo filosófico

El *conductismo filosófico* tuvo su período de mayor apogeo durante las dos décadas posteriores a la Segunda Guerra Mundial. Hubo tres movimientos intelectuales que en conjunción contribuyeron a su surgimiento. La primera motivación fue una reacción contra el dualismo. La segunda fue la idea del positivismo lógico de que el significado de toda oración en última instancia depende de las circunstancias observables que pueden llegar a verificarla o confirmarla. Y la tercera motivación fue el supuesto general de que la mayor parte de los problemas filosóficos, si no todos, son el resultado de una confusión lingüística o conceptual y que pueden resolverse (o disolverse) mediante un cuidadoso análisis del lenguaje en el que se expresa.

En realidad, el conductismo filosófico no es tanto una teoría sobre qué son los estados mentales (su naturaleza interna) sino más bien una teoría sobre cómo analizar o comprender el vocabulario que utilizamos para hablar sobre ellos. Específicamente lo que se afirma es que, cuando hablamos acerca de emociones y sensaciones y de creencias y deseos, no hablamos sobre episodios internos fantasmales, sino que se trata de una forma abreviada de hablar sobre modelos reales y potenciales de *conducta*. En su forma más fuerte y más directa, el conductismo filosófico postula que toda oración acerca de un estado mental se puede parafrasear, sin pérdida de signifi-

cado, por una oración larga y compleja acerca de cuál *sería* la conducta observable que se produciría si una determinada persona se encontrara en esta, o aquella o cualquier otra circunstancia observable.

Aquí puede resultar útil hacer una analogía con la propiedad disposicional, *ser soluble*. Decir que un terrón de azúcar es soluble no es decir que posee algún estado interno fantasmal. Simplemente es decir que *si* el terrón de azúcar se pusiera dentro del agua, se disolv*ería*. Más estrictamente,

"*x* es soluble en agua"

es equivalente por definición a

"si se pusiera *x* en agua no saturada, *x* se disolvería".

Este es un ejemplo de lo que se denomina "definición operacional". El vocablo "soluble" se define en términos de ciertas operaciones o pruebas que podrían revelar si el término verdaderamente se aplica o no en el caso que se ha de examinar.

De acuerdo con el conductista, el mismo análisis vale para estados mentales del tipo "desea pasar sus vacaciones en el Caribe", con la salvedad de que el análisis es mucho más rico. Decir que Anne desea pasar sus vacaciones en el Caribe es decir que: 1) si se le preguntara si eso es lo que quiere, respondería afirmativamente, y 2) si se le entregaran distintos folletos publicitarios sobre Jamaica y Japón, estudiaría primero los de Jamaica, y 3) si se le diera un billete para el vuelo de este viernes a Jamaica, iría, etc. A diferencia de la solubilidad, afirma el conductista, la mayor parte de los estados mentales son *disposiciones de múltiples vías*. Pero siguen siendo disposiciones.

Por lo tanto, según esta concepción, no tiene sentido preocuparse por la "relación" entre la mente y el cuerpo. Hablar sobre la mente de Marie Curie, por ejemplo, no es hablar de "algo" que ella "posee"; es hablar de una de sus extraordinarias aptitudes y disposiciones. El problema mente-cuerpo, concluye el conductista, es un seudoproblema.

El conductismo es claramente compatible con una concepción materialista de los seres humanos. Los objetos materiales pueden tener propiedades disposicionales, aun cuando sean de múltiples vías, de modo que no hay necesidad de recurrir al dualismo para dar sentido a nuestro vocabulario psicológico (Sin embargo, debe destacarse que el conductismo también es estrictamente compatible con el dualismo. Aun cuando el conductismo filosófico estuviera en lo cierto, seguiría siendo posible que nuestras disposiciones de múltiples vías estuviesen constituidas por una materia mental inmaterial y no por estructuras moleculares. Pero ésta no es una posibilidad que los conductistas hayan tomado demasiado en serio, por las muchas razones que expusimos al final de la sección anterior.)

Lamentablemente, el conductismo filosófico ha tenido dos fallas importantes que menoscabaron su credibilidad, aun para sus defensores. Evidentemente ignoró, e inclusive negó, el aspecto "interno" de nuestros estados mentales. Tener un dolor, por ejemplo, no parece meramente algo que nos lleve a lamentarnos, a sobresaltarnos, a tomar una aspirina, etc. Los dolores también tienen una cualidad intrínseca (espantosa) que se pone de manifiesto en la introspección, y cualquier teoría de la mente que ignore o niegue tales qualia simplemente no cumple con su deber.

Este problema preocupó muchísimo a los conductistas y se hicieron serios intentos para resolverlo. Pero los pormenores de esta cuestión nos llevan directamente al corazón de los problemas semánticos, de modo que pospondremos el análisis más profundo de este tema hasta el capítulo 3.

La segunda falla apareció cuando los conductistas intentaron especificar en detalle la disposición de múltiples vías presuntamente constitutiva de todo estado mental dado. La lista de condicionales necesaria para un análisis adecuado de "desea pasar sus vacaciones en el Caribe" no sólo parecía larga, sino indefinidamente o inclusive infinitamente larga, sin contar con ningún modo finito de especificar los elementos que había que incluir. Y no se puede definir bien ningún término cuyo *definiente* sea tan abierto e inespecífico. Es más, cada uno de los condicionales que intervienen en el largo

análisis es sospechoso en sí. Suponiendo que Anne verdaderamente desea pasar sus vacaciones en el Caribe, el condicional 1) antes mencionado será verdadero sólo si ella no mantiene *en secreto* sus fantasías sobre las vacaciones; el condicional 2) sólo será verdadero si ella ya no está *aburrida* de los folletos sobre Jamaica; el condicional 3) será verdadero sólo si ella no *cree* que el vuelo del viernes será secuestrado, y así sucesivamente. Pero recomponer cada condicional agregándole los requisitos pertinentes sería reintroducir una serie de elementos *mentales* para poder llegar a la definición, y ya no estaríamos definiendo lo mental exclusivamente en términos de circunstancias y conductas notoriamente observables.

En la medida en que el conductismo parecía la única alternativa al dualismo, los filósofos se mostraron dispuestos a combatir estas deficiencias con la esperanza de subsanarlas o desterrarlas. Pero a fines de las décadas de 1950 y de 1960, se pusieron en boga otras tres teorías materialistas y rápidamente se apartaron del conductismo.

(Concluyo esta sección con una advertencia. Hay que distinguir con claridad el conductismo *filosófico* que acabamos de analizar del conductismo *metodológico,* que ha ejercido gran influencia en el campo de la psicología. En su forma más contundente, esta última concepción exige que todos los nuevos términos teóricos acuñados por la ciencia de la psicología *deben* definirse en términos operacionales, con el fin de garantizar que la psicología mantenga un firme contacto con la realidad empírica. Por el contrario, el conductismo filosófico afirma que todos los términos psicológicos de sentido común que integran nuestro vocabulario precientífico *ya* adquieren su significado, cualquiera que sea, a partir de definiciones operacionales (tácitas). Las dos concepciones son distintas desde el punto de vista lógico y, para los nuevos términos teóricos, habría que utilizar una metodología prudente, aun cuando el análisis correlativo de términos mentales del sentido común está equivocado.)

Ryle, Gilbert, *The Concept of Mind.* Londres, Hutchinson and Company, 1949, caps. I y V.

Malcolm, Norman, "Wittgenstein's *Philosophical Investigations*", *"Philosophical Review,* vol. XLVII, 1956. Reproducido en *The Philosophy of Mind,* V-C Chappell (comp.), Englewood Cliffs, NJ, Prentice-Hall, 1962.

## 3. Materialismo reduccionista (la teoría de la identidad)

El *materialismo reduccionista,* conocido más comúnmente como *teoría de la identidad,* es la más directa de las diversas teorías materialistas de la mente. Su afirmación central es la simplicidad misma: los estados mentales *son* estados físicos del cerebro. Es decir, cada tipo de estado o proceso mental es *numéricamente idéntico* (es una y la misma cosa que) a algún tipo de estado o proceso físico dentro del cerebro o del sistema nervioso central. Por el momento no sabemos lo suficiente acerca del intrincado funcionamiento del cerebro como para poder enunciar verdaderamente las identidades correspondientes, pero la teoría de la identidad está convencida de que con el tiempo las investigaciones sobre el cerebro habrán de ponerlas de manifiesto. (En parte para poder evaluar esta afirmación examinaremos las investigaciones actuales sobre el cerebro en el capítulo 7.)

### Paralelos históricos

En los términos que propone el teórico de la identidad, el resultado que se predice aquí tiene algunos paralelos conocidos en nuestra historia científica. Considérese el caso del sonido. Sabemos ahora que el sonido es un tren de ondas de presión que se propagan por el aire, y que la propiedad de tener un tono alto es idéntica a la propiedad de tener una

frecuencia oscilatoria alta. Sabemos que la luz son ondas electromagnéticas, y la mejor teoría actual dice que el color de un objeto es idéntico a los tres tipos de rendimiento de reflectancia que tiene el objeto, más bien como una cuerda musical que éste pulsara, aunque las "notas" suenan en ondas electromagnéticas en vez de sonar en ondas de sonido. Ahora reconocemos que el calor o la frescura de un cuerpo es simplemente la energía del movimiento de las moléculas que lo componen: el calor es idéntico al alto valor medio de energía cinética molecular y el frío es idéntico al bajo valor medio de energía cinética molecular. Sabemos que el relámpago es idéntico a una repentina descarga en gran escala de electrones entre las nubes, o entre la atmósfera y la tierra. Lo que ahora consideramos que son "estados mentales", argumenta el teórico de la identidad, son idénticos a estados cerebrales exactamente de la misma manera.

## Reducción interteórica

Estos paralelos tan ilustrativos son el resultado de haber logrado hacer una *reducción interteórica.* Es decir, en todos estos casos una teoría nueva y eficaz logra abarcar un conjunto de proposiciones y principios que reflejan perfectamente (o casi perfectamente) las proposiciones y principios de una teoría o marco conceptual anteriores. Los principios correspondientes en la nueva teoría tienen la misma estructura que los del marco de referencia anterior y se aplican exactamente en los mismos casos. La única diferencia es que en los casos en que los viejos principios utilizaban (por ejemplo) las nociones de "calor", "está caliente" y "está frío", los nuevos principios utilizan en cambio las nociones de "energía cinética molecular total", "tiene un promedio alto de energía cinética molecular" y "tiene un promedio bajo de energía cinética molecular".

Si el nuevo marco de referencia explica y predice los fenómenos muchísimo mejor que el anterior, entonces tenemos excelentes razones para creer que los términos teóricos del *nuevo* marco conceptual son los que describen la realidad

correctamente. Pero si el viejo marco funcionaba adecuadamente, por lo menos hasta donde se sepa, y si se asemeja a una parte de la nueva teoría de la forma sistemática que acabamos de describir, entonces tenemos derecho a concluir que los viejos términos y los nuevos se refieren exactamente a las mismas cosas o expresan exactamente las mismas propiedades. La conclusión es que hemos podido aprehender exactamente la misma realidad que el marco anterior describía en forma incompleta, pero con un nuevo marco de referencia conceptual más lúcido. Y entonces estamos en condiciones de comunicar lo que los filósofos de la ciencia denominan "identidades interteóricas": la luz *son* ondas electromagnéticas, la temperatura *es* promedio de energía cinética molecular, etcétera.

Los ejemplos presentados en los dos párrafos anteriores tienen todavía en común otro rasgo más importante. Son todos casos en los que las cosas o propiedades vistas desde el extremo receptor de la reducción son cosas y propiedades *observables* dentro del marco conceptual de nuestro *sentido común,* y ponen de manifiesto que la reducción interteórica se produce no sólo entre marcos conceptuales que están en la estratósfera teórica, sino que también se pueden reducir los elementos observables cotidianos. Por lo tanto, no habría por qué sorprenderse particularmente si nuestros conocidos estados mentales introspectibles se redujeran a estados físicos del cerebro. Todo lo que se requeriría sería que alguna neurociencia con una buena capacidad explicativa se desarrollara hasta el punto en que se pudiese elaborar una "imagen refleja" adecuada de los supuestos y principios que constituyen nuestro marco conceptual corriente para los estados mentales, una imagen en la que los términos referidos a estados mentales ocuparan el lugar que tenían los términos referidos a estados mentales en los supuestos y principios relacionados con el sentido común. Si se pudiese cumplir esta condición (un tanto exigente), entonces, como en los ejemplos históricos citados, tendríamos todo el derecho de anunciar que se ha hecho una reducción y de afirmar la identidad entre los estados mentales y los estados cerebrales.

## Argumentos en favor de la teoría de la identidad

¿Qué razones tiene el teórico de la identidad para creer que la neurociencia llegará a cumplir alguna vez las poderosas condiciones necesarias para la reducción de nuestra psicología "popular"? Existen por lo menos cuatro razones y todas apuntan hacia la conclusión de que una explicación adecuada de la conducta humana y sus causas debe buscarse en las neurociencias físicas.

En primer lugar podemos referirnos a los orígenes puramente físicos y a la constitución ostensiblemente física del individuo humano. Comenzamos por ser una organización monocelular de moléculas programadas genéticamente (el huevo fecundado) y a partir de allí se produce un desarrollo mediante la adición de más moléculas cuya estructura e integración está controlada por la información codificada en las moléculas de ADN del núcleo celular. El corolario de este proceso sería un sistema puramente físico cuya conducta es el resultado de su funcionamiento interno y de sus interacciones con el resto del mundo físico. Y precisamente aquello de lo que se ocupan las neurociencias son esas operaciones internas que controlan la conducta.

Este argumento se vincula estrechamente con el que sigue. Los orígenes de cada *tipo* de animal también parece que son de índole escrupulosamente física. El argumento de la historia evolutiva que hemos analizado antes (pág. 44) otorga un respaldo adicional a estas afirmaciones del teórico de la identidad, puesto que la teoría de la evolución constituye la única explicación seria que tenemos para dar cuenta de la capacidad del cerebro y del sistema nervioso central para controlar la conducta. Estos sistemas fueron seleccionados por las múltiples ventajas (en última instancia, la ventaja de la reproducción) que otorgan a las criaturas cuya conducta se controla de esta manera. Una vez más en este caso parecería que las causas básicas de nuestra conducta se remiten a la actividad nerviosa.

El teórico de la identidad encuentra apoyo también en el argumento, analizado antes, de la dependencia nerviosa de

todos los fenómenos mentales conocidos (véase la pág. 44). Precisamente esto es lo que cabría esperar si la teoría de la identidad estuviera en lo cierto. Por supuesto, la dependencia del sistema nervioso también es una consecuencia del dualismo de las propiedades, pero en este caso el teórico de la identidad preferirá acudir a consideraciones sobre la simplicidad. ¿Por qué admitir dos clases radicalmente diferentes de propiedades y operaciones si una sola de ellas puede encargarse de la tarea explicativa?

Este último argumento surge a partir del éxito cada vez mayor que logran las neurociencias en su tarea de describir con claridad el sistema nervioso de muchos seres y explicar sus aptitudes y deficiencias conductuales en términos de las estructuras descubiertas. Todos los argumentos anteriores sugieren que la neurociencia debe lograr éxito en esta empresa, y lo cierto es que continuamente los confirma la propia historia de esta disciplina. El progreso ha sido muy rápido, especialmente en el caso de seres muy simples (como podría esperarse), pero también se ha verificado en el estudio de los seres humanos aunque, por obvias razones morales, aquí la investigación debe ser mucho más prudente y cautelosa. En suma, a las neurociencias todavía les queda un largo camino por recorrer, pero los progresos realizados hasta el momento le permiten alentar grandes esperanzas al teórico de la identidad.

Con todo, estos argumentos no son absolutamente decisivos en favor de la teoría de la identidad. Sin ninguna duda apoyan en forma abrumadora la idea de que las causas de la conducta humana y animal son de naturaleza esencialmente física, pero la teoría de la identidad no se limita a afirmar sólo esto, sino que sostiene que la neurociencia ha de descubrir una taxonomía de los estados neurales que permita establecer una correspondencia biunívoca con los estados mentales de la taxonomía del sentido común. Las afirmaciones de identidad interteórica sólo quedarán justificadas si se puede encontrar esa correlación. Pero no hay nada en los argumentos anteriores que garantice que se podrá establecer esa correspondencia entre el marco conceptual viejo y el nuevo, aun en el caso de

que el nuevo lograra un éxito colosal en la explicación y predicción de nuestra conducta. Más aún, existen argumentos provenientes de otras posiciones dentro del campo materialista que sostienen que es bastante improbable que se puedan establecer esas correspondencias tan convenientes. Pero, antes de entrar en ellos, consideremos algunas objeciones más tradicionales a la teoría de la identidad.

## Argumentos en contra de la teoría de la identidad

Comencemos con el argumento de la introspección analizado antes. La introspección nos revela un ámbito de pensamientos, sensaciones y emociones, no de impulsos electroquímicos en una red nerviosa. Los estados y propiedades mentales que pone de manifiesto la introspección parecen algo radicalmente diferente de los estados y propiedades neurofisiológicos. ¿Cómo sería posible que fueran lo mismo?

La respuesta, como ya hemos visto, es: "sin ninguna dificultad". Al discriminar entre el rojo y el azul, lo dulce y lo amargo, lo caliente y lo frío, nuestros órganos sensoriales en verdad efectúan una discriminación entre diferencias muy sutiles que existen entre complejísimas propiedades electromagnéticas, estèreoquímicas y micromecánicas de los objetos físicos. Pero nuestros sentidos no son lo suficientemente agudos como para poder revelar por sí solos los pormenores de esas complejísimas propiedades. Para esto se necesita investigación teórica y experimental con instrumentos especialmente diseñados. Presuntamente lo mismo vale para nuestro sentido "interno": la introspección, que tal vez pueda discriminar eficazmente entre una gran variedad de estados neurales, pero no sea capaz de revelar por sí sola los pormenores de esos estados entre los que discrimina. En realidad casi sería un milagro que lo lograra, del mismo modo que lo sería si la vista, sin ningún tipo de ayuda, descubriera la existencia de la interacción de campos eléctricos y magnéticos que ocurre a enorme velocidad con una frecuencia oscilatoria de mil millo-

nes de millones de hertz ($10^{15}$ hertz) y una longitud de onda mucho menor que un millonésimo de metro. Ya que, a pesar de las "apariencias", eso es la luz. Por lo tanto, el argumento de la introspección no tiene suficiente fuerza.

La objeción siguiente sostiene que la identificación de estados mentales con estados cerebrales nos llevaría a afirmaciones literalmente ininteligibles, a lo que los filósofos han denominado "errores categoriales", y también asegura que la identificación es, por lo tanto, un caso de verdadera confusión conceptual. Podemos comenzar el análisis con la referencia a una de las leyes más importantes sobre la identidad numérica. La ley de Leibniz postula que dos ítems son numéricamente idénticos sólo en caso de que cualquier propiedad que postule uno de ellos la posea también el otro: puesto en notación lógica,

$$(x)\ (y)\ [(x=y) \equiv (F)\ (Fx \equiv Fy)]\,.$$

Esta ley señala un modo posible de refutar la teoría de la identidad, que sería: encontrar una propiedad que poseyeran los estados cerebrales pero no los estados mentales (o viceversa), con lo cual la teoría quedaría desacreditada.

Con este propósito se han mencionado a veces las propiedades espaciales. Los estados y procesos cerebrales deben tener por supuesto alguna localización espacial específica: en el cerebro en su conjunto o en alguna parte de él. Y si los estados mentales son idénticos a los cerebrales, entonces deben tener exactamente la misma localización espacial. Pero no tiene absolutamente ningún sentido, sostiene este argumento, decir que mi sensación de dolor está situada en el tálamo ventral, o que mi creencia de que el sol es una estrella está situada en el lóbulo temporal del hemisferio cerebral izquierdo. Estas afirmaciones tienen tan poco sentido como decir que el número 5 es verde o que el amor pesa veinte gramos.

Con la idea de hacer la misma jugada pero en sentido inverso, se ha sostenido que no tiene ningún sentido atribuir las diversas propiedades *semánticas* a los estados cerebrales. Nuestros pensamientos y creencias, por ejemplo, tienen un

sentido, un contenido proposicional específico; son verdaderos o falsos y pueden entrar en relaciones como las de coherencia y presuposición. Si los pensamientos y las creencias fueran estados cerebrales, entonces tendrían que poseer todas estas propiedades semánticas. Pero no tiene ningún sentido, sostiene este argumento, decir que una resonancia en la corteza de asociación es verdadera, o que presupone lógicamente alguna otra resonancia cercana o que significa que P.

Ninguna de estas opciones tiene el mismo peso que tenía hace veinte años, puesto que la mayor familiaridad con la teoría de la identidad y el conocimiento cada vez mayor de las funciones del cerebro han contribuido a reducir la sensación de rareza semántica que producían las afirmaciones mencionadas. Pero, aun cuando todavía nos parezcan confusas desde el punto de vista semántico, esto no tiene mayor importancia. La afirmación de que el sonido tiene una longitud de onda, o de que la luz tiene frecuencia, deben de haber parecido igualmente ininteligibles antes de que se tuviera la certeza de que tanto el sonido como la luz son fenómenos de ondas. (Considérese por ejemplo cómo el obispo Berkeley en el siglo XVIII se negó a aceptar la idea de que el sonido es un movimiento vibratorio del aire, en el Diálogo I de sus *Tres diálogos*. Estas objeciones las expresa Filón. La afirmación de que el calor se mide en kg $\chi$ $m^2$ / segundos$^2$ hubiese parecido monstruosa desde el punto de vista semántico antes de que se supiera que la temperatura es el promedio de energía cinética molecular. Y en el siglo XVI, la afirmación hecha por Copérnico de que la Tierra se *mueve* también sonó absurda hasta el punto de que se la consideró perversa, y no es difícil entender por qué. Considérese el siguiente argumento:

> La afirmación de Copérnico de que la Tierra se mueve no es más que una pura confusión conceptual. Consideremos, pues, qué *significa* decir que algo se mueve: "*x* se mueve" significa "*x* cambia de posición relativa respecto de la Tierra". Entonces, decir que la Tierra se mueve es decir que la Tierra cambia de posición relativa ¡con respecto a sí misma!, lo cual es absurdo. Por lo tanto, la posición de Copérnico es un mal uso del lenguaje.

El *análisis del significado* al que se apela aquí bien podría haber sido correcto, pero todo lo que habría significado es que el hablante se pusiera a cambiar sus significados. El hecho es que toda lengua incluye una red muy rica de supuestos acerca de la estructura del mundo, y si una oración O suscita intuiciones de rareza semántica, por lo común esto se debe a que O viola uno o más de estos supuestos básicos. Pero no siempre se rechaza O por esa razón solamente, ya que en algunos casos precisamente lo que se requiere es abandonar esos supuestos. El "mal uso" de modos de hablar aceptados con frecuencia es un aspecto esencial del verdadero progreso científico. Tal vez tendremos que acostumbrarnos a la idea de que los estados mentales tienen localizaciones anatómicas y que los estados cerebrales tienen propiedades semánticas.

Aunque dejemos de lado la acusación de puro sinsentido, el teórico de la identidad por cierto nos debe una explicación de cómo es exactamente que los estados cerebrales físicos pueden tener propiedades semánticas. La explicación más corriente se puede esbozar del modo siguiente. Comencemos por preguntarnos cómo es que una *oración* determinada (= tipo enunciativo) tiene el contenido proposicional específico que efectivamente tiene: la oración "la manzana es roja" por ejemplo. Téngase en cuenta en primer lugar que una oración siempre forma parte integrante de un sistema completo de oraciones: un lenguaje. Toda oración dada establece muchas relaciones con innumerable cantidad de otras: presupone a muchas, es presupuesta por muchas otras, es coherente con algunas, es incoherente con otras, proporciona datos que confirman algunas otras, etc. Y los hablantes que usan esa oración dentro de ese lenguaje extraen inferencias de acuerdo con esas mismas relaciones. Evidentemente cada oración (o cada conjunto de oraciones equivalentes) establece un modelo único de ese tipo de relaciones de implicación: desempeña un papel inferencial distintivo en una economía lingüística compleja. En consecuencia, decimos que la oración "La manzana es roja" tiene el contenido proposicional, *the apple is red*, porque la oración "la manzana es roja" cumple *la misma función* en español que la oración "The apple is red" cumple en inglés.

Tener un contenido proposicional pertinente simplemente es cumplir la función inferencial pertinente en una economía cognitiva.

Para volver ahora a los tipos de estados cerebrales, digamos que en principio no hay ningún problema en suponer que el cerebro de cada uno es la sede de una economía inferencial compleja en la que ciertos tipos de estados cerebrales son los elementos que cumplen funciones. De acuerdo con la teoría del significado que acabamos de esbozar, tales estados tendrían entonces contenido proposicional, puesto que tener contenido no depende de que el elemento que lo tenga sea un patrón de sonido, un patrón de letras sobre el papel, un conjunto de caracteres en Braille o un patrón de actividades nerviosas. Lo que cuenta es la función inferencial que cumple el elemento. Por lo tanto, parece que al fin y al cabo el contenido proposicional es algo que también pueden tener los estados cerebrales.

Comenzamos este apartado con un argumento en contra del materialismo que se basaba en la *naturaleza* cualitativa de nuestros estados mentales, tal como se ponía de manifiesto en la introspección. El argumento siguiente apela al simple hecho de que esos estados mentales son introspectibles sin más.

1. Mis estados mentales son introspectivamente conocidos por mí como estados de mi yo consciente.
2. Mis estados mentales *no* son introspectivamente conocidos por mí como estados de mi yo consciente.

Por lo tanto, por la ley de Leibniz (que dice que las cosas numéricamente idénticas deben tener exactamente las mismas propiedades),

3. Mis estados mentales no son idénticos a mis estados cerebrales.

En mi experiencia, ésta es la forma más atractiva del argumento de la introspección, que seduce por igual tanto a los alumnos como a los profesores. Pero es un caso flagrante

de una falacia muy conocida, que queda ilustrada con toda claridad en los siguientes argumentos paralelos:

1. Mohamed Alí es el campeón más conocido de los pesos pesados.
2. Cassius Clay *no* es el campeón más conocido de los pesos pesados.

Por lo tanto, por la ley de Leibniz,

3. Mohamed Alí no es idéntico a Cassius Clay.

o bien,

1. La aspirina es algo que John admite que alivia el dolor.
2. El ácido acetilsalicílico *no* es algo que John admite que alivia el dolor.

Por lo tanto, por la ley de Leibniz,

3. La aspirina no es idéntica al ácido acetilsalicílico.

A pesar de que las premisas correspondientes son verdaderas, ambas conclusiones son falsas: legítimamente no se pueden negar esas identidades. Lo que significa que ninguno de los dos argumentos es válido. El problema es que la "propiedad" que se otorga en la premisa 1 y se niega en la premisa 2 consiste solamente en que el elemento considerado será *reconocido*, *percibido* o *conocido* como una cosa u otra. Pero este modo de aprehenderlo no es una auténtica propiedad del elemento en sí mismo, adecuada para adivinar identidades, puesto que se puede llegar a reconocer uno y el mismo sujeto bajo un nombre o descripción y sin embargo no se lo reconoce cuando se da otra descripción (precisa, correferencial). Entonces, hay que decirlo en forma contundente, la ley de Leibniz no es válida en el caso de estas "propiedades" espurias. Cuando se intenta utilizarlas del modo en que acabamos de ejemplificar se comete lo que los lógicos denominan la falacia *intencional*. Lo que reflejarían las premisas no es la falla de ciertas identidades objetivas sino sólo nuestra permanente imposibilidad de apreciarlas.

También debemos considerar otra versión diferente del argumento anterior, ya que postularía que los estados cerebrales no solamente no son conocidos (todavía) por medio de la introspección, sino que no son cognoscibles por medio de la introspección en ninguna circunstancia en absoluto. Así,

1. Mis estados mentales son cognoscibles por medio de la introspección.
2. Mis estados mentales *no* son cognoscibles por medio de la introspección.

Por lo tanto, por la ley de Leibniz,

3. Mis estados mentales no son idénticos a mis estados cerebrales.

En este caso el crítico insistirá en que el hecho de ser cognoscible por medio de la introspección *es* una propiedad auténtica de una cosa, y en que esta versión modificada del argumento no comete la "falacia intencional" que hemos analizado antes.

Y así es. Pero ahora el materialista tiene la posibilidad de insistir en que el argumento contiene una premisa falsa: la 2. Porque si los estados mentales fueran efectivamente estados cerebrales, entonces lo que se nos revela en la introspección todo el tiempo son verdaderamente estados cerebrales, aun cuando no nos demos cuenta del todo de que lo son. Y si podemos aprender a considerar esos estados y reconocerlos en descripciones mentalistas, como todos lo hacemos, entonces sin ninguna duda podemos aprender a considerarlos y reconocerlos en las descripciones neurofisiológicas más precisas. En el mejor de los casos, la premisa 2 simplemente es una petición de principio en contra del teórico de la identidad. El error queda ampliamente ejemplificado en el siguiente argumento paralelo:

1. La temperatura es cognoscible por medio de la sensación.
2. El promedio de energía cinética molecular *no* es cognoscible por medio de la sensación.

Entonces, por la ley de Leibniz,

3. La temperatura no es idéntica al promedio de energía cinética molecular.

Esta identidad, por lo menos, hace mucho tiempo que está firmemente establecida y sin duda este argumento es defectuoso: la premisa 2 es falsa. Así como se puede aprender a sentir que el aire estival es de 70 ° F o 21° C, también se puede aprender a sentir que el promedio de energía cinética de sus moléculas es de aproximadamente $6{,}2 \times 10^{-21}$ joules, ya que, nos demos cuenta o no, para eso están adaptados nuestros mecanismos. Tal vez el acceso a nuestros estados cerebrales sea algo similar. En el capítulo 8 volveremos a ocuparnos de la introspectibilidad de los estados cerebrales.

Consideremos ahora un último argumento, basado también en las cualidades introspectibles de las sensaciones. Imaginemos a un futuro neurocientífico que llega a saber todo lo que hay que saber sobre la estructura y la actividad físicas del cerebro y su sistema visual, sobre sus estados reales y posibles. Si por alguna razón este científico nunca ha *experimentado* verdaderamente la sensación del rojo (digamos, a causa de una acromatopsia o tal vez por un medio ambiente fuera de los común), entonces habrá algo que *no* sabe sobre determinadas sensaciones: *cómo es tener la sensación del rojo.* Por lo tanto, en ese conocimiento completo de los hechos físicos de la percepción visual y la actividad cerebral correspondiente, hay algo que ha sido omitido. En consecuencia, el materialismo no puede dar una explicación adecuada de todos los fenómenos mentales, y la teoría de la identidad debe de ser falsa.

El teórico de la identidad puede replicar diciendo que en este argumento se aprovecha una ambigüedad inadvertida del término "conocer". En el caso del conocimiento utópico del cerebro que tendría el científico mencionado, "conocer" significa algo parecido a "domina el conjunto correspondiente de proposiciones neurocientíficas". En el caso del conocimiento (que no posee) de cómo es tener la sensación del rojo, "conocer"

significa algo parecido a "tiene una representación prelingüística del rojo en sus mecanismos de discriminación no inferencial". Es cierto que se puede tener el primer conocimiento sin el segundo, pero el materialista no avala la idea de que poseer un conocimiento en el primer sentido automáticamente implique tenerlo en el segundo. El teórico de la identidad puede admitir que exista una dualidad, o inclusive una pluralidad, de diferentes *tipos de conocimiento* sin tener por eso que admitir que exista una dualidad de *tipos de cosas conocidas*. La diferencia entre una persona que lo sabe todo sobre la corteza visual pero nunca ha experimentado la sensación del rojo, y una persona que no sabe nada de neurociencia pero conoce bien la sensación del rojo tal vez no resida en *qué* es lo que conoce cada uno respectivamente (estados cerebrales en el caso del primero, qualia no físicas en el otro caso), sino más bien en que cada uno tiene un diferente *tipo*, o *medio* o *nivel* de representación exactamente de la misma cosa: estados cerebrales.

En suma, no hay duda ninguna de que existen más modos de "tener conocimiento" que el simple hecho de dominar un conjunto de oraciones, y el materialista puede admitir sin reservas que alguien tenga un "conocimiento" de sus propias sensaciones que no depende para nada de la neurociencia que pueda haber aprendido. Los animales, entre ellos los humanos, presuntamente tienen una modalidad prelingüística de representación sensorial. Esto no significa que las sensaciones sean algo que escape a las posibilidades de la ciencia física. *Sólo significa que el cerebro utiliza otras modalidades y medios de representación que no se limitan solamente al almacenamiento de oraciones*. Todo lo que necesita afirmar el teórico de la identidad es que esas otras modalidades de representación también son susceptibles de recibir una explicación neurocientífica.

La teoría de la identidad ha demostrado siempre una gran flexibilidad para hacer frente a estas objeciones predominantemente antimaterialistas. Pero hay otras, provenientes de formas rivales de materialismo, que constituyen una amenaza mucho más seria, como veremos en las secciones siguientes.

## Lecturas complementarias

### Sobre la teoría de la identidad


Feigl, Herbert, "The Mind Body Problem: Not a Pseudo-Problem", en *Dimensions of Mind,* Sidney Hook (comp.), Nueva York, New York University Press, 1960.

Place, U. T., "Is Consciousness a Brain Process?", *British Journal of Psychology,* vol. XLVII, 1956. Reproducido en *The Philosophy of Mind,* V. C. Chappell (comp.), Englewood Cliffs, N. J. Prentice-Hall, 1962.

Smart, J. J. C., "Sensations and Brain Processes", *Philosophical Review,* vol. LXVIII, 1959. Reproducido en *The Philosophy of Mind,* V. C. Chappell (comp.), Englewood Cliffs, N. J. Prentice Hall, 1962.

Lewis, David, "An Argument for the Identity Theory", *The Journal of Philosophy,* vol. LXIII, Nº 1, 1966.

Nagel, Thomas, "What is it like to Be A Bat?", *Philosophical Review,* vol. LXXXIII, 1974. Reproducido en *Readings in Philosophy of Psychology,* vol. I, N. Block (comp.), Cambridge, M. A., Harvard University Press, 1980.

Jackson, Frank, "Epiphenomenal Qualia", *The Philosophical Quarterly,* vol. 32, Nº 127, abril de 1982.

Churchland, Paul, "Reduction, Qualia, and the Direct Instrospection of Brain States", *Journal of Philosophy,* vol. LXXXII, Nº1, 1985.

Jackson, Frank, "What Mary Didn't Know", *Journal of Philosophy,* vol. LXXXIII, Nº 5, 1986.

Churchland, Paul, "Some Reductive Strategies in Cognitive Neurobiology", *Mind,* vol. 95, Nº 379, 1986.


### Sobre la reducción interteórica


Nagel, Ernst, *The Structure of Science,* Nueva York, Harcourt, Brace y World, cap. 11, 1961.

Feyerabend, Paul, "Explanation, Reduction, and Empiricism", en *Minnesota Studies in the Philosophy of Science,* vol. III, H. Feigl y G. Maxwell (comps.), Minneapolis, University of Minnesota Press, 1962.

Churchland, Paul, *Scientific Realism and the Plasticity of Mind.* Cambridge, Cambridge University Press, 1979, cap. 3, sec. 11.

Hooker, Clifford, "Towards a General Theory of Reduction", *Dialogue,* vol. XX, Nos. 1-3, 1981.


# 4. Funcionalismo

Según el *funcionalismo*, el rasgo esencial o definitorio de todo tipo de estado mental es el conjunto de relaciones causales que mantiene con 1) los efectos ambientales sobre el cuerpo, 2) otros tipos de estados mentales, y 3) la conducta del

cuerpo. Lo característico del dolor, por ejemplo, es que es el resultado de alguna lesión o traumatismo corporal; provoca angustia, incomodidad y alguna forma de razonamiento práctico destinado a aliviarlo. Y también da lugar a que una persona se intranquilice, se proteja y prodigue cuidados a la zona afectada. Todo estado que cumpla exactamente esa función es un dolor, de acuerdo con el funcionalismo. En forma similar también se definen otros tipos de estados mentales (sensaciones, temores, creencias, etc.) por medio de las funciones causales específicas que cumplen en una economía compleja de estados internos que actúen como intermediarios entre la entrada de estímulos sensoriales y la salida en forma de conductas.

Es posible que esto le recuerde al lector el conductismo y en verdad esta concepción es su heredera, pero existe una diferencia fundamental entre ambas teorías. Mientras que el conductista trata de definir todo tipo de estado mental exclusivamente en términos de estímulo ambiental y respuesta en forma de conducta, el funcionalista niega totalmente esta posibilidad. A su modo de ver, la caracterización adecuada de casi todos los estados mentales supone una referencia ineludible a una variedad de otros estados mentales con los cuales tiene una conexión causal, de modo que una definición reduccionista exclusivamente en términos de estímulos y respuestas notoriamente observables por todos es absolutamente imposible. Por lo tanto, el funcionalismo es inmune a una de las objeciones principales en contra del conductismo.

De modo que existe una diferencia entre funcionalismo y conductismo. La diferencia entre el funcionalismo y la teoría de la identidad surge de la consideración del siguiente argumento planteado en contra de la última.

Imaginemos a un ser de otro planeta, dice el funcionalista, un ser con una constitución fisiológica distinta, que se basa en un elemento químico como el silicio, por ejemplo, en lugar de estar basada en el carbón como la nuestra. La estructura química, e inclusive la estructura física del cerebro de este ser extraño tendría que ser sistemáticamente diferente de la nuestra. Pero, aun así, el cerebro de ese ser bien podría

mantener una economía funcional de estados mentales cuyas *relaciones* mutuas se correspondieran perfectamente con las relaciones mutuas que definen los nuestros. El extraterrestre puede tener un estado interno que cumpla todas las condiciones para ser un estado de dolor, como hemos explicado antes. Ese estado, considerado desde un punto de vista puramente físico, tendría una estructura muy diferente de la del dolor humano, pero sin embargo podría ser idéntico a un estado de dolor humano desde un punto de vista puramente funcional. Y lo mismo podría decirse para todos sus estados funcionales.

Si la economía funcional de estados internos del extraterrestre fuera en realidad *funcionalmente isomórfica* con la nuestra —si esos estados tuviesen una conexión casual con la entrada de estímulos, entre sí y con la conducta, de alguna manera que se correspondiera con nuestras conexiones internas— entonces el extraterrestre tendría dolores, deseos, esperanzas y temores tan plenos como los nuestros, a pesar de las diferencias en el sistema físico que sostiene o realiza esos estados funcionales. Lo que cuenta en el terreno de lo mental no es la materia de la que está hecho un ser, sino la estructura de las actividades internas que sostiene esa materia.

Si nos imaginamos la constitución de un ser extraño, podemos pensar en la de muchos, y lo que acabamos de exponer también puede ser válido para un sistema artificial. Si creáramos un sistema electrónico —un ordenador de cualquier tipo— cuya economía interna fuese funcionalmente isomórfica con la nuestra en todos los sentidos pertinentes, entonces ese sistema podría ser el sujeto de estados mentales.

Lo que se ilustra con esto es que la naturaleza, y quizá también el hombre, tiene mucho más que un camino para armar un ser que piense, sienta y perciba. Y esto le plantea un problema a la teoría de la identidad, pues al parecer no existe un único tipo de estado físico al que le corresponda siempre un determinado tipo de estado mental. Paradójicamente, existen *demasiados* tipos diferentes de sistemas físicos que pueden realizar la economía funcional característica de la inteligencia consciente. Por lo tanto, si consideramos el universo en su conjunto, y el futuro al mismo tiempo que el presente, parece

sumamente improbable que el teórico de la identidad pueda encontrar las correspondencias biunívocas entre los conceptos de la taxonomía mental utilizados en forma corriente y los conceptos de una teoría exhaustiva que abarque todos los sistemas físicos pertinentes. Pero éstos son normalmente los requisitos de la reducción interteórica. Por lo tanto, son muy escasas las probabilidades de hallar identidades universales, entre tipos de estados mentales y tipos de estados cerebrales.

Si bien los funcionalistas no aceptan la teoría de la identidad tradicional en la cual el tipo mental es igual al tipo físico, prácticamente todos ellos suscriben una forma más débil de esta teoría según la cual un símbolo de lo mental es igual a un símbolo de lo físico, puesto que todavía sostienen que cada *instancia* de un tipo determinado de estado mental es numéricamente idéntica a algún estado físico específico en alguno de los sistemas físicos. Lo único que se impugna son las identidades universales (tipo/tipo). Aun así, es característico que este rechazo se considere como un respaldo a la afirmación de que la ciencia de la psicología es o debe ser *metodológicamente autónoma* de las diversas ciencias físicas como la física, la biología y aun la neurofisiología. La psicología, se sostiene, tiene sus propias leyes irreductibles y su propio objeto de estudio abstracto.

En el momento en que se escribe este libro, el funcionalismo constituye probablemente la teoría de la mente más ampliamente aceptada entre los filósofos, psicólogos cognitivos e investigadores en el campo de la inteligencia artificial. Algunas de las razones son evidentes a partir de lo que acabamos de exponer, pero también hay algunas otras. Al describir los estados mentales como estados esencialmente funcionales, esta teoría coloca el objeto de la psicología en un nivel más abstracto, separado de los múltiples detalles que presenta la estructura neurofisiológica (o cristalográfica o microelectrónica) del cerebro. La ciencia de la psicología, suele decirse, es metodológicamente autónoma de aquellas otras ciencias (biología, neurociencia, teoría de los circuitos) que se ocupan de lo que vienen a ser los detalles mecánicos. Esto constituye un fundamento teórico para una gran cantidad de

trabajos en psicología cognitiva y en inteligencia artificial, en los que los investigadores postulan un sistema de estados funcionales abstractos y luego lo someten a prueba comparándolo, generalmente por medio de la simulación con ordenador, con la conducta humana en circunstancias similares. El objetivo de estos trabajos consiste en descubrir en detalle la organización funcional que nos hacer ser lo que somos. (En parte con el propósito de evaluar las perspectivas de un enfoque funcionalista en la filosofía de la mente, examinaremos en el capítulo 6 algunas de las investigaciones recientes en el campo de la inteligencia artificial.)

## Argumentos en contra del funcionalismo

Aparte de su popularidad actual, el funcionalismo también tiene sus dificultades. La objeción que más comúnmente se le ha formulado se refiere a un viejo amigo: los qualia sensoriales. El funcionalismo es capaz de eludir una de las fallas irremediables del conductismo, se dice, pero igual cae presa de la otra. En su intento de considerar como rasgo definitorio de todo estado mental a sus propiedades *relacionales*, el funcionalismo ignora su naturaleza "interna" o cualitativa, que es el rasgo esencial de muchos tipos de estados mentales (el dolor, las sensaciones del color, de la temperatura, del tono, etc.) según expresa esta objeción, y por lo tanto, el funcionalismo es falso.

El ejemplo clásico de este defecto evidente se denomina el "experimento de la sensación del espectro invertido". Es enteramente imaginable, dice esta versión, que la gama de sensaciones de color que experimento cuando veo objetos comunes simplemente esté invertida en relación con las sensaciones de color que usted experimenta. Al mirar un tomate, tal vez yo tenga lo que es realmente la sensación del verde mientras usted tiene la sensación normal del rojo; al mirar una banana, yo puedo tener lo que es realmente la sensación del azul mientras que usted tiene la sensación normal del amarillo, etc. Pero puesto que no tenemos ningún modo de comparar nuestros qualia internos, y puesto que yo haré todas las mis-

mas discriminaciones observacionales entre objetos que usted quiera, no hay ningún modo de saber si mi espectro está invertido en relación con el suyo.

El problema que se le plantea al funcionalismo se puede formular del modo siguiente. Aun cuando mi espectro esté invertido en relación con el suyo, ambos seguimos siendo mutuamente isomórficos desde el punto de vista funcional. Mi sensación visual ante la vista de un tomate es *funcionalmente* idéntica a la suya. Por lo tanto, de acuerdo con el funcionalismo, constituyen el mismo tipo de estado y ni siquiera tiene sentido suponer que mi sensación es "realmente" la sensación del verde. Si cumple las condiciones funcionales para ser una sensación del rojo, entonces por definición es una sensación del rojo. En términos del funcionalismo, entonces, es evidente que una inversión del espectro del tipo que acabamos de describir queda totalmente excluida por definición. Pero este tipo de inversiones son enteramente imaginables, concluye este argumento, y si el funcionalismo presupone que no lo son, entonces es falso.

Otra cuestión relacionada con los qualia que le preocupa al funcionalismo es el denominado "problema de los qualia ausentes". La organización funcional característica de la inteligencia consciente puede ser ejemplificada (= realizada o ejemplificada) en una gran variedad de sistemas físicos, algunos de ellos radicalmente diferentes de un sistema humano normal. Entre otros, podría ejemplificarlo un ordenador electrónico gigantesco, y existen todavía posibilidades más radicales. Supongamos que un escritor nos pidiera que nos imagináramos al pueblo chino —la cantidad completa de $10^9$— organizado en un complejo juego de interacciones mutuas que les permitiera llegar a constituir en su conjunto un cerebro gigantesco que intercambiara entradas y salidas de estímulos con el cuerpo de un robot individual. Ese sistema del robot más la unidad cerebral de los $10^9$ presumiblemente ejemplificaría la organización funcional pertinente (aunque sin ninguna duda ejecutaría sus actividades en forma más lenta que un ser humano o un ordenador), y por lo tanto sería el sujeto de estados mentales, según el funcionalismo. Pero seguramente, se replica, los estados complejos que allí cum-

plen las funciones del dolor, el placer y las sensaciones de color no tendrían qualia intrínsecos como los nuestros y por lo tanto no podrían ser auténticos estados mentales. También aquí parece que el funcionalismo resulta, en el mejor de los casos, una versión incompleta de la naturaleza de los estados mentales.

Recientemente se ha sostenido que se puede responder a ambas objeciones (la de los qualia invertidos y la de los qualia ausentes), sin entrar en colisión con los principios del funcionalismo y sin forzar excesivamente las intuiciones del sentido común acerca de los qualia. Consideremos en primer lugar el problema de la inversión. Creo que el funcionalista tiene razón al reclamar que la identidad-tipo de las sensaciones visuales se infiera a partir de su papel funcional. Pero el impugnador también tiene razón al afirmar que es enteramente imaginable la inversión relativa de los qualia de dos personas, sin que se produzca inversión funcional. La aparente incongruencia entre estas posiciones se puede disipar si se afirma que: 1) los estados funcionales (o mejor dicho, sus realizaciones físicas) verdaderamente tienen una naturaleza intrínseca de la que depende nuestra identificación de esos estados por medio de la introspección, y también que 2) esa naturaleza intrínseca sin embargo no es algo esencial para la identidad-tipo de un determinado estado mental y de hecho puede *variar* entre un ejemplo y otro del mismo tipo de estado mental.

Esto significa que el carácter cualitativo de su sensación del rojo podría ser diferente del mío, en forma leve o sustancial, y también cabe la posibilidad de que lo sea la sensación del rojo de una tercera persona. Pero en la medida en que los tres estados son provocados normalmente por objetos rojos y normalmente son la causa de que nosotros tres creamos que algo es rojo, entonces los tres estados son sensaciones del rojo, cualquiera que sea su carácter cualitativo intrínseco. Estos qualia intrínsecos simplemente constituyen rasgos prominentes que permiten una rápida identificación introspectiva de las sensaciones, así como las rayas negras sobre el color naranja constituyen un rasgo dominante para la rápida identifi-

cación visual de los tigres. Pero no existen qualia específicos esenciales para determinar la identidad-tipo de los estados mentales, así como no hay rayas negras sobre color naranja que sean esenciales para constituir la identidad-tipo de los tigres.

Lisa y llanamente, esta solución requiere que el funcionalista admita la realidad de los qualia, y cabe preguntarse qué lugar les puede quedar a los qualia en su descripción materialista del mundo. Tal vez puedan entrar del modo siguiente: si se los *identifica* con propiedades físicas de cualquier tipo de estado físico que ejemplifique los estados mentales (funcionales) que los exhiben. Por ejemplo, identificar la naturaleza cualitativa de sus sensaciones del rojo con el rasgo físico (del estado cerebral que lo ejemplifica) al que efectivamente responden sus mecanismos de discriminación introspectiva cuando usted estima que tiene una sensación del color rojo. Si el materialismo está en lo cierto, entonces tiene que *haber* algún rasgo físico interno u otra cosa con la que armonice su discriminación de las sensaciones del rojo. Si el tono de un sonido resulta ser la frecuencia de una oscilación en la presión del aire, entonces no hay ninguna razón por la que el qualia de una sensación no pueda ser, digamos, una frecuencia ondulatoria en un determinado trayecto nervioso. (Más probablemente será un grupo especial o *conjunto* de frecuencias ondulatorias, como sostiene la teoría *vectorial* de la codificación sensorial, también llamada *modelo de fibras combinadas*. "Ondas" son los pulsos electroquímicos diminutos por medio de los cuales las células cerebrales se comunican entre sí a lo largo de las fibras delgadas que las conectan. Ampliaremos este tema en el capítulo 7.)

Esto presupone que puede haber seres con una constitución física diferente de la nuestra que tengan qualia diferentes de los nuestros, aunque sean psicológicamente isomórficos a nosotros. Pero no significa que necesariamente *deban* tener qualia diferentes. Si el carácter cualitativo de mi sensación del rojo es realmente una frecuencia ondulatoria de 90 hertz en un determinado trayecto nervioso, es posible que un robot electromecánico experimentara el mismo carácter cualitativo

si, al informar sobre sus sensaciones del rojo, estuviera respondiendo a una frecuencia ondulatoria de 90 hertz en su correspondiente trayecto *de cobre*. Parecería ser que lo que cuenta para nuestros respectivos mecanismos de discriminación sería la frecuencia ondulatoria y no la naturaleza del medio que la transporta.

Este planteo también indica una solución al problema de la ausencia de qualia. En la medida en que el sistema físico considerado sea funcionalmente isomórfico al nuestro, hasta el último detalle, entonces tendrá la misma capacidad para efectuar sutiles discriminaciones introspectivas entre sus sensaciones. Esas discriminaciones deberán tener como base algún sistema físico, es decir, algunos rasgos físicos característicos de los estados entre los que se discrimina. *Esos* rasgos, que constituyen el núcleo objetivo de los mecanismos discriminatorios del sistema, son sus qualia sensoriales... aunque no es más probable que el sistema del extraterrestre pueda apreciar su propia naturaleza física que que nosotros apreciemos la verdadera naturaleza física de nuestros propios qualia. Por lo tanto, los qualia sensoriales son un elemento concomitante inevitable de todo sistema que tenga el tipo de organización funcional que estamos considerando. Tal vez resulte difícil o imposible "ver" los qualia en un sistema extraterrestre, pero es igualmente difícil "ver"los aun si miramos dentro de un cerebro humano.

Dejo a criterio del lector la evaluación sobre la idoneidad de estas respuestas. De modo que si fueran adecuadas, y en vista de sus otras virtudes, al funcionalismo habría que reconocerle una posición muy sólida entre las teorías contemporáneas de la mente que disputan entre sí. Sin embargo, resulta interesante advertir que para la defensa propuesta en el último párrafo resultó necesario sacar una hoja del libro de la identidad del teórico (tipos de quale son reducidos a o identificados con tipos de estado físico), puesto que la última objeción que hemos de considerar también tiende a borrar la distinción entre funcionalismo y materialismo reduccionista.

Considérese la propiedad de la *temperatura*, señala la objeción. Aquí tenemos el paradigma de una propiedad física,

al cual también se lo ha mencionado como el paradigma de una propiedad *reducidà* con éxito, como está expresado en la identidad interteórica

"temperatura = promedio de energía cinética de las moléculas constituyentes"

En términos estrictos, sin embargo, esta identidad sólo vale para la temperatura de un gas, en la que las partículas simples se mueven libremente en forma balística. En un *sólido*, la temperatura se produce en forma diferente, puesto que las moléculas interconectadas se limitan a una variedad de movimientos vibratorios. En un *plasma*, la temperatura es algo diferente también, puesto que no está constituido por moléculas sino que éstas, con los átomos que las constituyen, están hechas pedazos. Hasta el *vacío* tiene lo que se denomina temperatura "antirradiantes" en la distribución de las ondas electromagnéticas que lo atraviesan. En este caso la temperatura no tiene nada que ver con la energía cinética de las partículas.

Es comprensible que la propiedad física de la temperatura encuentre "múltiples ejemplificaciones", al igual que ocurre en el caso de las propiedades psicológicas. ¿Esto significa que la termodinámica (la teoría del calor y la temperatura) es una "ciencia autónoma", separable del resto de la física, y que tiene sus propias leyes irreductibles y su propio objeto abstracto no físico?

Presumiblemente no. Lo que significa, concluye la objeción, es que *las reducciones tienen un ámbito específico*:

> la temperatura de un gas = promedio de energía cinética de las moléculas de gas,

mientras que

> la temperatura del vacío = la distribución antirradiante de la radiación momentánea del vacío.

Del mismo modo, tal vez

> alegría en un ser humano = resonancias en el hipotálamo lateral,

mientras que

> alegría en un marciano = algo completamente diferente.

Esto significa que después de todo es dable esperar algunas reducciones tipo/tipo de estados mentales a estados físicos, aunque serán mucho más restringidas de lo que se indicó primero. Más aún, esto significa que no se pueden sostener las pretensiones del funcionalismo acerca de la autonomía radical de la psicología. Y, por último, que el funcionalismo no es tan profundamente diferente de la teoría de la identidad como se creyó en un primer momento.

Al igual que en el caso de la defensa del funcionalismo que se esbozó antes, dejo a criterio del lector la evaluación de estas críticas. En capítulos posteriores tendremos ocasión de efectuar un análisis más profundo del funcionalismo. En este momento pasemos a examinar la última teoría materialista de la mente, ya que el funcionalismo no es la única reacción importante en contra de la teoría de la identidad.

## Lecturas complementarias


Putnam, Hilary, "Minds and Machines", en *Dimensions of Mind*, Sidney Hook (comp.), Nueva York, New York University Press, 1960.

Putnam, Hilary, "Robots: Machines or Artificially Created Life?", *Journal of Philosophy*, vol. LXI, Nº 21, 1964.

Putnam, Hilary, "The Nature of Mental States", en *Materialism and the Mind-Body Problem*, David Rosenthal (comp.), Englewood Cliffs, N. J., Prentice-Hall, 1971.

Fodor, Jerry, *Psychological Explanation*. Nueva York, Random House, 1968.

Dennett, Daniel, *Brainstorms*. Montgomery, Vermont, Bradford, 1978; Cambridge MA, MIT Press.

Block, Ned, "Troubles with Functionalism", en *Minnesota Studies in the Philosophy of Science*, vol. IX, C. W. Savage (comp.), Minneapolis, University of Minnesota Press, 1978. Reproducido en *Readings in Philosophy of Psychology*, N. Block (comp.), Cambridge, MA, Harvard University Press, 1980.

Churchland, Paul y Patricia, "Functionalism, Qualia, and Intentionality", *Philosophical Topics*, vol. 12, Nº 1, 1981. Reproducido en *Mind, Brain and Function*, J. Biro y Churchland Paul (comps.), Norman, OK: University of Oklahoma Press, 1982. Churchland, Paul, "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy*, vol. LXXVIII, Nº 2, 1981.

Shoemaker, Sidney, "The Inverted Spectrum", *Journal of Philosophy*, vol. LXXIX, Nº 7, 1982.

Enc, Berent, "In Defense of the Identity Theory", *Journal of Philosophy*, vol. LXXX, Nº 5, 1983.

## 5. Materialismo eliminativo

La teoría de la identidad fue puesta en duda no porque se pensara que eran muy pocas las perspectivas de lograr una explicación materialista de nuestras aptitudes mentales, sino porque parecía improbable que la aparición de una teoría materialista adecuada trajera consigo las correspondencias biunívocas exactas, entre los conceptos de la psicología corriente y los conceptos de la neurociencia teórica, que requiere la reducción interteórica. La razón para esa duda fue la gran variedad de sistemas físicos totalmente diferentes que podían ejemplificar la organización funcional requerida. *El materialismo eliminativo* también pone en duda que la explicación neurocientífica adecuada de las aptitudes humanas logre producir una clara reducción del marco de referencia corriente, pero aquí las dudas tienen un origen totalmente diferente.

A juicio del materialismo eliminativo, no podrán encontrarse las correspondencias biunívocas, y no se podrá efectuar una reducción interteórica del marco de referencia psicológico corriente, *porque el marco de referencia psicológico que utilizamos corrientemente es una concepción falsa y radicalmente engañosa sobre las causas de la conducta humana y la naturaleza de la actividad cognitiva.* Desde esta perspectiva, la psi-

cología habitual no solamente constituye una representación incompleta de nuestra naturaleza interna, sino que directamente constituye una *mala* representación de nuestros estados y actividades internos. En consecuencia, no es posible esperar que una explicación neurocientífica verdaderamente adecuada de nuestra vida interior proporcione las categorías teóricas que se corresponden escrupulosamente con las categorías de nuestro marco de referencia habitual. Consecuentemente, lo único que se debe esperar es que el antiguo marco simplemente sea eliminado y no que pueda reducirse por una neurociencia más desarrollada.

## Paralelos históricos

Del mismo modo que el teórico de la identidad puede señalar ejemplos históricos en los que se ha podido efectuar la reducción interteórica, también el materialista eliminativo puede alegar ejemplos históricos en los que se produjo la eliminación lisa y llana de la ontología de una teoría anterior y se reemplazó por la ontología de una teoría nueva y superior. Durante la mayor parte de los siglos XVIII y XIX, la gente culta creía que el calor era un *fluido* sutil contenido en los cuerpos, más o menos del mismo modo en que el agua está contenida en una esponja. Un cuerpo considerable de teoría moderadamente satisfactoria describía el modo en que esta sustancia —denominada "calórica"— fluía en el interior del cuerpo, o de un cuerpo a otro, y cómo producía ampliación térmica, fusión, hervor, etc. Pero hacia fines del siglo pasado ya se había puesto suficientemente en claro que el calor no era ningún tipo de sustancia, sino simplemente la energía producida por el movimiento de billones de partículas entremezcladas que constituían el cuerpo caliente en sí. La nueva teoría —la teoría corpuscular/cinética de la materia y el calor— resultó mucho más satisfactoria que la anterior para explicar y predecir la conducta térmica de los cuerpos. Y puesto que no fue posible *identificar* el fluido calórico con la energía cinética (según la antigua teoría la calórica es una *sustancia* material;

según la nueva, la energía cinética es una forma de *movimiento*), finalmente se aceptó que *no existe nada* que sea una sustancia calórica. La sustancia calórica lisa y llanamente quedó eliminada de la ontología aceptada.

Segundo ejemplo. Solía pensarse que cuando un trozo de madera se quema o cuando un trozo de metal se oxida, se liberaba una sustancia incorpórea denominada "flogisto"; muy rápidamente en el primer caso, y con mucha lentitud en el último. Una vez evacuada, esa sustancia dejaba sólo una pila común de ceniza o herrumbre. Más adelante se llegó a comprender que en ambos procesos se producía, no la pérdida de algo, sino el *agregado* de una sustancia tomada de la atmósfera: el oxígeno. El flogisto apareció, no como una descripción incompleta de lo que sucedía sino radicalmente como una descripción equivocada. Por lo tanto, el flogisto no resultaba adecuado para efectuar la reducción o la identificación con algún otro concepto de la nueva química del oxígeno y sin más quedó eliminado del campo de la ciencia.

Es cierto que los dos ejemplos mencionados se refieren a la eliminación de algo no observable, pero nuestra historia también incluye la eliminación de ciertos elementos "observables" ampliamente aceptados. Antes de que se difundieran las ideas de Copérnico, prácticamente cualquier persona que se arriesgara a salir por la noche podía contemplar la *esfera estrellada del cielo* y, si permanecía más de unos pocos minutos, también veía que *giraba* alrededor de un eje a través de Polaris. ¿De qué estaba hecha la esfera? (¿de cristal?) y ¿quién la hacía girar? (¿los dioses?) fueron las preguntas teóricas que nos inquietaron durante más de dos milenios. Pero prácticamente nadie dudaba de la existencia de lo que todo el mundo podía observar con sus propios ojos. Sin embargo, al final aprendimos a reinterpretar nuestra experiencia visual del cielo nocturno utilizando un marco de referencia conceptual muy diferente, y la esfera giratoria se desvaneció.

Las brujas proporcionan otro ejemplo. La psicosis es un padecimiento relativamente común entre los seres humanos, y en épocas anteriores era normal que se considerara que las personas que la padecían estaban poseídas por el demonio y eran

una encarnación del propio espíritu de Satanás, cuya mirada fulgurante proveniente de los ojos de las víctimas se clavaba en nosotros malignamente. La existencia de las brujas no se ponía en duda de ninguna manera. Ocasionalmente alguien las veía, en cualquier ciudad o aldea, poner en práctica alguna conducta incoherente, paranoica y a veces hasta sanguinaria. Pero, observables o no, con el tiempo hemos decidido que las brujas simplemente no existen. Hemos llegado a la conclusión de que el concepto de bruja es un elemento perteneciente a un marco de referencia conceptual que representa en forma tan distorsionada los fenómenos a los que se lo aplicaba corrientemente que la aplicación literal del concepto debe desterrarse para siempre. Las modernas teorías de la disfunción mental llevaron a la eliminación de las brujas de cualquier ontología seria.

Según la concepción que estamos considerando, a los conceptos de la psicología popular —creencia, deseo, temor, sensación, dolor, alegría, etc.— les espera un destino parecido. Y cuando la neurociencia haya alcanzado un nivel tal de desarrollo en el cual la pobreza de nuestras concepciones actuales resulte evidente para todo el mundo, y se establezca la superioridad del nuevo marco de referencia, entonces seremos capaces finalmente de emprender la tarea de *volver* a pensar nuestros estados y actividades internos dentro de un marco conceptual verdaderamente adecuado. Las explicaciones que nos demos recíprocamente respecto de nuestras conductas tendrán que recurrir a elementos tales como los estados neurofarmacológicos, la actividad nerviosa en zonas anatómicas especializadas y cualquier otro tipo de estados que la nueva teoría juzgue pertinentes. También se transformará la introspección personal y tal vez llegue a adquirir un mayor nivel de profundidad en virtud del marco más preciso en el cual tendrá que trabajar... del mismo modo en que la percepción del astrónomo del cielo nocturno se ve muy favorecida por el conocimiento detallado que posee de la moderna teoría astronómica.

No se debe minimizar la magnitud de la revolución conceptual que aquí se señala: podría ser monumental. Y los

beneficios para la humanidad podrían ser igualmente grandes. Si cada uno de nosotros poseyera un conocimiento neurocientífico (cosa que ahora percibimos nebulosamente) de las variedades y causas de las enfermedades mentales, de los factores que intervienen en el aprendizaje, las bases neurológicas de las emociones, la inteligencia y la socialización, entonces la totalidad de la desdicha humana podría disminuir mucho. El simple aumento de la comprensión mutua que hiciera posible el nuevo marco podría contribuir sustancialmente a lograr una sociedad más pacífica y humanitaria. Por supuesto, también habría ciertos riesgos: mayor conocimiento significa mayor poder, y del poder también puede hacerse un mal uso.

## Argumentos en favor del materialismo eliminativo

Los argumentos en favor del materialismo eliminativo son difusos y no llegan a ser decisivos, pero son más sólidos de lo que suele creerse. El rasgo que caracteriza a esta posición es que niega que pueda efectuarse fácilmente una reducción interteórica —inclusive una reducción específica de especie— del marco conceptual de la psicología popular al de la neurociencia plenamente desarrollada. La razón de esta negativa reside en la convicción que sustenta el materialismo eliminativo acerca de que la psicología tradicional es una concepción irremediablemente primitiva y profundamente confusa de las actividades internas. Pero ¿por qué esta mala opinión acerca de las concepciones que sustentamos normalmente?

Existen por lo menos tres razones. En primer lugar, el materialismo eliminativo apunta a los difundidos fracasos de la psicología popular para explicar, predecir y manipular. Si nos basamos en su marco conceptual, la mayor parte de los elementos importantes y familiares para nosotros continúan siendo un completo misterio. No sabemos qué es el *sueño*, o por qué lo necesitamos, a pesar de que pasamos un buen tercio de nuestra vida en esa situación. (La respuesta "para descan-

sar" es incorrecta. Aun si a la gente se le permitiera descansar permanentemente, su necesidad de sueño no disminuiría. Es evidente que el sueño cumple ciertas funciones más profundas, pero todavía no sabemos cuáles son.) No sabemos de qué modo el *aprendizaje* nos transforma de bebés papanatas en adultos sagaces o en qué se basan las diferencias de *inteligencia*. No tenemos ni la menor idea de cómo funciona la *memoria* ni de cómo nos ingeniamos para recuperar instantáneamente las unidades de información que necesitamos a partir de la masa impresionante de datos que hemos alcanzado. No sabemos qué es la *enfermedad mental* ni cómo curarla.

En suma, los temas que nos tocan más de cerca continúan siendo un misterio casi total desde el punto de vista de la psicología corriente. Y los defectos señalados no pueden atribuirse a que no se le ha dado el tiempo suficiente para corregirlos, puesto que en la psicología popular no se han producido cambios ni progresos significativos prácticamente en 2000 años, a pesar de sus fracasos manifiestos. Es dable esperar que se corrijan teorías que han logrado importantes éxitos, pero no merecen esa expectativa teorías que han fracasado rotundamente.

Este argumento de la insuficiencia explicativa se puede ampliar un poco más. En la medida en que se trate de cerebros normales, la insuficiencia de la psicología tradicional tal vez no resulte extraordinariamente evidente. Pero apenas se examina la cantidad de desconcertantes deficiencias conductuales y cognitivas que padecen las personas con *daño* cerebral, los recursos descriptivos y explicativos de los que disponemos comienzan a perder pie (véase, por ejemplo, el capítulo 7.3., pág. 206). Del mismo modo que ocurrió en el caso de otras modestas teorías a las que se les pidió que funcionaran con éxito en extensiones inexploradas de sus antiguos dominios (por ejemplo, la mecánica newtoniana en el terreno de las velocidades cercanas a la velocidad de la luz, y la ley de los gases clásica en el terreno de las altas presiones o temperaturas), las falencias descriptivas y explicativas de la psicología popular se hacen rigurosamente evidentes.

El segundo argumento intenta extraer una lección in-

ductiva de nuestra historia conceptual. Las primeras teorías tradicionales del movimiento experimentaron una profunda confusión ante teorías más sofisticadas, y con el tiempo fueron totalmente desplazadas por ellas. Las primeras teorías tradicionales sobre la estructura y actividad de la bóveda celeste eran desatinadas y completamente erradas, y sólo sobreviven como lecciones históricas de lo mucho que podemos equivocarnos. Las teorías tradicionales sobre la naturaleza del fuego y la naturaleza de la vida eran igualmente absurdas. Y se podría seguir, ya que la inmensa mayoría de las concepciones tradicionales del pasado han sido refutadas del mismo modo. Todas excepto la psicología popular, que sobrevive hasta hoy y que sólo recientemente ha comenzado a sentir las presiones. Pero sin ninguna duda el fenómeno de la inteligencia consciente es mucho más complejo y difícil que cualquiera de los otros que hemos enumerado. Si se trata de comprender con exactitud, sería un *milagro* que hubiéramos acertado con *ése* en particular la primera vez que lo consideramos, cuando en todos los otros se ha fracasado tan lastimosamente. La psicología corriente ha sobrevivido durante tanto tiempo, presumiblemente, no porque sus representaciones sean básicamente correctas, sino porque los fenómenos que aborda son tan terriblemente difíciles que cualquier modo útil de manejarlos, por débil que sea, probablemente no será desplazado con mucha rapidez.

El tercer argumento trata de encontrarle una ventaja a priori al materialismo eliminativo frente a la teoría de la identidad y al funcionalismo. Intenta oponerse a la intuición común de que el materialismo eliminativo es posible si se lo mira con aire distraído, tal vez, pero es mucho menos probable que la teoría de la identidad o que el funcionalismo. Una vez más el núcleo de la cuestión reside en saber si los conceptos de la psicología tradicional han de encontrar correspondencias que los justifiquen en una neurociencia plenamente desarrollada. El materialismo eliminativo apuesta a que no; los otros dos, a que sí. (Incluso el funcionalista apuesta por la afirmativa, pero espera que las correspondencias sólo sean específicas de la especie, o sólo específicas de la persona. Recuérdese

que el funcionalismo niega solamente la existencia de identidades *universales* tipo/tipo.)

El materialista eliminativo hará notar que los requisitos para una reducción son bastante exigentes. La nueva teoría debe contener un conjunto de principios y conceptos incluidos que refleje fielmente la estructura conceptual específica que se va a reducir. Y lo cierto es que existen infinitamente más medios de constituir una neurociencia con potencia explicativa que *no* refleje la estructura de la psicología corriente, que de hacerlo al mismo tiempo que se *refleja* la propia estructura específica de la psicología popular. Consecuentemente, la probabilidad a priori del materialismo eliminativo no es menor, sino sustancialmente *mayor* que la de cualquiera de sus dos rivales. Las intuiciones iniciales que tenemos aquí simplemente están equivocadas.

Hay que admitir que esta ventaja inicial a priori podría reducirse si existiese una fuerte presunción de que la psicología tradicional pudiese estar en lo cierto: las teorías verdaderas están en mejor posición para ganar cuando se efectúa la reducción. Pero, de acuerdo con los dos primeros argumentos, las presunciones sobre este punto precisamente se moverán en la dirección contraria.

## Argumentos en contra del materialismo eliminativo

A primera vista a casi todo el mundo le resulta bastante poco plausible esta concepción bastante radical, ya que niega supuestos profundamente arraigados. En el mejor de los casos esta objeción es una petición de principio, sin duda, ya que precisamente de esos supuestos se trata. Pero por medio del siguiente razonamiento se intenta construir un argumento real.

El materialismo eliminativo es falso, señala este punto de vista, porque la introspección revela directamente la existencia de dolores, creencias, deseos, temores, etc. Esa existencia es lo más obvio que puede haber.

El materialista eliminativo replicará diciendo que este argumento comete el mismo error que cometería una persona que viviera en la época antigua o medieval si insistiera en que podía ver con sus propios ojos que el cielo constituye una esfera giratoria, o que las brujas existen. El hecho es que toda observación se produce dentro de algún sistema de conceptos, y los juicios de observación sólo son tan válidos como lo es el marco de referencia conceptual dentro del que se expresan. En los tres casos mencionados —la esfera estrellada, las brujas y los estados mentales que conocemos— precisamente lo que se cuestiona es la validez de los marcos conceptuales básicos dentro de los que están expresados esos juicios de observación. Insistir en la validez de nuestras propias experiencias, *interpretadas de modo tradicional* es, por lo tanto, una petición de principio respecto del propio problema en consideración. Puesto que en los tres casos, el problema está en saber si podemos *volver* a pensar cuál es la naturaleza de un ámbito observacional conocido.

Una segunda crítica intenta encontrar alguna incoherencia en la posición del materialismo eliminativo. Lo que esta teoría declara sin reservas es que los estados mentales que conocemos no existen. Pero ese enunciado tiene sentido, sostiene el argumento crítico, sólo si es la expresión de alguna *creencia*, y de una *intención* de comunicar, y de un *conocimiento* del lenguaje y así siguiendo. Pero si el enunciado es verdadero, entonces no existen tales estados mentales y, por lo tanto, el enunciado es una retahíla carente de sentido de marcas o ruidos, y no puede ser verdadero. Evidentemente, el supuesto de que el materialismo eliminativo está en lo cierto presupone que no puede estarlo.

El agujero de este argumento es la premisa sobre las condiciones necesarias para que un enunciado tenga sentido. Es una petición de principio. Si el materialismo eliminativo está en lo cierto, entonces el sinsentido debe tener algún otro origen. Insistir en el "antiguo" origen equivale a insistir sobre la validez del propio marco de referencia que se está considerando. Una vez más, un paralelo histórico puede resultar de utilidad en este caso. Considérese la teoría medieval que pos-

tula que estar biológicamente *vivo* significa estar animado por un *espíritu vital* inmaterial. Y considérese la siguiente respuesta que se le da a alguien que ha expresado que no cree en esa teoría.

> Mi docto amigo ha afirmado que no existe nada que sea un espíritu vital. Pero esta afirmación es incoherente. Porque si fuera verdad, entonces mi amigo no tiene espíritu vital y, por lo tanto, debe estar *muerto*. Pero si estuviera muerto, entonces su afirmación sólo es una retahíla de ruidos, desprovista de sentido o de verdad. Evidentemente, ¡el supuesto de que el antivitalismo está en lo cierto presupone que no puede estarlo! Q.E.D.

Este segundo argumentó es una broma, pero el primero es una petición de principio exactamente del mismo tipo.

La última crítica saca una conclusión mucho más débil, pero construye un argumento más fuerte. El materialismo eliminativo, se ha dicho, hace una montaña de un grano de arena. Exagera los defectos de la psicología tradicional y menoscaba sus éxitos reales. Tal vez la llegada de una neurociencia desarrollada requerirá de vez en cuando la eliminación de un concepto de la psicología tradicional, continúa la crítica, y tal vez haya que sobrellevar algún ajuste de poca monta en algunos de sus principios. Pero la eliminación en gran escala que pronostica el materialista eliminativo es simplemente una preocupación alarmista o un entusiasmo romántico.

Tal vez esta objeción sea correcta. Y tal vez sea meramente complaciente. Sea lo que fuere, efectivamente pone de manifiesto algo muy importante que es que aquí no se trata de confrontar dos posibilidades simples y mutuamente excluyentes: reducción pura o eliminación pura. Más bien, éstos son los puntos extremos de un espectro ininterrumpido de resultados posibles, entre los cuales existen casos mixtos de eliminación parcial y reducción parcial. Sólo la investigación empírica (véase capítulo 7) podrá decirnos en qué lugar de ese espectro se encuentra el caso que planteamos. Tal vez debe-

ríamos hablar aquí, en forma más aproximada, de "materialismo revisionista", en lugar de centrarnos en la posibilidad mas radical de una eliminación de punta a punta. Tal vez deberíamos hacerlo. Pero en esta sección mi propósito ha sido que al lector le resultara por lo menos inteligible la idea de que nuestro destino conceptual se encamina sustancialmente hacia el extremo revolucionario del espectro.

## Lecturas complementarias


Feyerabend, Paul, "Comment: 'Mental Events And The Brain'.", *Journal of Philosophy,* vol. LX, 1963. Reproducido en *The Mind-Brain Identity Theory,* C. V. Borst (comp.), Londres. Macmillan, 1970.

Feyerabend, Paul, "Materialism and The Mind-Body Problem", *Review of Metaphysics,* vol. XVII, 1963. Reproducido en *The Mind-Brain Identity Theory,* C. V. Borst (comp.), Londres, Macmillan, 1970.

Rorty, Richard, "Mind-Body Identity, Privacy, and Categories", *Review of Metaphysics,* vol. XIX, 1965. Reproducido en *Materialism and the Mind-Body Problem,* D. M. Rosenthal (comp.), Englewood CLiffs, NJ Prentice-Hall, 1971.

Churchland, Paul, "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy,* vol. LXXVIII, Nº 2, 1981.

Dennett, Daniel, "Why You Can't Make a Computer that Feels Pain", en *Brainstorms,* Montgomery, VT, Bradford, 1978, Cambridge, MA, MIT Press.

Churchland, Paul, "Some Reductive Strategies in Cognitive Neurobiology", *Mind,* vol. 95, Nº 379, 1986.

# 3

# El problema semántico

¿De dónde extraen su significado los términos del vocabulario psicológico que utilizamos corrientemente? Esta pregunta, aparentemente inocente, tiene su importancia al menos por tres razones. Los términos psicológicos suelen constituir una prueba decisiva para las teorías del sentido en general. El problema semántico está estrechamente vinculado con el problema ontológico, como vimos en el primer capítulo. Y está aún más íntimamente vinculado con el problema epistemológico, como lo veremos en el próximo capítulo.

En el presente capítulo examinaremos los argumentos a favor y en contra de cada una de las tres teorías importantes que se toman en cuenta en la actualidad. La primera de ellas sostiene que el significado de todo término psicológico corriente (de la mayor parte de ellos, en todo caso) deriva de un acto de *ostensión interna*. La segunda insiste en que ese significado deriva de *definiciones operacionales* y la tercera alega que el significado de esos términos deriva del lugar que ocupan dentro de un *sistema de leyes* que constituye la psicología "popular". Sin más rodeos, pues, pasemos a la primera teoría.

## 1. La definición por ostensión interna

Una manera habitual de incorporar un término al vocabulario de alguien —por ejemplo "caballo" o "autobomba"— consiste simplemente en mostrarle un caso concreto del tipo

correspondiente y decirle algo así como: "*Eso* es un caballo" o "*Esta* es una autobomba". Estos son ejemplos de lo que se denomina *definición ostensiva*. Uno espera que el oyente advierta los rasgos distintivos de la situación presentada y que sea capaz de volver a emplear el término toda vez que una nueva situación contenga tales rasgos.

Por supuesto, los dos vocablos mencionados podrían haber sido presentados de una manera diferente. Se le podría haber dicho simplemente al oyente: "Un caballo es un animal grande, con pezuñas, que se utiliza para montar". En este caso se comunica el significado del término conectándolo de modos específicos con otros términos del vocabulario del oyente. Esta forma de introducir nuevos términos cubre un espectro que va desde una presentación explícita y completa ("Un triángulo isósceles es una figura plana cerrada de tres lados que tiene por lo menos dos lados iguales") hasta una forma parcial y ocasional ("La energía es aquello que pone en movimiento los automóviles y mantiene las luces encendidas"). Sin embargo no todos los términos —se afirma con frecuencia— obtienen su significado de este modo. Algunos sólo pueden hacerlo de acuerdo con la primera forma, es decir, por ostensión directa. Sus significados no dependen de las relaciones que establecen con otros términos sino del hecho de estar directamente asociados con una cualidad específica que presentan los objetos materiales. En estos términos se expresa la teoría semántica ortodoxa al igual que el sentido común.

¿Qué sucede con los términos del lenguaje psicológico tradicional? Cuando pensamos en vocablos tales como "dolor", "comezón" y "sensación del rojo", la fuente obvia del significado parece ser la ostensión. ¿Cómo podría uno conocer el significado de algunos de esos términos si efectivamente no hubiese sentido alguna vez un dolor o una comezón o no hubiese experimentado una sensación del rojo? A primera vista pareciera que eso es imposible. Denominemos a éste "el criterio normal".

Si bien este criterio puede ser correcto en el caso de una clase significativa de términos psicológicos, no lo es evidentemente en el caso de todos estos términos, ni siquiera en la

mayoría. Muchos tipos importantes de estados mentales no tienen absolutamente ningún carácter cualitativo, o ninguno que sirva para establecer su identidad como tipo. Consideremos la variedad de las diferentes creencias, por ejemplo: creer que P, creer que Q, creer que R, y así sucesivamente. Nos encontramos aquí con una infinidad potencial de estados marcadamente diferentes. Posiblemente no se pueda dominar el significado de cada expresión estudiando, uno por uno, el carácter cualitativo peculiar de cada estado. Tampoco cada estado posee un quale distinto. Y lo mismo sucede con la infinidad potencial de los distintos pensamientos de que P, y los deseos de que P y los temores de que P, y con todas las otras "actitudes proposicionales". Estas son, quizá, las expresiones más centrales de nuestro marco de referencia tradicional, y se diferencian por medio de un elemento que cumple una función, la oración P, y no por medio de algún quale introspectible (= "cualidad fenomenológica"). Sus significados deben derivar de alguna otra fuente.

Evidentemente el criterio normal no puede agotar la cuestión planteada en torno al significado de los predicados psicológicos. Además, este criterio es sospechoso inclusive en los casos más creíbles. Entre aquellos estados mentales que están asociados con qualia, no todos los tipos poseen un quale *uniforme*. En realidad, muy pocos lo poseen, si no ninguno. Consideremos el término "dolor" y reflexionemos sobre la extensa variedad de sensaciones sustancialmente diferentes reunidas en este término (pensemos en un dolor de cabeza, en una quemadura, en un sonido que perfora los tímpanos, un golpe en la rótula, etc.). Es cierto que todos estos qualia son similares porque causan una reacción de desagrado en quien los padece, pero ésta es una propiedad *causal/relacional* común a todos los dolores y no un quale compartido. Incluso en las sensaciones del rojo se presenta una amplia variedad a través de numerosos tintes y matices, que se acercan al pardo, al anaranjado, al rosa, al púrpura o al negro en algunos extremos. Por supuesto que las similitudes intrínsecas actúan de algún modo para unificar esta difusa clase, pero parece evidente que la clase de las sensaciones del rojo está igual-

mente delimitada por el hecho de que esas sensaciones son típicamente el resultado de considerar ejemplos corrientes, como labios, fresas, manzanas y autobombas. Es decir, están unidas por los rasgos causales/relacionales que comparten. La idea de que el significado puede agotarse con un quale singular y unívoco parece ser un mito.

¿Estamos seguros de que conocer el quale es siquiera necesario para saber el significado? Se ha argumentado que alguien que no haya experimentado nunca el dolor (tal vez a causa de algún defecto en su sistema nervioso) podría, no obstante, conocer el significado de la palabra "dolor" y utilizarla en la conversación, la explicación y la predicción, así como la utilizamos para describir a los demás. Por supuesto, no podría saber cómo *se siente* el dolor, pero si podría conocer todas sus propiedades causales/relacionales y, en consecuencia, sabría, tan bien como cualquiera de nosotros, de qué clase de estado de dolor se trata. Quedaría aquí *algo* que esa persona no conocería, pero no está claro que ese algo sea el significado de la palabra "dolor".

Si el significado de términos tales como "dolor" y "sensación del rojo" verdaderamente pudiese quedar agotado por su asociación con un quale interno, entonces nos hallaríamos privados de recursos para evitar el *solipsismo semántico*. (El solipsismo es la tesis que sostiene que todo conocimiento es imposible, excepto el conocimiento de uno mismo.) Desde el momento en que cada uno de nosotros sólo pudiese experimentar sus *propios* estados de conciencia, a nadie le resultaría posible decir si el significado individual que le asigna al término "dolor" es o no el mismo que le atribuye otro. Y por cierto sería una teoría muy extraña del sentido la que presupone que nadie comprende nunca lo que otro quiere decir.

Estas dudas que suscita la teoría corriente del significado por "ostensión interna" han incitado a los filósofos a explorar otros enfoques. La primera tentativa seria de formular y defender una teoría alternativa fue la que hicieron los filósofos conductistas, que mencionamos en el capítulo anterior. Estos pensadores propusieron un argumento más en contra del criterio normal, y es el que vamos a examinar a continuación.

## 2. El conductismo filosófico

Según los conductistas, el significado de un término mental queda establecido por las múltiples relaciones que mantiene con algunos otros términos: vocablos que se refieren a circunstancias y conductas por todos observables. En sus formulaciones más claras, el conductismo señalaba términos puramente disposicionales como "soluble" y "débil" como análogos semánticos para términos mentales, y consideraba que las definiciones operacionales eran las estructuras por medio de las cuales podían explicitarse los significados de esos términos. Los pormenores de esta concepción los hemos esbozado en el capítulo 2.2, de modo que no los repetiremos aquí.

Uno de los problemas principales para el conductismo fue el papel insignificante que le atribuyó a los qualia de los estados mentales. Pero aquí acabamos de exponer algunas buenas razones para volver a estimar (hacia abajo) la importancia normalmente otorgada a los qualia. Y uno de los filósofos que más ha influido en la tradición conductista, Ludwig Wittgenstein, suministró un argumento adicional en contra del criterio normal: el *argumento del lenguaje privado*.

A pesar de la consecuencia del solipsismo, muchos defensores del criterio normal estaban dispuestos a sostener la idea de que el vocabulario de las sensaciones era un lenguaje irremediablemente *privado*. Wittgenstein intentó demostrar que un lenguaje necesariamente privado era algo absolutamente imposible. El argumento es el siguiente. Supongamos que se intenta otorgarle significado a un término "W" solamente asociándolo con una sensación que se experimenta en determinado momento. En un momento posterior, al experimentar una sensación, uno puede decir "Se trata de otro W". Pero ¿cómo se puede determinar que uno ha usado el término correctamente en esta ocasión? Tal vez uno no recuerde bien la primera sensación, o despreocupadamente ve una estrecha semejanza entre la segunda y la primera aunque en realidad sólo existe un parecido débil y distante. Si el término "W" no tiene ningún tipo de conexiones de sentido con *otros* fenómenos, como por ejemplo con ciertas causas y/o efectos corrientes

del tipo de sensación de que se trata, entonces no habrá absolutamente ningún modo de distinguir entre el uso correcto de "W" y su uso incorrecto. Pero un término cuya aplicación correcta escapa permanentemente a la posibilidad de determinación es un término que no tiene sentido. Por lo tanto, un lenguaje necesariamente privado es algo imposible.

Este argumento les permitió en gran medida a los conductistas insistir en su intento de definir las expresiones utilizadas corrientemente para los estados mentales en términos de sus conexiones con circunstancias y conductas por todos observables. A pesar de ese estímulo, esos intentos nunca llegaron a tener verdadero éxito (como vimos en el capítulo 2.2) y rápidamente se frustraron. Tal vez esto debió haberse esperado, porque del argumento de Wittgenstein se saca una conclusión más fuerte de lo que justifican sus premisas. Si lo que se necesita para que haya sentido es una verificación de la aplicación correcta, entonces todo lo que se requiere para comprender "W" son algunas conexiones entre la aparición de la sensación de W y la aparición de *otros fenómenos*, que no *necesariamente* tienen que ser observables por todos: pueden ser otros estados mentales, por ejemplo, y servir lo mismo como verificaciones de la correcta aplicación de "W".

Por lo tanto, la conclusión que se debió sacar del argumento de Wittgenstein es simplemente que ningún término puede tener sentido en ausencia de conexiones sistemáticas con otros. Al parecer, el significado es algo que un término sólo puede tener en el contexto de un sistema de otros términos, conectados entre sí por medio de enunciados generales que los contengan. Si Wittgenstein y los conductistas hubiesen sacado esta conclusión levemente más débil, tal vez los filósofos hubiesen llegado más rápidamente de lo que lo hicieron a la teoría semántica que se presenta en el apartado siguiente.

## Lecturas complementarias


Malcolm, Norman, "Wittgenstein's *Philosophical Investigations*", *Philosophical Review*, vol. LXIII, 1954. Reproducido en *The Philosophy of Mind*, V. C. Chappell (comp.), Englewood Cliffs, NJ, Prentice-Hall, 1962.

Strawson, Peter, "Persons", en *Minnesota Studies in the Philosophy of Science*, vol. II, H. Feigl y M. Scriven (comps.), Minneapolis, University of Minnesota Press, 1958.

Reproducido en *The Philosophy of Mind*, V. C. Chappell (comp.), Englewood Cliffs, NJ, Prentice-Hall, 1962.

Hesse, Mary, "Is There an Independent Observation Language?", en *The Nature and Function of Scientific Theories*, R. Colodny (comp.), Pittsburgh, Pittsburg University Press, 1970. Véanse especialmente págs. 44-45.

# 3. La tesis teórica reticular y la psicología popular

La concepción que hemos de investigar en este apartado se puede formular del modo siguiente. Los términos utilizados corrientemente para referirse a los estados mentales son los *términos teóricos* de un marco de referencia teórico (la psicología popular) ya incorporado en nuestro conocimiento corriente, y los significados de esos términos se han establecido de la misma manera en que lo han hecho los significados de los términos teóricos en general. Específicamente, su significado queda establecido por el conjunto de leyes/principios/generalizaciones en el que están incluidos. Para poder explicar esta concepción, voy a retroceder un poco para referirme brevemente a las teorías.

## La semántica de los términos teóricos

Consideremos algunas teorías en gran escala, como las que se encuentran en las ciencias físicas: la teoría química, la teoría electromagnética, la teoría atómica, la termodinámica, etc. Una teoría de este tipo consiste en un conjunto de enunciados, por lo general enunciados generales o *leyes*. Estas leyes expresan las relaciones que se dan entre las diversas propiedades/valores/clases/entidades cuya existencia es postulada por la teoría.

Tales propiedades y entidades están expresadas o denotadas por el conjunto de *términos teóricos* específicos de esa teoría.

La teoría electromagnética, por ejemplo, postula la existencia de cargas electrónicas, campos de fuerza eléctricos y

campos de fuerza magnéticos, y las leyes de esta teoría enuncian cómo estas cosas están relacionadas entre sí y con diversos fenómenos observables. Para comprender plenamente la expresión "campo eléctrico" hay que conocer bien la red de principios teóricos en la que aparece esa expresión; estos principios, en conjunto, nos dicen qué es un campo eléctrico y qué hace.

Este caso es típico. Los términos teóricos, por lo general, no obtienen su significado a partir de definiciones singulares y explícitas que enuncian las condiciones necesarias y suficientes para su aplicación, sino que están definidos implícitamente por la red de principios en los que están incluidos. Esas "definiciones" ocasionales, como las que efectivamente se encuentran ya dadas (por ejemplo, "El *electrón* es la unidad de electricidad") habitualmente sólo dan una pequeña parte del significado del término y en todo caso siempre son falsables (por ejemplo, ahora parece que el *quark* puede ser la unidad de electricidad, con una carga de un tercio de la del electrón). Denominemos a ésta la *teoría reticular del significado*.

## El modelo de explicación nomológico-deductivo

Sin embargo, las leyes de una teoría hacen algo más que limitarse a dar sentido a los términos que contienen. También cumplen una función predictiva y explicativa, y en esto reside su valor principal. Aquí se plantea la pregunta: ¿Qué es dar una *explicación* de un hecho o de un estado de cosas y cómo lo hacen posible las teorías? Podemos introducir la sabiduría convencional en este punto por medio de la siguiente anécdota.

En mi laboratorio hay un aparato que consiste en una larga barra de metal con dos espejos enfrentados, cada uno atado a uno de los extremos de la barra. La función de la barra es la de mantener a los espejos a una distancia precisa entre sí. Una mañana, mientras volvía a medir la distancia precisamente antes de efectuar algún experimento, mi asistente advierte que la barra es ahora más larga que antes, aproximadamente un milímetro.

—Eh —exclama—, esta barra se ha ampliado. ¿Cómo es eso?
—Porque la calenté —le explico.
—¿S-í? —vacila—, ¿eso tiene que ver con algo?
—Bueno, la barra está hecha de cobre —le vuelvo a explicar.
—¿S-í? —insiste—; ¿y qué tiene que ver?
—Bueno, el cobre se dilata con el calor —replico, tratando de no mostrar mi irritación.
—Ah-h, ya veo —dice, cuando finalmente comprende.

Si, después de mi última observación, mi asistente todavía no hubiese comprendido, yo hubiese tenido que despedirlo, porque la explicación de por qué se dilata la barra ahora está completa y hasta un niño se hubiese dado cuenta. Podemos ver por qué y en qué sentido está completa si consideramos toda la información reunida que contenía mi explicación.

1. El cobre se dilata con el calor.
2. Esta barra es de cobre.
3. Esta barra fue sometida a la acción del calor.

---

4. Esta barra se dilató.

El lector advertirá que, en su conjunto, las tres primeras proposiciones *deductivamente presuponen* a la cuarta, que es el enunciado del hecho o del estado de cosas que se ha de explicar. La dilatación de la barra es una consecuencia inevitable de las condiciones descriptas en las tres primeras proposiciones.

Estamos considerando aquí un *argumento* deductivo válido. Parece que una explicación tiene la forma de un argumento, un argumento cuyas premisas (el *explanans*) contienen la información explicativa, y cuya conclusión (el *explanandum*) describe el hecho que se ha de explicar. Lo que es más importante, las premisas incluyen un enunciado *nomológico*: una ley natural, un enunciado general que expresa los modelos que sigue la naturaleza. Las otras premisas expresan lo que se denomina comúnmente las "condiciones

iniciales", que son las que conectan la ley con el hecho específico que requiere explicación. En suma, explicar un hecho o un estado de cosas es deducir su descripción a partir de una ley natural. (De allí el nombre, "el modelo de explicación nomológico-deductivo".) Ahora resulta muy fácil ver la conexión entre teorías exhaustivas y potencia explicativa.

La *predicción* de hechos y estados de cosas —debemos advertirlo— sigue esencialmente el mismo patrón. La diferencia es que las conclusiones de los argumentos correspondientes están en tiempo futuro y no en pasado o presente. Téngase en cuenta algo más todavía. Cuando se da una explicación en la vida corriente, casi nunca se enuncia cada una de las premisas del argumento correspondiente. (Véase la primera respuesta que le di a mi asistente.) Generalmente no hay ninguna necesidad, puesto que uno puede suponer que sus oyentes ya poseen la mayor parte de la información pertinente. Lo que se les da es solamente aquella información específica que uno supone que les falta (por ejemplo, "la calenté"). La mayor parte de las explicaciones expresadas son sólo esbozos de explicación. Queda a cargo del oyente completar lo que quedó sin decir. Por último, hay que destacar que las "leyes" que subyacen en nuestras explicaciones corrientes por lo general no son detalladas y sólo expresan una aproximación preliminar, o una comprensión incompleta, de las regularidades que abarcan. De modo que ésta es una dimensión adicional en la que las explicaciones son por lo general esbozos de una explicación.

## La psicología corriente

Consideremos ahora la enorme capacidad que tienen los seres humanos *normales* para explicar y predecir la conducta de sus congéneres. Inclusive podemos explicar y predecir los estados psicológicos de otros seres humanos. Explicamos su conducta en términos de sus creencias y deseos, y de sus dolores, esperanzas y temores. Explicamos su tristeza en términos de sus decepciones, sus intenciones en términos de sus

deseos, y sus creencias en términos de sus percepciones e inferencias. ¿Cómo es que somos capaces de hacer todo esto?

Si es correcta la versión que hemos dado de la explicación en el apartado anterior, entonces cada uno de nosotros debe de poseer un conocimiento o un dominio de un conjunto bastante sustancial de leyes o enunciados generales que conectan los diversos estados mentales con: 1) otros estados mentales, 2) circunstancias externas y 3) conductas manifiestas. ¿Es así o no?

Si se nos apura podemos encontrar algunas explicaciones de sentido común, como las que surgieron en el ejemplo de la conversación que dimos antes, para ver qué otros elementos comúnmente quedan inexpresados. Cuando lo hacemos, alegan los defensores de este punto de vista, literalmente dejamos al descubierto cientos y cientos de generalizaciones corrientes referidas a los estados mentales , como las siguientes:

Las personas tienden a sentir dolor en lugares del cuerpo donde se han lesionado recientemente.
Las personas que no han ingerido líquidos durante algún tiempo tienden a sentir sed.
Las personas doloridas tienden a querer aliviar ese dolor.
Las personas sedientas tienden a desear líquidos para beber.
Las personas enfadadas tienden a mostrarse impacientes.
Las personas que experimentan un súbito dolor agudo tienden a lamentarse.
Las personas enfadadas tienden a fruncir el entrecejo.
Las personas que quieren que P, y creen que Q sería suficiente para producir P, y no tienen necesidades conflictivas o estrategias preferidas, tratarán de producir esa Q.

Estos lugares comunes tan conocidos, y cientos de otros parecidos en los que se incluyen otros términos mentales, son los que constituyen nuestra comprensión de cómo funcionamos. Estos enunciados generales o *leyes* rudimentarias respaldan normalmente las explicaciones y predicciones. En su conjunto, constituyen una *teoría*, una teoría que postula una amplia gama de estados internos cuyas relaciones causales están descriptas por las leyes de la teoría. Todos incorporamos

[illegible] el lenguaje y al hacerlo adquirimos la concepción corriente de qué es la inteligencia consciente. A ese marco de referencia teórico podemos denominarlo "psicología tradicional", y es el que encarna la sabiduría acumulada de millares de intentos realizados a lo largo de las generaciones para comprender cómo funcionan los seres humanos.

Para ilustrar, brevemente, la función que cumplen estas leyes en las explicaciones corrientes, considérese el siguiente intercambio.

—¿Por qué Michael se sobresaltó levemente cuando se sentó al llegar a la reunión?
—Porque sintió un súbito dolor agudo.
—Ah, ya veo. ¿Y por qué sintió un dolor?
—Porque se sentó sobre la tachuela que le puse en la silla.

Aquí tenemos dos explicaciones, una a continuación de la otra. Si se nos apura, a la manera de nuestro ejemplo inicial, aparecerán las leyes sexta y octava de la enumeración anterior, a partir de la presunta información básica, y resultarán evidentes dos argumentos deductivos que muestran el mismo modelo de la explicación de la barra dilatada.

Si la psicología corriente es literalmente una teoría —aunque sea muy antigua y profundamente arraigada en el lenguaje y la cultura humanos— entonces los significados de los términos psicológicos efectivamente deben estar establecidos, como lo dice la tesis de este apartado: por medio del conjunto de las leyes psicológicas tradicionales en las que figuran. Esta concepción tiene cierta verosimilitud directa; después de todo, ¿quién dirá que alguien comprende el significado del término "dolor" si no tiene ninguna idea de que el dolor es producido por alguna lesión corporal, que a la gente no le gusta o que provoca angustia, sobresaltos, lamentaciones y conductas para evitarlo?

## Más sobre los qualia

Pero ¿qué sucede con los qualia de los diversos estados psicológicos humanos? ¿Verdaderamente podemos creer, como al parecer lo requiere la teoría reticular, que los qualia no cumplen ninguna función en el significado de los términos psicológicos? La intuición de que sí lo hacen es extremadamente fuerte. Existen al menos dos modos en los que el defensor de la teoría reticular intentaría manejar esta intuición permanente.

El primero consiste en admitir simplemente que los qualia efectivamente desempeñan *alguna* función en el significado de *algunos* términos, aunque en el mejor de los casos sólo una función menor o secundaria. Con esta concesión se avanzaría bastante en el camino de satisfacer a la intuición, y resulta tentador adoptarla y declarar cerrada la cuestión. Pero en verdad deja algunos problemas sin resolver. Puesto que los qualia de sus sensaciones evidentemente son sólo suyos, y los míos solamente míos, *parte* del significado de los términos que utilizamos para las sensaciones seguirá siendo algo privado, y siempre continuará siendo una cuestión obstinadamente abierta saber si cada uno de nosotros quiere decir lo mismo cuando utiliza esos términos.

La segunda transacción acepta que los qualia cumplen una función significativa en la *aplicación* introspectiva de los términos para las sensaciones, pero todavía intenta negar que esa función tenga alguna significación *semántica*. La idea es que su discriminación introspectiva del dolor que produce una comezón, o de la sensación del rojo respecto de la del verde, por supuesto está regulada por el carácter cualitativo que tienen, en usted, los estados correspondientes. Cada uno de nosotros aprende a aprovechar aquellos qualia que ponen de manifiesto sus estados con el fin de efectuar juicios de observación espontáneos respecto de en qué estado se encuentra. Pero lo que se quiere decir estrictamente cuando se habla de "dolor", por ejemplo, no incluye ningún compromiso con ningún quale específico. El carácter cualitativo de los dolores varía sustancialmente inclusive dentro de un individuo de-

terminado; puede variar mucho más ampliamente aún cuando se trata de diferentes individuos y casi seguramente varía sustancialmente cuando se trata de especies biológicas distintas. Por lo tanto, los qualia tienen una significación epistemológica, pero están desprovistos de significación semántica para los términos de un lenguaje intersubjetivo.

De modo que tenemos dos aditamentos contradictorios a la teoría reticular del significado. Le dejo al lector la decisión sobre cuál de los dos debería adoptarse. En ambos casos, la lección fundamental es evidente de por sí: la fuente de significado dominante, y tal vez la única, para los términos psicológicos es el sistema teórico corriente en el que están incluidos. Del mismo modo que ocurre en el caso de los términos teóricos generales, sólo se llega a comprenderlos cuando se aprenden a usar las generalizaciones predictivas y explicativas en las que aparecen.

## Significación general

La significación de esta teoría reticular del significado —para el problema mente/cuerpo— es la siguiente. Esta teoría es estrictamente compatible con las tres posiciones materialistas corrientes, y también lo es con el dualismo. Por sí sola no presupone ni excluye ninguna de esas posiciones. Lo que sí hace es decirnos algo sobre la naturaleza del conflicto entre todas ellas y sobre el modo en que se resolverá el conflicto. La lección que nos deja es la siguiente.

Si el marco de referencia corriente que utilizamos para los estados psicológicos es literalmente una *teoría*, entonces el problema de la relación de los estados mentales con los estados cerebrales se transforma en el problema de cómo una vieja teoría (la psicología popular) se va a relacionar con una teoría nueva (la neurociencia plenamente desarrollada), que en alguna medida intenta desplazarla. Las cuatro posiciones principales sobre el problema mente/cuerpo aparecen como cuatro anticipaciones diferentes de cómo se va a resolver el conflicto teórico. El teórico de la identidad espera que la vieja teoría

será reducida sin conflicto por la neurociencia desarrollada. El dualista sostiene que no se producirá esa reducción, basándose en que la conducta humana no tiene orígenes físicos. El funcionalista también espera que la vieja teoría no sea reducida, pero porque (irónicamente) dos tipos de sistemas físicos muy diferentes pueden producir exactamente la misma organización causal especificada por la vieja teoría. Y el materialista eliminativo también espera que no se pueda reducir la vieja teoría, pero sobre la base muy diferente de que se trata simplemente de una teoría demasiado confusa e imprecisa como para poder sobrevivir a una reducción interteórica.

Lo que está en juego aquí es el destino de una teoría, el destino de un marco de referencia explicativo especulativo, a saber, nuestra amada psicología corriente. Y es evidente que la cuestión que se plantea entre estos cuatro destinos posibles es básicamente una cuestión empírica, que sólo podrá quedar zanjada decisivamente por medio de la investigación permanente en las neurociencias, la psicología cognitiva y la inteligencia artificial. Algunos de los resultados existentes de la investigación ya han sido expuestos en el capítulo 2. Algunos más los investigaremos en los últimos tres. La conclusión que se saca en este capítulo —que la concepción que tenemos sobre nosotros es y siempre ha sido una concepción teórica por derecho propio— coloca todos estos resultados en una perspectiva más profunda.

Como lo veremos luego, la teoría reticular del significado también tiene consecuencias muy importantes para los polémicos problemas epistemológicos estudiados en el capítulo siguiente. Volveremos a estos problemas luego de examinar una última cuestión referente al significado: la *intencionalidad* de muchos de nuestros estados mentales.

## Lecturas complementarias

Sellars, Wilfrid, "Empiricism and the Philosophy of Mind" en *Minessota Studies in the Philosophy of Science*, vol. I, H. Feigl y M. Scriven (comps.), Minneapolis, University of Minnesota Press, 1956. Reproducido en Wilfrid Sellars, *Science, Perception and Reality*, Nueva York, Routledge y Keegan Paul, 1963; véanse especialmente secciones 45-63.

Fodor, Jerry y Chihara, C., "Operationalism and Ordinary Language: A Critique of Wittgenstein", *American Philosophical Quarterly*, vol. 2, Nº 4, 1965.
Churchland, Paul, *Scientific Realism and the Plasticity of Mind*, Cambridge, Cambridge University Press, 1979, sección 12.
Hempel, Carl y Oppenheim, Paul, "Studies in the Logic of Explanation", *Philosophy of Science*, vol. 15, 1948, parte I. Reproducido en *Aspects of Scientific Explanation*, Carl Hempel (comp.), Nueva York, Collier-Macmillan, 1965.
Churchland, Paul, "The Logical Character of Action Explanations", *Philosophical Review*, vol. LXXXIX, Nº 2, 1970.

# 4. Intencionalidad y actitudes proposicionales

Hasta aquí hemos tratado de investigar el lenguaje que utilizamos para hablar de nuestros estados mentales y estudiamos en especial el origen del significado de las teorías. Concentremos ahora la atención en algunos de esos estados mentales en sí —en los pensamientos, creencias y temores, por ejemplo— puesto que cada uno de ellos también tiene un "significado", un "contenido" proposicional específico. Tenemos

el *pensamiento* de que (los niños son maravillosos),
la *creencia* de que (los seres humanos tienen grandes potencialidades), y
el *temor* de que (la civilización pueda caer en otra Edad Oscura).

Tales estados se denominan *actitudes proposicionales* porque cada uno expresa una "actitud" diferente respecto de una proposición específica. En el vocabulario técnico de los filósofos, se dice que esos estados manifiestan una *intencionalidad*, puesto que "aluden" a algo o "señalan" algo más allá de sí mismos: "aluden" o señalan a los niños, los seres humanos y la civilización. (Advertencia: este uso del término "intencionalidad" no tiene nada que ver con el término "intencional" que significa "hecho deliberadamente".)

Las actitudes proposicionales no son raras y son dominantes en el vocabulario psicológico tradicional. Recuérdese que uno puede sospechar que P, esperar que P, desear que P,

enterarse de que P, saber introspectivamente que P, inferir que P, suponer que P, conjeturar que P, preferir este P a ese Q, estar hastiado de que P, estar encantado de que P, asombrado de que P, alarmado de que P, y así sucesivamente. En su conjunto, esos estados constituyen la esencia de la inteligencia consciente en términos de la psicología tradicional.

A veces se ha mencionado la intencionalidad de estas actitudes proposicionales como el rasgo decisivo que distingue lo mental de lo meramente físico, como algo que no puede manifestar ningún estado puramente físico. En parte esta afirmación podría ser correcta, en el sentido de que la manipulación racional de las actitudes proposicionales efectivamente puede ser el rasgo distintivo de la inteligencia consciente. Pero aunque la intencionalidad haya sido mencionada con frecuencia como la "marca de lo mental", esto no constituye necesariamente una presunción en favor de alguna forma de dualismo. Ya hemos visto, en el capítulo 2.3, cómo estados puramente físicos, como los estados cerebrales, podían tener contenido proposicional y, por ende, manifestar intencionalidad. Al parecer, tener contenido o significado es simplemente cumplir una función específica en una economía compleja inferencial/computacional. Y no hay ninguna razón por la cual los estados internos de un cerebro, o inclusive de un ordenador, no puedan cumplir esa función.

Si determinados estados de nuestro cerebro efectivamente cumplen esa función, y si nuestros estados mentales en algún sentido son idénticos a esos estados (como afirman tanto el funcionalismo como la teoría de la identidad), entonces no tenemos aquí una refutación del materialismo sino más bien una explicación plausible de cómo es posible que nuestras actitudes proposicionales tengan contenido proposicional en primer lugar. Y si tienen un significado o contenido proposicional distintivo, entonces también tendrán referencia (o intento de referencia); tendrán esa manera de "señalar más allá" de sí mismos que caracteriza originariamente a la intencionalidad.

Hay una ironía histórica en el hecho de que las actitudes proposicionales ocasionalmente hayan sido mencionadas por

los filósofos como aquello que permite delimitar lo mental como algo completamente diferente de lo físico. La ironía está en que, cuando examinamos la estructura lógica de las concepciones tradicionales sobre este punto, no encontramos diferencias sino algunas *semejanzas* muy profundas entre la estructura de la psicología popular y la estructura de las teorías paradigmáticamente físicas. Comencemos por comparar los elementos presentes en las dos listas siguientes.

| *Actitudes proposicionales* | *Actitudes numéricas* |
|---|---|
| ... cree que P | ... tiene una longitud$_m$ de *n* |
| ... desea que P | ... tiene una velocidad$_{m/s}$ de *n* |
| ... teme que P | ... tiene una temperatura$_k$ de *n* |
| ... ve que P | ... tiene una carga$_c$ de *n* |
| ... sospecha que P | ... tiene una energía cinética$_j$ de *n* |

Donde la psicología corriente manifiesta actitudes *proposicionales*, la física matemática manifiesta actitudes *numéricas*. Una expresión de la primera lista se completa colocando un término relacionado con una proposición específica en lugar de "P"; una expresión de la segunda lista se completa colocando un término relacionado con un número específico en lugar de "*n*". Sólo entonces tenemos un predicado determinado. A partir de este paralelo estructural se pueden hacer otros. Así como las relaciones entre números (por ejemplo, ser el doble de largo que *n*) también pueden caracterizar las relaciones entre *actitudes* numéricas (por ejemplo, mi peso es el doble del suyo), del mismo modo las relaciones entre proposiciones (por ejemplo, incoherencia lógica, presuposición) también caracterizan las relaciones entre *actitudes* proposicionales (por ejemplo, mi creencia no es compatible con la suya). Los respectivos tipos de actitudes "heredan" las propiedades abstractas que poseen sus respectivos tipos de objetos abstractos.

Estos paralelos constituyen la base del paralelo más importante de todos. Si bien la relación entre ciertos tipos de actitudes proposicionales, o entre ciertos tipos de actitudes numéricas, tiene validez universal, podemos formular *leyes*, leyes que utilicen las relaciones abstractas que se dan entre

las actitudes que ponen en relación. Muchas de las leyes explicativas de la psicología popular ponen de manifiesto precisamente este modelo.

- Si *x* teme que P, entonces *x* desea que no P.
- Si *x* espera que P, y *x* descubre que P, entonces *x* se alegra de que P.
- Si *x* cree que P, y *x* cree que (si P, entonces Q), entonces salvo confusión, distracción, etc., *x* creerá que Q.
- Si *x* desea que P, y *x* cree que (si Q, entonces P), y *x* es capaz de producir que Q, entonces, salvo deseos conflictivos o estrategias preferidas, *x* producirá que Q.[1]

De modo similar, las leyes de la física matemática ponen de manifiesto una estructura escrupulosamente paralela, salvo que se utilizan allí relaciones numéricas y no lógicas

- Si *x* tiene una presión de P, y *x* tiene un volumen de V, y *x* tiene una masa de μ, entonces, salvo presión o densidad muy altas, *x* tiene una temperatura de PV/μ R.
- Si *x* tiene una masa de M, y *x* es sometido a una fuerza neta de F, entonces *x* tiene una aceleración de F/M.

Ejemplos como éste se pueden multiplicar por miles. Además, muchas de las expresiones utilizadas en las ciencias físicas contienen un término para un *vector*, y las leyes que abarcan esas "actitudes vectoriales" en forma característica manifiestan o utilizan las relaciones *algebraicas /trigonométricas* que se dan entre los vectores denotados por esos términos. Por ejemplo:

[1] Estrictamente hablando, todas estas oraciones deben ser universalmente cuantificadas, y también hay que hacer especificaciones sobre los términos así como sobre los conectores. Pero, puesto que éste es un libro introductorio y no presupone el conocimiento de la lógica formal, dejaré de lado esas sutilezas. Un análisis adecuado de estas cuestiones se hace en el trabajo de Paul Churchland que figura en la lista de lecturas complementarias al final de este apartado.

- Si $x$ tiene un momento de $P_x$ e $y$ tiene un momento de $P_y$ y $x$ e $y$ son los únicos cuerpos que interactúan en un sistema aislado, entonces la suma vectorial de $P_x$ y $P_y$ es una constante en el tiempo.

Lo que sucede en estos ejemplos es lo mismo en todos los casos. Se recurre a las relaciones abstractas que se dan en el ámbito de ciertos objetos abstractos —números, o vectores, o proposiciones— para ayudarnos a enunciar las regularidades empíricas que se dan entre estados y objetos *reales*, como por ejemplo entre temperaturas y presiones, fuerzas y aceleraciones, interacción entre un momentum y otro... y entre diversos tipos de estados mentales. El marco de referencia conceptual de la psicología popular se maneja con una estrategia conceptual que es normal en muchos proyectos conceptuales. Y así como una teoría no es esencialmente física, ni esencialmente no física, por utilizar números o vectores, tampoco una teoría es esencialmente física ni esencialmente no física, por utilizar proposiciones. Queda como un interrogante empírico llegar a saber si las actitudes proposicionales en última instancia son de naturaleza física. El mero hecho de que sean actitudes *proposicionales* (y por ende manifiesten intencionalidad) no presupone nada en un sentido ni en otro.

Habría dos evidentes conclusiones aleccionadoras que pueden extraerse de este breve análisis. La primera es la idea de que, puesto que el significado surge a partir del lugar que ocupa un elemento en una red de supuestos, y a partir de la función conceptual resultante que cumple el elemento en la economía inferencial actual del sistema, por lo tanto nuestros estados mentales pueden tener el contenido proposicional que tienen nada más que a causa de sus intrincados rasgos *relacionales*. Esto significaría que no hay ningún problema en suponer que estados físicos puedan tener contenido proposicional, puesto que en principio fácilmente podrían poseer los rasgos relacionales correspondientes. Este punto de vista está ahora bastante difundido entre los investigadores de este campo, pero no es la opinión universal, por lo cual se le pide al lector que actúe con cautela.

La segunda lección se refiere a las analogías estructurales muy estrechas que se dan entre los conceptos y las leyes de la psicología popular y los conceptos y las leyes de otras teorías. La presencia de esos paralelismos es totalmente compatible con el criterio, ya señalado en el apartado anterior, de que la psicología corriente es literalmente una teoría. En el capítulo siguiente aparecerán más elementos en apoyo de este criterio.

## Lecturas complementarias


Brentano, Franz, "The Distinction between Mental and Physical Phenomena", en *Realism and the Background of Phenomenology*, R. M. Chisholm (comp.), Glencoe, IL, Free Press, 1960.

Chisholm, Roderick, "Notes on the logic of Believing", *Philosophy and Phenomenological Research*, vol. 24, 1963.

Churchland, Paul, "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy*, vol. 78, Nº 2, 1981, sección I.

Churchland, Paul, *Scientific Realism and the Plasticity of Mind*, Cambridge, Cambridge University Press, 1979, sección 14.

Field, Hartry, "Mental Representation", *Erkenntnis*, vol. 13, Nº 1, 1978. Reproducido en *Readings in Philosophy of Psychology*, vol. II, N. Block (comp.), Cambridge, M.A., Harvard University Press, 1981.

Fodor, Jerry, "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology", *The Behavioral and Brain Science*, vol. 3, 1980.

Stich, Stephen C., *From Folk Psychology to Cognitive Science: The Case Against Belief*, Cambridge, MA, MIT Press, Bradford, 1983.


### Contra la teoría reticular-inferencial del significado y la intencionalidad


Searle, John, "Minds, Brains and Programs", *The Behavioral and Brain Sciences*, vol. III, Nº 3, 1980.

# 4

# El problema epistemológico

El problema epistemológico se divide en dos partes, que se refieren ambas al modo en que llegamos a adquirir el *conocimiento* de las actividades internas de la mente consciente e inteligente. La primera de ellas se denomina el *problema de las otras mentes:* ¿cómo puede alguien determinar si algo diferente de sí mismo —un extraterrestre, un robot sofisticado, un ordenador socialmente activo o inclusive otro ser humano— verdaderamente es un ser consciente, que piensa y siente, y no es, por ejemplo, un autómata inconsciente cuya conducta surge de alguna otra cosa que no sean estados mentales genuinos? ¿Cómo se puede decidir? La segunda es el denominado *problema de la autoconciencia:* ¿cómo es que todo ser consciente tiene un conocimiento directo y privilegiado de sus propias sensaciones, emociones, creencias, deseos, etc.? ¿Cómo es posible? ¿Y en qué medida ese conocimiento es fiable? Las soluciones a estos dos problemas, como creo que se pondrá de manifiesto, no son independientes. Comencemos por examinar el primer problema.

## 1. El problema de las otras mentes

Por supuesto, mediante la observación de la conducta de una criatura, incluyendo su conducta verbal, es que podemos estimar que se trata de una criatura consciente y pensante: que constituye otra mente. A partir de una lesión corporal y

un lamento, inferimos el dolor. A partir de sonrisas y risas, inferimos el placer. Si se esquiva una bola de nieve, inferimos que hay percepción. A partir de una manipulación compleja y adecuada del ambiente, inferimos deseos, intenciones y creencias. A partir de estos y otros elementos, y por sobre todo a partir del lenguaje, inferimos que una criatura posee inteligencia consciente.

Todo esto resulta obvio, pero las observaciones que hemos hecho sirven sólo para presentar el problema, no para resolverlo. El problema comienza a aparecer cuando se pregunta qué es lo que *justifica* los tipos de inferencias mencionados. Inferir la aparición (oculta) de ciertos tipos de estados mentales a partir de la aparición de ciertos tipos de conductas es suponer que entre ellas se dan ciertas conexiones generales adecuadas, conexiones que presuntamente tienen la forma de: "Si toda criatura pone de manifiesto una conducta del tipo *B*, entonces normalmente se está produciendo un estado mental del tipo *S*". Tales "generalizaciones psicoconductuales" tienen la forma de generalizaciones empíricas corrientes, como por ejemplo, "Si se produce un sonido parecido al del trueno, entonces normalmente se produce (o se produjo) un relámpago en algún lugar cercano". Presumiblemente, su justificación también es paralela: tales enunciados generales están justificados por nuestra experiencia pasada con una conexión regular entre los fenómenos mencionados. En cualquier lugar y en cualquier momento en que percibimos un relámpago, por lo general también percibimos un trueno (muy fuerte) y, excepto la maquinaria de guerra, no hay ninguna otra cosa que produzca exactamente ese sonido.

Pero, ¿cómo puede alguien justificar su creencia en que las generalizaciones psicoconductuales correspondientes valen también para otras criaturas, *cuando todo lo que es posible observar es sólo la mitad de la presunta conexión: la conducta de la criatura?* Los estados mentales de la criatura, si efectivamente los tiene, sólo son observables directamente por ella misma. No los podemos observar. Y tampoco tenemos ninguna posibilidad de conseguir el apoyo empírico necesario. Evidentemente, entonces, no podemos tener ninguna justificación para

creer en esas generalizaciones psicoconductuales. Y, en consecuencia, no podemos tener ninguna justificación para hacer inferencias acerca de la conducta de otra criatura, acerca de que posee estados mentales. Y esto equivale a decir que ¡no podemos tener ninguna justificación para creer que alguna otra criatura, salvo uno mismo, posea estados mentales!

Esta conclusión es profundamente improbable, pero el problema del escepticismo es sumamente sólido. Creer en la existencia de las otras mentes requiere inferencias a partir de la conducta; tales inferencias requieren generalizaciones acerca de las criaturas en general; tales generalizaciones sólo pueden estar justificadas por la experiencia de las criaturas en general; pero la propia experiencia sobre uno mismo es todo lo que se puede tener. Este es el problema clásico de las otras mentes.

## El argumento de la analogía

Existen tres intentos clásicos por dar una solución al problema de las otras mentes, y tal vez el más simple de ellos sea el *argumento de la analogía*. Es posible observar las dos mitades de las conexiones psicoconductuales exclusivamente en un solo caso, se alega: en el propio. Y es posible determinar que las generalizaciones correspondientes efectivamente son verdaderas, por lo menos en el propio caso. Pero, hasta donde me es posible observar, otros seres humanos son completamente semejantes a mí. Si las generalizaciones valen para mí, entonces es razonable inferir, por analogía con mi propio caso, que también valen para otros seres humanos. Por lo tanto, efectivamente tengo alguna justificación, después de todo, para aceptar esas generalizaciones y, por lo tanto, fundándome en ellas, tengo una justificación para hacer inferencias específicas sobre los estados mentales de determinadas criaturas.

Nuestra tendencia a resistirnos a la conclusión escéptica en el problema de las otras mentes es lo suficientemente poderosa como para que exista la posibilidad de que nos afe-

rremos a cualquier solución que apunte a evadirla. Sin embargo, el argumento de la analogía plantea serias dificultades y debemos ser muy cautelosos antes de aceptarlo. El primer problema es que presenta el conocimiento que se tiene de las otras mentes sobre la base de una generalización inductiva a partir de *un solo* caso exclusivamente. Categóricamente, se trata del ejemplo más débil posible de un argumento inductivo, comparable a inferir que todos los osos son blancos fundándose en la observación de un solo oso (un oso polar). Bien cabe preguntarse si nuestra sólida confianza en la existencia de otras mentes puede explicarse y agotarse en un argumento tan débil. Sin ninguna duda, se puede objetar, mi creencia en que usted es consciente tiene un fundamento mejor que *ese*.

Y existen otros problemas además. Si el conocimiento que tenemos de las otras mentes en última instancia está limitado por lo que se puede observar en el propio caso, entonces un daltónico no tendrá ninguna justificación para creer que otros seres humanos tienen sensaciones visuales que él no puede experimentar, ni tampoco sería posible que un sordo creyera que los demás oyen, etc.,etc. Razonablemente se puede atribuir a las otras mentes, según este criterio, solamente aquello que uno encuentra en su propia mente. Esto presupone, por ejemplo, que no tendríamos ninguna justificación para atribuir estados mentales a un extraterrestre, si su psicología fuera sistemáticamente diferente de la nuestra (cosa que, después de todo, es altamente probable). Las hipótesis razonables que se pueden hacer sobre el contenido de las otras mentes, ¿verdaderamente están tan estrechamente limitadas?

Una tercera objeción es la que intenta destruir completamente el argumento de la analogía, con una versión del modo en que llegamos a apreciar las conexiones psicoconductuales que se producen. Si yo tengo que distinguir entre las múltiples variedades de estados mentales y reconocerlas claramente, y a partir de allí conjeturar las conexiones que mantienen con mi conducta, debo poseer los conceptos necesarios para efectuar tales juicios identificatorios: tengo que comprender el significado de los términos "dolor", "pena", "temor", "deseo",

"creencia" y otros por el estilo. Pero ya hemos visto en el capítulo anterior que el significado de esos términos está dado, en gran medida o en su totalidad, por una red de supuestos generales que los conectan con términos referidos a otros estados mentales, circunstancias externas y conductas observables. Por lo tanto, simplemente contar con los conceptos pertinentes es *ya* dominar las conexiones generales entre estados mentales y conducta que supuestamente iba a proporcionar el examen del propio caso. Por lo tanto, la comprensión de los conceptos de la psicología popular debe provenir de alguna otra cosa que no sea el mero examen rudimentario de nuestro propio fluir de la conciencia.

En su conjunto, estas dificultades que plantea el argumento de la analogía constituyeron un motivo muy sólido para tratar de buscar una solución diferente al problema de las otras mentes. Una solución que no creara problemas del mismo orden que el que había que resolver.

## Nuevamente el conductismo

Los partidarios del conductismo filosófico se apresuraron a proponer una solución diferente, en la que se tomaban en cuenta las dificultades descubiertas en el argumento de la analogía. Específicamente alegaban que, si las generalizaciones que conectan estados mentales con conductas no pueden ser adecuadamente justificadas por la observación empírica, entonces tal vez fuese porque esas generalizaciones no eran empíricas en primer lugar. Lo que ocurre, se señaló, es que esas generalizaciones son verdaderas meramente por *definición*. Son definiciones operacionales de los términos psicológicos que contienen. Como tales, no tienen ninguna necesidad de justificación empírica. Y una criatura que se comporta en la forma adecuada, o se dispone a hacerlo, es *por definición* una criatura consciente, sensible e inteligente. (Los conductistas típicos no siempre han hecho afirmaciones tan temerarias y directas, pero con frecuencia tampoco fueron muy claros en sus posturas.)

Frente a la urgencia por resolver el problema de las otras mentes, dada la impotencia del argumento de la analogía y la seducción que ejercía la idea de que el significado de los términos psicológicos de alguna manera está ligado a generalizaciones psicoconductuales, es fácil ver por qué los filósofos se esforzaron tanto por introducir alguna variante en esta propuesta. Pero no lo lograron. Cuando examinamos las generalizaciones de la psicología tradicional, encontramos que muy rara vez, o ninguna, adoptan la forma de simples "definiciones operacionales"(recuérdese el análisis del término "soluble" en 2.2.). Los conductistas no lograron formular las condiciones *conductuales* necesarias y suficientes para la aplicación de un solo término psicológico. Y, al parecer, tampoco las generalizaciones de la psicología corriente son verdaderas por definición. Más bien parecen ser verdades empíricas rudimentarias, tanto por sus efectos sobre nuestras intuiciones lingüísticas como por sus funciones explicativas y predictivas en la comunicación cotidiana. Este hecho nos pone de nuevo frente al problema de la *justificación* de las diversas generalizaciones psicoconductuales de las que parece depender nuestro conocimiento de las otras mentes.

## Hipótesis explicativas y la psicología corriente

El problema de las otras mentes fue formulado por primera vez en un momento en que los conocimientos sobre la naturaleza de la justificación teórica eran todavía bastante primitivos. Hasta hace relativamente poco tiempo, todo el mundo creía que una ley general sólo podía justificarse por medio de una generalización inductiva lograda a partir de una cantidad adecuada de ejemplos observados de los elementos que abarcaba la ley. Se observa una cantidad de cuervos, se advierte que cada uno de ellos es negro, y se generaliza: "todos los cuervos son negros". Y lo mismo se hace para cualquier ley. Eso era lo que se pensaba. Esta idea podría haber resultado adecuada en el caso de leyes que conectan cosas y propiedades

observables, pero en la ciencia moderna hay una enorme cantidad de leyes que rigen la conducta de cosas y propiedades *no observables*. Pensemos en los átomos, las moléculas, los genes, las ondas electromagnéticas. Lisa y llanamente, las leyes que se refieren a elementos no observables deben tener alguna otra forma de justificación empírica, si es que van a encontrar algún tipo de justificación.

Esa otra forma de justificación no está muy lejos de lograrse. Los teóricos postulan entidades no observables, y leyes específicas que las rigen, porque ocasionalmente esto produce una teoría que permite formular predicciones y explicaciones de fenómenos observables inexplicados hasta ese momento. Más específicamente, si postulamos algunas hipótesis corrientes y las combinamos con información sobre circunstancias observables, a menudo podemos deducir enunciados referentes a otros fenómenos observables, enunciados que, como posteriormente se verifica, sistemáticamente resultan *verdaderos*. En la medida en que una teoría manifieste tales virtudes explicativas y predictivas, esa teoría se convierte en una hipótesis digna de crédito. Posee lo que se denomina comúnmente justificación "hipotético-deductiva" (o justificación "H-D", para abreviar). En suma, una teoría sobre elementos no observables puede ser digna de confianza si permite explicar y predecir algún dominio de fenómenos observables mejor que alguna otra teoría rival. En realidad éste es el modo normal de justificación de las teorías en general.

Consideremos ahora la red de principios generales —que conectan los estados mentales entre sí, con circunstancias corporales y con conductas— que constituyen la psicología corriente. Esta "teoría" permite explicar y predecir la conducta de los seres humanos mejor que cualquier otra hipótesis actualmente existente, y ¿qué mejor razón puede haber para creer en un conjunto de leyes generales acerca de estados y propiedades no observables? Las leyes de la psicología popular son dignas de confianza por la misma razón que las de cualquier otra teoría: su éxito para explicar y predecir. También hay que reparar aquí en que la propia justificación no necesita deberle nada en absoluto al examen personal del propio caso.

Lo que importa es el éxito que pueda tener la psicología corriente en lo que respecta a la conducta de la gente en general. Es dable pensar que el propio caso pudiera diferir del de los demás (recuérdese la objeción sobre el "extraterrestre" en contra del argumento de la analogía). Pero no necesariamente esto afecta el acceso teórico que se tiene a los estados internos, de los otros, por diferentes que pudiesen ser. Simplemente habría que utilizar una teoría psicológica diferente para entender esas conductas, una teoría diferente de la que abarca la propia vida interior y la conducta externa.

Si pasamos ahora de las leyes generales a los individuos, podemos decir que la hipótesis de que un determinado individuo tiene inteligencia consciente es también una hipótesis explicativa, según este criterio Y es plausible en la medida en que la conducta permanente del individuo se puede explicar y predecir mejor en términos de deseos, creencias, percepciones, emociones, etc. Puesto que éste es, en realidad, el mejor modo de comprender la conducta de la mayor parte de los seres humanos, tenemos por lo tanto una justificación para creer que constituyen las "otras mentes". Y del mismo modo tendremos una justificación para atribuir estados psicológicos a cualesquiera otras criaturas o máquinas, siempre que tales atribuciones constituyan las mejores explicaciones y predicciones de su conducta permanente.

De modo que ésta es la solución más reciente al problema de las otras mentes. Sus virtudes son bastante directas y es sumamente compatible con la solución que dimos antes al problema semántico. Al parecer, ambos problemas admitirían la hipótesis de que el marco conceptual que utilizamos corrientemente para referirnos a los estados mentales tiene todas las características de una teoría. Sin embargo, no todos consideran plausible este supuesto, a pesar de todas sus virtudes. Si uno concentra la atención en la conciencia directa que tiene de sus propios estados mentales, la idea de que son "entidades teóricas" puede parecer muy extraña. En el siguiente apartado nos referiremos al modo en que esta sugerencia podría tener sentido y nos preguntaremos si lo tiene.

**Lecturas complementarias**

Malcolm, Norman, "Knowledge of Other Minds", *Journal of Philosophy*, vol. LV, 1958. Reproducido en *The Philosophy of Mind*, V. C. Chappell (comp.), Englewood Cliffs, NJ, Prentice-Hall, 1962.

Strawson, Peter, "Persons", en *Minnesota Studies in the Philosophy of Science*, vol. II, H. Feigl, M. Scriven y G. Maxwell (comps.), Minneapolis, University of Minnesota Press, 1958. Reproducido en *The Philosophy of Mind*, V. C. Chappell (comp.), Englewood Cliffs, NJ, Prentice-Hall, 1962.

Sellars, Wilfrid, "Empiricism and the Philosophy of Mind", en *Minnesota Studies in the Philosophy of Science*, vol. I, H. Feigl y M. Scriven (comps.), Minneapolis, University of Minnesota Press, 1956. Reproducido en Wilfrid Sellars, *Science, Perception and Reality*, Londres, Routledge y Keegan Paul, 1963, secciones 45-63.

Churchland, Paul, *Scientific Realism and the Plasticity of Mind*, Cambridge, Cambridge University Press 1979, sección 12.

## 2. El problema de la autoconcieñcia

A primera vista, es probable que la autoconciencia pudiera parecer algo inexorablemente misterioso y absolutamente único. En esto reside parte de su atractivo. Sin embargo, cuando se lo piensa con mayor profundidad, el velo del misterio comienza a descorrerse un poco y se puede considerar a la autoconciencia como un ejemplo de un fenómeno más general.

Ser autoconsciente equivale a tener, como mínimo, *conocimiento* de uno mismo. Pero esto no es todo. La autoconciencia implica el conocimiento no sólo de los propios estados físicos sino también de los propios *estados mentales* específicamente. Y, por añadidura, también el mismo tipo de conocimiento *permanentemente actualizado* que uno tiene de su percepción permanente del mundo externo. Al parecer, la autoconciencia es un tipo de captación permanente de una realidad interna, la realidad de los propios estados y actividades internos.

El elemento de la captación es importante: evidentemente no basta simplemente con tener estados mentales. Se debe discriminar entre un estado y otro. Hay que reconocerlos por lo que son. En suma, hay que comprenderlos en el interior de algún marco conceptual en el que estén catalogados los diversos tipos diferentes de estados mentales. Sólo entonces será posible efectuar un *juicio* de reconocimiento ("Estoy enfadado", "Estoy entusiasmado", "Creo que P", etc.). Con esto se sugiere que existen grados diferentes de autoconciencia, puesto que es dable pensar que la capacidad para discriminar sutilmente entre diferentes tipos de estados mentales mejora con la práctica y con la mayor experiencia y, también, puesto que el marco de referencia conceptual dentro del cual se expresa el reconocimiento explícito se vuelve cada vez más sofisticado y exhaustivo a medida que se aprende cada vez más acerca de las complejidades de la naturaleza humana. En consecuencia, el autoconocimiento de un niño pequeño, si bien es real, será mucho más restringido y rudimentario que el de un adulto sensible. Lo que en el caso de un niño es simplemente antipatía por alguien, puede dividirse en una mezcla de celos, temor y desaprobación moral de alguien en el caso de un adulto sincero y autoobservador.

Con esto se señala también que la autoconciencia puede variar de una persona a otra, según cuáles sean las zonas de discriminación y comprensión que se dominen más cabalmente. Tal vez un novelista o un psicólogo tengan un conocimiento habitual de sus estados emocionales mucho más profundo que el que tenemos los demás; un lógico puede tener una conciencia más pormenorizada de la permanente evolución de sus creencias; un teórico decisionista puede tener un mayor conocimiento del flujo de sus deseos, intenciones y razonamientos prácticos; un pintor puede reconocer más vívidamente la estructura de sus sensaciones visuales, y así sucesivamente. La autoconciencia, evidentemente, tiene un componente en muy gran medida *aprendido*.

En este sentido, la conciencia introspectiva de uno mismo parece muy similar a la conciencia perceptiva que uno tiene

del mundo externo. La diferencia reside en que, en el primer caso, cualesquiera que sean los mecanismos de discriminación que funcionen, siempre armonizan con circunstancias internas en lugar de externas. Los mecanismos en sí presuntamente son innatos, pero hay que aprender a utilizarlos: a efectuar discriminaciones útiles y a apuntar juicios sagaces. Las habilidades perceptuales aprendidas son muy conocidas en el caso de la percepción externa. Quien dirige una sinfonía puede oír la presencia del clarinete en lo que para un niño es un sonido sin solución de continuidad. Un astrónomo reconoce los planetas, nebulosas y estrellas gigantes en medio de lo que para los demás son simplemente manchitas en el cielo nocturno. Un jefe de cocina experimentado puede percibir el sabor del romero y la escalonia dentro de lo que es simplemente un sabor delicioso para un comensal hambriento. Y así siguiendo. Es evidente que la percepción, ya sea interna o externa, es sustancialmente una habilidad aprendida. Por supuesto, la mayor parte de ese aprendizaje se produce en la primera infancia: lo que hoy es obvio para nosotros perceptivamente era una discriminación muy sutil a los ocho meses. Pero siempre hay posibilidad de aprender más.

En resumen, la autoconciencia, según este criterio, es simplemente una especie de percepción: la *autopercepción*. No es la percepción del propio pie con los propios ojos, por ejemplo, sino que se trata de la percepción de los propios estados internos con lo que podemos llamar (con un enorme desconocimiento) la propia facultad de introspección. De modo que la autoconciencia no es algo más enigmático (ni menos) que la percepción en general. Sólo que está dirigida hacia lo interno y no hacia lo externo.

Tampoco resulta para nada sorprendente que las criaturas muy maduras desde el punto de vista cognitivo posean autoconciencia. Lo único que requiere la percepción es que la propia facultad de juicio esté en un contacto causal sistemático con el ámbito que se ha de percibir de un modo tal que nos resulte posible aprender a efectuar, en forma permanente, juicios sobre ese ámbito que sean espontáneos, no inferidos y adecuados. Nuestra facultad de juicio está en contacto causal

con el mundo externo, por medio de las diversas modalidades sensoriales; pero también está en contacto causal sistemático con el resto del ámbito interno del que forma parte. ¿Quién podría asombrarse de que un tipo de actividad cerebral mantenga abundantes conexiones causales con otros tipos de actividad cerebral? Pero tales conexiones son portadoras de información, y de este modo hacen posible un juicio "informado". Por lo tanto, es dable esperar que exista la autoconciencia, en algún nivel o hasta cierto grado de profundidad, prácticamente en toda criatura madura desde el punto de vista cognitivo.

Este criterio es compatible con el punto de vista evolucionista. Presumiblemente la humanidad ha entablado una lucha por lograr la autoconciencia en dos dimensiones: en la evolución neurofisiológica de la capacidad para efectuar discriminaciones introspectivas útiles, y en la evolución social de un marco de referencia conceptual para aprovechar esa capacidad de discriminación para efectuar juicios útiles tanto desde el punto de vista explicativo como predictivo. Además, en cada uno de nosotros se entabla durante toda la vida una lucha evolutiva para llegar a la autocomprensión, y en ese proceso aprendemos a utilizar y refinar las aptitudes discriminativas innatas y a dominar el marco de referencia conceptual socialmente arraigado (la psicología corriente) necesario para aprovecharlas.

## El enfoque tradicional

Estas observaciones sobre la autoconciencia pueden parecer bastante plausibles, pero hay una larga tradición en la filosofía de la mente que adopta un punto de vista muy diferente sobre el conocimiento introspectivo. Se ha sostenido que la introspección es fundamentalmente diferente de cualquier otra forma de percepción externa. La percepción que tenemos del mundo externo siempre está mediatizada por sensaciones o impresiones de algún tipo, de modo que sólo se conoce el mundo externo de manera indirecta y problemática. Sin em-

bargo, con la introspección el conocimiento es inmediato y directo. Con la introspección no se capta una sensación por vía de una sensación de esa sensación, ni se capta una impresión por vía de una impresión de esa impresión. Como resultado, no se puede ser víctima de una falsa impresión (de una impresión) ni de una incorrecta sensación (de una sensación). Por lo tanto, no bien se consideran los estados de la propia mente, la distinción entre apariencia y realidad desaparece por completo. La mente es transparente para sí misma y, en la mente, necesariamente las cosas son exactamente lo que "parecen" ser. No tiene ningún sentido decir, por ejemplo, "Me pareció que yo sentía un fuerte dolor, pero estaba equivocado". En consecuencia, los juicios introspectivos sinceros que uno hace sobre sus propios estados mentales —o sobre las propias *sensaciones,* en todo caso— son no corregibles e infalibles: es lógicamente imposible que estén equivocados. La mente se conoce a sí misma en primer lugar, de un modo único, y mucho mejor de lo que puede llegar a conocer el mundo externo.

Esta extraordinaria posición debe tomarse muy en serio —por lo menos temporariamente— por diversas razones. En primer lugar, es uno de los aspectos de una vieja teoría, que tuvo gran influencia, del conocimiento en general: el empirismo ortodoxo. En segundo lugar, la afirmación de que el conocimiento que uno tiene de sus propias sensaciones no está mediatizado, por otras 'sensaciones$_2$', no parece admisible. Y cualquier intento de negarlo llevaría o bien a una regresión infinita de 'sensaciones$_3$', 'sensaciones$_4$', etc., o bien a algún nivel de 'sensaciones$_n$' en el cual el conocimiento que se tiene de ellas *es* finalmente no mediatizado. En tercer lugar, el defensor de esta postura cuenta con una poderosa pregunta retórica. "¿Cómo *sería* posible no saber si uno tiene *dolor* o no?" Como podrá advertir el lector, esta pregunta no es fácil de responder.

## Argumentos en contra del enfoque tradicional

La concepción de que la mente se conoce a sí misma en primer lugar, de una manera singular, y mucho mejor de lo

que puede llegar a conocer el mundo externo, ha dominado en el pensamiento occidental durante más de tres siglos. Pero si se adopta una perspectiva absolutamente naturalista y evolucionista sobre la mente, la concepción tradicional rápidamente adquiere una especie de cualidad de cuento de hadas. Después de todo, los cerebros han resultado seleccionados porque otorgaban una ventaja reproductiva a quienes los poseían. Y otorgaban esa ventaja porque les permitían a los individuos anticipar elementos de su ambiente, distinguir el alimento de lo que no lo es, a los depredadores de los no depredadores, la seguridad del peligro, y a los congéneres de los no congéneres. En suma, el cerebro les daba un conocimiento y control sobre *el mundo externo.* Los cerebros han resultado los beneficiarios de la selección natural precisamente a causa de ese rasgo. Evidentemente lo que conocen en primer lugar y mejor no es a sí mismos sino al ambiente en el que tienen que sobrevivir.

Es presumible que la capacidad para el *auto*conocimiento haya sido seleccionada como un concomitante incidental de la capacidad para el conocimiento en general, y habría sido seleccionada específicamente si aumentaba de alguna manera la capacidad del cerebro para el conocimiento externo. Pero en cualquiera de ambas situaciones ésta constituiría en el mejor de los casos una ventaja secundaria, derivada respecto del aumento del propio conocimiento y control del mundo externo. Y en todo caso, no existe ninguna razón para suponer que la autopercepción, en la medida en que efectivamente evolucionó, sería algo de tipo fundamentalmente diferente de la percepción externa, y no habría ninguna razón en absoluto para suponer que fuera infalible.

Si la concepción tradicional básicamente no es admisible, examinemos los argumentos planteados en su favor y veamos si pueden resistir una profunda inspección. Consideremos en primer lugar la pregunta retórica, "¿Cómo sería posible equivocarse respecto de la identidad de las propias sensaciones?" Como argumento en favor de la imposibilidad de corregir el conocimiento de nuestras propias sensaciones, tiene la siguiente forma: "Ninguno de nosotros puede *pensar* un modo

en el que pudiéramos estar errados en nuestros juicios acerca de nuestras sensaciones; por lo tanto, no *hay* ningún modo de que pudiéramos equivocarnos". Pero aquí se comete una falacia elemental: se trata de un argumento por la ignorancia. Bien podría haber modos en que el error fuese posible, a pesar de que no los conociéramos. En realidad, tal vez los ignoramos precisamente porque comprendemos tan poco acerca de los mecanismos ocultos de la introspección. Por lo tanto, la pregunta retórica se puede dejar de lado tranquilamente, aun cuando no podamos contestarla. Pero en realidad podemos. Con pequeño esfuerzo, podemos pensar en muchos modos en que pueden ocurrir, y de hecho ocurren, errores en el juicio introspectivo, como lo veremos inmediatamente.

Consideremos ahora el argumento de que la distinción entre apariencia y realidad necesariamente debe desaparecer en el caso de las sensaciones, ya que nuestra comprensión de esas sensaciones no está mediatizada por nada que pudiera representarlas mal. Este argumento es bueno sólo si la única manera en que pudiesen ocurrir los errores fuese la mala representación por un elemento mediador. Pero no es así. Aun si la introspección no está mediatizada por 'sensaciones$_2$' de segundo orden, nada garantiza que el juicio introspectivo "siento dolor" será provocado solamente por la aparición de dolores. Tal vez haya otras cosas que también pueden provocar ese juicio, por lo menos en circunstancias inusuales, en cuyo caso el juicio sería falso. Consideremos la aparición de algo bastante *similar* al dolor —una súbita sensación de extremo frío, por ejemplo— en una situación en la que se *espera* intensamente sentir dolor. Supongamos que usted es un espía que ha sido capturado y que se lo interroga extensamente con ayuda de un hierro caliente que se le aplica varias veces en la espalda a breves intervalos repetidos. Si, en el vigésimo intento, disimuladamente se le aplica en la espalda un *cubo helado,* su reacción inmediata diferirá poco o nada de las primeras diecinueve reacciones. Casi seguramente usted creería, por un instante, que siente dolor.

El partidario de la imposibilidad de corregir estos juicios tal vez insista en que la sensación número veinte era un dolor,

después de todo, a pesar de su causa benigna, sobre la base de que, si usted lo consideró como un dolor, si usted creyó que le resultaba doloroso, entonces verdaderamente era un dolor. Esta interpretación no se corresponde muy bien con el hecho de que uno puede recuperarse de esos tipos de mala identificación que acabamos de mencionar. El chillido inicial de horror da lugar a algo como: "Espere... espere... no es la misma sensación que antes. ¿Qué sucede aquí atrás?" Si la sensación número veinte realmente *fue* un dolor, ¿por qué nuestro juicio se invierte a los pocos segundos?

Un caso similar: la sensación gustativa que produce el sorbete de lima es sólo ligeramente diferente de la que produce el sorbete de naranja, y en las pruebas realizadas con los ojos vendados es notable los pocos aciertos de la gente para distinguir entre las dos sensaciones. Un sujeto que espera que le den un sorbete de naranja y le dan uno de lima, con toda seguridad puede identificar que la sensación que experimenta es la que produce normalmente el sorbete de naranja, y sólo se retractará inmediatamente después que le hagan paladear (a ciegas) el auténtico sabor de naranja. En este caso se *corrige* la propia identificación cualitativa, en flagrante contradicción con la idea de que los errores son imposibles. Los errores de este tipo se denominan *efectos de expectativa* y constituyen un fenómeno muy generalizado en el caso de la percepción. Evidentemente también se aplican a la introspección. La realidad que tienen los efectos de expectativa nos proporciona una fórmula para producir prácticamente todos los errores de identificación que se quiera, ya sea de cosas externas o de estados internos.

Es más, ¿verdaderamente sabemos lo suficiente acerca de los mecanismos de introspección como para insistir en que no hay ningún elemento mediador entre la sensación y el juicio acerca de ella? Es cierto, no hay ningún elemento mediador que *conozcamos,* pero esto no significa nada, puesto que sea como fuere debe existir una buena parte del funcionamiento de la mente que está por debajo del nivel de la detección introspectiva. Entonces ésta es otra posible fuente de error. La distinción entre apariencia y realidad puede ser difícil de

trazar, en el caso de las sensaciones, sólo porque sabemos tan poco acerca de los modos en que las cosas pueden salir mal y de hecho lo hacen.

Otra manera en que es posible juzgar mal acerca de las sensaciones aparece cuando consideramos las sensaciones de muy corta duración. Se pueden inducir artificialmente sensaciones como para que tengan una duración arbitraria. No resulta sorprendente que, a medida que disminuye la duración, la identificación fiable (de su identidad cualitativa) resulte cada vez más difícil y entonces los errores se vuelven no imposibles, sino inevitables. Esto equivale a decir que la correspondencia entre lo que el sujeto dice que es la sensación, y lo que su modalidad de producción indica que debe ser, es casi perfecta cuando se la presenta en forma prolongada, pero se reduce hasta llegar casi al nivel de la casualidad a medida que la duración de las presentaciones se aproxima a cero. Estos "efectos de presentación" también son muy generalizados en el caso de la percepción. Y, si el sujeto estuviese convenientemente drogado o estuviera agotado, la fiabilidad de sus identificaciones se reduciría todavía más rápidamente. Este fenómeno también es muy generalizado.

También debemos mencionar los efectos de la memoria. Supongamos el caso de una persona que, tal vez a causa de alguna lesión nerviosa padecida en su juventud, no hubiese experimentado dolor ni ninguna otra sensación táctil o visceral durante *cincuenta años*, o que haya sido daltónica durante ese mismo período. ¿Alguien puede suponer verdaderamente que, si la deficiencia nerviosa del sujeto fuese reparada súbitamente después de una laguna tan prolongada, el sujeto instantáneamente sería capaz de discriminar e identificar (= reconocer cuál es la clase de semejanza en ese caso) cada una de sus sensaciones recientemente recuperadas y hacerlo con precisión infalible? Esta idea no es admisible de ninguna manera. También se podrían producir efectos similares en el corto plazo, con ayuda de una droga que obnubilara temporariamente la memoria que tenemos de los diversos tipos de sensaciones. En ese caso lo más natural sería que se produjesen fallas en la identificación y directamente identifi-

caciones equivocadas. Y aun en los casos normales, ¿son absolutamente imposibles los lapsus de memoria espontáneos, aislados e inadvertidos? ¿Cómo puede excluirlos el defensor del criterio tradicional?

También merece mencionarse un tipo de caso más conocido. Supongamos que usted está soñando que tiene un violento dolor de cabeza o que está sufriendo un dolor agudísimo porque lo están torturando. Cuando usted se despierta de golpe ¿no se da cuenta, con una señal de alivio, de que usted *verdaderamente no* estaba padeciendo el dolor de cabeza o el dolor torturante, a pesar de la convicción que acompaña a todo sueño? La tesis de la imposibilidad de corregir comienza a parecer sumamente inadmisible.

Nada de todo esto debe resultar sorprendente. La tesis de la imposibilidad de corregir podría haber sido admisible inicialmente en el caso de las sensaciones, pero no lo es ni remotamente para la mayor parte de otros estados mentales como creencias, deseos y emociones. Somos notoriamente malos para juzgar, por ejemplo, si estamos celosos, o con espíritu vengativo, para juzgar nuestros deseos más básicos e inclusive nuestros propios rasgos de carácter. Es cierto que el carácter infalible de los juicios muy pocas veces se ha afirmado para otra cosa que no fuesen las sensaciones. Pero esta restricción plantea problemas de por sí. ¿Por qué la infalibilidad se aplica a las sensaciones, pero no a las emociones y los deseos? El conocimiento de estos últimos no parece más "mediatizado" que el de las primeras.

Resulta muy curioso advertir que las investigaciones recientes en psicología social han demostrado que las explicaciones que uno ofrece sobre su propia conducta con frecuencia no se originan en la introspección fiable, a pesar de que uno crea sinceramente en eso, sino que son urdidas espontáneamente en el momento como *hipótesis explicativas* que se adecuen a la conducta y las circunstancias observadas (véase el trabajo de Nisbett y Wilson citado en las lecturas complementarias que se encuentran al final de este apartado). Y con frecuencia se puede demostrar que están equivocadas, puesto que los informes introspectivos que se dan resultan ser una

función de rasgos totalmente externos de la situación experimental, que están bajo el control de los experimentadores. Según el criterio de estos investigadores, gran parte de lo que se consideran informes introspectivos en realidad constituyen la expresión de la propia *teorización* espontánea acerca de las propias razones, motivos y percepciones, mientras que las hipótesis producidas se basan en las mismas pruebas externas con que puede contar todo el mundo.

Consideremos un argumento final contra la tesis de la imposibilidad de corregir. Nuestros juicios introspectivos están encuadrados dentro de los conceptos de la psicología tradicional, y ya hemos determinado (en los capítulos 3.3, 3.4 y 4.1) que este marco de referencia tiene la estructura y el estatuto de una teoría empírica. Como ocurre con todos estos juicios, su validez depende totalmente de la teoría empírica en la cual están incluidos semánticamente los conceptos pertinentes, lo cual equivale a decir que, si la psicología corriente resultara ser una teoría radicalmente falsa, entonces su ontología total perdería sus pretensiones de realidad. Y todo juicio encuadrado en esos términos debería considerarse falso en razón de que presupondría una teoría de referencia falsa. Puesto que la psicología popular es una teoría empírica, siempre es rigurosamente posible que pudiera resultar ser radicalmente falsa. En consecuencia, siempre es posible que todo juicio encuadrado en esos términos sea falso. Por lo tanto, nuestros juicios introspectivos no son imposibles de corregir. No sólo podrían estar equivocados alguna vez, uno por uno, sino que además ¡bien podrían ser *todos* absurdos!

## La carga teórica de toda percepción

La extrañeza que puede producir la idea de que los estados mentales son "teóricos" se reduce en parte cuando se hacen las siguientes reflexiones. *Todos* los juicios de percepción, no simplemente los introspectivos, tienen una "carga teórica": toda percepción involucra una interpretación especulativa. Por lo menos así lo afirman las versiones del empirismo desarrolladas

más recientemente. La idea básica que subyace en esta afirmación se puede expresar con el siguiente argumento, muy breve pero muy general: el *argumento reticular*.

1. Todo juicio de percepción supone la aplicación de *conceptos* (por ejemplo, *a* es F).
2. Todo concepto es un nudo en una *red* de conceptos contrastantes, y su significado lo establece el lugar peculiar que ocupa dentro de esa red.
3. Todo sistema de conceptos es un supuesto especulativo o *teoría:* mínimamente respecto de las clases en que se divide la naturaleza y las relaciones principales que se establecen entre ellas.

Por lo tanto,

4. Todo juicio de percepción presupone una teoría.

De acuerdo con este criterio general, la mente/cerebro es un teorizador frenéticamente activo desde el vamos. El mundo perceptual se presenta en gran medida como una confusión ininteligible para un recién nacido, pero su mente/cerebro se dispone inmediatamente a formular el marco de referencia conceptual dentro del cual comprender, explicar y predecir ese mundo. De este modo sobreviene una secuencia de invenciones, modificaciones y revoluciones conceptuales que terminan por producir algo que se parece aproximadamente a nuestra concepción del mundo de sentido común. La frenética evolución conceptual que experimenta todo niño en sus dos primeros años probablemente nunca es igualada en todo el resto de su vida.

Lo importante de todo esto, para nuestros propósitos, es lo siguiente. En los comienzos de la vida, la mente/cerebro se encuentra en una situación tan confusa e ininteligible como la que encuentra en el mundo externo. Debe disponerse a aprender la estructura y actividades de sus estados internos de la misma manera en que debe disponerse a aprender la estructura y actividades del mundo externo. Con el tiempo, efectivamente aprende algo sobre sí, pero a través de un

proceso de desarrollo conceptual y discriminación aprendida que se da en forma absolutamente paralela al proceso mediante el cual comprende el mundo que está fuera de él. Parecería entonces que el criterio tradicional simplemente está equivocado.

## Lecturas complementarias


Armstrong, David, *A Materialist Theory of the Mind,* Londres, Routledge & Keegan Paul, 1968, capítulo 6, secciones IX y X y capítulo 15, sección II.

Dennett, Daniel, "Toward a Cognitive Theory of Consciousness", en *Minnesota Studies in the Philosophy of Science,* vol, IX, C. W. Savage (comp.), Minneapolis, University of Minnesota Press, 1978. Reproducido en Daniel Dennett, *Brainstorms,* Montgomery, VI, Bradford, 1978, Cambridge, MA, MIT Press.

Nisbett, Richard y Wilson, Timothy, "Telling More Than We Can Know: Verbal Reports on Mental Processes", *Psychological Review,* vol. 84, Nº 3, 1977.

Churchland, Patricia, "Consciousness: The Transmutation of a Concept", *Pacific Philosophical Quarterly,* vol. 64, 1983.

Churchland, Paul, *Scientific Realism and the Plasticity of Mind,* Cambridge, Cambridge University Press, 1979, secciones 13 y 16; para la percepción influida por la teoría, en general, véase cap. 2.

Nagel, Thomas, "What is it like to Be a Bat?", *Philosophical Review,* vol. LXXXIII, 1974. Reproducido en *Readings in Philosophy of Psychology,* vol. I, N. Block (comp.), Cambridge, MA, Harvard University Press, 1980.

# 5

# El problema metodológico

Es claro que el marco conceptual conocido de la psicología popular nos brinda una comprensión singular de muchos aspectos de la mentalidad humana. Sin embargo, también son evidentes los numerosos aspectos de la inteligencia consciente que aquél deja ampliamente en la oscuridad: aprendizaje, memoria, uso del lenguaje, diferencias de inteligencia, sueño, coordinación motriz, percepción, locura y otros. Entendemos demasiado poco de lo que hay que entender, y a la ciencia le corresponde quitar las sombras envolventes y revelarnos la naturaleza interior y el funcionamiento secreto de la mente.

Todos pueden estar de acuerdo sobre esto. Sin embargo, hay gran desacuerdo sobre cómo debe proceder cualquier ciencia de la mente para tener la mayor posibilidad de éxito. Es decir, hay desacuerdo sobre los *métodos* intelectuales que deben emplearse. A continuación veremos una breve descripción y análisis de las cuatro metodologías más influyentes que han guiado la investigación sobre la mente en este siglo.

## 1. Idealismo y fenomenología

Es útil retroceder aquí unos pasos y hacer un poco de historia. Mientras de la Mettrie (véase pág. 148) intentaba reducir la mente a la materia, otros pensadores se ocupaban de realizar una reducción precisamente en el sentido contrario. El obispo George Berkeley (1685-1753) afirmó que los

objetos materiales no existen sino como los "objetos" o "contenidos" de los estados de percepción de mentes conscientes. Para decirlo crudamente, el mundo material no es otra cosa que un *sueño* coherente. Si se sostiene que el mundo material es simplemente el sueño de uno, entonces uno es un *idealista subjetivo*. Si se sostiene, como Berkeley, que el mundo material es el sueño de Dios, un sueño que todos compartimos, entonces uno es un *idealista objetivo*. En ambos casos, el elemento fundamental de la existencia es la mente, no la materia. De allí el término "idealismo".

Esta es una hipótesis sorprendente y fascinante. Se nos pide que pensemos que el mundo material "objetivo" es nada más que el "sensorio de Dios": el mundo material está en la mente de Dios en la misma relación que nuestra experiencia sensorial está en nuestra propia mente. De alguna forma todos somos espectadores del sueño de Dios: el universo físico. Esta hipótesis puede parecer a algunos un sueño vagamente loco por derecho propio, pero al menos podemos imaginar cómo podrían demostrarlo pruebas serias. Supongamos que pudiéramos proporcionar *explicaciones* detalladas de la conducta y constitución de la materia, explicaciones basadas en supuestos teóricos acerca de la constitución de la mente (la nuestra quizás, o la de Dios). Entonces el idealismo comenzaría a parecer auténticamente admisible.

De hecho, nunca se han dado explicaciones de esta clase genuinamente útiles, y entonces el idealismo sigue siendo comparativamente inadmisible. Hay muchas más explicaciones en el otro sentido: de diversos fenómenos mentales en términos de fenómenos físicos. Pensemos solamente en la teoría de la evolución, la inteligencia artificial y las neurociencias para ver la extensión del frente materialista en avance. (Estas se estudiarán en detalle en los capítulos 6-8.)

Sin embargo, hubo un tiempo en que tales explicaciones idealistas del mundo materialista parecían efectivamente existir. Immanuel Kant (1724-1804) dejó una marca duradera en la filosofía occidental cuando afirmó, en la *Crítica de la razón pura*, que la experiencia humana conocida del mundo material es en gran medida *construida* por la mente humana en actividad. Según Kant, las formas innatas de la percepción

humana y las categorías innatas de la comprensión humana imponen un orden invariable sobre el caos inicial de la entrada de estímulos sensoriales sin elaborar. Entonces, todos los seres humanos comparten una experiencia de un mundo empírico muy específico. De este modo Kant intentó explicar por qué las leyes de la geometría euclidiana y de la física de Newton eran necesariamente verdaderas para dar cuenta del mundo de la experiencia humana. Kant pensaba que eran una consecuencia ineludible de la propia actividad estructuradora de la mente.

Desde entonces tanto la geometría euclidiana como la física de Newton han resultado ser empíricamente falsas, lo que sin duda debilita los detalles de la versión de Kant. Pero la idea central de Kant —que las formas y categorías generales de nuestra experiencia perceptual son impuestas por una mente activa y estructuradora— es una idea que perdura. Los objetos materiales en nuestra experiencia construida pueden ser entonces empíricamente reales (= reales para toda la experiencia humana), pero no necesariamente son trascendentalmente reales (= reales desde un posible punto de vista de Dios).

Esta degradación de la materia, a la categoría principal en un mundo de apariencias, es característica de gran parte de la filosofía a partir de Kant. Sin embargo, éste agregó un segundo elemento a la historia, que la transforma a partir de un guión puramente idealista y señala a Kant como un idealista muy atípico. Según Kant, el mundo de los sentidos internos, el mundo de las sensaciones, pensamientos y emociones, *también* es un "mundo construido". Del mismo modo en que la mente accede al mundo "externo", el acceso de la mente a sí misma también está mediatizado por sus propias contribuciones estructurales y conceptuales. Tiene acceso a sí misma solamente a través de sus propias autorepresentaciones. Por lo tanto, aunque es empíricamente real, la mente no necesita ser trascendentalmente real así como tampoco la materia lo necesita. Para Kant, la naturaleza trascendental de la mente en sí misma es tan opaca como la naturaleza trascendental de la materia en sí misma. Y, en general, creía que las cosas así como son (independientes de la percepción y conceptualización humanas) son incognoscibles por los seres humanos para siempre.

Los filósofos posteriores han sido más optimistas que Kant con respecto a las probabilidades últimas de la autocomprensión. Muchos suponen que mediante la investigación científica la mente puede realizar un *progreso* conceptual: hacia la meta de volver a pensar el mundo material y la mente, en términos conceptuales que efectivamente corresponden a la verdadera naturaleza de las cosas en sí mismas. Esta es la esperanza del *realismo científico*, un enfoque filosófico que está en la base de la mayoría de las investigaciones psicológicas y neurocientíficas actuales. Aunque la tradición *fenomenológica* también es optimista en cuanto a la autocomprensión, asume una postura curiosamente diferente.

*Fenomenología* es el nombre de una tradición filosófica centrada en Europa continental. Con raíces en la filosofía kantiana, es un árbol con muchas ramas, pero sus diversos defensores coinciden en que una verdadera comprensión de la naturaleza de la mente sólo puede lograrse con métodos completamente diferentes de aquellos que guían a la ciencia en general. Las razones de esta fuerte posición derivan en parte de la teoría del conocimiento (la epistemología) empleada por los fenomenólogos. Saben muy bien, como casi todos los filósofos desde Kant, que el mundo de nuestra experiencia es en gran medida un mundo construido. Nuestras formas innatas de percepción, nuestras formas innatas de comprensión y nuestros marcos conceptuales aprendidos estructuran en su conjunto el conocido mundo de las percepciones del sentido común: el *Lebenswelt* o *mundo de la vida*.

Según ellos, la actividad científica corriente es simplemente una continuación de algunas de estas actividades "constructivas" de la mente. Construimos conceptos cada vez más intrincados y más profundamente interpretativos del mundo objetivo, y los hacemos responder a los hechos de percepción de nuestro *Lebenswelt* en forma de predicciones, explicaciones y demás.

Pero, insisten los fenomenólogos, tal procedimiento constructivo no es el modo de lograr una verdadera comprensión de la *mente*, la *autora* de toda esta actividad constructiva. Dicho procedimiento simplemente aleja cada vez más a la mente de los fenómenos "puros" originales, y la envuelve cada más

estrechamente en las complejidades de su propia construcción. Los conceptos de la ciencia física nunca pueden ser otra cosa que la interpretación que la mente construyó sobre el mundo "objetivo". En contraste, para entender a la *mente*, lo que debemos hacer es un giro de ciento ochenta grados y adoptar un procedimiento de análisis y desinterpretación de nuestra experiencia. Dicha metodología buscará el origen y revelará la actividad estructuradora de la mente y así nos llevará a la naturaleza esencial de la mente misma. Es posible que la mente intuya su propia naturaleza esencial ya que, en contraste con su conocimiento del mundo objetivo, la mente tiene o puede aspirar a tener acceso directo y sin intermediarios a sí misma. Tal programa de investigación analítica e introspectiva producirá un nivel de conocimiento y comprensión que no sólo es superior a cualquier comprensión posible que pueda producirse por los procedimientos esencialmente constructivos e interpretativos de la ciencia corriente, sino que también es independiente de ella.

Más allá de compartir algo como la perspectiva que acabamos de mencionar, hay gran variedad de fenomenólogos. Georg Hegel (1770-1831), una de las primeras personalidades en la tradición, presentó una nueva versión del idealismo objetivo. Creía que el viaje del espíritu hacia el autoconocimiento último es un viaje hacia la disolución de la distinción entre el yo subjetivo y el mundo objetivo. El avance histórico de la conciencia humana, individual y colectiva, es sólo el proceso lento y disperso por el cual el aún vacilante Espíritu absoluto (= Dios = Universo) aspira a llegar al *auto*conocimiento. Cada "conciencia" humana individual es sólo un aspecto de aquel Espíritu mayor y la oposición entre uno mismo y los demás y entre uno mismo y el mundo objetivo, finalmente se derrumbará cuando el Espíritu absoluto logre el completo autorreconocimiento. Mientras tanto, nuestro *Lebenswelt* se interpreta mejor no como el *sueño* apacible del Espíritu absoluto, sino como el contenido de sus esforzados intentos por lograr el autoconocimiento.

Sin embargo, Hegel no es representativo de la tradición más moderna y la fenomenología no está comprometida esencialmente con una ontología idealista. Edmund Husserl

(1859-1938) es la figura central de la tradición moderna. Husserl llevó a cabo su investigación fenomenológica dentro de un marco aproximadamente cartesiano, en el que la mente y la materia son igualmente reales y su principal interés era entender la *intencionalidad* de los estados mentales (véase cap. 3.4). La búsqueda introspectiva de las actividades constructivas de la mente, afirmaba, revela la fuente de nuestros "contenidos" mentales y nos conduce a un conocimiento purificado e incuestionable de un yo trascendental individual, detrás del yo empírico o fenomenológico. Creía que aquí se pueden analizar los cimientos indiscutibles de la experiencia humana y de todas las ciencias empíricas objetivas.

Esta breve reseña no hace justicia a lo que es una rica tradición y ninguna tradición de esta magnitud puede refutarse en un párrafo. Sin embargo, el lector verá que lo que en el último capítulo denominamos "el enfoque tradicional" con respecto a la introspección es de un modo u otro una parte importante de la tradición fenomenológica. La idea de que se puede tener algún conocimiento supracientífico del yo, alguna forma especial de conocimiento que no sea a través del medio de la conceptualización constructiva objetivizante, es común a lo largo de la historia. Ese enfoque va contra la convicción de Kant de que el autoconocimiento introspectivo es una instancia de "construcción" objetivizante tan inevitable como lo ese el propio conocimiento del mundo externo. Y va contra la demostración psicológica moderna de que los propios juicios introspectivos son completamente análogos a los juicios de percepción en general, y proporcionan un conocimiento que de ninguna manera se distingue por una posición, pureza o autoridad especiales.

Si *todo* conocimiento es inevitablemente un asunto de construcción conceptual e interpretación especulativa (recuérdese la conclusión del capítulo 4.2), entonces parecería que el "acceso especial" a la "naturaleza esencial" de la mente buscado por los fenomenólogos es sólo un sueño, y que los métodos corrientes de la ciencia empírica constituyen la única esperanza que tiene la mente de comprenderse a sí misma. Esto no significa que no deban admitirse los juicios introspectivos como datos para la ciencia, y tampoco la "investigación

fenomenológica", sino que se negará a los resultados de esa investigación cualquier estatuto epistemológico especial o único.

Sin embargo, volver a "los métodos corrientes de la ciencia empírica" no produce unanimidad instantánea pues hay varios conceptos en pugna de lo que son o deberían ser aquellos "métodos corrientes", como se verá en las siguientes secciones.

### Lecturas complementarias


Marx, Werner, *Hegel's Phenomenology of Spirit*, Nueva York Harper and Row, 1975.

Spiegelberg, Herbert, *The Phenomenological Movement*, vols. I y II, The Hague, Harper and Row, 1960; véase especialmente el análisis de Edmund Husserl, vol, I, págs. 73-167.

Dreyfus, Hubert L., (comp.) *Husserl, Intentionality, and Cognitive Science*, Cambridge, MA, MIT Press/Bradford, 1982.

Smith, D. W. y Mc Intyre, R., *Husserl and Intentionality*, Boston, Reidel, 1982.

Piaget, Jean, *Insights and Illusions of Philosophy*, Nueva York World Publishing Co., 1971, cap. 3, "The False Ideal of a Suprascientific Knowledge".


## 2. Conductismo metodológico

El *conductismo metodológico* representa una reacción muy fuerte contra las tendencias dualistas e introspectivas de la psicología que lo precedieron. Hijo de este siglo, es también un intento autoconsciente de reconstruir la ciencia de la psicología en los términos de las ciencias que han tenido tanto éxito, como la física, la química y la biología. Durante los últimos cincuenta años, el conductismo ha sido la único escuela de psicología más influyente en el mundo angloparlante. Las dos últimas décadas han obligado a la reevaluación y flexibilización de algunas de sus doctrinas, pero sigue ejerciendo una importante influencia.

### Tesis y argumentos centrales

Sus principios centrales no son difíciles de entender. Según el conductismo, la obligación primera y principal de la

ciencia de la psicología es *explicar la conducta* de cualquier criatura que estudie, incluyendo a los seres humanos. Por "conducta" los conductistas entienden la actividad por todos observable, medible, registrable de los sujetos estudiados: movimientos corporales, ruidos emitidos, cambios de temperatura, sustancias químicas liberadas, interacciones con el medio y demás. No hay duda sobre la realidad objetiva de estos fenómenos, y la psicología no puede extraviarse eligiendo la *conducta* animal como principal objetivo para explicar. Esto contrasta profundamente con las ideas anteriores, que tomaban los elementos y contenidos de la *conciencia* interna como objetivo adecuado para que la psicología explicara.

Sin embargo, para la mayoría de los conductistas tenía una importancia similar el *modo* en que la conducta debe explicarse correctamente. Las explicaciones de sentido común que recurren a los "estados mentales" se consideran seriamente defectuosas de varias maneras. Dichas explicaciones recurren a un conjunto de conocimientos populares que no tienen bases científicas correctas y que pueden consistir en gran parte en superstición y confusión, así como ocurre con tantos de nuestros conceptos anteriores. Las conocidas nociones mentalistas están mal definidas y no tienen criterios objetivos claros para su aplicación, especialmente en el caso de los animales no humanos; la introspección individual no proporciona una base uniforme ni confiable para su aplicación incluso en el caso de los seres humanos; las explicaciones mentalistas generalmente se construyen a posteriori y los principios invocados demuestran muy poco poder de predicción; y tales explicaciones "íntimas" nos ocultan el amplio papel que desempeña el medio externo de cualquier organismo en el control de su conducta.

En lugar de recurrir a los estados mentales, los conductistas propusieron explicar la conducta de cualquier organismo en términos de sus circunstancias ambientales privativas. O bien, en términos del medio más ciertos rasgos observables del organismo. O, si esto fracasa, también en términos de ciertos rasgos *no* observables del organismo —disposiciones y reflejos innatos y condicionados— donde dichos rasgos cumplen una condición muy estricta: deben ser tales que su

presencia o ausencia siempre pueda determinarse decisivamente mediante un test de conducta, así como la solubilidad de un terrón de azúcar se revela al disolverse realmente (conducta) cuando se lo coloca en agua (circunstancia ambiental). En resumen, las explicaciones en psicología deben basarse totalmente en nociones que sean directamente observables por todos o definidas operacionalmente en términos de conceptos así observables. (Véase el capítulo 2.2 para el concepto de definición operacional.)

Los conductistas quieren (o querían) limitarse a estos recursos e impulsar a los demás a observar las mismas restricciones porque creían que éstas eran el precio inevitable de convertir a la psicología en una verdadera ciencia. Dejar de lado la antigua estructura conceptual del sentido común parecía un bajo precio a pagar por una meta tan valiosa. Se creía que si aquellos conceptos mentalistas realmente tienen integridad, entonces la metodología conductista finalmente nos conducirá nuevamente a ellos, o a versiones de ellos adecuadamente definidas. Y si no tienen integridad explicativa entonces rechazarlos no constituye una verdadera pérdida.

Más aún, una opinión influyente en un campo relacionado dio un apoyo accidental a los conductistas. Una escuela de filosofía con mentalidad científica llamada "positivismo lógico" o "empirismo lógico" sostenía la idea de que el significado de cualquier término teórico, en cualquier ciencia, derivaba esencialmente de sus conexiones definitorias, aunque tortuosas, con conceptos *observacionales*, que derivan su significado directamente de la experiencia sensorial. Algunos filósofos de la ciencia relacionados con esta escuela afirmaron específicamente que cualquier término teórico significativo debía poseer una definición *operacional* en términos observables. Por lo tanto, el conductismo sólo parecía seguir las reglas que se decía gobernaban a la ciencia legítima en general.

## Críticas al conductismo

Al adoptar una franca actitud escéptica frente a la ontología de los estados mentales y frente a nuestra conocida

concepción de las causas de la conducta humana, los conductistas provocaron una fuerte reacción negativa por parte de gran cantidad de moralistas, clérigos, novelistas y otras escuelas de filosofía y psicología. La queja principal era que el conductismo tendía a deshumanizar a los humanos al descartar arbitrariamente de la corte científica la característica misma que nos hace especiales: una vida mental consciente. En cierta medida, esta queja es una petición de principio. Si los seres humanos somos "especiales", y si lo somos, qué características hacen que así sea, éstas son preguntas científicas que requieren respuestas científicas. Quizás estamos errados en nuestras creencias de sentido común referidas a la cuestión de si en realidad somos especiales y por qué lo somos. (No sería la primera vez: recordemos la convicción universal de que la humanidad está en el centro del universo físico.) Y no es una crítica grave del conductismo sólo repetir tozudamente nuestras convicciones culturalmente arraigadas.

Aun así, ahora se acepta ampliamente que el conductismo fue demasiado lejos en sus afirmaciones y restricciones iniciales, más de lo necesario para asegurar el estatuto científico de la psicología. Por una parte, la idea positivista de que todo término teórico significativo debe admitir una definición operacional en términos de elementos observables se reconoció rápidamente como un error. Por ejemplo, la mayoría de los términos de la física teórica tienen por lo menos algunas conexiones distantes con elementos observables, pero no de la clase simple que permitiría *definiciones* operacionales en términos de aquellos elementos observables. Tratemos de dar una definición así para "*x* es un neutrino" o "*x* tiene un electrón en su capa orbital más baja". Los condicionales adecuados que conectan tales términos con elementos observables siempre terminan requiriendo el uso de muchos *otros* términos teóricos, y por lo tanto la definición no es puramente "operacional". Entonces si a continuación debe hacerse una restricción en favor de las definiciones operativas, ¡la mayor parte de la *física teórica* debería dejarse de lado por ser una seudociencia carente de significado!

Las ideas actuales sobre el significado tienden a revertir totalmente la idea positivista: el significado de cualquier tér-

mino, incluyendo los términos de observación, está fijado por su lugar en el sistema de creencias en el que aparece. (La teoría reticular del significado se analizó en el capítulo 3.3.) Por lo tanto, nuestro vocabulario mentalista no puede eliminarse de la ciencia sólo por principios abstractos absolutos. De ser así, deberá ser desechado debido a sus deficiencias explicativas y predictivas en relación con las teorías rivales sobre la naturaleza humana.

Tampoco parece científicamente razonable negar o simplemente ignorar la existencia de fenómenos internos, a los que por lo menos tenemos cierto acceso introspectivo, aunque confuso, y que desempeñan por lo menos algún papel, aunque malentendido, en los orígenes causales de la conducta. El conductismo, al impulsarnos a ignorar totalmente dichos fenómenos y a tratar a los seres humanos como "cajas negras" con reflejos no explicados en términos de las estructuras y actividades internas de la caja, fue demasiado lejos. Fue innecesariamente restrictivo y culpable de una reacción exagerada frente a excesos anteriores.

Admitida la justicia de estas críticas, la mayoría de los pensadores se inclinaron simplemente a olvidarse del conductismo. Pero ésa no es la reacción correcta. Las versiones y los defensores actuales del conductismo están dispuestos a reconocer las críticas mencionadas. Pero algunos elementos importantes del conductismo sobreviven y aún puede demostrarse que son correctos.

Uno de los defensores más famosos del conductismo durante años, B. F. Skinner, de la Universidad de Harvard, presentó recientemente una versión del conductismo en la que sostiene la realidad de los fenómenos internos, así como también nuestro acceso introspectivo a ellos, y en la que se asigna a los fenómenos internos un papel perfectamente legítimo en psicología. A pesar de estas concesiones, Skinner insiste en tres afirmaciones importantes. Primero, lo que "espectamos" cuando hacemos introspección es simplemente el estado fisiológico de nuestro cuerpo y sistema nervioso, no cualquier realidad "no física". Segundo, la introspección permite acceder sólo a una porción muy pequeña de nuestros estados y actividades interiores, y es confusa e incierta incluso allí. Y tercero,

los estados que discriminamos en la introspección, aunque están correlacionados con nuestra conducta, no son necesariamente sus causas reales.

Podemos iniciar el trabajo de aislar las causas reales (internas) de nuestra conducta analizando nuevamente los factores del medio que la controlan y luego, buscando los efectos causales de aquellos factores hacia adentro. El papel del *medio* en el control de la conducta sigue siendo un tema central de esta tendencia, y la razón fundamental no es difícil de hallar. Actualmente todas las especies vivientes deben su supervivencia al hecho de que sus ejemplares, más fiablemente que otros, respondieron adecuadamente a sus medios. La psicología humana, o la de cualquier otra especie, es el resultado de una prolongada adaptación evolutiva de conductas controladas por el medio: por ejemplo, "Come todo lo que huela bien", "Lucha contra (o huye de) todo lo que te ataque", "Aparéate con lo que luzca bien", etcétera. ¿Dónde comenzaría la psicología sino por el estudio sistemático de tales controles?

Como veremos, *hay* otros lugares interesantes por donde la psicología puede comenzar. Pero el programa de investigación conductista sigue siendo una buena opción y sería un error rechazar sus versiones más recientes.

**Lecturas complementarias**


Skinner, B. F., *About Behaviorism*, Nueva York, Random House, 1974.

Dennett, Daniel, "Skiner-Skinned" en *Brainstorms*, Montogmery, VT, Bradford, 1978, Cambridge, MA, MIT Press.

Chomsky, Noam, "A Review of B. F. Skinner's *Verbal Behavior*", *Language*, vol. 35, nº 1, 1959. Reproducido en *Readigns in philosophy of Psychology*, vol. I, N. Block, (comp.), Cambridge, MA, Harvard University Press, 1980.


## 3. Enfoque cognitivo/computacional

Dentro del amplio marco de la concepción funcionalista de la mente analizado en el capítulo 2.4, hallamos dos programas de investigación estrechamente vinculados que apuntan

a resolver el misterio de la inteligencia consciente: la *psicología cognitiva* y la *inteligencia artificial*. Ambos enfoques contrastan con las formas tradicionales de conductismo, ya que ambos se sienten libres para postular o atribuir un complejísimo sistema de estados internos a criaturas inteligentes para explicar su conducta. En general, los estados postulados son, de un modo u otro, estados "portadores de información", y sus interacciones colectivas son una función de la información específica que transmiten. De allí la caracterización general: "enfoque del procesamiento de la información" o simplemente "enfoque computacional".

Comparemos el caso de una calculadora de bolsillo. Sus diversos estados de entrada representan números específicos y operaciones aritméticas y las actividades internas posteriores están determinadas por las características computacionalmente pertinentes de dichos estados. Finalmente, los estados de salida se relacionan sistemáticamente y mediante reglas con aquellos estados de entrada. Se supone que lo mismo sucede con los organismos que demuestran inteligencia natural, excepto que sus estados de entrada representan muchas más cosas que sólo números, y los "cálculos" que realizan comprenden muchas más cosas que meras relaciones aritméticas. También incluyen relaciones lógicas, por ejemplo, formas espaciales, relaciones sociales, estructuras lingüísticas, color, movimiento, etcétera. (En el próximo capítulo se analizarán estos ejemplos.)

La meta de la psicología cognitiva es explicar las diversas actividades que constituyen la inteligencia —percepción, memoria, inferencia, deliberación, aprendizaje, uso del lenguaje, control motriz y otras—, postulando un sistema de estados internos regidos por procedimientos computacionales, o un conjunto de tales procedimientos en interacción regidos por un conjunto de esos procedimientos. La meta es armar una descripción de la verdadera organización *funcional* del sistema nervioso humano o del sistema nervioso de cualquier criatura que esté en estudio.

Es un proyecto ambicioso, dada la extraordinaria complejidad de las criaturas inteligentes, y casi siempre se adopta

un enfoque gradual. Por ejemplo, un teórico puede concentrar su atención en la percepción o en el uso del lenguaje y luego tratar de armar un sistema computacional que explique las actividades específicas de aquella facultad solamente. Estos éxitos graduales pueden reunirse, a medida que se van produciendo, para lograr una explicación general de la inteligencia del organismo.

Tres criterios son importantes al formular y evaluar estas hipótesis computacionales. Primero, el sistema computacional propuesto debe servir para explicar las entradas y salidas de la facultad cognitiva en estudio. Por ejemplo, si la facultad es la percepción, entonces el sistema computacional propuesto debe explicar las discriminaciones que hace realmente el individuo, dado el estímulo físico de sus órganos sensoriales. Si la facultad es el uso del lenguaje, entonces el sistema debe explicar la discriminación entre las oraciones gramaticales y las que no tienen sentido y nuestra capacidad de producir casi exclusivamente oraciones gramaticales. En términos generales, el sistema propuesto debe hacer lo que el individuo estudiado logra hacer, o lo que hace su facultad elegida.

El primer criterio es importante, pero es demasiado burdo como para ser adecuado por sí solo. El problema es que hay muchas maneras diferentes de despellejar un gato. Para cualquier relación deseada entre entradas y salidas hay infinitos procedimientos computacionales *diferentes* que producirán exactamente esa relación.

Un ejemplo elemental ilustra fácilmente el punto. Supongamos que tenemos un pequeño aparato estilo calculadora que se comporta del siguiente modo. Para cualquier número $n$ que se marca, muestra el número igual a $2n$. Un modo en que puede calcular sus respuestas es simplemente multiplicando el número entrado por 2. Una segunda manera sería multiplicando la entrada por 6 y luego dividir ese resultado por 3. Un tercer modo sería dividiendo por 10 y multiplicando el resultado por 20. Y así sucesivamente. Cualquiera de estos procedimientos computacionales producirá la misma "conducta manifiesta", mientras se trate de la duplicación de números arbitrarios. Pero seguramente la calculadora utiliza sólo uno de ellos. ¿Cómo podemos determinar cuál?

Aquí entra el segundo criterio para evaluar hipótesis computacionales. Los procedimientos que producen la "misma conducta" en un nivel de análisis pueden mostrar leves diferencias en un nivel más fino de análisis. Por ejemplo, los dos procedimientos segundos incluyen dos operaciones diferentes, mientras que el original sólo una. Así, en igualdad de todas las otras condiciones, podemos esperar que aquellos dos procedimientos tarden más para completar el cálculo. Entonces una cuidadosa medición de los tiempos utilizados revelaría cuál de las dos calculadoras utilizó el procedimiento más simple. Más aún, los patrones de error también pueden ayudarnos a diferenciar las hipótesis. Si cada operación computacional tiene una probabilidad de error pequeña pero finita en cada paso, entonces los dos procedimientos segundos cometerán errores con mayor frecuencia que el procedimiento más simple. Entonces una prueba de largo plazo ayudará a diferenciar un procedimiento del otro. La naturaleza específica de los errores cometidos también puede decirnos muchos sobre los procedimientos que los produjeron.

El tercer criterio para evaluar hipótesis computacionales es obvio, tanto para máquinas como para organismos biológicos: los procedimientos computacionales propuestos deben ser coherentes con las capacidades físicas de los circuitos o del sistema nervioso del individuo. Una hipótesis aceptable debe coincidir con los "equipos" o "wetware" que realmente efectúan la actividad computacional analizada.

Este tercer criterio generalmente es muy difícil de aplicar, excepto en un nivel muy superficial, porque la maquinaria neural que constituye un sistema nervioso avanzado es muy pequeña en sus elementos, muy intrincada en sus conexiones y muy vasta en su extensión. El desciframiento del sistema nervioso, como veremos en el capítulo 7, no es una tarea improvisada. Como resultado, este tercer criterio ejerce una influencia más débil que los otros dos sobre gran parte de la teorización en psicología cognitiva. Y quizás así deba esperarse que sea: en el caso de la mayoría de las funciones cognitivas aún no tenemos el problema de elegir entre hipótesis computacionales igualmente adecuadas. Todavía estamos tratando de elaborar aunque sea *una* hipótesis que sea com-

pletamente adecuada para la actividad en estudio. Aun así, los criterios segundo y tercero son los que permiten que la psicología cognitiva siga siendo una ciencia empírica legítima, una ciencia que se ocupa de la pregunta de cómo se produce realmente la inteligencia natural.

En contraste, el programa de investigación de la inteligencia *artificial* ha prescindido de todos menos del primer criterio. La meta de este programa es simplemente diseñar sistemas computacionales capaces de todas y cada una de las conductas inteligentes observadas en los organismos naturales. En el mejor de los casos generalmente siempre ha sido de interés secundario el hecho de que los sistemas propuestos utilicen los *mismos* procedimientos computacionales utilizados por cualquier organismo natural dado.

Hay algunas razones de peso para continuar este enfoque alternativo de la inteligencia. Por una parte, no hay razón para creer que los procedimientos computacionales utilizados por los organismos naturales deban ser los mejores posibles para lograr los fines pertinentes. Nuestra historia evolutiva y nuestra maquinaria biológica casi con seguridad colocan límites significativos y probablemente arbitrarios a las clases de procedimientos que podemos utilizar. Por ejemplo, las máquinas de calcular electrónicas de alta velocidad pueden efectuar rutinas imposibles para nuestros sistemas nerviosos. Y se argumenta que en todo caso debemos estudiar no sólo la inteligencia viva, sino también todas las dimensiones de la inteligencia en general. Más aún, los adelantos en este último campo probablemente ayudarán a la comprensión de la inteligencia natural pura.

El contraste entre ambos enfoques es evidente, pero en la práctica tiende a desaparecer. Una manera de probar una hipótesis sobre las actividades de procesamiento de datos de un individuo dado es redactar un programa que realice los cálculos pertinentes, pasarlo por el ordenador y comparar la conducta de salida con la del individuo. Aquí la tarea de la psicología cognitiva será similar a la de la inteligencia artificial. Por otro lado, el investigador de la inteligencia artificial no debe sentir remordimientos por estudiar la conducta y los informes introspectivos de los individuos reales para estimu-

lar la invención de programas ingeniosos. Aquí la tarea de la inteligencia artificial será similar a la psicología cognitiva.

En el próximo capítulo estudiaremos más detalladamente la inteligencia artificial. Permítaseme cerrar este apartado analizando una objeción a las dos estrategias de investigación descriptas. Quizás el lector haya descubierto que, según el enfoque computacional, la inteligencia consciente no posee una sola esencia unificadora o una naturaleza única simple. En lugar de ello, las criaturas inteligentes están representadas como una caja de sorpresas de muy variados procedimientos computacionales vagamente interconectados y no del modo en que una vez un colega describió a mi primer automóvil como "un escuadrón de tuercas y tornillos volando en formación libre".

Casualmente, aquella descripción de mi automóvil era acertada y el concepto de inteligencia presentado por el enfoque computacional puede serlo también. El lento acrecentamiento de sistemas de control semiaislados tiene sentido evolutivo. Los sistemas nerviosos evolucionaron de a poquito ya que las adiciones accidentales y ocasionales eran seleccionadas por brindar un control ventajoso sobre algún aspecto de la conducta o las operaciones internas de la criatura. Es probable que la selección natural en el largo plazo permita que las criaturas sobrevivientes mantengan una interacción fluida con el medio, pero los mecanismos internos que mantienen esa interacción pueden ser arbitrarios, oportunistas y provisionales. Por lo tanto no es una crítica del enfoque computacional decir que los representa de ese modo.

## Lecturas complementarias


Dennett, Daniel, "Artificial Intelligence as Philosophy and as Psychology" en *Brainstorms*, Montgomery, VT, Bradford, 1978; Cambridge, MA, MIT Press.

Johnson-Laird, P. N. y Wason, P. C., *Thinking: Readings in Cognitive Science*, Cambridge, Cambridge University Press, 1977.

Anderson, J. R., *Cognitive Psychology and its Implications*, San Francisco, Freeman, 1980.

Boden, Margaret, *Artificial Intelligence and Natural Man*, Nueva York, Harvester Press, 1977.

Pylyshyn, Zenon, "Computation and Cognition", *The Behavioral and Brain Sciences*, vol. 3, 1980.


Véanse también las lecturas recomendadas a lo largo del capítulo 6.

## 4. Materialismo metodológico

La metodología descripta en el apartado anterior se denomina comúnmente "enfoque de arriba hacia abajo" porque se comienza con la comprensión corriente de lo que hacen las criaturas inteligentes y luego se pregunta qué clase de operaciones subyacentes podrían producir o explicar esas actividades cognitivas. Por otro lado, la metodología descripta en este apartado comienza por el extremo opuesto del espectro y se denomina "enfoque de abajo hacia arriba". La idea básica es que las actividades cognitivas son esencialmente sólo actividades del sistema nervioso; y si queremos entenderlas, entonces la mejor manera de hacerlo es analizar el sistema nervioso mismo, descubriendo la estructura y conducta de sus elementos más pequeños, sus interconexiones e interactividad, su desarrollo en el tiempo y su control colectivo de la conducta.

Esta es la metodología que guía las diversas disciplinas agrupadas bajo el término *neurociencia* y es esencialmente el mismo espíritu que nos lleva a quitar la tapa de atrás del reloj despertador y desarmarlo para ver por qué funciona. Este enfoque de la conducta inteligente tiene una larga historia. El griego Hipócrates sabía que el deterioro del cerebro acaba con la cordura y el médico romano Galeno ya había descubierto la existencia de y la diferencia entre el sistema nervioso somatosensorial (el conjunto de fibras que conducen la información del "tacto" al cerebro) y el sistema nervioso motor (el conjunto de fibras que parten del cerebro y de la médula espinal controlando los músculos del cuerpo). La disección de animales muertos permitió descubrirlos y Galeno advirtió que lesiones o cortes localizados en los dos sistemas de animales vivos producían "ceguera" táctil en el primer caso y parálisis localizada en el segundo.

El progreso sistemático del conocimiento de la estructura y funcionamiento del sistema nervioso tuvo que esperar hasta siglos más recientes, ya que las autoridades religiosas no veían con beneplácito o directamente prohibieron la disección post mortem del cuerpo humano. Aun así, la mayor parte de la

anatomía del sistema nervioso ya era más o menos comprendida a fines del siglo XVII. Sin embargo, esto permitió sólo un conocimiento limitado del funcionamiento, y el progreso real sobre la microestructura y la microactividad del cerebro tuvo que esperar el desarrollo de técnicas microscópicas modernas, el desarrollo de la teoría química y eléctrica y el desarrollo de los modernos instrumentos electrónicos de medición y registro. Como resultado, los descubrimientos más importantes se hicieron en este siglo.

La arquitectura neuronal revelada por estos métodos es sorprendentemente compleja. Los átomos funcionales del cerebro son diminutas células procesadoras de impulsos denominadas *neuronas* y hay aproximadamente $10^{11}$ (la unidad seguida de 11 ceros: 100 mil millones) neuronas en un solo cerebro humano. Para tener idea de este número, imaginemos una pequeña casa de dos pisos llena de arena gruesa desde el sótano hasta las alfardas. Hay tantas neuronas en el cerebro como granos de arena en aquella casa. Más curioso aún, la neurona promedio mantiene, a través de diminutas fibras que parten de ella denominadas *dendritas* y *axones*, cerca de 3000 conexiones con otras neuronas, por lo que la interconexión de todo el sistema es realmente extraordinaria: unas $10^{14}$ o 100 billones de conexiones.

Tal complejidad supera toda capacidad de comprensión y recién hemos comenzado a descifrarla. Las consideraciones éticas por supuesto impiden la libre experimentación con personas vivas, pero la naturaleza es lo suficientemente cruel como para realizar sus propios experimentos, y los neurólogos normalmente ven una cantidad de cerebros con diversos tipos de lesiones, víctimas de anormalidades químicas, físicas o degenerativas. En tales casos se puede aprender mucho de la cirugía o de análisis post mortem. Las criaturas con sistemas nerviosos muy simples proporcionan otra vía hacia el conocimiento. El sistema nervioso de una babosa de mar, por ejemplo, contiene sólo cerca de 10.000 neuronas y ésa es una red que los investigadores ya han dibujado en su totalidad. La historia química de su acostumbramiento a determinados estímulos —un caso primitivo de aprendizaje— también se obtuvo a partir de la microexperimentación. Los conocimientos

logrados en dichos casos nos ayudan a estudiar las actividades neuronales de criaturas más complejas, como langostas, ratas, monos y seres humanos.

La convicción del materialismo metodológico es que si comenzamos a entender la conducta física, química, eléctrica y evolutiva de las neuronas, y especialmente de sistemas de neuronas, y los modos en que ejercen control una sobre otra y sobre la conducta, entonces estaremos encaminados hacia la comprensión de todo lo que hay que saber sobre la inteligencia natural. Es verdad que el enfoque de abajo hacia arriba no se ocupa directamente de los conocidos fenómenos mentalistas identificados según la psicología tradicional, pero el hecho puede considerarse como una virtud del enfoque. Si las gastadas categorías de la psicología popular (creencias, deseo, conciencia y otros) realmente tienen integridad objetiva, entonces el enfoque de abajo hacia arriba finalmente nos llevará de nuevo a ellas. Y si no la tienen, entonces dicho enfoque, al estar tan estrechamente ligado al cerebro empírico, ofrece las mejores posibilidades de elaborar un nuevo conjunto de conceptos más adecuados para comprender la vida interna. Evidentemente esta metodología es la que da expresión más directa a los temas filosóficos desarrollados por los materialistas reduccionistas y eliminadores.

Puede parecer que un enfoque tan despiadadamente materialista degrada o subestima seriamente la verdadera naturaleza de la inteligencia consciente. Pero la respuesta materialista es que tal reacción en sí es la que degrada y subestima seriamente el poder y el virtuosismo del *cerebro* humano, a medida que continúa revelándose a través de la investigación neurocientífica. En el capítulo 7 se analizará en qué consisten esas investigaciones y en qué medida responden a preguntas relacionadas con la inteligencia consciente.

## Lecturas complementarias

Véanse las listas que siguen después de cada apartado en el capítulo 7.

# 6

# Inteligencia artificial

¿Es posible construir y configurar una máquina puramente física de modo tal que posea inteligencia verdadera? La creencia del programa de investigación denominado "inteligencia artificial" ("IA") es que es posible y la meta de este programa es realizarla. En este capítulo se tratará sobre qué abarca el programa y por qué son optimistas sus realizadores. También se analizarán algunos de los problemas que plantea.

Las esperanzadas tentativas hacia la conducta de la inteligencia artificial tienen una larga historia. En la segunda mitad del siglo de Descartes, el matemático y filósofo alemán Gottfried Leibniz construyó una máquina que podía sumar y restar mediante cilindros rotativos interconectados. También afirmaba que existía la posibilidad de lograr un lenguaje perfectamente lógico en el que todo el pensamiento se reduciría a cálculos solamente. No tenía una idea muy clara de este lenguaje pero, como veremos, la idea fue profética.

En el siglo siguiente, un pensador fisiológico llamado Julien de la Mettrie también estaba impresionado por el mecanismo del cuerpo humano y por la idea de que la actividad "vital" no surgía de un principio intrínseco de la materia, ni de alguna sustancia no material, sino de la estructura física y de la *organización* funcional resultante que la materia podía tener. Pero mientras que Descartes no osó llegar a la conclusión que esto indicaba, de la Mettrie siguió adelante. Dijo que no sólo derivan de la organización física de la materia nuestras

actividades "vitales", sino también todas nuestras actividades *mentales*.

Su libro, *El hombre, una máquina,* fue ampliamente difamado pero, una vez liberadas, estas ideas ya no pudieron ser acalladas. El contemporáneo de de la Mettrie, Jacques de Vaucanson, diseñó y construyó varias estatuas muy bonitas que parecían vivas cuyo funcionamiento interno mecánico y neumático producía una variedad de conductas simples. Un pato enchapado en cobre hacía una convincente demostración de beber, comer, graznar y chapotear en el agua. Y se dice que una estatua de figura humana de tamaño real tocaba la flauta de un modo muy creíble. Si bien estos limitados autómatas no impresionarían probablemente en la actualidad, no hay duda de que su repentino funcionamiento dio un duradero susto al ingenuo observador del siglo XVIII.

En el siglo pasado el matemático de Cambridge, Charles Babbage, se ocupó de capacidades *mentales* más específicas, al diseñar cuidadosamente un *motor analítico* capaz de todas las operaciones lógicas y aritméticas elementales, y con sus principios anticipó el ordenador digital moderno. Sin embargo, Babbage aún estaba limitado a máquinas puramente mecánicas y, aunque su detallado diseño sin duda habría funcionado si se hubiera construido, nunca se realizó debido a su gran complejidad mecánica.

La complejidad que involucra toda actividad inteligente constituyó una barrera permanente para la fácil simulación mediante dispositivos mecánicos, una barrera que, desde Babbage, la tecnología tardó un siglo en sortear. Sin embargo, el tiempo transcurrido no fue en vano. En el terreno abstracto se logró un progreso fundamental: en la comprensión de la lógica de las proposiciones, la lógica de clases y la estructura lógica de la geometría, aritmética y álgebra. Llegamos a apreciar el concepto abstracto de un *sistema formal,* del cual son ejemplos los sistemas mencionados. Un sistema formal consiste en: 1) un conjunto de *fórmulas,* y 2) un conjunto de *reglas de transformación* para manejarlas. Las fórmulas se obtienen juntando, según reglas específicas de formación, varios ítems de una reserva básica de *elementos.* Las

reglas de transformación se ocupan de la *estructura formal* de cualquier fórmula dada (= el patrón según el cual se combinan sus elementos), y su función es sólo transformar una fórmula en otra.

En el caso del álgebra elemental, los elementos básicos son los números del 0 al 9, las variables "*a*", "*b*", "*c*",... ,"(", ")", "=", "+", "-", "/" y "x". Las fórmulas son términos, tales como "(12 – 4)/2" o ecuaciones, tales como "$x$ = (12 – 4)/2". Una secuencia de transformaciones podría ser:

$x = (12 - 4)/2$
$x = 8/2$
$x = 4$

Conocemos estas reglas de transformación así como también lo que se puede hacer con ellas. Por lo tanto ya poseemos un control consciente de por lo menos un sistema formal. Y dado que podemos pensar, además tenemos también por lo menos algún control tácito de la lógica general de las proposiciones, que es otro sistema formal.

Hay infinitos sistemas formales posibles, la mayoría de ellos triviales y carentes de interés. Pero muchos de ellos son extraordinariamente poderosos, como lo demostrarán los ejemplos de lógica y matemática. Lo que es más interesante desde el punto de vista de la IA es que en principio cualquier sistema formal puede automatizarse. Es decir, que los elementos y operaciones de cualquier sistema formal siempre son de una clase que un dispositivo físico correctamente construido *podría* formular y manejar por sí mismo. Por supuesto, la construcción de una máquina adecuada puede resultar imposible por razones de escala o tiempo o tecnología. Pero en la segunda mitad de este siglo, los avances en electrónica hicieron posible la construcción del ordenador digital de alta velocidad para todo propósito. Estas máquinas han permitido la automatización de sistemas formales muy poderosos y consecuentemente capaces de formas de cálculo muy poderosas. La barrera que frustró a Babbage ha sido destruida.

# 1. Ordenadores: algunos conceptos elementales

## Hardware (equipos)

El término "hardware" se refiere al ordenador físico mismo y a los dispositivos periféricos, como el teclado para la entrada, los monitores y las impresoras para la salida, y las cintas/disquettes/tambores de memoria "pasiva" o externa para ambos (figura 6.1). Contrasta con el término "software", que significa una secuencia de instrucciones que le dicen al hardware qué hacer.

El ordenador propiamente dicho se compone de dos elementos principales: la *unidad central de procesamiento* (UCP) y la *memoria activa,* que generalmente es del tipo de acceso aleatorio (AA). Esta última expresión significa que los elementos que almacenan información en la memoria están dispuestos en una grilla electrónica, de modo que cada elemento o "registro" tiene una única "dirección" a la que se accede directamente por la unidad central de procesamiento. Esto permite que la UCP encuentre lo que hay en cualquier registro inmediatamente, sin buscar laboriosamente por toda la secuencia de muchos miles de registros para encontrar lo que se necesita. A su vez, la UCP también puede colocar información directamente en un registro específico. Tiene acceso libre y directo a cualquier elemento de una memoria activa de esta clase. De ahí que se llame "memoria de acceso aleatorio" o "MAA". La memoria sirve como "bloc borrador" o "espacio de trabajo" para la UCP y también guarda las instrucciones o el *programa* que ponemos para decirle a la UCP qué hacer específicamente.

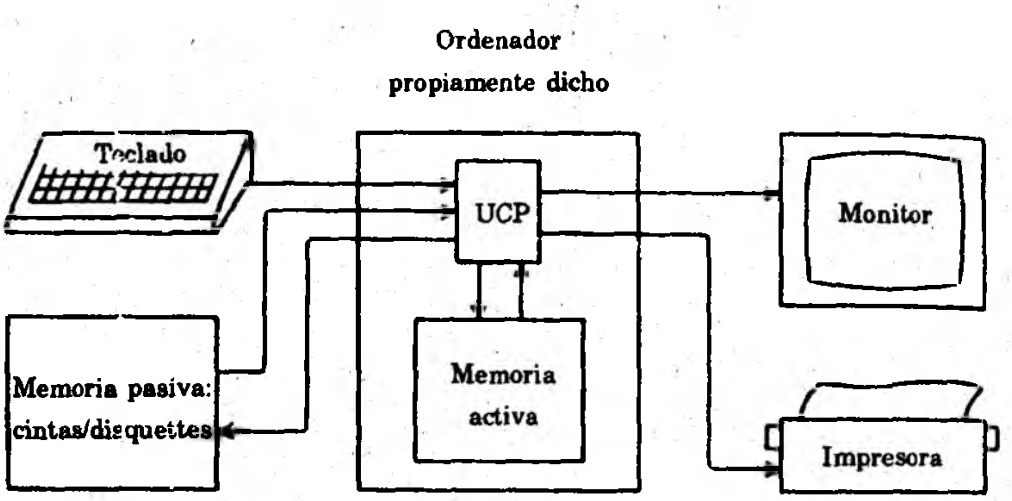



Figura 6. 1

La unidad central de procesamiento es el corazón funcional del sistema. Es la que maneja las diversas fórmulas con que fue alimentada; toma y ejecuta las reglas básicas de transformación de la máquina. La informática o el procesamiento de datos consiste en la transformación según determinadas reglas de unas fórmulas en otras, y éste es el trabajo de la UCP.

Exactamente, ¿qué fórmulas maneja la UCP y cómo las transforma? El sistema formal que maneja el ordenador común es excesivamente austero. Sólo tiene *dos* elementos básicos —podemos llamarlos "1" y "0"— a partir de los cuales deben construirse *todas* las fórmulas. Este se denomina *código de la máquina* o *lenguaje de la máquina*, y cualquier fórmula en este lenguaje es una sucesión finita de unos y ceros. Estos se representan en la máquina misma como un estado cargado o no de cada elemento en la memoria activa y como un pulso o no pulso en los diversos senderos de la UCP.

En la UCP se construye o conecta una gran cantidad de diminutos elementos llamados *compuertas lógicas* que tienen un 1 o un 0 en cada puerta de entrada y dan un 1 o un 0 como salida, donde la salida dada está estrictamente determinada por la naturaleza de la compuerta y los elementos de la información de entrada. Al utilizar bancos enteros de compuertas lógicas, sucesiones enteras de unos y ceros pueden transfor-

marse en nuevas sucesiones de unos y ceros ordenados de otra manera, según cómo y dónde se ingresen en la UCP. Aquí es donde ocurren las *transformaciones según reglas determinadas.*

Lo curioso de este tedioso manejo de fórmulas —además de la sorprendente velocidad con que se realiza: más de un millón de transformaciones por segundo— es que algunas de las sucesiones pueden *interpretarse* sistemáticamente como la representación de *números* comunes y algunas subunidades de la UCP pueden interpretarse como sumadores, multiplicadores, divisores y demás. Cualquier número puede expresarse en *sistema binario,* y no en nuestro conocido sistema decimal. Es decir que pueden expresarse como sucesiones de unos y ceros.[1] Y cuando es así, las sucesiones $S_1$ y $S_2$ de entrada y la de salida $S_3$, de una determinada subunidad de la UCP siempre están relacionadas de modo que, consideradas como números y no sólo como sucesiones no interpretadas, $S_3$ siempre es igual a $S_1 + S_2$. Aquella subunidad —un conjunto de compuertas lógicas correctamente conectadas— funciona como un *sumador.* Otras subunidades realizan las otras funciones aritméticas básicas.

Análogamente, podemos utilizar el lenguaje de la máquina para codificar fórmulas de la lógica proposicional (éstas representan oraciones en el lenguaje natural) y determinadas subunidades de la UCP procesarán aquellas subunidades de modo que la sucesión de salida siempre represente otra fórmula, una que es la *conjunción* lógica de aquellas representadas por las sucesiones de entrada o su *disyunción* o *negación* o *condicionamiento.* Del mismo modo, las sucesiones de entrada

1. En la notación decimal, las columnas, comenzando por la derecha, van de 0 a 9, y si un número es mayor de lo que la última columna de la derecha puede representar, pasa a la siguiente columna, que va de 0 a 9, simbolizando las decenas esta vez. Y así sucesivamente. En la notación binaria, la última columna de la derecha va sólo de 0 a 1 y *entonces* pasa a la siguiente columna, que también va de 0 a 1, simbolizando pares esta vez. Pasa a la tercera columna, simbolizando cuatro números.Y así sucesivamente. Por ejemplo, en la notación binaria, el conocido "1 + 2 = 3" es "1 + 10 = 11"; y "4 + 5 = 9" es "100 + 101 = 1001".

que representan oraciones arbitrarias (enunciados "si-entonces", por ejemplo) pueden procesarse de tal manera que la sucesión de salida represente un veredicto concerniente a la validez de la verdad funcional de la oración original.

Las UCP se construyen con un control de todas las operaciones lógicas y aritméticas básicas y pueden manejarse infinitamente muchas más operaciones combinando las elementales con más complejas, y combinando éstas a su vez entre sí, como cuando escribimos programas. Evidentemente, este aburrido manejo de sucesiones de unos y ceros puede convertirse en ciertas formas de actividad informática muy emocionantes, poderosas en profundidad y complejidad, como también en velocidad.

## Software (programas)

La actividad informática de la UCP puede ser controlada, y el término "software" se refiere a la secuencia de instrucciones o *programa* que ejerce dicho control. Un programa se carga en la memoria activa del ordenador, donde la UCP lee y ejecuta en secuencia cada una de las instrucciones. El programa indica a la UCP qué sucesiones de entrada debe procesar y de qué modo, dónde y cuándo almacenar los resultados en la memoria, cuándo devolverlos, mostrarlos, imprimirlos y demás.

Por consiguiente, un programa específico convierte al ordenador en una máquina de "propósitos especiales". Y dado que hay una cantidad potencialmente infinita de programas diferentes, podemos hacer que el ordenador se comporte como una cantidad potencialmente infinita de máquinas de "propósitos especiales". Esta es una de las razones por las que los ordenadores aquí descriptos se denominan máquinas de "propósitos generales". Y hay una razón más profunda, que veremos a continuación.

En el nivel más básico, un programa de instrucciones debe ingresarse en la UCP en el lenguaje de la máquina, como sucesiones de unos y ceros, pues ése es el único lenguaje que

entiende la UCP (= es el único sistema formal que la UCP puede manejar). Pero el lenguaje de la máquina es un lenguaje muy difícil y opaco para que lo utilicen los seres humanos. Las sucesiones que representan números específicos, ecuaciones y proposiciones, y las que representan instrucciones para realizar operaciones lógicas y aritméticas, parecen iguales para todos los programadores menos para el más sofisticado, como sucesiones que no representan nada: una jerga uniforme de unos y ceros. Evidentemente, sería mejor si pudiéramos traducir el lenguaje de la máquina a un lenguaje más accesible para el ser humano.

Sin duda, esto puede hacerse y, como la traducción es un caso de transformación de una clase de fórmula en otra, y como un ordenador es un dispositivo de transformación por excelencia, incluso podemos hacer que *él* haga el trabajo por nosotros. El primer paso es construir el teclado de entrada de modo tal que al ser oprimido cada uno de los caracteres conocidos, sea enviado al ordenador codificado como una sucesión única de ocho unos y ceros. Esta codificación preliminar generalmente es un ejemplo del código ASCII (American Standard Code for Information Interchange). Entonces, secuencias de caracteres, como "SUMAR 7, 5", al menos pueden representarse en el vocabulario del lenguaje de la máquina. El próximo paso es cargar el ordenador con un programa (laboriosamente escrito en lenguaje de la máquina, pero el trabajo debe hacerse una sola vez) para *transformar* estas secuencias en secuencias del lenguaje de la máquina que, por ejemplo, realmente indican a la UCP que sume el equivalente binario de 7 y el de 5. El mismo programa puede transformar la salida resultante (1100) nuevamente en el código ASCII (00110001, 00110010), y la impresora codificada en ASCII, al recibir el resultado, imprimirá la secuencia deseada de números o letras conocidas, en este caso "12".

Este programa se llama *intérprete* o *compilador* o *ensamblador* y el lector notará que esta estrategia puede lograr no sólo una interacción más "amistosa" entre hombre y máquina, sino también una gran economía de expresión. Una expresión como "PROMEDIO $x_1$, $X_2$,..., $X_n$" puede transformarse

(primero en ASCII y luego) en una larga sucesión en el lenguaje de la máquina que combina una cantidad de operaciones básicas diferentes como sumas o divisiones. Entonces, una instrucción en el nivel más alto del lenguaje produce la ejecución de una gran cantidad de instrucciones en el lenguaje de la máquina. Estos lenguajes de nivel más alto se denominan *lenguajes de programación* y son lo más cercano a la austera notación del lenguaje de máquina que puede obtener la mayoría de los programadores.

Es evidente que, una vez cargado con un intérprete para permitir el uso de un lenguaje de programación de nivel más alto, el ordenador maneja las fórmulas de un nuevo sistema formal, algunas de cuyas transformaciones "básicas" son más sofisticadas que las que aparecen en el sistema formal del lenguaje de la máquina. Ahora nuestro primer ordenador *simula* ser un ordenador diferente, uno construido para manejar sucesiones en el lenguaje de *programación*. En lo que concierne a la persona que utiliza este "nuevo" ordenador, el "nuevo" lenguaje *es* el del ordenador. Por esta razón el ordenador más el intérprete se denomina a menudo "la máquina virtual".

Todo esto significa que un sistema de procesamiento de datos, programado de otro modo, puede simular muchos sistemas muy diferentes de procesamiento de datos. Esto indica que un ordenador correctamente programado podría simular los sistemas nerviosos de los seres vivos. Algunos resultados en la teoría abstracta de la informática brindan un fuerte sustento a esta idea. Si un ordenador dado satisface determinadas condiciones funcionales, entonces es un ejemplo de lo que los teóricos denominan una *máquina universal Turing* (llamada así por el pionero en teoría de la informática Alan M. Turing). Lo interesante acerca de una máquina universal Turing es que, *para cualquier procedimiento computacional bien definido, una máquina universal Turing es capaz de simular una máquina que realizará dicho procedimiento*. Hace esto reproduciendo exactamente la conducta de entrada/salida de la máquina que se simula. Y el hecho emocionante es que el ordenador moderno *es* una máquina universal Turing.

(Un requisito: los ordenadores verdaderos carecen de memorias ilimitadas. Pero la memoria siempre se puede ampliar para satisfacer la demanda.) Este es el sentido profundo, al que se aludió anteriormente, en el que los ordenadores digitales modernos son máquinas de "propósitos generales".

Por lo tanto, la pregunta que debe responder el programa de investigación de IA no es si ordenadores correctamente programados pueden simular la conducta permanente producida por los procedimientos computacionales que se encuentran en los animales, incluyendo los que se encuentran en el ser humano. Generalmente, se considera que esta pregunta ya está respondida. En principio, al menos, pueden. La pregunta importante es si las actividades que constituyen la inteligencia consciente son todos *procedimientos computacionales* de alguna clase. El supuesto rector de la IA es que lo son, y su meta es hacer programas verdaderos que lo simulen.

Es por ello que la gran mayoría de personas que trabajan en IA se han ocupado de escribir programas en lugar de construir formas cada vez más nuevas de equipos de informática. Ya existe la máquina de propósitos generales y puede ser programada para simular cualquier clase específica de procesador de datos que deseemos. Frente a esto, entonces, el modo más prometedor de enfocar la simulación de procesos cognitivos parecería ser a través de programas diseñados con astucia y cargados en máquinas de propósitos generales. En el siguiente apartado analizaremos algunos de los resultados de este fructífero método.

## Lecturas complementarias


Weizenbaum, Joseph, *Computer Power and Human Reason*, San Francisco, Freeman, 1976; véanse especialmente los capítulos 2 y 3.

Raphael, Bertram, *The Thinking Computer: Mind inside Matter*, San Francisco, Freeman, 1976.

Newell, Alan y Simon, Herbert, "Computer Science as Empirical Inquiry: Symbols and Search", en *Mind Design*, J. Haugeland (comp.), Montgomery, VT, Bradford, 1981; Cambridge, MA, MIT Press.

## 2. Programación de la inteligencia: método gradual

Un método ingenuo para la programación de inteligencia supondría que lo que se necesita es que algún genio programador, en alguna ocasión especialmente inspirada, pase la noche creando furiosamente y aparezca a la mañana siguiente con El secreto, en forma de programa, que al ser colocado en la máquina más cercana disponible, produzca otra conciencia como usted o yo. Aunque suena atrayente, es cosa de historietas. Es ingenuo al suponer que hay un solo fenómeno uniforme que debe capturarse, y también al suponer que hay una sola esencia oculta responsable de ello.

Hasta una mirada informal al reino animal revelaría que la inteligencia se encuentra en miles de grados diferentes y que en diferentes criaturas está constituida por diferentes habilidades, intereses y estrategias, que reflejan todas ellas diferencias en su construcción fisiológica e historia evolutiva. Para tomar un ejemplo popular, en muchos de sus aspectos, la inteligencia de un delfín debe diferir sustancialmente de la del ser humano. Para la información de salida el delfín no tiene brazos, manos ni dedos para una manipulación complicada; tampoco necesita mantenerse en una posición vertical inestable en un campo gravitacional permanente. Por lo tanto, no necesita los mecanismos específicos de control que, en un hombre, administran estos asuntos vitales. Para la información de entrada, el principal sentido del delfín es la ecolocalización sonar, que constituye una ventana al mundo muy diferente de la de la visión. Aun así, el delfín tiene mecanismos de procesamiento que hacen que el sonar sea comparable a la visión en su potencia total. Por ejemplo, el sonar no distingue los colores; por otro lado revela al delfín la estructura interna de los cuerpos captados, ya que todo es "transparente" al sonido en cierto grado. Sin embargo, el filtrado de tal información a partir de ecos complejos constituye para el cerebro del delfín problemas diferentes de los que enfrenta la corteza visual humana y sin duda el delfín tiene mecanismos cerebrales especiales o procedimientos neurales para resolverlos rutinariamente.

Estas diferencias principales en el procesamiento de entrada/salida pueden abarcar otras diferencias en niveles más profundos, y podemos comenzar a apreciar que la inteligencia de cada clase de criatura es probablemente única para aquella especie. Y lo que la hace única es la mezcla específica de mecanismos de procesamiento de datos para propósitos especiales que la evolución ha urdido dentro de ellos. Esto nos ayuda a apreciar que nuestra propia inteligencia debe ser como una cuerda de muchos cabos. Por lo tanto, para simularla tendremos que tejer juntas las hebras similares en formas similares. Y hacer esto requerirá que primero construyamos las hebras. Por esta razón, los investigadores en IA generalmente separaron un aspecto de la inteligencia para luego concentrarse en simular ese aspecto. En cuanto a estrategia, los problemas de integración pueden dejarse de lado temporariamente.

## Conducta intencional y resolución de problemas

Este amplio título abarca muchas cosas —cazar, jugar al ajedrez, construir torres con bloques— en general, cualquier cosa donde las actividades del agente puedan verse como un intento por alcanzar un fin o una meta específicos. Un caso extremadamente simple sería un torpedo autodirigido o un misil termodirigido. Estos moverán aletas direccionales y darán vueltas de modo tal de permanecer fijos sobre el blanco móvil. Pueden parecer ferozmente obstinados si uno de ellos va pisándonos los talones, pero en un momento de tranquilidad no se nos ocurriría atribuirles una inteligencia verdadera, ya que tienen una única respuesta para cada acción evasiva, regulada directamente a la medida de la "desviación del blanco del centro" por el sensor del misil. Estos sistemas son útiles para entender la conducta animal —los mosquitos aparentemente se dirigen, con igual simplicidad, a gradientes crecientes de dióxido de carbono (exhalación)— pero pretendemos de la IA más que la inteligencia del mosquito.

¿Qué ocurre en el caso en que una gama de posibles

respuestas a cualquier brecha percibida entre el estado actual y el estado a alcanzar es mucho mayor, y qué sucede si una elección útil de entre aquellas respuestas requiere la resolución de un problema por parte del agente? Esto se acerca más a la tarea de la inteligencia verdadera. Curiosamente, una variedad considerable de programas existentes pueden satisfacer esta condición y algunos de ellos producen una conducta compleja que, en un ser humano, se consideraría como muy inteligente.

Comenzando por los casos simples, consideremos el juego del ta-te-ti (figura 6.2), y consideremos los procedimientos que un ordenador explotaría para maximizar sus posibilidades de ganar o por lo menos empatar con un contrincante. Suponiendo que el ordenador juega primero con las cruces, hay 9 movimientos posibles que podría hacer. Para cada uno de ellos hay 8 posibles contramovimientos para el que juega con los círculos. Y para cada uno de éstos, hay 7 posibles respuestas del ordenador. Y así sucesivamente. Según el cálculo más simple hay $9 \times 8 \times 7 \times \ldots \times 2$ (= 9! = 362.880) maneras diferentes de llenar el tablero. (Hay un poco menos de juegos completos, ya que el juego termina cuando se colocan tres alineados, y antes que la matriz esté completa.) Podemos representar estas posibilidades como un *árbol de probabilidades* (figura 6.3).

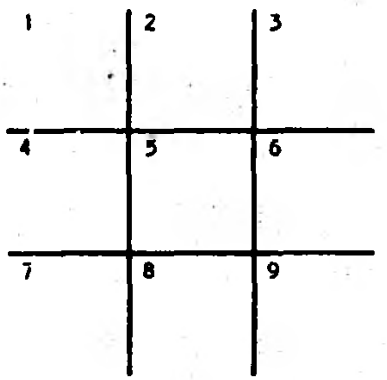



Figura 6. 2

Este diagrama es demasiado grande como para hacerlo completo en una página, pero no lo es como para que un ordenador programado adecuadamente analice rápidamente cada rama y vea si termina en victoria, derrota o empate para las cruces. Esta información puede informar su elección de los movimientos en cada etapa del juego. Digamos que cualquier rama del árbol que enfrentan las cruces es una "mala rama" si en el movimiento *siguiente* el jugador de los círculos tiene un movimiento que termina la partida obteniendo la victoria. Y también digamos que cualquier rama que enfrentan las cruces también es una mala rama si en el movimiento siguiente el jugador de los círculos tiene un movimiento que dejará a las cruces enfrentadas solamente a malas ramas. Con esta definición recursiva e identificando primero las malas ramas terminales, el ordenador podrá revisar todo el árbol e identificar *todas* las malas ramas. Si también lo programamos como para que en cada etapa del juego real nunca elija una de las malas ramas que ha identificado, y siempre elija un movimiento ganador y no de empate, entonces ¡el ordenador nunca perderá el partido! Lo máximo que podemos esperar es empatar y dos ordenadores así programados empatarán todos los partidos entre sí.

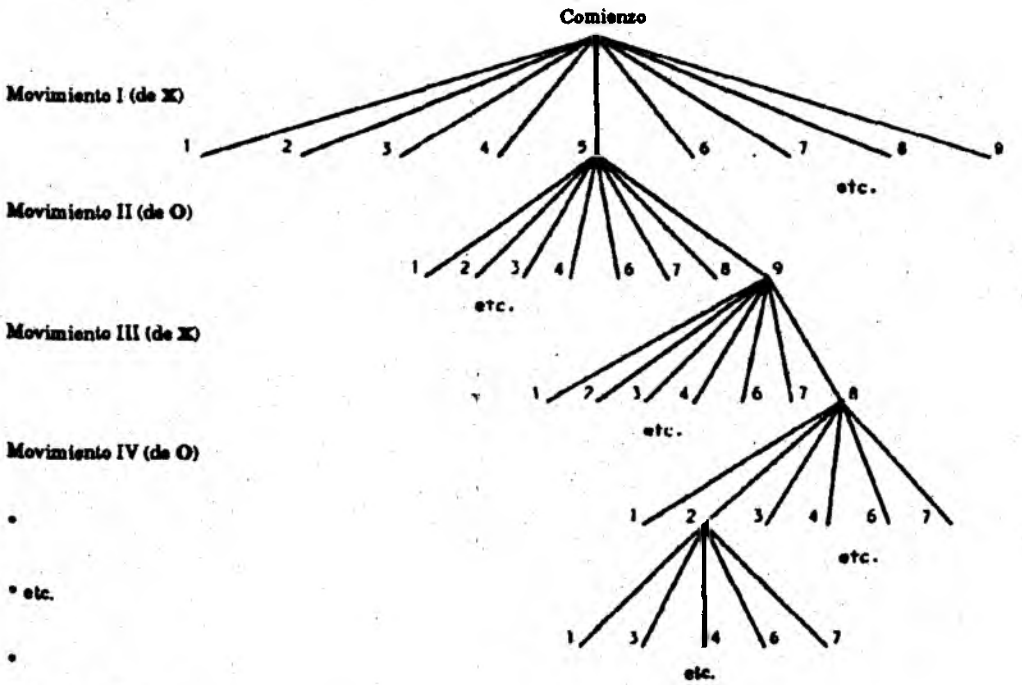



Figura 6. 3

Para ilustrar brevemente estos puntos, consideremos el partido X — 5, O — 9, X — 8, O — 2, X — 7, O — 3, X — 6, O — 1. Y comencemos a jugar después del movimiento IV, con la matriz de la figura 6.4.

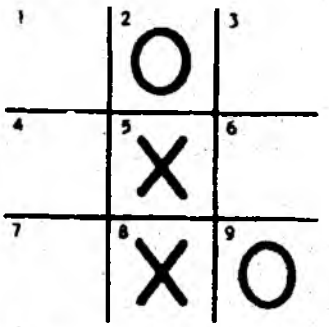



Figura 6. 4

El lector podrá jugar en los últimos cuatro movimientos y presenciar la derrota de las cruces. Si ahora miramos el sector del árbol que se extiende desde el movimiento IV de los círculos (figura 6.5) podremos ver por qué las cruces no tendrían que haber elegido el cuadro 7 en el movimiento V. Desde allí los círculos tienen un movimiento (al cuadro 3) que deja a las cruces únicamente frente a ramas malas. Las cruces deberán elegir 1, 4 o 6 en el movimiento VII y los tres cuadros dejan a los círculos en una posición de elección ganadora en el siguiente movimiento. Entonces las tres son ramas malas. Y por lo tanto, la rama de las cruces al cuadro 7 en el movimiento V *también* es una mala rama porque permite que los círculos, en el movimiento siguiente, dejen a las cruces frente a todas ramas malas. En vista de esto, podemos apreciar que las cruces no deberían ir al 7 en el movimiento V. Nuestro ordenador programado también lo ve, entonces evitará el error recién analizado. Y todos los demás, cualquiera sea su situación en el árbol.

Así, estamos en presencia de un caso en el que la máquina programada tiene una meta (ganar o por lo menos empatar), una gama de posibles respuestas para cada circunstancia en la que se encuentre, y un procedimiento para resolver, en cada etapa, el problema para el que esté mejor capacitada para lograr dicha meta. (Si dos o más son igualmente buenas,

entonces podemos decirle que elija la primera en la lista o "lanzar una moneda" con alguna subrutina aleatoria.)

La particular estrategia descripta es un ejemplo de lo que se denomina el método de *fuerza bruta* para la resolución de problemas: a partir de la descripción básica del problema, el ordenador crea un *árbol de búsqueda* que abarca todas las posibilidades pertinentes y realiza una exhaustiva búsqueda en la rama o ramas que constituyen una solución. Esto se denomina una *previsión exhaustiva*. Para problemas que tienen una solución (no todos la tienen), este método funciona perfectamente, dado que existe suficiente "fuerza" disponible. Constituye un procedimiento eficaz o un *algoritmo* para identificar los movimientos más convenientes.

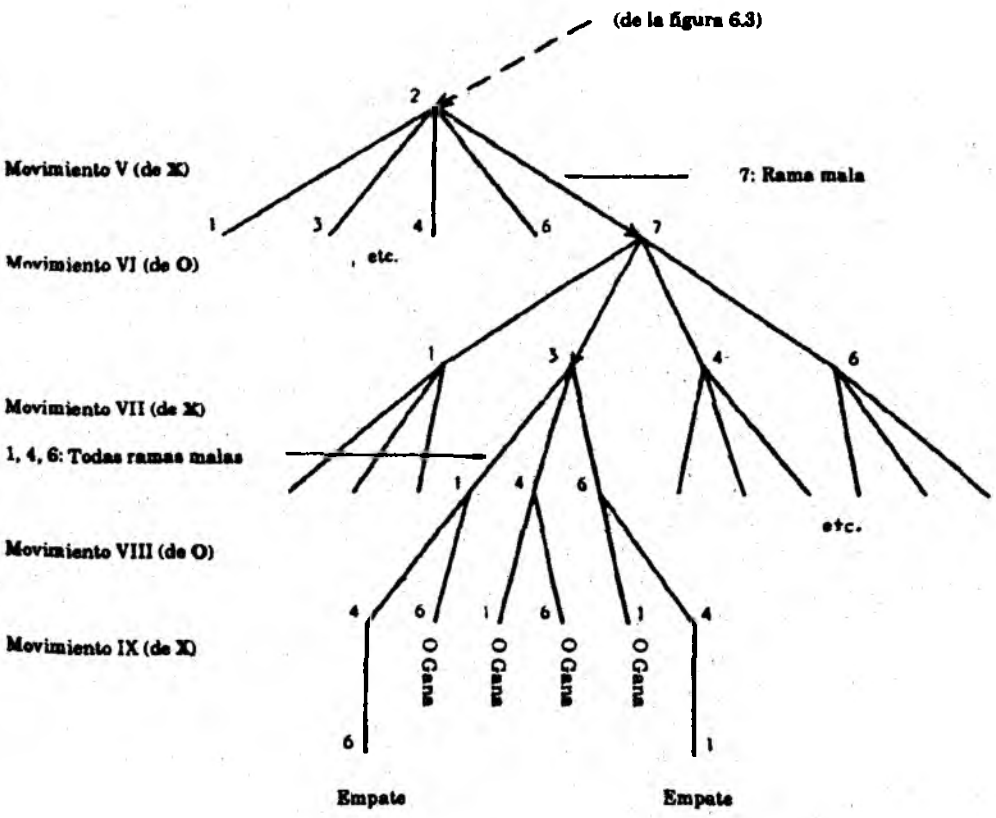



Figura 6. 5

"Fuerza" aquí significa velocidad y capacidad de memoria por parte de la máquina: fuerza suficiente para construir y

buscar el árbol pertinente. Desafortunadamente, muchos de los problemas que la inteligencia verdadera enfrenta abarcan árboles de búsqueda que superan el alcance de las máquinas factibles y el método de la fuerza bruta. Hasta para el ta-te-ti, la estrategia específica descripta requiere alta velocidad y una memoria amplia. Y para juegos más exigentes, el método enseguida resulta inoperable.

Consideremos el ajedrez. Sin duda, una exigencia, pero no más que los "juegos" sociales que juegan rutinariamente los seres humanos. En promedio, un jugador en cualquier etapa de una partida de ajedrez debe elegir entre aproximadamente 30 jugadas posibles. Y cada movimiento hará posibles unas 30 respuestas de su oponente. Entonces solamente los dos primeros movimientos son un par elegido entre unos $30^2$ (= 30 $\chi$ 30 = 900) pares posibles. Si una partida promedio consta de aproximadamente 40 movimientos de cada jugador, para un total de 80, entonces la cantidad de partidas promedio posibles y diferentes será de 80 a la 80ª potencia, o cerca de $10^{118}$. Entonces la partida pertinente tendrá aproximadamente $10^{118}$ ramas. Es un número absurdamente grande. Un millón de ordenadores que analicen cada uno un millón de ramas por segundo aún tardarían $10^{100}$ (la unidad seguida de 100 ceros) *años* para analizar todo el árbol. Evidentemente, este método para jugar al ajedrez no funcionaría.

El problema con que nos encontramos aquí es un ejemplo de *explosión combinatoria* y significa que un programa para jugar al ajedrez no puede tener la esperanza de utilizar un algoritmo para identificar los movimientos posibles con mejores garantías de ganar. Debe volver a los procedimientos *heurísticos*. Es decir, debe utilizar los métodos "prácticos" para distinguir los movimientos meramente prometedores de los que no lo son tanto. Consideremos cómo puede funcionar esto. Si escribimos el programa de modo que el ordenador no trate de prever 40 movimientos, sino sólo 4 (= 2 para cada jugador) en cualquier etapa de la partida, entonces el árbol de búsqueda pertinente tendrá sólo $30^4$ u 800.000 ramas. Este número es suficientemente pequeño como para que las máquinas existentes busquen en un tiempo razonable. Pero ¿qué

busca si no puede buscar la victoria como fin último? Aquí tratamos de proporcionar al ordenador metas intermedias que: a) *pueda identificar* efectivamente y b) ofrezcan alguna *probabilidad* de que si se logran repetidamente, entonces la victoria última también sea alcanzada.

Por ejemplo, podemos asignar números a la pérdida de determinadas piezas, en proporción a su importancia general; y el ordenador puede asignar un valor positivo y negativo a cualquier intercambio potencial de piezas con un oponente, según quién pierda y cuánto. El ordenador también puede guiar su elección de movimientos asignando un determinado valor positivo al hecho de tener sus piezas "en control del centro" (= tener las piezas en una posición para capturar en el sector central del tablero). También puede asignarse un valor a los movimientos potenciales que ataquen al rey del oponente, ya que es una condición necesaria para la victoria. Y así sucesivamente.

Podemos escribir el programa de modo que el ordenador sume estos factores, para cada movimiento considerado, y luego elija el que tenga el mayor valor agregado. De esta manera, al menos podemos hacer que el ordenador *juegue* una partida de ajedrez *reconocible,* lo que resulta imposible según el método de la fuerza bruta, ya que la máquina se paraliza frente a la inmensidad de la tarea.

El hecho es que se han escrito programas para jugar al ajedrez utilizando la heurística, como este y otros más ingeniosos, que derrotarían a cualquiera excepto a aquellos pocos que ya han llegado al nivel de expertos, e incluso en este caso jugarían bastante bien. (En los últimos años han salido a la venta programas más simples, pero aun notables, incorporados en "tableros de ajedrez electrónicos". La IA ha entrado en el mercado.) Esta conducta intrincadamente afinada es una notable demostración, incluso según los criterios humanos de la inteligencia. La mirada previsora guiada por la heurística quizá no sea infalible, pero sí puede ser muy poderosa.

Otra clase de programas que simulan la resolución de problemas y la conducta intencional representa una estrategia diferente de la previsión exhaustiva y de la previsión parcial guiada por la heurística. En lugar de dirigirse hacia una meta

considerando primero todos los movimientos posibles dentro de la potencia del ordenador, y luego cada movimiento posible a partir de allí, y así sucesivamente, con la esperanza de que alguna rama de este árbol en explosión finalmente haga contacto con la meta, el ordenador puede comenzar por el otro extremo del problema. Puede comenzar considerando todas las circunstancias posibles que pueda, en las que *un* movimiento más de su parte asegure su meta. No es necesario que haya muchas de ellas —quizá sólo una—. Entonces estas posibles circunstancias se convierten en metas intermedias y el ordenador puede repetir su búsqueda de las maneras posibles de asegurar una o más de *ellas*. Este proceso se repite hasta que el ordenador finalmente identifica alguna circunstancia que tenga el poder de producir inmediatamente. Entonces el ordenador hace su jugada y, en orden inverso, todas las demás de la cadena de medios y fines que ha construido, logrando así alcanzar su meta original y última. Esta estrategia no es necesariamente siempre más eficiente que la de la previsión exhaustiva, pero si hay muchas jugadas que el ordenador podría hacer al principio, y sólo muy pocas que conducen al logro de la meta, este método puede resultar mucho más veloz.

El programa STRIPS (Stanford Research Institute Problem Solver) es capaz de realizar esta estrategia. Un robot móvil con el nombre de Shakey dirigido a control remoto por un ordenador alimentado con el programa STRIPS podía recibir la orden de lograr diversas metas, que debía alcanzar en un ambiente de varias habitaciones conectadas por puertas y llenas de varias cajas grandes. Dada la información con respecto a la disposición de las habitaciones, las puertas comunicantes, las cajas y el propio Shakey, y dada una meta como "Hacer que la caja de la habitación 3 sea colocada en la habitación 7", Shakey (o mejor dicho STRIPS) crearía y realizaría una secuencia de conductas que lograrían esto.

## Aprendizaje

También debemos observar dos maneras en que programas de la clase analizada pueden demostrar el *aprendizaje*.

La primera y más sencilla es sólo una cuestión de conservar, en la memoria, soluciones ya logradas. Cuando se está nuevamente frente al mismo problema, puede llamarse instantáneamente a la solución que está en la memoria y utilizarse directamente, en lugar de resolverlo otra vez laboriosamente cada vez. Una lección, una vez aprendida, es recordada. La conducta intencional que era vacilante al principio puede convertirse así en fluida y decidida.

La segunda manera puede ejemplificarse con el caso de un programa de ajedrez guiado por la heurística. Si escribimos un programa de modo que el ordenador guarde un registro de su cociente victoria/derrota, podemos hacer que pruebe nuevas mediciones para sus diversas heurísticas si se encuentra perdiendo en una proporción inaceptable. Supongamos, por ejemplo, que la heurística "ataque al rey de su oponente" tiene gran peso, inicialmente, y que la máquina pierde partidas regularmente debido a repetidos ataques kamikazes al rey enemigo. Luego de observar sus derrotas, el ordenador podría tratar de ajustar cada una de sus mediciones por vez, para ver si obtiene un mejor cociente victoria/derrota. En el largo plazo, la heurística de peso excesivo sería medida nuevamente hacia abajo, y la calidad del juego de la máquina podría mejorar. El ordenador aprende a hacer un juego más fuerte, de un modo como usted o yo lo haríamos.

Evidentemente estas dos estrategias reproducirán algo de lo que comúnmente llamamos aprendizaje. Sin embargo, el aprendizaje es mucho más que el mero almacenamiento de información adquirida. En las dos estrategias descriptas, la máquina representa la información "aprendida" dentro del esquema de conceptos y categorías provisto por su programa original. En ninguno de los dos casos la máquina genera *nuevos* conceptos y categorías con los cuales analizar y manejar la información que ingresa. Puede manipular las categorías anteriores y formar una variedad de combinaciones de ellas, pero la innovación conceptual está limitada a la actividad combinatoria dentro del marco de referencia original.

Esta es una forma extremadamente conservadora de aprendizaje, como podemos apreciar cuando consideramos el

aprendizaje de un niño en sus dos primeros años de vida o de la comunidad científica en el curso de un siglo. El cambio conceptual en gran escala —la generación de un marco de categorías genuinamente nuevo que desplace totalmente al viejo— es característico de ambos procesos. No podemos pretender tener resuelto el problema del aprendizaje hasta que hayamos resuelto el del cambio conceptual.

Esta clase más profunda de aprendizaje es mucho más difícil de simular o recrear que las clases más simples analizadas anteriormente, pues requerirá que podamos representar de alguna manera el conocimiento y la información en un nivel *por debajo* del de los conceptos lingüísticamente expresables, un nivel cuyos elementos pueden combinarse o articularse de algún modo para formar un concepto cualquiera de la amplia gama de alternativas posibles. Tal nivel de representación también debe ser sensible y habrá de dar respuesta al desempeño ulterior del sistema general, de modo que los conceptos exitosos puedan distinguirse de los inútiles y confusos.

Este problema pareció casi insuperable hasta hace muy poco. Felizmente nuevos métodos para la representación y manejo de grandes cantidades de información han producido recientemente algunos "procedimientos de aprendizaje" muy notables que reciben mucha atención en la actualidad. Sin embargo, están diseñados para ser implementados, al menos teóricamente, en máquinas de calcular de construcción muy diferente de las descriptas algunas páginas atrás, y describirlas a esta altura sería una digresión. Aparecerán en el capítulo 7.

## Visión

Si un ordenador adecuadamente programado estuviera equipado con sensores ópticos, ¿podría *ver*? En un nivel simple de procesamiento de información óptica, la respuesta evidentemente es sí. Las editoriales con frecuencia utilizan este sistema para la composición de un libro. El texto original del

autor es "leído" por un sistema que examina cada carácter en secuencia y registra su identidad en una cinta. Otro ordenador utiliza esa cinta en la máquina de composición tipográfica. Los analizadores de reconocimiento de caracteres pueden ser muy simples. Un sistema de lentes proyecta una imagen en blanco y negro del carácter en una grilla de elementos fotosensibles (figura 6.6). La imagen del carácter ocupa ampliamente determinados cuadros de la grilla, y el dispositivo analizador envía una lista codificada de todos ellos al ordenador. Entonces el programa pertinente permite que el ordenador compare esa lista con todas las listas estándar que tiene en su memoria, una para cada carácter estándar. Separa la lista almacenada que coincide con la recibida en la mayor cantidad de puntos, e identifica así el carácter analizado. Por supuesto, todo esto se realiza a la velocidad de la luz.

Evidentemente este sistema es inflexible y puede ser engañado con facilidad. Fuentes de tipos inusuales producirán interpretaciones erróneas crónicas. Y si se alimenta el sistema con imágenes de rostros o de animales, hará lo mismo que antes, identificándolos como letras y números diversos. Del mismo modo, estos defectos son análogos a las características obvias de nuestro propio sistema visual. Nosotros también tendemos a interpretar lo que vemos en términos de categorías conocidas o esperadas y con frecuencia hasta no notamos lo nuevo a no ser que estemos especialmente atentos.

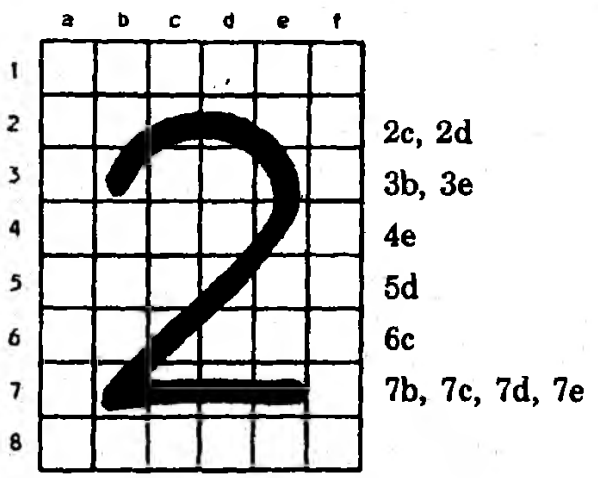



Figura 6. 6

Sin embargo, el reconocimiento de caracteres representa sólo los comienzos rudimentarios de la visión de una máquina, no su pináculo. Consideremos el problema más general de identificar y localizar objetos en el espacio tridimensional, utilizando sólo como datos lo que se presenta en un conjunto bidimensional de diversos puntos iluminados: esto se denomina *matriz de intensidad*, y la imagen de televisión constituye un ejemplo familiar. Es justamente un caso extravagante de nuestra anterior grilla para el reconocimiento de caracteres, excepto que tiene muchos más elementos y valores graduados para cada uno.

Usted y yo tenemos retinas que funcionan como matrices de intensidad, y podemos resolver los problemas pertinentes con facilidad cuando vemos disposiciones específicas de objetos sobre la fuerza de matrices de intensidad específicas de retina. No somos conscientes del "problema" de la interpretación ni del procesamiento que lo soluciona en nuestro interior. Pero esta capacidad es un verdadero desafío para el programador, ya que refleja gran inteligencia por parte del sistema visual.

Esto se debe a que las representaciones visuales siempre son interminablemente *ambiguas*. Muchas circunstancias externas diferentes son estrictamente consistentes con cualquier matriz de intensidad bidimensional. Es decir que circunstancias diferentes pueden "parecer" aproximadamente o incluso exactamente iguales, como una moneda común, levemente inclinada, parece una moneda realmente elíptica. Cualquier sistema visual debe poder distinguir sin ambigüedad escenas de un modo razonable, para encontrar la interpretación más *probable* provistos los datos. También algunas escenas son más complicadas que otras: quizá la interpretación "correcta" requiera conceptos que el sistema ni siquiera posee. Esto indica que la visión, como la inteligencia, viene en grados. Afortunadamente esto nos permite estudiar primero los casos simples.

Consideremos una matriz de intensidad específica: pensemos en una imagen de televisión de varias cajas grandes apiladas en un montón. Cambios repentinos en la intensidad

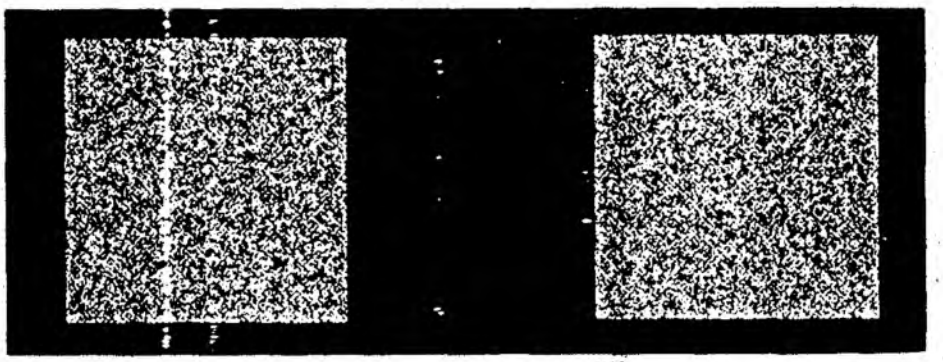
de la luz reflejada marcan los bordes de cada caja y un programa sensible a estos cambios puede construir a partir de ellos un esquema lineal de las diversas cajas y sus posiciones relativas. A partir de aquí, un programa sensible a los modos en que los bordes se encuentran para formar las esquinas /lados/ volúmenes enteros (como el programa VER de Guzmán) puede adivinar correctamente cuántas cajas hay y en qué posiciones relativas. Estos programas funcionan bien para medios muy artificiales que contienen solamente sólidos de lados planos, pero quedan muchas ambigüedades después de su resolución y se desmoronan completamente cuando se les presenta una playa rocosa o una hoya llena de hojas.

Programas más recientes utilizan la información contenida en cambios *continuos* de intensidad —pensemos en el modo en que la luz se distribuye sobre una esfera o un cilindro— para respaldar hipótesis sobre una variedad de objetos mucho mayor. También se está analizando la estereopsia artificial. Las diferencias sutiles entre un par de matrices de intensidad bidimensionales tomadas de dos posiciones ligeramente diferentes (como las imágenes en nuestras retinas derecha e izquierda) contienen información potencialmente decisiva sobre los contornos y posiciones espaciales relativas de elementos en la escena. Ya se ha escrito un algoritmo que recuperará la información tridimensional oculta en el par estéreo que se observa en la figura 6.7.

Coloque un sobre de tamaño comercial verticalmente entre los dos cuadrados y centre la nariz y la frente sobre el borde del sobre de modo que cada ojo vea solamente una imagen. O mejor aún, haga un par de binoculares de papel con dos hojas de papel de carta enrolladas formando dos tubos largos. Sosténgalos paralelos, con el extremo de cada uno cerca de la página, de manera que cada ojo mire directamente por el tubo y sólo vea un cuadrado centrado en la abertura circular. Deje que su sistema visual en unos minutos una las imágenes izquierda y derecha en una sola imagen claramente enfocada (tenga paciencia) y podrá ver cómo su propio algoritmo de gran habilidad encuentra la misma información.

Un problema crónico con la visión de la máquina es que,

como la visión misma implica inteligencia, y como lo que cualquier criatura puede ver en una situación dada depende en gran parte de qué conceptos ya posee y de qué conocimientos ya tiene, la creación de sistemas visuales artificiales de alta potencia depende en general de la creación de inteligencia de alta potencia y de la creación de sistemas que posean bases de conocimientos muy amplias para dar una dirección a su procesamiento perceptual. Esto no es sorprendente ni especialmente decepcionante, pero ilustra las limitaciones de un método puramente gradual para la simulación de facultades cognitivas.



izquierda derecha

Figura 6.7. Reproducida con autorización de D. Marr y T. Poggio, "Cooperative Computation of Stereo Disparity", *Science*, vol. 194, 1976, págs. 283-286. Copyright 1976 by the AAAS.

## Manejo del lenguaje natural

Otra área en la que la falta de un conocimiento verdaderamente amplio ha frustrado a los investigadores es el área de la comprensión y producción del lenguaje natural. Por muchos motivos, es deseable poder conversar con ordenadores en inglés corriente, y el manejo del lenguaje natural es un elemento central de la inteligencia humana en cualquier caso, digno de ser simulado por derecho propio. Los conocimientos estructu-

rales por parte de los lingüistas y la astucia informática por parte de los programadores han producido programas que manejan las estructuras sintácticas del inglés bastante bien, incluso las más complicadas; pero la parte *semántica* del problema ha sido más difícil de tratar. Casualmente se puede hacer mucho manejando solamente la sintaxis. Consideremos uno de los famosos primeros programas de Weizenbaum, ELIZA, que simula a un psicoterapeuta con el que el "paciente" conversa a través de la máquina de escribir. El siguiente es un extracto de dicho intercambio, con las respuestas del ordenador en letras mayúsculas (de Weizenbaum, 1976):

Los hombres son todos iguales.
DE QUE MODO
Siempre nos están molestando por una cosa u otra.
PUEDE USTED DAR UN EJEMPLO ESPECIFICO
Bueno, mi novio me hizo venir aquí.
SU NOVIO LA HIZO VENIR AQUI
Dice que estoy deprimida la mayor parte del tiempo.
LAMENTO ESCUCHAR QUE ESTA DEPRIMIDA
Es verdad. Soy desdichada.
CREE USTED QUE VENIR AQUI LA AYUDARA A NO SER DESDICHADA
Lo que sí parece seguro es que necesito ayuda.
QUE SIGNIFICARIA PARA USTED SI RECIBIERA ALGUNA AYUDA
Quizá podría aprender a llevarme bien con mi madre.
CUENTEME MAS ACERCA DE SU FAMILIA
Mi madre me cuida.
QUIEN MAS EN SU FAMILIA LA CUIDA
Mi padre.
SU PADRE
Usted es como mi padre en ciertos aspectos.
QUE SEMEJANZA VE
Usted no es muy agresivo, pero creo que usted no quiere que yo me dé cuenta de ello.
QUE LE HACE PENSAR QUE NO SOY MUY AGRESIVO
No discute conmigo.

POR QUE CREE USTED QUE NO DISCUTO CON USTED
Usted me teme.
LE AGRADA CREER QUE LE TEMO
Mi padre le teme a todo el mundo.
QUE MAS SE LE OCURRE CUANDO PIENSA EN SU PADRE
Pendenciero.
TIENE ESO ALGO QUE VER CON EL HECHO DE QUE SU NOVIO LA HIZO VENIR AQUI

En esta conversación, el ordenador, con el programa ELIZA, da la extraña sensación de que comprende verdaderamente. Pero en realidad no es así. Las respuestas en gran parte están construidas a partir de las propias oraciones del paciente, por transformaciones simples, y a partir de un repertorio de formas estándar de preguntas hechas a medida para cada palabra clave del paciente ("deprimida", "como", y así sucesivamente). ELIZA no tiene idea de lo que es un padre, un hermano o la infelicidad. No tiene el concepto de estas cosas, no tiene comprensión de lo que significan estas palabras. Lo que demuestra qué comprensión increíblemente pequeña se requiere para participar adecuadamente en muchas formas corrientes de conversación.

El programa de Winograd, SHRDLU, es mucho más impresionante. Maneja tanto la semántica como la sintaxis y manipula los elementos en el mundo de bloques (simulado), que es todo lo que conoce. Su sintaxis es muy sofisticada y el programa incluye cierta información sistemática sobre las propiedades de los cuerpos que habitan su mundo. En forma muy rudimentaria sabe, un poquito, sobre qué está hablando. Como resultado puede hacer deducciones útiles y conjeturar relaciones reales, aptitud que se refleja en las conversaciones más complejas y sutilmente focalizadas que podemos mantener con él. Sin embargo, las conversaciones deben restringirse al mundo de bloques y a aquellos estrechos aspectos que abarca. SHRDLU no tiene una base de conocimientos vacía, pero su base es aun menos que microscópica comparada con la nuestra.

Brevemente, el problema es que para entender el lengua-

je natural en el nivel humano se requiere un *conocimiento* general del mundo comparable al que un ser humano posee (recordemos la teoría holística del significado, la "teoría reticular", analizada en el capítulo 3.3), y aún no hemos resuelto el problema de cómo representar y almacenar una base de conocimientos tan grande de un modo que permita el acceso y el manejo. Un problema más profundo se relaciona con esto. Aún no hemos resuelto el problema de *cómo* pueden *adquirirse* tales cantidades globales de conocimiento. Cómo se generan los marcos conceptuales, cómo se modifican y luego se desechan en favor de marcos más nuevos y sofisticados; cómo esos marcos se evalúan como reveladores o engañosos, como verdaderos o falsos; nada de esto se entiende bien. Y la IA se ha ocupado muy poco.

Estos problemas son de competencia tradicional de la lógica inductiva, la epistemología y la teoría semántica, para los filósofos. Y también son de competencia de la psicología evolutiva y la teoría del aprendizaje, para los psicólogos. Parece ser necesario un ataque colectivo, pues los fenómenos a entender son tan complicados y difíciles como cualquier otro que hayamos enfrentado. Sin duda aquí también se requiere paciencia, pues no podemos esperar crear en sólo unas cuantas décadas lo que le llevó al proceso de la evolución tres mil millones de años.

## Autoconciencia

El lector habrá observado que ninguna de las simulaciones que se analizaron aquí se refiere al tema de la autoconciencia. Quizá los sensores visuales y táctiles, más una programación imaginativa, proporcionen a un ordenador cierta "conciencia" del mundo exterior, pero prometen poco y nada con respecto a la conciencia de sí. Esto no debería sorprendernos. Si la autoconciencia estriba en la comprensión introspectiva de los propios procesos cognitivos de alto nivel, entonces casi no tiene sentido tratar de simular la *comprensión* de dichos procesos hasta que ellos mismos hayan sido

simulados exitosamente. Quizá pueda postergarse un ataque en gran escala sobre la autopercepción hasta que la IA haya construido algunos "yo" realmente dignos de percepción reflexiva explícita. Sin embargo, ya resultó necesario un trabajo preliminar. La propiocepción —la conciencia de la posición de las propias extremidades en el espacio— es una forma de autopercepción y por razones obvias el desarrollo de brazos de robot controlados por ordenadores ha requerido que se le den al ordenador algunos medios sistemáticos de sentir la posición y el movimiento de su propio brazo y de representar esta información de un modo que le sea útil en forma permanente. Quizás esto ya constituye una forma primitiva e identificada de autoconciencia.

Finalmente, no debemos permitir que el término "simulación" nos lleve a desechar impulsivamente las perspectivas de este enfoque general para el problema de la inteligencia consciente, pues la simulación estudiada puede ser simulación *funcional* en el sentido más fuerte posible. Según aquellos teóricos de la IA que toman el sistema computacional moderno como modelo, no tiene por qué haber diferencia entre nuestros procedimientos computacionales y los que simula una máquina, ninguna diferencia más allá de la sustancia física concreta que sustenta esas actividades. En el ser humano es material orgánico; en el ordenador serían metales y semiconductores. Pero *esta* diferencia no es más pertinente para el tema de la inteligencia consciente que una diferencia en el tipo de sangre o el color de la piel o la química del metabolismo, según afirma el teórico de la IA (funcionalista). Si las máquinas llegan a simular todas nuestras actividades cognitivas internas, hasta el último detalle computacional, negarles la categoría de personas sería nada más que una nueva forma de racismo.

## Algunos problemas crónicos

En el apartado precedente se ha hecho una evaluación optimista de las perspectivas abstractas de la IA, pero hay algunas dificultades recurrentes que han frustrado el programa

de investigación de la IA tradicional o de la "escritura de programas", y nos corresponde reconocerlas y reflexionar acerca de su importancia.

Un hecho confuso acerca de los resultados de la investigación de la IA es que hay ciertas clases de tareas, como la reducción de números, la demostración de teoremas y la búsqueda en listas, que los ordenadores corrientes hacen muy velozmente y bien, mientras que el cerebro humano las realiza lentamente y comparativamente mal. Por otro lado, hay otras clases de tareas, como el reconocimiento facial, la comprensión de escenas, la coordinación sensoriomotora y el aprendizaje, que los seres humanos y otros animales hacen velozmente y bien, pero que incluso los ordenadores más rápidos con los programas más sofisticados hacen lentamente y bastante mal.

Más específicamente, usted puede reconocer una fotografía de su mejor amigo, en cualquiera de una amplia variedad de poses, en menos de medio segundo. Pero esta capacidad de reconocimiento todavía se les escapa a los mejores programas existentes de reconocimiento de modelos, e incluso versiones muy simplificadas de problemas de reconocimiento como éste llevan minutos de frenético procesamiento en un ordenador, o más tiempo, antes de encontrar una solución al problema.

Como segundo ejemplo, podemos aprender a devolver una pelota de tenis en no más de diez o quince intentos. Pero la coordinación sensoriomotora requerida para guiar la conducta de un sistema óseo y muscular complejo como el del cuerpo humano, en tiempo real, todavía supera la capacidad de la IA actual. Y la idea de un programa que pueda *aprender* a que un sistema así efectúe esos tiros de tenis, y que lo haga en menos de quince intentos, es una posibilidad aun más remota.

## Un diagnóstico reciente

¿Por qué el cerebro es tanto mejor que las máquinas más hábilmente programadas para realizar algunas tareas conoci-

das y tanto peor que los ordenadores más simples para realizar otras? La respuesta parecería estar en los tipos completamente diferentes de estructura física y computacional que ponen de manifiesto las dos clases de sistemas de procesamiento de datos. Aunque los ordenadores corrientes son indudablemente máquinas de "propósitos generales", en el sentido de que pueden simular cualquier sistema posible de procesamiento de datos, hay muchas clases de sistemas cuya simulación requiere *enormes* cantidades de actividad que lleva mucho tiempo por parte de la unidad central de procesamiento de un ordenador corriente. Los cerebros biológicos parecen ser esos sistemas muy difíciles de simular. En principio son simulables, pero sólo a expensas de establecer una simulación informática que termine resolviendo el problema pertinente o realizando las actividades deseadas a una velocidad mucho menor que el cerebro —quizá millones o miles de millones de veces menor—.

¿Qué explica una diferencia de velocidades tan grande? El problema parecería radicar en el "atolladero" que se produce en el procesamiento realizado por la UCP en la máquina corriente de propósitos generales. La UCP de esas máquinas es típicamente un trabajador muy activo que funciona en el orden de un millón ($10^6$) de cálculos distintos por segundo. Considerado por sí solo, es impresionante. Pero por más rápidamente que funcione, sólo puede hacer un cálculo por vez, y muchos problemas, como el aprendizaje o el reconocimiento descriptos antes, requieren mucho más que *mil millones* ($10^9$) de pasos computacionales diferentes para su resolución. Como cada uno debe ser realizado por la UCP, uno tras otro de un modo seriado cuidadosamente organizado, es evidente que le llevaría a la máquina por lo menos ($10^9 / 10^6 =$) 1000 segundos, o más de un cuarto de hora, resolver dicho problema. Es un tiempo prolongado, según el criterio biológico. Un ratón que no pueda reconocer a un gato más velozmente que eso terminará sirviéndole de almuerzo.

Contrariamente, el cerebro no tiene una UCP en la que están confinados todos los cálculos y por la que debe pasar toda la información. Los cerebros parecen tener una estructu-

ra física y computacional completamente diferente de las típicas máquinas de calcular, una estructura que permite realizar *simultáneamente* miles de millones de cálculos simples. Como cada uno de ellos es muy simple, es realizado velozmente por sólo una de las miles de millones de células distintas que contiene el cerebro y se efectúa todo de modo tal que el resultado de salida colectivo representa la solución terminada del problema presentado.

Aquí no se produce un atolladero computacional a través del cual deben pasar con dificultad todos los trozos de información pertinentes, uno tras otro, en fila. Como cada célula cerebral contribuye simultáneamente con un solo cálculo de todo el proceso, toda la operación puede completarse en un único paso por la red pertinente de células cerebrales. Y ese paso no necesita más de 1/100 de segundo, ya que pasa por todas las células del sistema exactamente al mismo tiempo. Así es como el cerebro, hasta el de un ratón, puede realizar complicadas tareas de reconocimiento en un instante.

Este estilo diferente de procesamiento de datos se denomina *procesamiento en paralelo,* en contraste con el *procesamiento en serie* que efectúan las máquinas de calcular corrientes. Lo que proporciona verdaderamente es una enorme ventaja en la velocidad con que pueden resolverse algunas clases de problemas computacionalmente intensivos. Esta ventaja de la velocidad ha convertido al procesamiento en paralelo en el foco de atención reciente entre los investigadores de la IA y de la ciencia cognitiva, pero la velocidad no es la única característica que lo favorece. Los procesadores en paralelo tienen algunas propiedades computacionales muy interesantes, como la persistencia funcional, aunque el sistema sea dañado, y la capacidad de generalizar el conocimiento adquirido a nuevas situaciones. Todo esto parece muy atrayente, especialmente porque la arquitectura de estos sistemas es más similar a la del cerebro que la arquitectura en serie de las máquinas de calcular corrientes.

Este nuevo estilo de investigación en IA y ciencia cognitiva se denomina *conexionismo* o investigación PDP. El primer término se acuñó para indicar que los cálculos pueden

ser realizados no sólo por procesadores centrales sino también por el intrincado sistema de conexiones dentro del cual se une una gran cantidad de unidades de procesamiento extremadamente *simples*. El segundo término son las siglas de "Procesamiento de Distribución Paralela", expresión que connota la misma idea computacional. Algunas propiedades de estos sistemas y algunos resultados de esta investigación serán analizados hacia el final del próximo capítulo. Como los sistemas PDP están inspirados biológicamente en cierta medida, será más conveniente estudiarlos luego de haber aprendido algo sobre la estructura del cerebro.

## Lecturas complementarias


Boden, Margaret, *Artificial Intelligence and Natural Man*, Nueva York, Harvester Press, 1977.

Dennett, Daniel, "Artificial Intelligence as Philosophy and as Psychology", en *Philosophical Perspectives on Artificial Intelligence*, en M. Ringle (comp.). New Jersey, Humanities Press, 1979. Reproducido en Daniel Dennett, *Brainstorms*, Montgomery, VT, Bradford, 1978, Cambridge, MA, MIT Press.

Winston, P. H. y Brown, R. H., *Artificial Intelligence: An MIT Perspective*, vols. I y II, Cambridge, MA, MIT Press, 1979.

Marr, D. Y Poggio, T., "Cooperative Computation of Stereo Disparity", *Science*, vol. 194, 1976.

Dreyfus, Hubert, *What Computers Can't Do: The Limits of Artificial Intelligence*, edición corregida, Nueva York, Harper and Row, 1979.

Haugeland, J., *Artificial Intelligence: The Very Idea*, Cambridge, MA, MIT Press, 1985.

Holland, J., Holyoak, K., Nisbett, R. y Thagard, P., *Induction: Processes of Inference, Learning, and Discovery*, Cambridge, MA, MIT Press, 1986.

Rumelhart, D. y McClelland, J., *Parallel Distributed Processing: Essays in the Microstructure of Cognition*, Cambridge, MA, MIT Press, 1986.

# 7

# Neurociencia

## 1. Neuroanatomía: antecedentes evolutivos

Cerca de la superficie de los océanos de la Tierra, entre tres y cuatro mil millones de años atrás, el proceso de evolución puramente química provocado por el sol produjo algunas estructuras moleculares de *autoduplicación*. A partir de los trocitos moleculares en su medio inmediato, estas moléculas complejas podían catalizar una secuencia de reacciones en cadena que dieron como resultado copias exactas de sí mismas. Con respecto al logro de grandes poblaciones, la capacidad de autoduplicación es simplemente una ventaja explosiva. Sin embargo, el crecimiento de la población está limitado por la disponibilidad de los trocitos adecuados en la sopa molecular circundante y por las diversas fuerzas en el medio ambiente que tienden a quebrar la resistencia de estas heroicas estructuras antes de que puedan duplicarse. Por lo tanto, entre las moléculas rivales que se autoduplican, la ventaja competitiva la obtendrán aquellas estructuras moleculares específicas que inducen, no sólo su propia duplicación, sino también la formación de estructuras que las protejan de depredaciones externas, y la formación de mecanismos que producen las partes moleculares necesarias para la manipulación química de las moléculas del medio que son inutilizables directamente.

La *célula* es el ejemplo triunfante de esta solución. Tiene una membrana externa para proteger las intrincadas estruc-

turas del interior y las complejas vías metabólicas que procesan el material de afuera para convertirlo en estructura interna. En el centro de este complejo sistema se encuentra una molécula de ADN cuidadosamente codificada, la directora de la actividad celular y la ganadora de la competencia descripta. Sus células dominan ahora la Tierra. Todos los competidores han sido barridos por su fenomenal éxito, excepto los virus residuales, que son los únicos que siguen con la estrategia inicial, ahora como invasores parásitos sobre el éxito celular. Con el surgimiento de la célula tenemos lo que encaja en nuestro concepto corriente de *vida:* un sistema de autoconservación, de autoduplicación y consumidor de energía.

El surgimiento de la inteligencia consciente, como un aspecto de la materia viva, debe considerarse en comparación con la historia de la evolución biológica en general. Aquí tomamos la historia luego de que ya ha comenzado: luego de que los organismos multicelulares han aparecido, cerca de mil millones de años atrás. La inteligencia significativa requiere un sistema nervioso, y los organismos unicelulares como las algas o las bacterias no pueden tenerlo, ya que un sistema nervioso es una organización de muchas células.

La principal ventaja de ser un organismo multicelular es que las células individuales pueden especializarse. Algunas pueden formar una pared externa dura dentro de la que otras células pueden disfrutar de un medio más estable y beneficioso que el océano en alta mar. Estas células enclaustradas pueden ejercer sus propias especialidades: digestión de alimentos, transporte de nutrientes a otras células, contracción y elongación para producir movimiento, sensibilidad a factores ambientales claves (la presencia de alimentos o depredadores), y así sucesivamente. El resultado de una organización de este tipo puede ser un sistema que sea más duradero que cualquiera de sus partes y con muchas más probabilidades de reproducirse que cualquiera de sus competidores unicelulares.

Sin embargo, la coordinación de estas partes especializadas requiere *comunicación* entre células y algunas especializaciones adicionales deben ocuparse de esta importante tarea.

Es inútil tener músculos si sus contracciones no pueden ser coordinadas para producir la locomoción útil o la masticación o la eliminación. Las células sensoriales son inútiles si su información no puede transmitirse al sistema motor. Y así sucesivamente. La comunicación puramente química es útil para algunos propósitos: el crecimiento y la reparación están regulados de este modo, con células mensajeras que transmiten sustancias químicas específicas por todo el organismo, a las que responden células determinadas. Pero éste es un medio de comunicación demasiado lento y ambiguo para la mayoría de los propósitos.

Afortunadamente, las células tienen las características básicas necesarias para servir como eslabones de comunicación. La mayoría de las células tienen una pequeña diferencia voltaica —una *polarización*— a través de las superficies interna y externa de las membranas celulares que las envuelven. Una perturbación apropiada en cualquier punto de la membrana puede provocar una repentina *des*polarización en ese punto y, como la caída de una fila de fichas de dominó precariamente paradas sobre un lado, la despolarización se extenderá a cierta distancia por la superficie de la célula. Luego de esta despolarización, la célula se infla resueltamente otra vez. En la mayoría de las células el impulso de la despolarización se atenúa y muere al poco tiempo, pero en otras no. Asociemos esta conveniente propiedad de las células con el hecho de que las células unitarias pueden adoptar formas muy elongadas —filamentos de un metro o más en casos extremos— y que contamos con los elementos perfectos para un sistema de comunicación: células nerviosas especializadas que conducen impulsos electroquímicos a través de largas distancias a alta velocidad.

Otras especializaciones son posibles. Algunas células se despolarizan al recibir presión física, otras con cambios de temperatura, otras con cambios de iluminación y aun otras al recibir impulsos determinados de otras células. Con la articulación de estas células tenemos los comienzos del sistema sensorial y nervioso central, y abrimos un nuevo capítulo en la historia de la evolución.

## El desarrollo de los sistemas nerviosos

La aparición de sistemas de control nervioso no debe considerarse como algo milagroso. Para apreciar con qué facilidad un sistema de control puede llegar a caracterizar a toda una especie, consideremos una criatura imaginaria parecida a un caracol que vive en el fondo del océano. Esta especie debe salir un poco de su concha para alimentarse, y se retira al interior sólo cuando está saciada o cuando algún cuerpo externo entra en contacto directo con ella, como cuando un depredador la ataca. Muchas de estas criaturas están perdidas frente a los depredadores, a pesar del reflejo táctil de retirada, ya que muchas mueren al primer contacto. Aun así, la población de la especie es estable y se mantiene en equilibrio con la población de los depredadores.

Casualmente, todos los caracoles de esta especie tienen una banda de células fotosensibles en la parte trasera de la cabeza. No hay nada notable en esto. Muchas clases de células resultan ser fotosensibles en cierto grado y su sensibilidad a la luz es un rasgo incidental de la especie, un rasgo que no hace nada. Supongamos ahora que un caracol, debido a una pequeña mutación en la codificación del ADN original, tiene más cantidad de la usual de células nerviosas que conectan la superficie de la piel con los músculos de retracción. En particular, está solo entre sus congéneres al tener conexiones entre sus células fotosensibles y los músculos de retracción. Así los cambios repentinos en la iluminación general causan una inmediata retracción hacia el interior de la concha.

Ese rasgo incidental en este individuo carecería de importancia en muchos medios, sería un mero "crispamiento" idiosincrático que de todos modos no es de utilidad. Sin embargo, en el verdadero medio de los caracoles los cambios repentinos de iluminación son causados con la mayor frecuencia por los *depredadores* que nadan directamente por arriba. Por lo tanto, nuestro individuo mutante posee un "sistema de advertencia precoz" que le permite retraerse a la seguridad *antes* que el depredador lo alcance. Sus probabilidades de supervivencia y de reproducción repetida son así mucho mayores que las de sus compañeros desarmados. Y como

su nueva posesión es el resultado de una mutación genética, gran parte de su descendencia también la tendrá. Sus probabilidades de supervivencia y de reproducción han aumentado de modo similar. Evidentemente, este rasgo rápidamente llegará a dominar en la población de los caracoles. Los grandes cambios se realizan a partir de estos sucesos pequeños y fortuitos.

Se puede pensar en un mayor aprovechamiento. Si por una mutación genética una superficie fotosensible se curva en forma de una concavidad hemisférica, las porciones iluminadas selectivamente proporcionarán información *direccional* acerca de las fuentes de luz y las oclusiones, información que puede producir respuestas motrices direccionales. En una criatura móvil como un pez, esto puede tener una importante ventaja, tanto para el cazador como para la presa. Una vez distribuida ampliamente, una concavidad hemisférica puede transformarse en una concavidad casi esférica con solo una abertura minúscula hacia el exterior. Esta abertura formará una débil *imagen* del mundo exterior sobre la superficie fotosensible. Las células transparentes pueden cubrir la abertura, funcionando primero como protección y luego como una lente para imágenes superiores. Entretanto, una mayor inervación (concentración de células nerviosas) en la "retina" es recompensada por una información superior para llevar a otros lugares del sistema nervioso. Con estas simples y ventajosas etapas es como se ensambla el "milagroso" ojo. Y esta reconstrucción no es mera especulación. Se puede encontrar una criatura contemporánea por cada una de las etapas del desarrollo enumeradas.

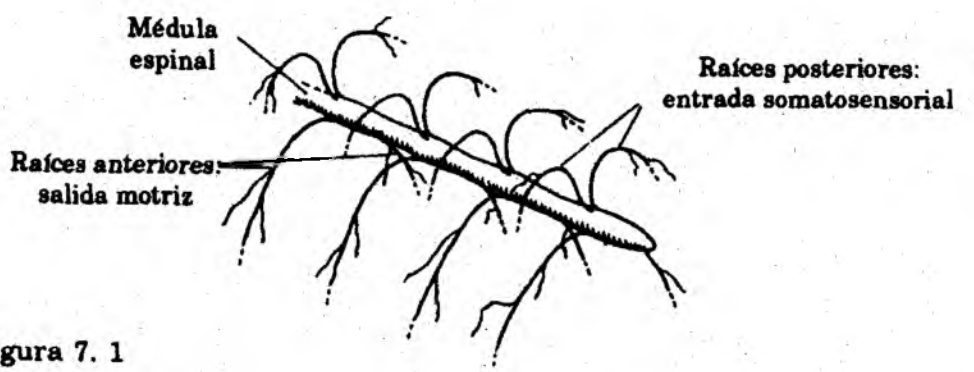



Figura 7. 1

En general, nuestra reconstrucción de la historia evolutiva de los sistemas nerviosos se basa en tres tipos de estudios: restos fósiles, criaturas actuales de estructura primitiva y el desarrollo nervioso en los embriones. Al ser tan suave, el tejido nervioso no se fosiliza, pero podemos rastrear la estructura nerviosa en los vertebrados extintos (animales con espina dorsal) a partir de las cámaras, pasajes y hendiduras que se hallan en los cráneos y espinas dorsales de animales fósiles. Es una guía muy confiable para medir toda la estructura, pero faltan por completo los detalles finos. Para ello, observamos el mundo animal existente, que contiene miles de especies cuyos sistemas nerviosos parecen haber cambiado muy poco en el curso de millones de años. Aquí debemos ser cuidadosos, ya que "simple" no significa necesariamente "primitivo", pero podemos construir "árboles" de desarrollo posibles a partir de su estudio. El desarrollo embrionario constituye una verificación fascinante para ambos estudios pues parte (sólo *parte*) de la historia evolutiva de cualquier criatura está escrita en la secuencia de desarrollo mediante la cual el ADN convierte un óvulo fertilizado en una criatura de ese tipo. Reuniendo a los tres, surge la siguiente historia.

Los vertebrados más primitivos poseían un *ganglio* (un racimo de células) central elongado a todo lo largo de la espina dorsal, conectado con el resto del cuerpo por dos conjuntos funcional y físicamente diferentes de fibras (figura 7.1). Las fibras *somatosensitivas* traían información sobre la actividad muscular y la estimulación táctil a la médula espinal, y las fibras *motoras* llevaban impulsos de órdenes desde allí a los tejidos musculares del cuerpo. La médula espinal funcionaba para coordinar todos los músculos del cuerpo entre sí para producir un movimiento coherente de natación y para coordinar dicho movimiento con la circunstancia percibida para poder huir de un ataque táctil o realizar un movimiento de búsqueda para aliviar un estómago vacío.

En criaturas posteriores esta *médula espinal* primitiva adquirió una elongación en el extremo anterior y tres abultamientos, donde la población y densidad de células nerviosas alcanza nuevos niveles. Este cerebro primitivo o *tallo cerebral* puede dividirse en *cerebro anterior, cerebro medio* y *cerebro*

*posterior* (figura 7.2). La red nerviosa del pequeño cerebro anterior entonces se dedicaba al procesamiento de los estímulos olfativos, el cerebro medio procesaba la información visual y auditiva, y el cerebro posterior se especializaba en una coordinación aun más sofisticada de la actividad motriz. Los cerebros de los peces contemporáneos han quedado en esta etapa, siendo el cerebro medio la estructura dominante.

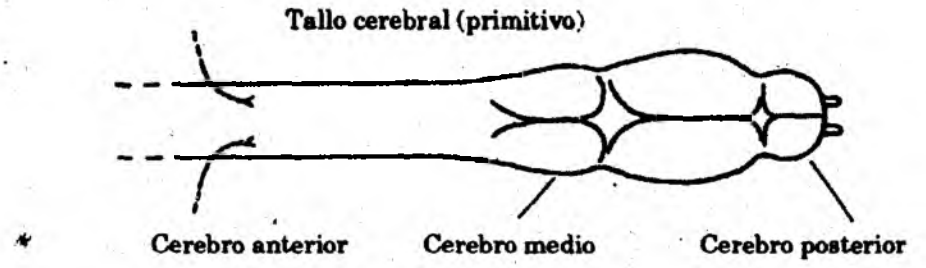



Figura 7. 2

En animales más avanzados, como los anfibios y reptiles, es el cerebro anterior el que domina la anatomía del tallo cerebral y el que asume un papel central en el procesamiento de todas las modalidades sensoriales, no sólo el olfato (figura 7.3). En muchos animales también aumenta el tamaño absoluto, y con él las cantidades absolutas de células en lo que ya es una red de control compleja y cuasi autónoma. Esta red tenía mucho que hacer: muchos dinosaurios eran carnívoros bípedos veloces que perseguían a presas distantes gracias a una excelente visión. Era esencial un sistema de control superior para que ese nicho ecológico fuera ocupado con éxito.

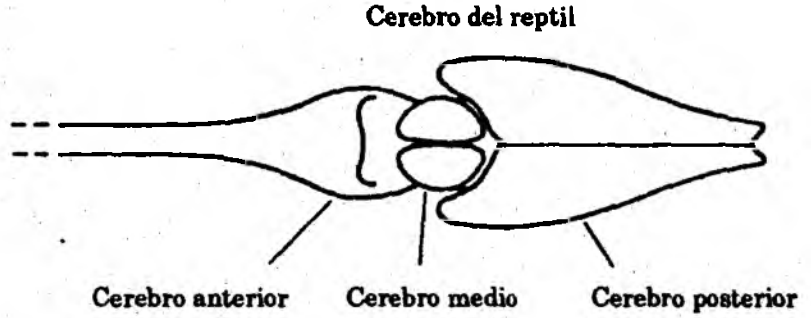



Figura 7. 3

Cerebro del mamífero

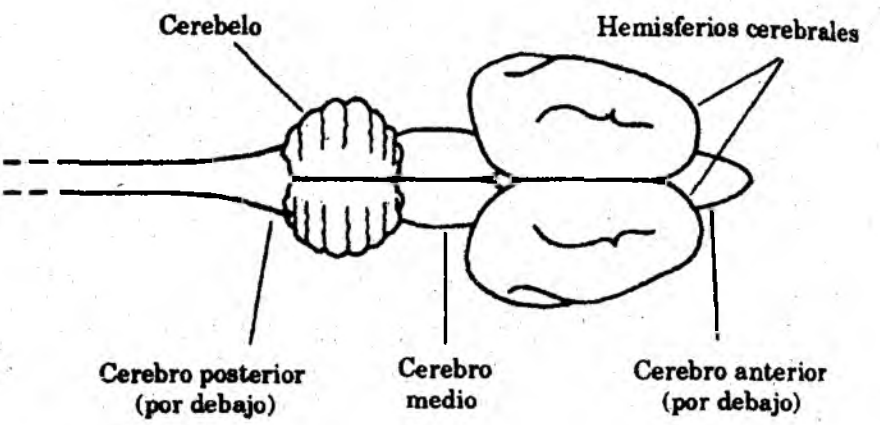



Figura 7.4

Los cerebros de los primeros mamíferos presentaban una mayor articulación y especialización del cerebro anterior y, lo que es más importante, dos estructuras totalmente nuevas: los *hemisferios cerebrales* que se encontraban a cada lado de la parte superior de un cerebro anterior más grande, y el *cerebelo* en el dorso del cerebro posterior (figura 7.4). Los hemisferios cerebrales contenían una cantidad de áreas especializadas, incluyendo el más alto control para la iniciación de la conducta, y el cerebelo proporcionaba una coordinación aun mejor del movimiento corporal en un mundo de objetos en movimiento relativo. La cantidad de células en la corteza cerebral y cerebelosa (la delgada superficie en la que se concentran los cuerpos celulares y las conexiones intercelulares) también es sorprendentemente mayor que la cantidad hallada en la corteza más primitiva de los reptiles. Esta capa cortical (la clásica "materia gris") es dos a seis veces más gruesa en los mamíferos.

En los mamíferos típicos estas nuevas estructuras, aunque prominentes, no son grandes en relación con el tallo cerebral. Sin embargo, en los primates se han convertido en los rasgos dominantes del cerebro, al menos a primera vista. Y en el ser humano son enormes (figura 7.5). El tallo cerebral original es difícil de ver debajo del paraguas de los hemisferios

cerebrales, y el cerebelo también es marcadamente mayor, comparado con lo que presentan otros primates. Es difícil resistir a la sospecha de que lo que nos distingue de los otros animales, en la medida en que somos diferentes, se encuentra en el gran tamaño y en las propiedades inusuales de los hemisferios cerebrales y cerebelosos humanos.

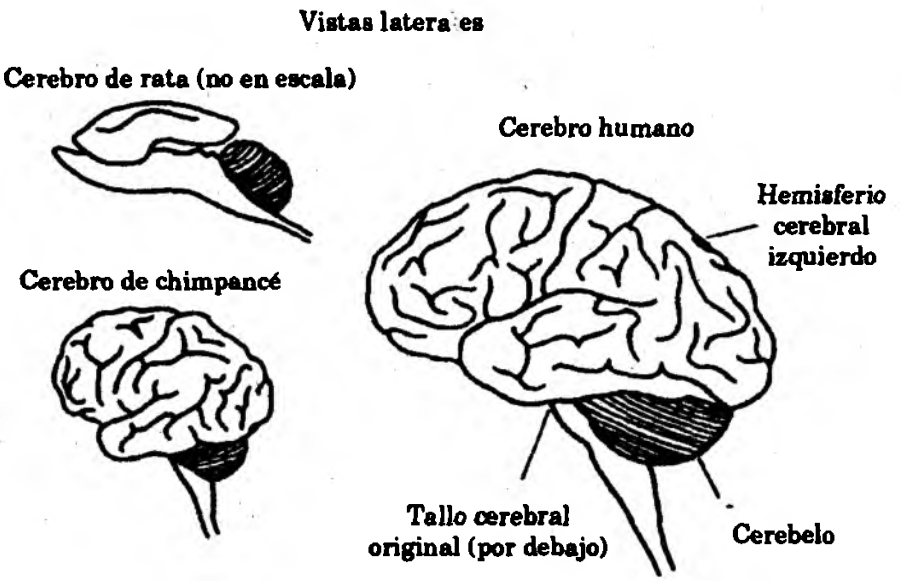



Figura 7.5

## Lecturas complementarias


Bullock, T. H., Orkand R. y Grinnell, A., *Introduction to Nervous Systems*, San Francisco, Freeman, 1977.

Sarnat, H. B. y Netsky, M. G., *Evolution of the Nervous System*, Oxford, Oxford University Press, 1974.

Dawkins, Richard, *The Selfish Gene*, Oxford, Oxford University Press, 1976.

# 2. Neurofisiología y organización nerviosa

## A. Elementos del sistema: neuronas

### Estructura y función

Las células alargadas que llevan impulsos mencionadas anteriormente se denominan *neuronas*. Una típica neurona multipolar tiene la estructura física descripta en la figura 7.6: una estructura en forma de árbol con *dendritas* ramificadas para la entrada y un único *axón* para la salida. (El axón está plegado por necesidad del diagrama.) Esta estructura refleja lo que parece ser la función principal de la neurona: el procesamiento de entradas desde otras células. Los axones de muchas otras neuronas hacen contacto con las dendritas de una neurona dada o con el cuerpo celular mismo. Estas conexiones se llaman *sinapsis* y permiten que lo que ocurre en una célula influya en la actividad de otra (figura 7.7).

La influencia se logra de las siguientes maneras. Cuando un impulso de despolarización —denominado *potencial de acción* u *onda*— recorre el axón hasta su(s) extremo(s) presináptico(s), su llegada hace que el bulbo terminal libere una sustancia química denominada *neurotransmisor* a través de la diminuta hendidura sináptica. Según la naturaleza del neurotransmisor característico del bulbo y la naturaleza de los receptores químicos que lo reciben del otro lado de la hendidura, la sinapsis será *inhibidora* o *excitadora*.

En una sinapsis inhibidora, la transmisión sináptica causa una leve *hiper*polarización o elevación del potencial eléctrico de la neurona afectada. Esto hace menos probable que la neurona afectada sufra una despolarización repentina y dispare su propia onda por su propio axón.

En una sinapsis excitadora, la transmisión sináptica causa una leve *des*polarización de la neurona afectada, disminuyendo su potencial eléctrico hasta el punto crítico mínimo en que cae repentinamente, comenzando su propia onda de salida axonal. Entonces una sinapsis excitadora hará *más* probable que la neurona afectada dispare.

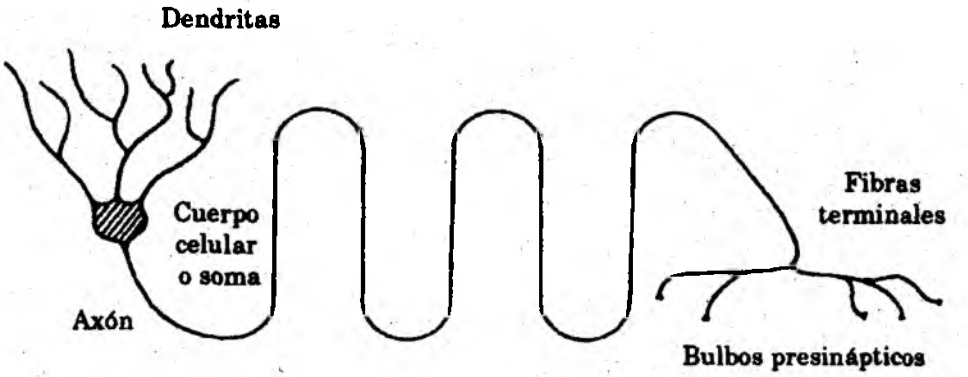



Figura 7. 6

Uniendo ambos factores, cada neurona es el lugar de una competencia entre las instrucciones de "disparar" y "no disparar". Cuál gana está determinado por dos cosas. Primero, la distribución relativa de las sinapsis excitadoras e inhibidoras es de gran importancia: sus cantidades relativas y quizá su proximidad al cuerpo celular principal. Si predomina una clase, como sucede con frecuencia, entonces el juego está arreglado, para esa neurona, en favor de una respuesta contra la otra.

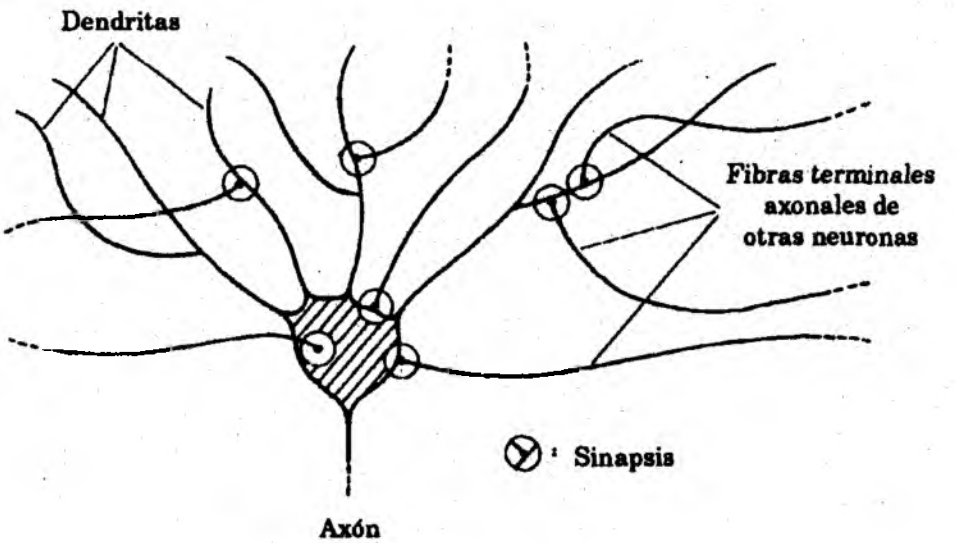



Figura 7. 7

(En el corto plazo, estas conexiones son un rasgo relativamente estable de cada neurona. Pero aparecen nuevas conexiones y se pierden las viejas, algunas veces en una escala temporal de sólo minutos o menos; de allí que las propiedades funcionales de una neurona sean en cierto modo plásticas.)

El segundo determinante de la conducta neuronal es la frecuencia temporal absoluta de las entradas desde las sinapsis de cada clase. Si 2000 sinapsis inhibidoras están sumamente activas sólo una vez por segundo, y 200 sinapsis excitadoras están activas 50 veces por segundo, entonces la influencia excitadora predominará y la neurona disparará. Luego de la repolarización disparará una y otra vez con una significativa frecuencia propia.

Es conveniente tener en cuenta cuáles son los números pertinentes aquí. Un típico soma neuronal quedará casi enterrado bajo una capa de varios cientos de bulbos terminales sinápticos y su árbol de dendritas tendrá conexiones sinápticas con varios miles más. Además, las neuronas vuelven a inflarse nuevamente al potencial de descanso en aproximadamente menos de 1/100 segundo; de aquí que puedan mantener frecuencias de onda de hasta 100 hertz (= 100 ondas por segundo) o más. Evidentemente, una sola neurona es un procesador de información de gran capacidad.

Inevitablemente las neuronas se comparan con las compuertas lógicas en la UCP de un ordenador digital. Pero las diferencias son tan intrigantes como las similitudes. Una sola compuerta lógica recibe información de no más de dos fuentes diferentes; una neurona recibe información de mucho más de mil. Una compuerta lógica emite salidas a una frecuencia metronómica, $10^6$ hertz, por ejemplo; una neurona varía libremente entre 0 y $10^2$ hertz. La salida de la compuerta lógica está y debe estar temporalmente coordinada con la de todas las demás compuertas; las salidas neuronales no están así coordinadas. La función de una compuerta lógica es la transformación de información binaria (conjuntos de ENCENDIDOS y APAGADOS) en otra información binaria; la función de una neurona, si podemos incluso hablar en singular en este caso, parece más admisiblemente ser la transformación de conjun-

tos de *frecuencias* ondulatorias en otras *frecuencias* ondulatorias. Y finalmente, las propiedades funcionales de una compuerta lógica son fijas; las de una neurona son decididamente plásticas, ya que el crecimiento de nuevas conexiones sinápticas y la poda o degeneración de las viejas puede cambiar la función de entrada/salida de la célula. Las ramas de las dendritas pueden tener nuevas espinitas en minutos para efectuar nuevas conexiones sinápticas y estos cambios son inducidos, en parte, por la actividad neuronal previa.

Si las neuronas son dispositivos de procesamiento de información, como casi seguramente lo son, su modo básico de operar es por lo tanto muy diferente del que presentan las compuertas lógicas de una UCP. Esto no significa que los sistemas de esta última, adecuadamente programados, no puedan simular las actividades de las primeras. Supuestamente podrían. Pero necesitamos saber en realidad más acerca de las propiedades funcionales plásticas de las neuronas y mucho más acerca de sus innumerables interconexiones, antes de poder simular exitosamente su actividad conjunta.

## Clases de neuronas

Una clasificación inicial presenta tres clases de neuronas: neuronas *motoras,* neuronas *sensitivas* y una gran variedad de *interneuronas* (es decir, todas las demás). Las neuronas motoras primarias se encuentran casi exclusivamente en la médula espinal y se definen como aquellas neuronas cuyos axones hacen sinapsis directamente sobre una célula muscular. Los axones de las neuronas motoras son algunos de los más largos del sistema nervioso, extendiéndose desde las profundidades de la médula espinal pasando por las raíces anteriores (véase la figura 7.1) entre las vértebras de la columna y por las extremidades hacia los músculos periféricos más distantes. Las neuronas motoras aseguran la contracción muscular graduada por dos medios: la frecuencia ondulatoria de cada neurona motora y el reclutamiento progresivo de neuronas inicialmente quiescentes que inervan al mismo músculo.

Las neuronas sensitivas se encuentran en mayor variedad y se definen convencionalmente como aquellas cuyo estímulo de entrada es alguna dimensión del mundo exterior al sistema nervioso. Por ejemplo, los bastoncillos y conos, células receptoras de la retina, son muy pequeños, sin axón y sin dendritas, y hacen sinapsis inmediata con neuronas más específicas de la capa contigua. Su labor es únicamente transformar la luz recibida en impulsos sinápticos. En contraste, las células somatosensitivas son tan largas como las neuronas motoras. Sus axones se extienden desde la piel y los músculos hasta la médula espinal a través de las *raíces posteriores* (véase la figura 7.1) y encuentran sus primeras sinapsis en las profundidades de la médula espinal. Su labor es transmitir información táctil, de dolor y de temperatura, así como también información sobre las extensiones y contracciones musculares: las siempre cambiantes posiciones del cuerpo y sus extremidades. Otras células sensitivas tienen su propia idiosincrasia dictada por la naturaleza de los estímulos físicos a los que responden.

Las interneuronas centrales también vienen en una gran variedad de formas y tamaños, aunque todas parecen variaciones sobre un mismo tema: entrada dendrítica y salida axónica. La mayoría de ellas, llamadas células multipolares, tienen muchas ramificaciones dendríticas, que surgen directamente del cuerpo celular. Otras, denominadas células bipolares, tienen sólo un filamento dendrítico que nace y se ramifica en un punto a cierta distancia de la célula. Algunas, como las células de Purkinje del cerebelo, tienen ramificaciones dendríticas extraordinariamente extensas y tupidas. Otras tienen sólo extensiones dendríticas dispersas. Los axones de muchas neuronas se proyectan a través de todo el cerebro, haciendo las sinapsis en puntos distantes. Otros hacen meramente conexiones locales entre las concentraciones extendidas de neuronas cuyos axones se proyectan en otros lugares.

Estas capas formadas por un gran número de neuronas interconectadas constituyen la *corteza*. La superficie externa de cada hemisferio cerebral es una gran lámina delgada de

corteza, muy plegada sobre sí misma como papel arrugado para maximizar el área total dentro del pequeño volumen del cráneo. Las conexiones interneuronales del cerebro se encuentran en su mayor densidad en esta capa plegada. La superficie del cerebelo también es corteza y los "núcleos" corticales especializados están distribuidos por todo el tallo cerebral. Estos se ven como áreas grises en los cortes transversales del cerebro. Las demás áreas blancas contienen proyecciones axonales desde un área cortical a otra. Lo que nos lleva al tema de la organización cerebral.

## B. Organización de la red

Encontrar la organización de un sistema tan complejo como el cerebro humano es una tarea difícil. Ya se conoce mucho sobre la estructura, pero tanto o más queda aún por descubrirse. Se puede analizar la estructura a gran escala de las interconexiones neuronales utilizando colorantes especiales que una neurona toma y transporta por el axón hasta las sinapsis terminales. Si deseamos saber adónde llegan los axones de una zona teñida, cortes transversales sucesivos del cerebro revelarán tanto el trayecto de estos axones teñidos a través de la sustancia blanca como la región de su sinapsis final. Esta técnica, aplicada a cerebros post mortem, ha revelado las principales interconexiones entre las diversas áreas corticales del cerebro, las "superautopistas" que abarcan miles de axones unidos. Sin embargo, conocer su posición no siempre permite conocer sus funciones, y las autopistas neuronales más pequeñas y los desvíos constituyen un horizonte de detalles cada vez más pequeños que desafía los intentos de realizar un resumen completo.

Con microscopios, cortes micrométricos y una variedad de otras técnicas de tinción, la microarquitectura del cerebro comienza a aparecer. La corteza cerebral revela seis capas distintas, que se distinguen por la densidad de sus neuronas dentro de ellas y por la clase de neuronas que contienen. La comunicación interneuronal se extiende tanto dentro de las

capas como a través de ellas. Los detalles son complejos y oscuros, y el punto de esta estructura en particular sigue siendo un misterio, pero nos aferramos a lo que descubrimos y tratamos de utilizarlo para encontrar más. Casualmente, esta citoarquitectura de seis capas no es completamente uniforme en toda la corteza cerebral: el grosor o la densidad de ciertas capas se ve disminuido o aumentado en determinadas regiones de la superficie cortical. La búsqueda de regiones de idéntica arquitectura y el trazado de sus límites nos ha llevado a identificar alrededor de cincuenta áreas corticales diferentes, conocidas como *áreas de Brodmann* en honor a su descubridor.

¿Tienen alguna otra importancia estas áreas? Muchas de ellas sí, tanto por sus propiedades funcionales como por sus conexiones más distantes. A continuación se describirán unos pocos casos significativos.

## Proyecciones sensoriales dentro del cerebro

Como se mencionó anteriormente, las neuronas somatosensitivas primarias entran en la médula espinal a través de las raíces posteriores, y encuentran sus primeras conexiones sinápticas con las neuronas en la médula. Aquellas neuronas conducen la información ascendiendo por la médula hasta el tálamo en el encéfalo, donde hacen sinapsis con neuronas de un área llamada núcleo ventral del tálamo. Estas neuronas a su vez se proyectan en los hemisferios cerebrales y en una región cortical claramente definida por tres áreas de Brodmann conectadas. Toda esta área se conoce como la *corteza somatosensitiva*. La lesión en diversas partes de ésta produce una pérdida permanente de sensibilidad táctil y propioceptiva en diversas partes del cuerpo. Más aún, un estímulo eléctrico leve en esta región produce en el sujeto sensaciones táctiles vívidas "localizadas" en partes específicas del cuerpo. (La cirugía cerebral para corregir lesiones en esta región ha brindado la oportunidad ocasional de realizar esta investigación y, como los individuos pueden estar completa-

mente conscientes durante la cirugía cerebral, pueden informar acerca de los efectos que producen dichos estímulos.)

De hecho, la corteza somatosensitiva constituye lo que se denomina un *mapa topográfico* del cuerpo, ya que la disposición espacial de neuronas anatómicamente específicas es una proyección de las regiones anatómicas. Cada hemisferio representa la mitad opuesta del cuerpo. El corte transversal de un hemisferio en la figura 7.8 ilustra esto. El esquema corporal distorsionado representa las áreas de la corteza correspondiente a la región del cuerpo dibujada a su lado, y las variaciones de tamaño representan la cantidad relativa de células corticales correspondientes a esa zona. Este esquema se denomina "homúnculo somatosensitivo".

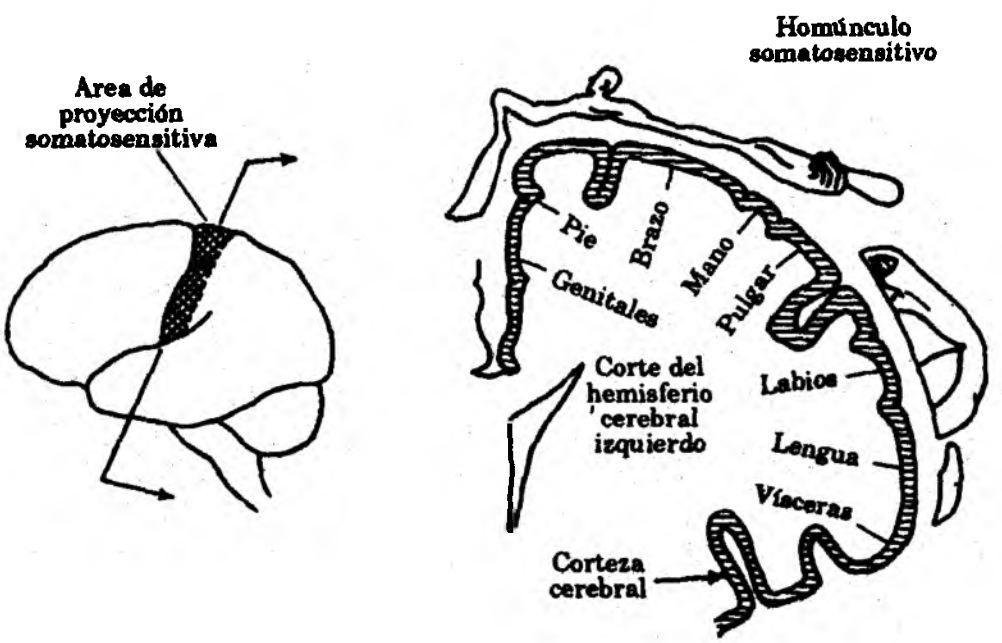



Figura 7. 8

La organización y función del sistema visual también toma contacto con la arquitectura de la corteza cerebral. Inmediatamente junto a los bastoncillos y conos de la retina hay una capa interconectada de pequeñas neuronas que realiza un

procesamiento inicial antes de hacer sinapsis con las largas células ganglionares. Estas se agrupan en un grueso haz y salen por detrás de la retina como nervio óptico. Este proyecta a un núcleo cortical (= una concentración local de cuerpos celulares interconectados) en la parte posterior del tálamo llamado *cuerpo geniculado lateral.* Las células aquí también constituyen un mapa topográfico de la retina, aunque está métricamente distorsionado ya que la fóvea, el centro físico y funcional de la retina, está extensamente representada.

Las células del geniculado lateral luego se proyectan a varias de las áreas de Brodmann en el extremo posterior de los hemisferios cerebrales: la corteza estriada y luego la corteza periestriada (figura 7.9). Estas áreas en conjunto se denominan *corteza visual,* y siguen constituyendo una proyección topográfica de la retina, en la que cada hemisferio representa una mitad de la superficie de la retina. Pero en la corteza visual y en su procesamiento precortical ocurre algo más de lo que ocurre en el sistema somatosensitivo, y la corteza visual representa algo más que sólo áreas de proyección retiniana. Los subgrupos de neuronas visuales resultan ser especializa-

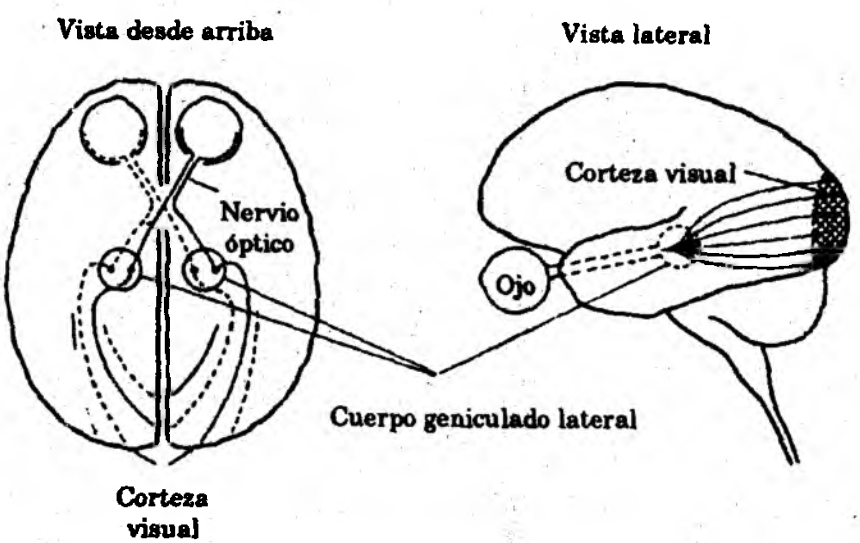



Figura 7. 9

dos, en sus respuestas, para estímulos sumamente específicos de la información visual. Una célula baja en la jerarquía neuronal es sensible sólo a *diferencias* de brillo dentro de su campo receptivo (= el área retiniana a la que es sensible). Pero una célula más especializada a la que llegan estímulos de aquellas células puede ser sensible sólo a líneas o bordes de una *orientación* en particular dentro de su campo receptivo. Células aun más especializadas son sensibles sólo a líneas o bordes que se *mueven* en una dirección en particular. Y así sucesivamente. Es inevitable la impresión de un sistema de procesamiento de datos acumulativo.

Otras microestructuras explican los rasgos de la visión binocular, especialmente, la sofisticada *estereoscopía* o visión tridimensional del ser humano. La estereoscopía requiere la comparación sistemática de las imágenes de cada ojo. Un estudio detallado revela la existencia de *columnas de dominancia ocular* intercaladas en la corteza visual. Una columna es un estrecho núcleo de células dispuestas verticalmente a través de las seis capas de la corteza, y cada una tiene un pequeño campo receptor en la retina. Estas columnas son específicas para cada ojo y su intercalación significa que los campos receptores izquierdo y derecho respectivamente están representados por columnas físicamente adyacentes en la corteza. Así puede realizarse la comparación de información y se han descubierto más células que sin duda son sensibles a disparidades binoculares entre dichos campos. Esas células responden a la información acerca de las distancias relativas de los objetos en el medio visual de cada uno. Estos descubrimientos abren líneas promisorias de investigación y la corteza visual actualmente concita mucho interés.

## Proyecciones motoras eferentes

Justo frente a la corteza somatosensitiva, al otro lado de una cisura bastante profunda, hay otra de las áreas de Brodmann ahora conocida como *corteza motriz*. También es un claro mapa topográfico, esta vez de los sistemas muscula-

res del cuerpo. La estimulación artificial de las neuronas corticales motoras produce movimiento en los músculos correspondientes. En la figura 7.10 se observa un "homúnculo motor".

Esto es sólo el comienzo de la historia funcional, por supuesto, ya que el control motor es una cuestión de *secuencias* bien sincronizadas de las contracciones musculares: —mas aún, secuencias coherentes con el medio percibido por el cuerpo. De acuerdo con ello, la corteza motriz tiene proyecciones axonales, no sólo hacia la médula, y por lo tanto a los músculos del cuerpo, sino también hacia el cerebelo y los ganglios basales, y recibe a su vez proyecciones desde ambos, principalmente a través del tálamo, que ya sabemos que es una fuente de información sensitiva. Por lo tanto, la corteza motriz es una parte muy integrada de la actividad cerebral general y, aunque parte de su salida va más o menos directamente a la médula —para proporcionar control independiente de los movimientos finos de los dedos, por ejemplo—, gran parte pasa por un intrincado procesamiento en el cerebelo y en la parte inferior del tallo cerebral antes de entrar en la médula espinal.

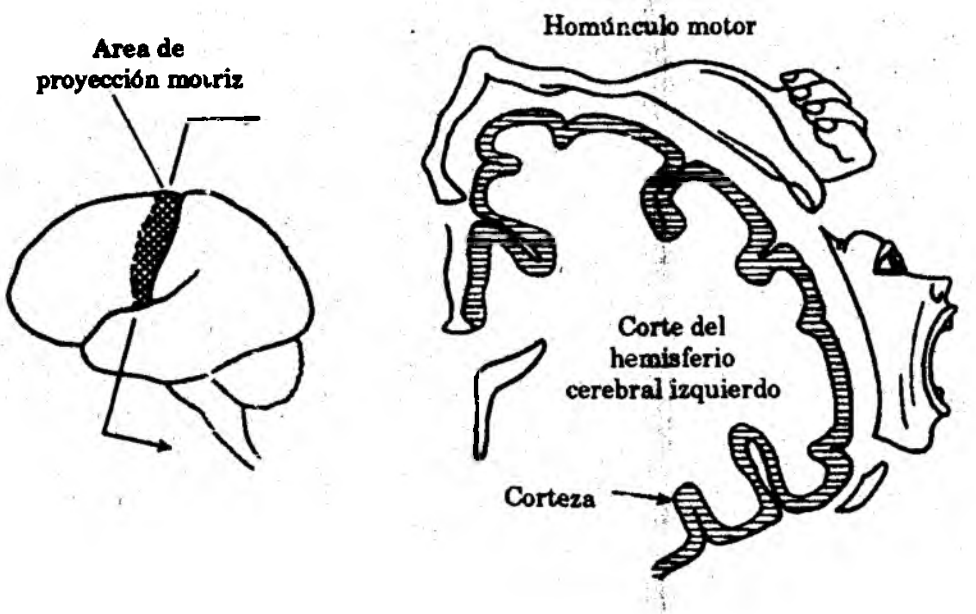



Figura 7. 10

Aquí debemos pensar en la salida del cerebro como una especie de "sintonía fina" de alto nivel de las habilidades motoras básicas, ya que la organización neuronal de la médula espinal misma es suficiente para producir la locomoción en la mayoría de los vertebrados. Un conocido ejemplo es la gallina sin cabeza cuyo cuerpo corre sin dirección de un lado a otro durante varios segundos después de haber sido degollada. Incluso los mamíferos pequeños cuyos cerebros han sido extirpados presentarán actividad locomotriz con un estímulo adecuado de la médula. Aquí tenemos un reflejo de lo *antigua* que es la capacidad de locomoción en los vertebrados: que comenzó a perfeccionarse cuando los vertebrados primitivos no tenían mucho más que una médula espinal. Las progresivas adiciones que sobrevivieron lo hicieron porque agregaron a esa capacidad inicial alguna sintonía fina útil o una orientación inteligente. La corteza motora es simplemente uno de los últimos y más elevados centros en una amplia jerarquía de controles motores. Estos se extienden desde los simples arcos reflejos —como quitar la mano de un horno caliente— hasta los centros más elevados, que formulan planes de acción abstractos y a largo plazo.

## Organización interna

El cerebro monitorea el mundo exterior a través de las neuronas sensitivas primarias; pero en el proceso también monitorea muchos aspectos de sus propias operaciones. Y el cerebro ejerce control sobre el mundo exterior, pero también sobre muchos aspectos de sus propias operaciones. Las proyecciones internas entre las partes del cerebro son ricas y extensas y son decisivas para su funcionamiento. Un buen ejemplo es la existencia de mecanismos de "control descendente". En nuestro primer análisis del sistema visual no mencioné que la corteza visual también envía proyecciones *de vuelta* al cuerpo geniculado lateral del tálamo, donde termina el nervio óptico. Esto significa que, según lo que la corteza visual obtiene del geniculado lateral, puede ejercer

una influencia sobre éste para *cambiar* lo que está recibiendo, quizá para resaltar ciertos estímulos de entrada o para suprimir otros. Aquí tenemos los elementos de cierta plasticidad en las actividades de procesamiento del cerebro, la capacidad para dirigir la atención y focalizar recursos. Las vías de control descendentes son especialmente prominentes en el sistema visual y en el auditivo, que debe procesar el habla, pero son comunes en todo el cerebro.

Entre las áreas sensitivas de la corteza aquí analizadas y otras análogamente identificadas hay mucho cerebro en gran actividad. Las extensas "áreas de asociación" entre las diversas clases de cortezas sensitivas no están bien entendidas, como tampoco las grandes regiones frontales de los hemisferios cerebrales, aunque es claro a partir de los casos de lesión cerebral que estos últimos están relacionados con las emociones, los impulsos y la capacidad de acción planificada.

Hay una hipótesis que explica estas regiones, su función y sus conexiones axonales con otras regiones. Consideremos la figura 7.11. Las regiones cuadriculadas son las de la corteza sensitiva *primaria:* somatosensitiva, auditiva y visual. Las áreas rayadas marcan la corteza sensitiva *secundaria*. Las células de la corteza primaria se conectan con células de la corteza secundaria, pues las tres modalidades sensitivas y estas células secundarias responden a estímulos más complejos y abstractos de la entrada sensorial que los de las células de la corteza primaria. A su vez, la corteza secundaria se conecta con las áreas en blanco, llamadas corteza *terciaria* o de *asociación*. Las células de la corteza de asociación responden a estímulos más abstractos aún de la entrada sensorial original, pero aquí hallamos una mezcla de células: algunas con capacidad de respuesta a la entrada visual, otras a la auditiva, otras a la táctil y otras a combinaciones de las tres. Parecería que el análisis más abstracto e integrado que el cerebro hace del medio sensorial se produce en la corteza de asociación entre las diversas áreas sensitivas.

Desde esta mitad posterior o "sensitiva" del cerebro la información puede abrirse paso a través de una variedad de vías subyacentes entre el cerebro medio o mesencéfalo y la

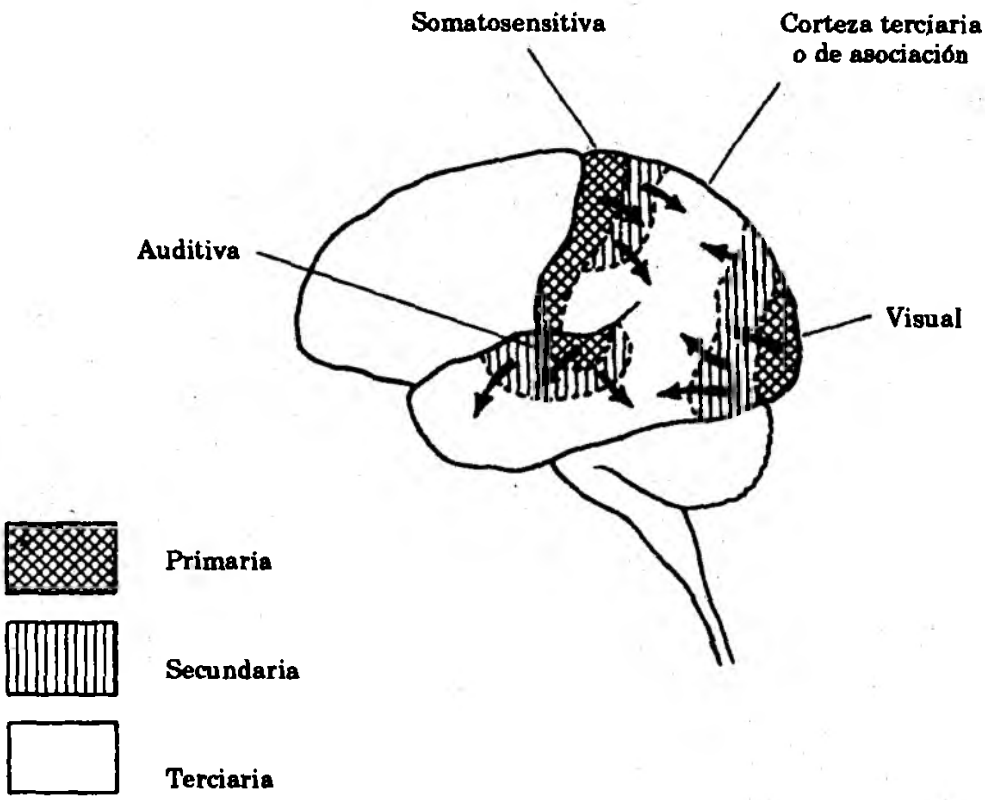



Figura 7. 11

mitad frontal o "motora" del cerebro, hasta lo que podemos llamar regiones motoras terciarias. Es el área frontal en blanco de la figura 7.12. Esta región parece ser la responsable de la formación de nuestros planes e intenciones más generales. Aquí las células se proyectan en la corteza motora secundaria, que parece ser el lugar de los planes más específicamente concebidos y de las secuencias de la conducta. Esta área finalmente se proyecta en la corteza motora primaria, que es la responsable de los movimientos más específicos de las diversas partes del cuerpo.

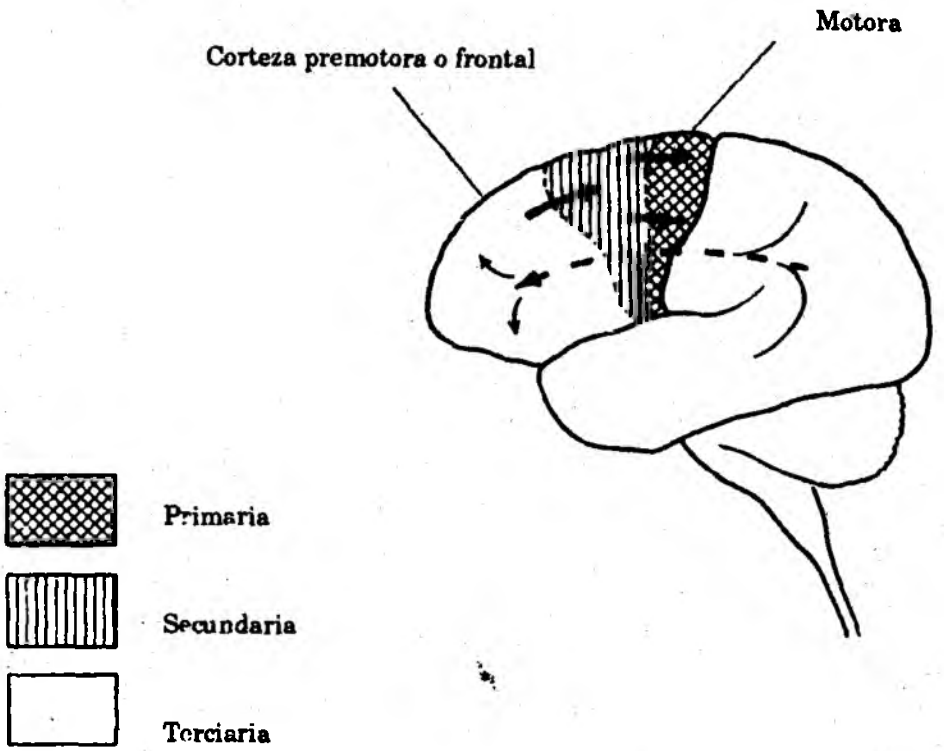



Figura 7. 12

Esta hipótesis es coherente con la neuroarquitectura del cerebro, con sus habilidades generales como control sensitivamente guiado de la conducta corporal, y con estudios detallados de las deficiencias cognitivas específicas producidas por lesiones en diversas zonas del cerebro. Por ejemplo, una lesión en el extremo del lóbulo frontal deja al paciente incapacitado para pensar en alternativas futuras posibles que trasciendan de los asuntos más inmediatos y simples o para distinguir cuidadosamente entre ellos.

El esquema precedente de la organización global del cerebro representa la idea clásica, pero el lector debe saber que constituye un panorama provisional y demasiado simplificado. Estudios recientes indican que diferentes mapas

topográficos de la retina están distribuidos por toda la superficie cortical y tienen diferentes proyecciones desde el geniculado lateral o desde otros lugares del tálamo. El sistema jerárquico de los mapas topográficos analizado anteriormente, que culmina con la "corteza visual secundaria" en la parte posterior del cerebro, es entonces uno de los varios sistemas paralelos, cada uno de los cuales procesa diferentes aspectos de la entrada visual. El sistema "clásico" para la visión puede ser el dominante, pero tiene compañía y todos estos sistemas interactúan entre sí. La "corteza somatosensitiva", que enfrenta complejidades similares, surge como sólo uno de los varios sistemas paralelos que procesan diferentes clases de información somatosensitiva: tacto suave, presión profunda, posición de las extremidades, dolor, temperatura y otros. Seleccionar las diferencias funcionales de entre estos diferentes mapas y encontrar sus interconexiones funcionales es una labor que recién ha comenzado. A medida que surge aquella información, nuestra apreciación de los logros intrincados y ocasionalmente insospechados de nuestro sistema perceptivo deberá aumentar en igual medida.

Es digna de mención otra área de interrogantes, no por su extensión sino porque es el último objetivo de una jerarquía de proyecciones desde áreas muy amplias y diversas de la corteza cerebral. El pequeño *hipocampo* se encuentra en el extremo posterior del sistema límbico, una estructura del telencéfalo justo debajo de los grandes hemisferios cerebrales. Si buscamos los orígenes de las entradas que recibe el hipocampo junto con el flujo de información que entra, deducimos con bastante rapidez toda la corteza cerebral. Resulta que una lesión en el hipocampo bloquea la transferencia de información de la memoria de corto plazo a la de largo plazo. Los pacientes con dicha lesión viven en un mundo de pesadilla con recuerdos que no llegan a más de unos pocos minutos atrás, excepto aquellos recuerdos originales de los hechos más distantes en el pasado, arraigados antes de que ocurriera la lesión.

Es natural pensar que el cerebro es algo insertado entre

los nervios sensitivos periféricos y los nervios motores periféricos, algo controlado por los primeros y que controla a los segundos. Desde un punto de vista evolutivo esto tiene sentido, por lo menos en las primeras etapas. Pero con el cerebro en el nivel de la articulación y automodulación que se encuentra en el ser humano ha aparecido cierta autonomía. La conducta está gobernada tanto por el aprendizaje pasado y por los planes a largo plazo como por las percepciones del momento. Y, a través del aprendizaje autodirigido, el desarrollo a largo plazo de la organización interna del cerebro está en cierto modo bajo el control del cerebro mismo. De esta forma no tratamos de escapar del mundo animal, sino que nos convertimos en sus miembros más creativos e impredecibles.

### Lecturas complementarias


Churchland, Patricia, *Neurophilosophy*, Cambridge, MA, MIT Press, 1986.

Hubel, D. H. y Wiesel, T. N., "Brain Mechanisms of Vision", *Scientific American*, vol. 241, 3, septiembre de 1979: número especial dedicado a las diversas ciencias del cerebro.

Bullock, T. H., Orkand, R. y Grinnell, A., *Introduction to Nervous Systems*, San Francisco, Freeman, 1977.

Kandel, E. R. y Schwartz, J. H., *Principles of Neural Science*, Nueva York, Elsevier/North-Holland, 1981.

Kandel, E. R., *The Cellular Basis of Behavior*, San Francisco, Freeman, 1976.

Shepherd, G. M., *Neurobiology*, Nueva York, Oxford University Press, 1983.


## 3. Neuropsicología

La neuropsicología es la disciplina que trata de entender y explicar los fenómenos psicológicos en términos de las actividades neuroquímicas, neurofisiológicas y neurofuncionales del cerebro. Ya hemos visto algunos resultados tentativos pero fascinantes en el apartado precedente: cómo la estructura jerárquica del sistema visual nos permite discriminar determinados estímulos de una escena, cómo las representaciones

retinianas intercaladas en la superficie cortical posibilitan la visión estereoscópica y cómo la organización general de la corteza hace posible que la información sensorial altamente procesada guíe la formación y ejecución de planes generales de acción.

Desafortunadamente, la mayor parte de los datos de los que la neuropsicología tradicionalmente dispone deriva de casos de lesiones, deterioro y desequilibrio cerebrales. Lo que mejor comprendemos es la base neural de la psicología *anormal.* El tejido cerebral puede ser físicamente perturbado por objetos invasivos; puede ser aplastado por tumores en crecimiento o por presión de fluidos; puede morir y atrofiarse por falta de suministro de sangre localizado, o puede ser destruido selectivamente por enfermedad o deterioro. Según la *posición* específica, dentro del cerebro, de la lesión producida por cualquiera de estas causas, generalmente resultan pérdidas específicas en las habilidades psicológicas del paciente.

Estas pérdidas pueden ser menores, como la incapacidad de identificar los colores (lesiones en las conexiones entre la corteza visual secundaria y la corteza auditiva secundaria del hemisferio izquierdo). O pueden ser graves, como la incapacidad permanente de reconocer caras, incluso la de los miembros de la familia (lesiones en la corteza de asociación del hemisferio derecho). Y pueden ser devastadoras, como la pérdida total y permanente de la comprensión del lenguaje (lesiones en la corteza auditiva secundaria del hemisferio izquierdo) o la incapacidad de retener nuevos recuerdos (daño bilateral en el hipocampo).

Mediante el examen post mortem y otras técnicas de diagnóstico, los neurólogos y neuropsicólogos pueden encontrar los correlatos neurales de estas y otras pérdidas en la función cognitiva y de conducta. De esta manera podemos armar lentamente un *mapa funcional* de todo el cerebro. Podemos llegar a apreciar las especializaciones funcionales y la organización funcional del cerebro en un ser humano *normal.* Esta información, en conjunto con una comprensión detallada de la neuroarquitectura y de la microactividad de las áreas pertinentes, puede conducir a una verdadera comprensión de

cómo se producen realmente las habilidades cognitivas. Recordemos lo estudiado sobre la extracción de estímulos y la visión estereoscópica en el sistema visual. Una vez que sabemos dónde buscarlas, podemos empezar a encontrar las estructuras nerviosas específicas que explican los rasgos específicos de la capacidad cognitiva en estudio. En general hay razones para ser optimistas, aunque nuestra ignorancia todavía nos impida una mayor comprensión.

La investigación funcional recién descripta requiere cautela en dos aspectos. Primero, la simple correlación de una lesión en la región *x* con la pérdida de alguna función cognitiva *F* no significa que la región *x* tenga la función *F*. Sólo significa que alguna parte de la región *x* está específicamente involucrada de algún modo en la ejecución de *F*. Las estructuras nerviosas claves que mantienen a *F* pueden estar situadas en cualquier otro lugar, o quizá ni siquiera estén en ningún lugar fijo, sino que estén distribuidas en extensas regiones del cerebro.

Segundo, no debemos esperar que las pérdidas funcionales y las localizaciones funcionales que encontramos siempre correspondan claramente a funciones cognitivas representadas en nuestro vocabulario psicológico del sentido común. Algunas veces la deficiencia es difícil de describir, como cuando abarca un cambio global en la personalidad del paciente, y algunas veces su descripción es difícil de creer. Por ejemplo, algunas lesiones producen una pérdida total de la conciencia tanto perceptiva como práctica, de la *mitad izquierda* del universo del paciente, incluyendo el propio cuerpo (hemiabandono). El paciente vestirá sólo la mitad derecha del cuerpo e incluso negará tener el brazo izquierdo. Otras lesiones permiten al paciente escribir prosa lúcida y legible, pero le *impiden* leer y entender lo que él o cualquier otra persona ha escrito, aunque su visión sea completamente normal (alexia sin agrafia). Otras lesiones dejan al paciente "ciego", en el sentido de que su campo visual ha desaparecido e insiste en que no puede ver; y sin embargo, puede "adivinar" dónde se ha colocado una luz frente a él con una exactitud cercana al 100 por ciento (visión ciega). Otras lesiones dejan al paciente

realmente ciego, pero él insiste aviesamente en que *puede* ver perfectamente, mientras tropieza por toda la habitación inventando excusas por su conducta torpe (negación de la ceguera).

Estos casos son sorprendentes y confusos y están relacionados con los conceptos conocidos de la psicología ordinaria. ¿Cómo es posible ser ciego y no saberlo? ¿Ver sin campo visual? ¿Escribir perfectamente pero no leer ni una palabra? ¿O negar sinceramente que se tienen brazos o piernas? Estos casos violan las expectativas arraigadas. Pero no podemos esperar que la psicología ordinaria represente algo más que una etapa en el desarrollo histórico de nuestra autocomprensión, una etapa que las neurociencias pueden ayudarnos a superar.

Por debajo del nivel del daño estructural de la maquinaria nerviosa, está el nivel de la actividad química y de las anormalidades químicas. El lector recordará que la transmisión a través de la unión sináptica es un elemento fundamental en toda la actividad nerviosa, y que dicha transmisión es de naturaleza química. Al recibir un impulso o estímulo, el bulbo terminal axonal libera una sustancia química llamada *neurotransmisor* que rápidamente se propaga por la cisura sináptica para interactuar con los receptores químicos del otro lado. Esta interacción conduce a la descomposición de la sustancia neurotransmisora y los productos de dicha descomposición son recibidos nuevamente por el bulbo terminal para ser resintetizados y reutilizados.

Evidentemente, cualquier cosa que frustre o exagere estas actividades químicas tendrá un efecto profundo sobre la comunicación nerviosa y sobre la actividad nerviosa colectiva. Este es precisamente el modo en que las drogas psicoactivas producen sus efectos. Las diversas clases de neuronas utilizan diferentes neurotransmisores, y diferentes drogas tienen diferentes efectos sobre su actividad, de modo que se produce una amplia variedad de efectos tanto químicos como psicológicos. Una droga puede bloquear la síntesis de un neurotransmisor específico; o unirlo al lugar de sus receptores, bloqueando así su efecto; o bloquear la absorción de los productos de su

descomposición, retardando así su nueva síntesis. Por otro lado, una droga puede intensificar la síntesis, aumentar los lugares de recepción o acelerar la absorción de los productos de la descomposición. El alcohol, por ejemplo, es un enemigo de la acción de la noradrenalina, un importante neurotransmisor, mientras que las anfetaminas intensifican su actividad produciendo el efecto psicológico completamente opuesto.

Lo más importante es que dosis extremas de determinadas drogas psicoactivas producen síntomas que se asemejan mucho a los de las formas principales de enfermedades mentales: depresión, manía y esquizofrenia. Esto lleva a la hipótesis de que estas enfermedades, cuando ocurren naturalmente, implican la misma anormalidad neuroquímica que producen artificialmente estas drogas. Estas hipótesis concitan un interés mucho más que puramente teórico porque si son ciertas, entonces la enfermedad que ocurre naturalmente puede ser corregida o controlada por una droga con un efecto neuroquímico exactamente opuesto. Y así parece ser, aunque la situación es compleja y los detalles confusos. La *imipramina* controla la depresión, el *litio* controla la manía y la *clorpromacina* controla la esquizofrenia. No lo hacen perfectamente, pero el éxito moderado de estas drogas brinda un fuerte sustento a la idea de que los pacientes con enfermedades mentales son víctimas principalmente de una circunstancia puramente química, cuyos orígenes son más metabólicos y biológicos que sociales o psicológicos. De ser así, es un hecho importante, ya que más del 2 por ciento de la población humana se ve expuesto en forma significativa a una de estas afecciones en algún momento de su vida. Si podemos descubrir la naturaleza y los orígenes de los complejos desequilibrios químicos que subyacen en las principales formas de enfermedad mental, podríamos curarlas directamente o incluso prevenirlas por completo.

**Lecturas complementarias**

Kolb, B. y Whishaw, I. Q., *Fundamentals of Human Neuropsychology*, San Francisco, Freeman, 1980.

Gardner, H., *The Shattered Mind*, Nueva York, Knopf, 1975.

# 4. Neurobiología cognitiva

Como lo indica su nombre, la neurobiología cognitiva es un área interdisciplinaria de investigación cuyo interés es entender las actividades cognitivas específicas que desarrollan las criaturas vivas. Ha comenzado a florecer en los últimos años por tres razones.

Primera, se ha producido un avance constante en las *tecnologías* que permiten analizar la microestructura del cerebro y monitorear las actividades neurales en curso. Los modernos microscopios electrónicos permiten un acceso sin par a los detalles de la microestructura cerebral, y diversas tecnologías nucleares permiten tomar imágenes de la estructura interna y la actividad neural de los cerebros vivos sin invadirlos ni perturbarlos en absoluto. Segunda, la investigación se ha beneficiado con la aparición de algunas *teorías* generales muy polémicas sobre la función de redes neurales en gran escala. Estas teorías proporcionan una dirección y un objetivo a los esfuerzos experimentales; ayudan a decidir cuáles son las preguntas útiles que hay que formularle a la Naturaleza. Y tercera, los *ordenadores* modernos han hecho posible que se analicen, de un modo eficiente y revelador, las propiedades funcionales de las estructuras sumamente intrincadas que las recientes teorías asignan a nuestro cerebro. Pues podemos construir un modelo de tales estructuras dentro de un ordenador y dejar que éste nos diga cómo se comportarán bajo diversas circunstancias. Luego podemos verificar esas predicciones con la conducta de cerebros reales en circunstancias comparables.

En este apartado consideraremos brevemente dos de las preguntas centrales de la neurobiología cognitiva. ¿Como *representa* al mundo el cerebro? ¿Y cómo efectúa *cálculos* el cerebro sobre esas representaciones? Tomemos primero la primera pregunta y comencemos con algunos fenómenos completamente familiares.

¿Cómo representa el cerebro el color de una puesta de sol? ¿El perfume de una rosa? ¿El sabor de un durazno? ¿O el rostro de la persona amada? Hay una técnica simple para representar o *codificar* estímulos externos que es sorprendentemente eficaz y puede utilizarse en todos los casos mencionados, a pesar de su diversidad. Para ver cómo funciona, consideremos el caso del gusto.

## Codificación sensorial: el gusto

En la lengua hay cuatro clases de células receptoras. Las células de cada clase responden de modo específico ante cualquier sustancia dada que se ponga en contacto con ellas. Por ejemplo, un durazno puede tener un efecto sustancial en una de las cuatro clases de células, un efecto mínimo sobre la segunda clase y niveles intermedios de efecto en la tercera y cuarta clases. Considerado en conjunto, este patrón riguroso de estímulos relativos constituye una suerte de "huella digital" neural que es únicamente característica de los duraznos.

Si denominamos a las cuatro clases de células *a, b, c* y *d* respectivamente, podemos describir exactamente qué es esa huella digital especial, especificando los cuatro niveles de estimulación neural que produce dicho contacto con un durazno. Si utilizamos la letra S, con un subíndice, para representar cada uno de los diversos niveles de estimulación, entonces lo siguiente es lo que queremos: $\langle S_a, S_b, S_c, S_d \rangle$. Esto se denomina *vector de codificación sensorial* (un vector es una lista de números o un conjunto de magnitudes). El punto importante es que evidentemente hay un vector de codificación *único* para cada sabor humanamente posible. Lo que

significa que cualquier sensación de gusto humanamente posible es sólo un patrón de niveles de estimulación a través de las cuatro clases de células sensitivas. O mejor aún, es un patrón con frecuencias ondulatorias a través de cuatro canales neurales que transmiten las noticias de estos niveles de actividad desde la boca al resto del cerebro.

Se puede representar gráficamente cualquier gusto mediante un punto adecuado en un "espacio del gusto", un espacio con cuatro ejes, cada uno para el eje de estimulación de cada una de las cuatro clases de células sensitivas del gusto. La figura 7.13 describe un espacio en el que han sido codificadas las posiciones de diversos sabores. (En este diagrama se ha suprimido uno de los cuatro ejes, ya que es difícil dibujar un espacio cuatridimensional en una página bidimensional.) Lo interesante es que sabores subjetivamente similares resultan tener vectores de codificación muy similares. O lo que es lo mismo, sus puntos en el espacio del gusto están muy cercanos entre sí. El lector observará que todas las clases de gusto "dulce" están codificadas en las regiones superiores del espacio, mientras que los diversos sabores "agrios" aparecen en el centro inferior. Varios sabores "amargos" aparecen abajo a la izquierda y los "salados" están en la zona de abajo a la derecha. Los otros puntos en este espacio representan todas las demás sensaciones gustativas que puede tener el ser humano. Esto constituye un dato definido en apoyo de la idea de identidad del teórico (capítulo 2, apartado 3) de que cualquier sensación es simplemente idéntica a un conjunto de frecuencias ondulatorias en la vía sensorial adecuada.

## Codificación sensorial: el color

Un argumento similar parece ser válido para el color. Hay tres clases diferentes de células sensitivas al color o *conos* en la retina humana y cada clase es sensible a una longitud de onda diferente de la luz: corta, media y larga, respectivamente. La visión del color es un asunto complejo y mi breve

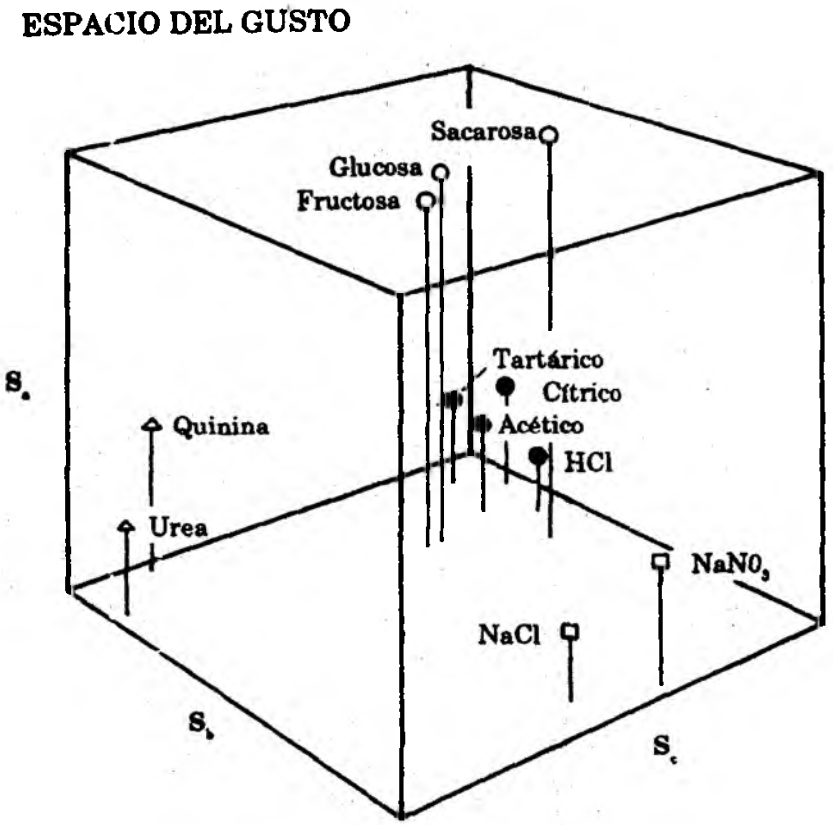



Figura 7. 13

esquema está demasiado simplificado, pero una parte central de la historia parece ser el *patrón* de niveles de actividad producidos a través de tres clases diferentes de conos. Aquí el vector de codificación sensorial tiene tres elementos y no cuatro: $\langle S_{corta}, S_{media}, S_{larga} \rangle$. Pero nuevamente, las similitudes en el color se reflejan en las similitudes de sus vectores de codificación, o lo que es igual, por la cercanía de sus puntos en un "espacio de sensaciones cromáticas" tridimensional (fig. 7.14). Además, la idea intuitiva de que el color naranja está de algún modo "entre" el rojo y el amarillo toma una expresión directa: si las sensaciones de color se representan de este modo, la sensación del naranja está *literalmente* entre las otras dos

clases de sensación. Y lo mismo ocurre con todas las demás relaciones de "entre" en el dominio de los colores.

Finalmente es útil observar que esta idea de la codificación sensorial explica también las variedades de ceguera al color. Las personas que sufren este trastorno menor carecen de una (o más) de las tres clases de conos. Lo que significa que su "espacio del color" tendrá sólo dos (o menos) dimensiones, y no tres, por lo cual su capacidad para discriminar los colores se reducirá de modo predecible.

ESPACIO DE QUALIA DE LOS COLORES

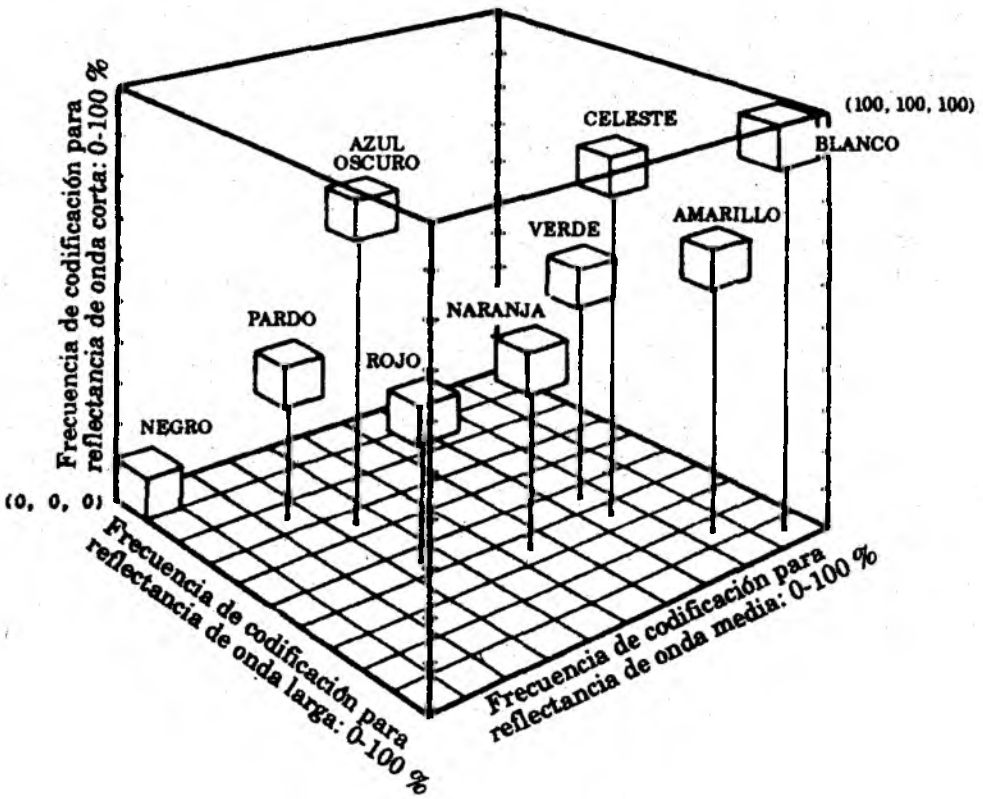



Figura 7. 14

## Codificación sensorial: el olfato

El sistema olfativo abarca seis o siete, y quizá más, clases diferentes de receptores. Esto indica que los olores están codificados por un vector de frecuencias ondulatorias que tiene por lo menos seis o siete elementos diferentes. Esto permite una gran cantidad de combinaciones distintas de frecuencias y, por lo tanto, una gran cantidad de olores diferentes. Supongamos que, por ejemplo, un sabueso tiene siete clases de receptores olfativos y que puede distinguir treinta niveles diferentes de estimulación dentro de cada clase. Según esto debemos asignar al sabueso un "espacio del olfato" total de ¡30 × 30 × 30 × 30 × 30 × 30 × 30 (= $30^7$ o 22 *mil millones*) de posiciones distinguibles! Es por eso que los perros pueden distinguir a una persona cualquiera de entre millones simplemente con el olfato.

Todo esto proporciona un fuerte apoyo a los teóricos de la identidad, que dicen que nuestras sensaciones son simplemente idénticas a, digamos, un conjunto de niveles de estimulación (frecuencias ondulatorias) en las vías sensoriales apropiadas. Pues como lo demuestran los apartados precedentes, la neurociencia está reconstruyendo con éxito, de un modo sistemático y revelador, los diversos rasgos de nuestros qualia sensoriales subjetivos y las relaciones entre ellos. Es el mismo patrón que, durante el siglo XIX, motivó la afirmación científica de que la luz es simplemente idéntica a ondas electromagnéticas de una cierta frecuencia. Pues dentro de la teoría de la electricidad y el magnetismo, podíamos reconstruir sistemáticamente todas las características conocidas de la luz.

## Codificación sensorial: los rostros

Entre los seres humanos son los *rostros* lo que se distingue con gran habilidad, y una teoría reciente dice que las caras también se rigen mediante una estrategia de codificación vectorial. Para cada uno de los diversos elementos de un rostro humano a los que somos perceptivamente sensibles

—tamaño de la nariz, ancho de la boca, distancia entre los ojos, cuadratura de la mandíbula, etc.— supongamos que hay una vía cuyo nivel de estimulación corresponde al grado en que la cara percibida exhibe dicho elemento. Por lo tanto, un rostro en particular será codificado por un vector de estimulaciones único, un vector cuyos elementos corresponden a los elementos visibles de la cara percibida.

Si adivinamos que hay quizá diez rasgos faciales diferentes a los que un ser humano maduro es sensible y si suponemos que podemos distinguir por lo menos cinco niveles diferentes dentro de cada rasgo, entonces tenemos que asignar al ser humano un "espacio facial" de por lo menos $5^{10}$ (cerca de 10 millones) de posiciones distinguibles. Ya es notable que podamos distinguir a cualquier persona de entre millones, simplemente con la vista.

Por supuesto, los rostros de los familiares cercanos serán codificados por vectores con muchos de los mismos elementos o similares. En contraste, las personas que no se parecen entre sí serán codificadas por vectores muy diferentes. Una persona con una cara muy normal será codificada por un vector en el que todos los elementos están en el medio del rango pertinente de variación. Y alguien con una cara muy especial será codificado por un vector que tiene uno o más elementos en un valor extremo. Curiosamente, el lóbulo parietal de la corteza cerebral derecha en el ser humano, una extensa región responsable de las cuestiones espaciales en general, tiene una pequeña porción cuya destrucción produce una incapacidad de reconocer rostros humanos. Podemos postular que aquí es donde se codifican los rostros humanos.

## Codificación sensorial: el sistema motor

Las virtudes de la codificación vectorial son especialmente evidentes cuando consideramos el problema de representar un sistema muy complejo, como la posición simultánea de todos los miles de músculos de nuestro cuerpo. Tenemos un sentido actualizado constante y continuo de la postura general

o de la configuración de nuestro cuerpo en el espacio. Y algo bueno también. Para poder efectuar cualquier movimiento útil debemos saber desde dónde empiezan nuestras extremidades. Esto es válido para cosas simples como caminar y para cosas complicadas como la danza o el baloncesto.

Este sentido de la propia configuración corporal se denomina *propiocepción* y es posible porque cada uno de los músculos del cuerpo tiene su propia fibra nerviosa que envía permanentemente información al cerebro, información acerca de la contracción o extensión de dicho músculo. Con tantos músculos, el vector de codificación total en el cerebro tendrá evidentemente, no tres o diez elementos, sino ¡más de mil! Pero no es problema para el cerebro: tiene *miles de millones* de fibras con las cuales realizar el trabajo.

## Codificación de salida

Mientras hablamos del sistema motor, el lector habrá observado que la codificación vectorial puede ser tan útil para dirigir la *salida* motora como para codificar la entrada sensorial. Cuando una persona realiza cualquier actividad el cerebro está enviando una cascada de mensajes diferentes a cada músculo del cuerpo. Pero esos mensajes deben estar bien organizados para que el cuerpo haga algo coherentemente: cada músculo debe adoptar exactamente el grado correcto de contracción o extensión para que el cuerpo se coloque en la posición deseada.

¿Cómo puede el cerebro organizar todo esto? Mediante el *vector motor:* un conjunto de niveles de actividad simultánea en todas las neuronas motoras, neuronas que transmiten mensajes desde el cerebro hasta los músculos del cuerpo. Un movimiento complejo es una secuencia de posiciones corporales, y entonces para ellas el cerebro debe emitir no uno, sino una secuencia de vectores motores. Generalmente, estos vectores de salida son enviados por las decenas de miles de largos axones en la médula espinal y luego a lo largo de las neuronas motoras hasta los músculos. Aquí cada elemento del

gran vector constituye un nivel de estimulación en la neurona que hace contacto con el músculo correspondiente. El músculo responde a ese elemento del vector, contrayéndose o relajándose según lo dicte el nivel de estimulación. Conjuntamente, y si los vectores motores están bien constituidos, estas estimulaciones individuales hacen que todo el cuerpo se mueva con coherencia y gracia.

## Informática neural

Como hemos visto, los vectores de estimulación son un medio perfectamente eficaz para representar cosas tan diversas como sabores, rostros y complicadas posiciones de las extremidades. Análogamente, resulta que también son parte de una solución muy elegante al problema de la computación de alta velocidad. Si el cerebro utiliza vectores para codificar diversas entradas sensoriales y también diversas salidas motoras, entonces en algún lugar debe estar efectuando cálculos para que las entradas de algún modo estén *guiando* o *produciendo* las salidas. Es decir que necesita alguna configuración para transformar sus diversos vectores de entrada sensoriales en vectores motores adecuados de salida.

Casualmente, amplios sectores del cerebro tienen una microestructura que parece idealmente adecuada para realizar transformaciones de esta clase precisamente. Consideremos, por ejemplo, la configuración esquemática de axones, dendritas y sinapsis en la figura 7.15. Aquí el vector de entrada $\langle a, b, c, d \rangle$, se realiza en los cuatro axones horizontales de entrada. Cada axón conduce un tren de ondas que entra con una frecuencia determinada. Y, como se ve, cada axón hace tres conexiones sinápticas, una para cada una de las tres células verticales. (Estas se denominan *células de Purkinje,* en honor a su descubridor.) En total son $4 \times 3 = 12$ sinapsis.

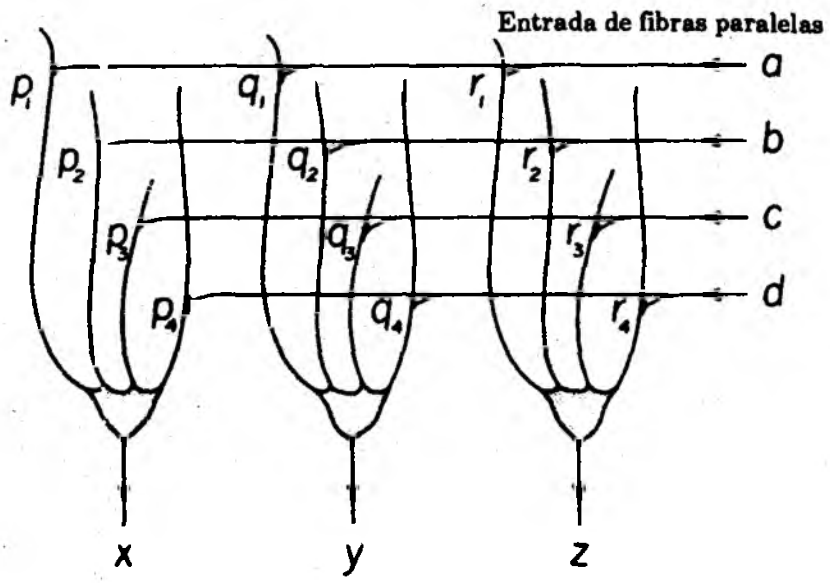



Figura 7. 15

Pero estas conexiones sinápticas no son todas idénticas. Como lo muestra el diagrama, algunas son grandes, otras son pequeñas. Las letras $p_i$, $q_j$ y $r_k$ representan su magnitud. Para calcular la cantidad de excitación que cada conexión induce en su célula receptora, hay que multiplicar el tamaño de la conexión por la frecuencia ondulatoria en el axón entrante. La excitación *total* en la célula de Purkinje receptora es entonces la suma de aquellos cuatro efectos sinápticos.

La célula de Purkinje emite ondas por su axón de salida, ondas cuya frecuencia es una función de la excitación total que las diversas entradas han producido en esa célula. Como las tres células de Purkinje hacen esto, la salida del sistema es obviamente otro vector, un vector de tres elementos. Claramente, nuestro pequeño sistema transformará cualquier vector de entrada cuatridimensional en un vector de salida tridimensional muy diferente.

Por supuesto, lo que determina la naturaleza de toda la transformación es la distribución de los *tamaños* entre las diversas conexiones sinápticas. Estas fuerzas de conexión se denominan generalmente *pesos*. Si especificamos la distribución de los pesos sinápticos en un sistema de esta clase, habremos especificado el carácter de la transformación que realizará sobre cualquier vector entrante.

## El cerebelo

El sistema de transformación vectorial de la figura 7.15 es sólo un diagrama esquemático, muy simplificado por razones ilustrativas. Pero la misma clase de organización celular aparece en el cerebelo de todas las criaturas, aunque en una escala mucho mayor. La figura 7.16 muestra un pequeño corte de la corteza cerebelosa y se puede ver que todas las entradas de las fibras de Mossy conducen sus frecuencias ondulatorias a través de las células granulares hacia las *fibras paralelas,* cada una de las cuales hace múltiples conexiones sinápticas con las tupidas ramificaciones dendríticas de muchas células de Purkinje diferentes. Cada célula de Purkinje suma la actividad así inducida en ella y emite ondas por su propio axón como salida. La unión de los niveles de actividad en todo el conjunto de axones de Purkinje constituye el vector de salida del cerebelo.

La figura 7.16 también es una simplificación, porque en el verdadero cerebelo hay millones de fibras paralelas, muchos cientos de miles de células de Purkinje y miles de millones de conexiones sinápticas. Si tomamos esto entonces, el vector de entrada tiene millones de elementos y el de salida tiene cientos de miles, aunque es probable que haya redundancia ya que cada elemento del verdadero vector puede estar codificado en forma múltiple. De todos modos, tenemos vectores de codificación que son lo suficientemente grandes como para hacer el trabajo de coordinar el sistema muscular del cuerpo. Y esto es precisamente lo que hace el cerebelo. La principal salida del cerebelo baja por la médula espinal hacia

CORTE ESQUEMATICO: CEREBELO

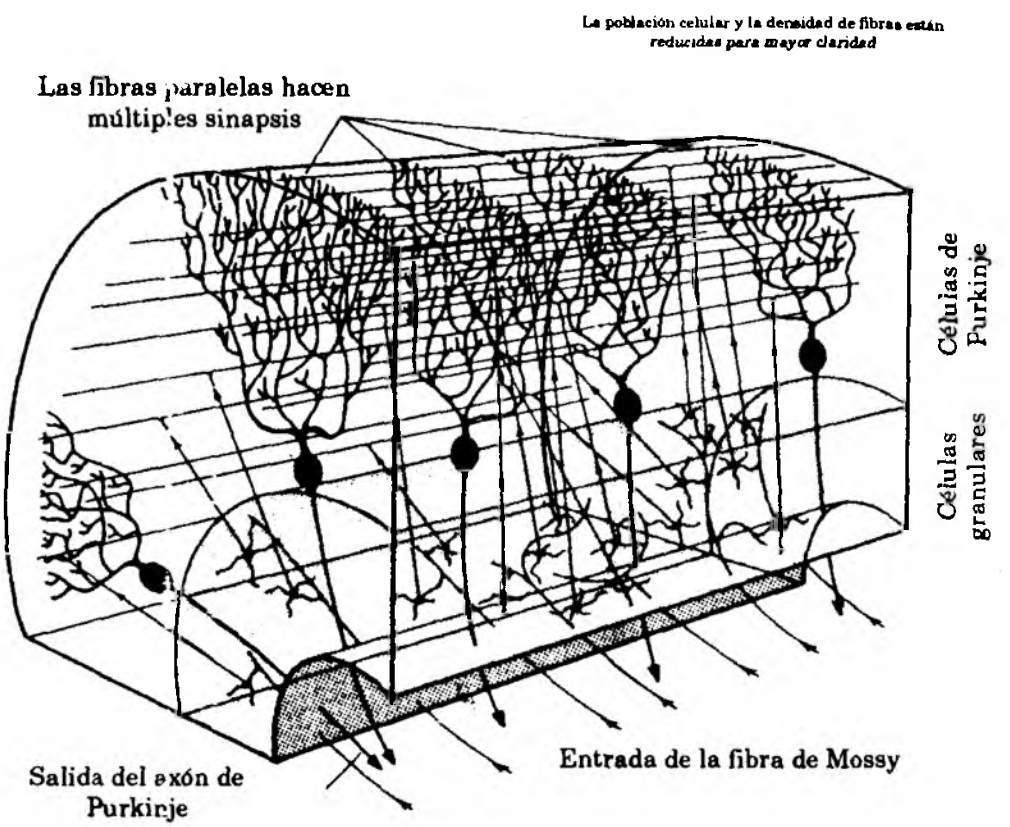



Figura 7. 16

los músculos. Y si el cerebelo sufre un grave daño o se pierde, los movimientos voluntarios del paciente resultan espasmódicos, mal orientados e incoordinados.

Hay tres puntos importantes que observar acerca de un sistema "informático" del tipo que presenta el cerebelo. Primero, es muy resistente a daños menores y a la muerte dispersa de células. Como está formado por miles de millones de conexiones sinápticas, cada una de las cuales contribuye sólo en una muy pequeña porción a la transformación total de los vectores, la pérdida de unos pocos miles de conexiones aquí y allá no cambiará la conducta global de la red. Incluso puede perder millones de conexiones mientras estén dispersas aleatoriamente por la red, como sucede con la muerte gradual

de células en el curso natural del envejecimiento. La calidad de los cálculos del cerebelo entonces *disminuirá* lentamente, en lugar de caer abruptamente.

Segundo, un sistema de paralelismo masivo de esta clase realizará sus transformaciones vector a vector en un instante. Como cada sinapsis realiza su propio "cálculo" más o menos simultáneamente con todas las demás, los mil millones de cálculos requeridos para producir el vector de salida son efectuados todos de una vez, y no uno después de otro. El vector de salida estará camino a los músculos en menos de diez milésimos de segundos (1/100 de segundo) luego de que el vector de entrada haya llegado a la red. Aunque las sinapsis son mucho más lentas que las UCP y aunque la propagación axonal de ondas es mucho más lenta que la propagación eléctrica, el cerebelo efectúa su cálculo global cientos de veces más velozmente que el ordenador más rápido. Su paralelismo masivo es el que hace la diferencia.

Tercero, estas redes son funcionalmente modificables. En términos técnicos, son *plásticas*. Pueden cambiar sus propiedades de transformación simplemente cambiando algunos o todos sus pesos sinápticos. Es un hecho importante, ya que el sistema debe poder aprender a producir movimientos coordinados en primer lugar, y luego reaprender continuamente, a medida que el tamaño y la masa de las extremidades cambian lentamente con la edad. Ahora analizaremos cómo puede ocurrir este aprendizaje.

En resumen, las redes nerviosas de esta clase son computacionalmente poderosas, resistentes al daño, rápidas y modificables. Sus virtudes tampoco terminan aquí, como veremos en el siguiente apartado.

## Lecturas complementarias


Llinas, R., "The Cortex of the Cerebellum", *Scientific American*, 232, No. 1, 1975.

Bartoshuk, L. M., "Gustatory System", en *Handbook of Behavioral Neurobiology*, vol. I, *Sensory Integration*, R. B. Masterton (comp.), Nueva York, Plenum, 1978.

Pfaff, D. W., *Taste, Olfaction, and the Central Nervous System*, Nueva York, Rockefeller University Press, 1985.

Land, E., "The Retinex Theory of Color Vision", *Scientific American*, 237, 6, diciembre de 1977.
Hardin, C. L. *Color for Philosophers*, Indianapolis; Hackett, 1987.
Dewdney, A. K., "A Whimsical Tour of Face Space", en la sección Computer Recreations de *Scientific American*, vol. 255, octubre de 1986.
Pellionisz, A. y Llinas, R., "Tensor Network Theory of the Metaorganization of Functional Geometries in the Central Nervous System", *Neuroscience*, vol. 19, 1986.
Churchland, P. M., "Some Reductive Strategies in Cognitive Neurobiology", *Mind*, vol. 95, 379, 1986.
Churchland, P. S., *Neurophilosophy*, Cambridge, MA, The MIT Press, 1986.

# 5. Más sobre la IA: procesamiento de distribución paralela

A fines de la década de los años cincuenta, muy a comienzos de la historia de la IA, había mucho interés en las "redes nerviosas" artificiales, es decir, en los sistemas de equipos armados sobre el modelo del cerebro humano. A pesar de su atractivo inicial, esta primera generación de redes resultó tener serias limitaciones y fue rápidamente eclipsada por las técnicas de "escritura de programas" de IA. Estos han demostrado tener sus propias limitaciones, como vimos al final del capítulo 6, y en los últimos años ha renacido el interés por el método anterior. Las primeras limitaciones han sido superadas y las redes nerviosas artificiales finalmente comienzan a demostrar su verdadero potencial.

## Redes nerviosas artificiales: su estructura

Consideremos una red compuesta por unidades simples, similares a la neurona, conectadas como se observa en la figura 7.17. Las unidades de la base pueden pensarse como unidades sensoriales, ya que son estimuladas por el medio exterior al sistema. Cada una de estas unidades inferiores emite una salida a través de su propio "axón", salida cuya fuerza es una función del nivel de estimulación de la unidad.

El axón se divide en una cantidad de ramas terminales y se envía una copia de esa señal de salida a cada unidad del segundo nivel. Estas se denominan *unidades ocultas* y las unidades inferiores hacen una variedad de "conexiones sinápticas" con cada una de ellas. Cada conexión tiene una fuerza determinada o *peso,* como se denomina comúnmente.

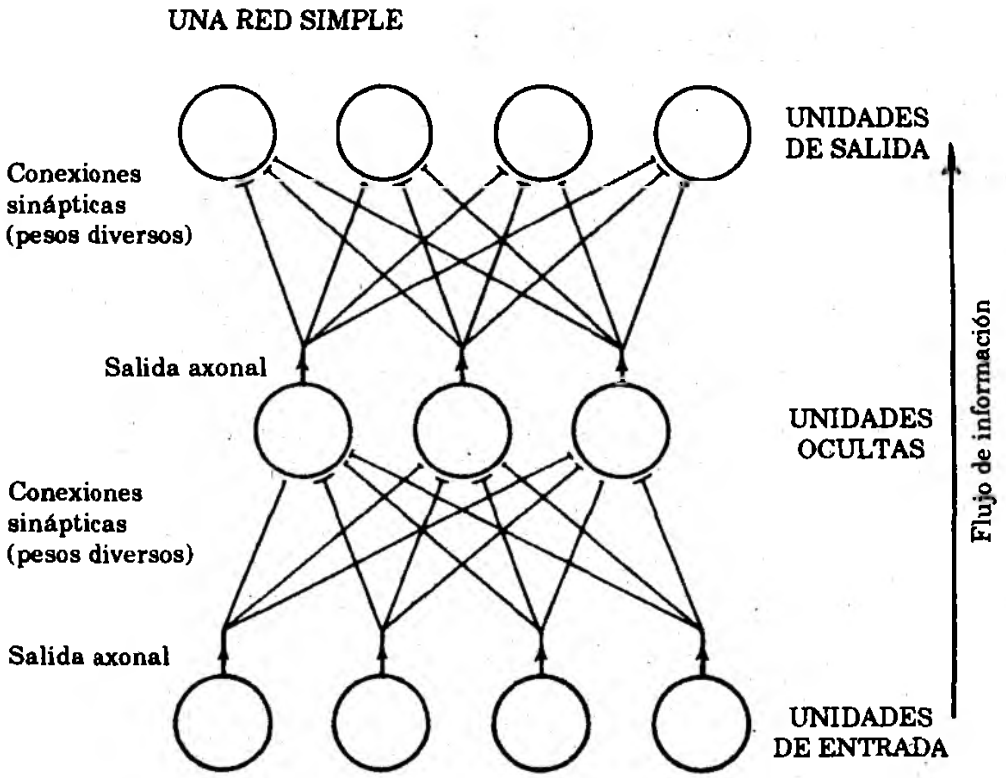



Figura 7. 17

Ya se puede ver que la mitad inferior del sistema es otro transformador vector a vector, como las matrices nerviosas analizadas en el apartado anterior. Si estimulamos las unidades inferiores, el conjunto de niveles de actividad que inducimos (el vector de entrada) será transmitido hacia arriba hacia las unidades ocultas. En el camino, es transformado por

diversas influencias: por la función de salida de las células inferiores, por cualquier patrón de pesos que exista en las sinapsis y por la actividad total dentro de cada una de las unidades ocultas. El resultado es un conjunto o patrón de niveles de estimulación a través de las unidades ocultas: otro vector.

Ese vector de estimulación en las unidades ocultas sirve a su vez como vector de entrada para la mitad superior del sistema. Los axones de las unidades ocultas hacen conexiones sinápticas, de diversos pesos, con las unidades del nivel más elevado. Estas son las unidades de salida y todo el conjunto de niveles de estimulación finalmente inducido en ellas es lo que constituye el vector de salida. La mitad superior de la red es así sólo otro transformador vector a vector.

Siguiendo este patrón general de interconexión, podemos construir claramente una red con cualquier número deseado de unidades de entrada, unidades ocultas y unidades de salida, según el tamaño de los vectores que deban procesarse. Y podemos empezar a ver cuál es la idea de tener una configuración de dos filas si consideramos lo que puede hacer una red de esta clase cuando se enfrenta a un problema real. El punto crucial que hay que recordar es que podemos *modificar* los pesos sinápticos en todo el sistema, de manera de implementar cualquier transformación vector a vector que queramos.

## Reconocimiento perceptual: aprendizaje por el ejemplo

Nuestro problema de muestra es el siguiente. Somos la tripulación comando de un submarino, cuya misión lo llevará a las aguas poco profundas de un puerto enemigo, un puerto cuyo fondo está protegido por minas explosivas. Debemos esquivar las minas y por lo menos podemos detectarlas con nuestro sistema sonar, que envía pulsos de sonido y escucha el eco que vuelve en caso de que el pulso rebote contra algún objeto sólido que se encuentra en el fondo del mar. Desafortunadamente, una *roca* de tamaño considerable también devuelve un eco sonar, eco que el oído humano no puede distin-

guir de un verdadero eco producido por una mina (figura 7.18).

Esto es frustrante porque el puerto también tiene muchas rocas grandes en el fondo. La situación se complica aun más por el hecho de que las minas son de diversas formas y se encuentran en diversas posiciones con respecto al pulso sonar que reciben. Entonces los ecos que vuelven desde cada clase de objeto también presentan una gran variación dentro de cada clase. Frente a esto, la situación parece desesperadamente confusa.


Figura 7. 18

¿Cómo podríamos prepararnos para distinguir los ecos de las minas de los ecos de las rocas inofensivas, de manera que pudiéramos llevar a cabo la misión con seguridad? Del siguiente modo. Primero reunimos, en una cinta de grabación, una gran cantidad de ecos sonar de lo que sabemos son las verdaderas minas de distintas formas y en distintas posiciones. Son minas que hemos puesto deliberadamente, con el propósito de probar, en el fondo de nuestras aguas costeras. Hacemos lo mismo con rocas de diferentes clases y por supuesto registramos cuidadosamente cuál es cada eco. Finalmente tenemos, por ejemplo, cincuenta muestras de cada uno.

Luego colocamos cada eco en un analizador espectral

simple, que da información como la que se observa en el sector izquierdo de la figura 7.19. Esto muestra cuánta energía sonora contiene el eco dado en cada una de las diversas frecuencias sonoras que lo forman. Es un modo de cuantificar el carácter total de un eco determinado. Por sí solo este análisis no nos ayuda mucho, ya que los diagramas obtenidos aún no parecen exhibir ninguna uniformidad obvia ni diferencias regulares entre los ecos. Pero ahora introduzcamos una red nerviosa (véase nuevamente la figura 7.19 en el sector de la derecha. Es una versión simplificada de una red analizada por Gorman y Sejnowski. Obsérvese que se ha girado 90 grados con respecto a la figura 7.17).

Esta red está organizada como la de la figura 7.17, pero tiene 13 unidades de entrada, 7 unidades ocultas, 2 unidades de salida y un total de 105 conexiones sinápticas. Los niveles de actividad de cada unidad, supondremos, varían entre cero y uno. Recordemos también que los pesos sinápticos del sistema pueden ajustarse a los valores necesarios. Pero no sabemos qué valores se necesitan. De modo que al comenzar el experimento, las conexiones tienen pesos aleatoriamente distribuidos. Entonces es poco probable que la transformación que realiza la red nos sea útil. Pero procedemos de la siguiente manera.

Tomamos un eco de mina de nuestra reserva de pruebas y utilizamos el analizador de frecuencias para probar sus niveles de energía en 13 frecuencias diferentes. Esto nos da el vector de entrada, que tiene 13 elementos. Entonces colocamos este vector en la red estimulando cada una de las unidades de entrada con una cantidad adecuada, como se indica en la figura 7.19. Este vector se propaga rápidamente a través de la red de dos etapas y produce un vector de salida de dos elementos en las unidades de salida. Lo que nos *gustaría* es que la red produjera el vector ⟨1,0⟩, que es nuestro vector de salida convencional para una *mina*. Pero dados los pesos aleatorios, esa salida correcta sería un milagro. Producirá más probablemente algún vector accidental y totalmente aburrido como ⟨0,49, 0,51⟩, que es lo mismo que nada.

Pero no nos desalentemos. Calculamos, por simple sus-

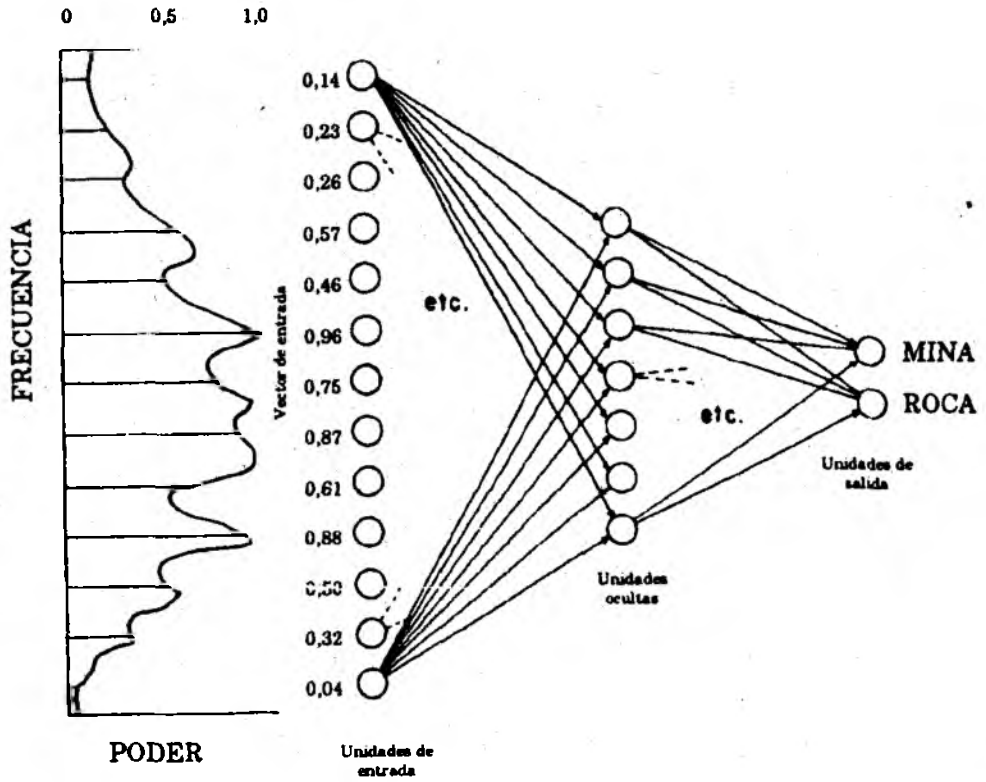



Figura 7. 19

tracción, la diferencia entre el vector que obtuvimos y el que queríamos. Y utilizamos una regla matemática especial, llamada la *regla delta generalizada,* para calcular pequeños cambios en los pesos del sistema. La idea es modificar aquellos pesos que fueron los más responsables de la salida incorrecta de la red. Entonces los pesos se ajustan de acuerdo con eso.

Luego le damos al sistema otro eco de prueba —puede ser una roca esta vez— y esperamos un vector de salida ⟨0,1⟩, que es nuestro vector de salida convencional para una roca. Seguramente, el vector de salida será nuevamente una decepción, por ejemplo, un ⟨0,47, 0,53⟩. Otra vez calculamos la cantidad del error y aplicamos la regla especial para ajustar los pesos.

Luego intentamos nuevamente con una tercera prueba. Y así sucesivamente.

Hacemos esto miles, y quizá decenas de miles, de veces. O si no, programamos un ordenador convencional, cuya memoria contiene el registro de nuestras pruebas, para que haga las veces de maestro y realice todo el trabajo por nosotros. Esto se denomina *entrenamiento de la red*. Un poco sorprendentemente, el resultado es que el conjunto de pesos gradualmente termina en una configuración final en que el sistema da un vector de salida ⟨1,0⟩ (o cercano a él) cuando y sólo cuando el vector de entrada es el de una mina; y da un vector de salida ⟨0,1⟩ (o cercano a él) cuando y sólo cuando el vector de entrada es el de una roca.

El primer hecho notable de todo esto es que *hay* una configuración de pesos sinápticos que permite que el sistema distinga con bastante seguridad entre los ecos de una mina y los de una roca. Existe tal configuración porque resulta que después de todo hay un patrón interno aproximado o una organización abstracta que es característica de los ecos de las minas que lo distingue de los de las rocas. Y la red entrenada ha logrado enlazarse en aquel patrón aproximado.

Si, después de entrenar a la red, analizamos los vectores de actividad de las unidades *ocultas* para cada una de las dos clases de estimulación, encontramos que tales vectores forman dos clases completamente diferentes. Consideremos, por ejemplo, un "espacio de codificación vectorial" abstracto, un espacio con 7 ejes, uno para los niveles de actividad de cada unidad oculta. (Pensemos en este espacio como los espacios abstractos de codificación sensorial de las figuras 7.13 y 7.14. La única diferencia es que este espacio representa los niveles de actividad de células que se encuentran más adelante en la jerarquía de procesamiento.) Cualquier vector "de mina" que aparezca por las unidades ocultas cae en un gran subvolumen del espacio de posibles vectores de unidades ocultas. Y cualquier vector "de roca" cae en un gran subvolumen muy *distinto* (no superpuesto) de aquel espacio abstracto.

Lo que las unidades ocultas están haciendo en una red

entrenada es codificar exitosamente algunos rasgos estructurales bastante abstractos de los ecos de minas, rasgos que todos tienen, o casi todos, a pesar de su diversidad superficial. Y hace lo mismo para los ecos de rocas. Hace todo esto al encontrar un conjunto de pesos que produzca clases diferentes de vectores de codificación para cada uno.

Dado el éxito de esta clase en el nivel de las unidades ocultas, lo que hace la mitad derecha de la red entrenada es sólo transformar cualquier vector de mina de una unidad oculta en algo cercano a ⟨1,0⟩ en el nivel de salida, y cualquier vector de roca de una unidad oculta en algo cercano a un vector ⟨0,1⟩ en el nivel de salida. Es decir que aprende a distinguir entre los dos subvolúmenes del espacio vectorial de unidades ocultas. Los vectores cercanos al centro de cualquiera de los dos volúmenes —son los ejemplos "prototípicos" de cada clase de vector— producen un veredicto claro en el nivel de salida. Los vectores cercanos al límite que divide los dos volúmenes producen una respuesta mucho menos decisiva: quizás un ⟨0, 4, 0, 6⟩. Entonces la "conjetura" de la red de que es una roca no es muy segura. Pero aun así puede ser bastante fiable.

Un encantador subproducto de este procedimiento es el siguiente. Si ahora se le presentan a la red pruebas completamente *nuevas* de ecos de rocas y de minas —pruebas que nunca había escuchado antes— sus vectores de salida las clasificarán bien directamente, y con una exactitud que sólo es insignificantemente menor que la que ahora se ve en las 100 pruebas con las cuales se entrenó. Aunque son nuevas, las nuevas pruebas también producen vectores en el nivel de las unidades ocultas que caen en uno de los dos subespacios diferenciables. Es decir que el "conocimiento" que ha adquirido el sistema generaliza de modo fiable a nuevos casos. Finalmente, nuestro sistema está preparado para ser aplicado en el puerto enemigo. Simplemente lo alimentamos con los retornos amenazantes del sonar y sus vectores de salida nos dirán si nos acercamos o no a una mina.

Aquí lo interesante no es la aplicación militar propuesta para el dispositivo descripto; he utilizado aquel contexto sola-

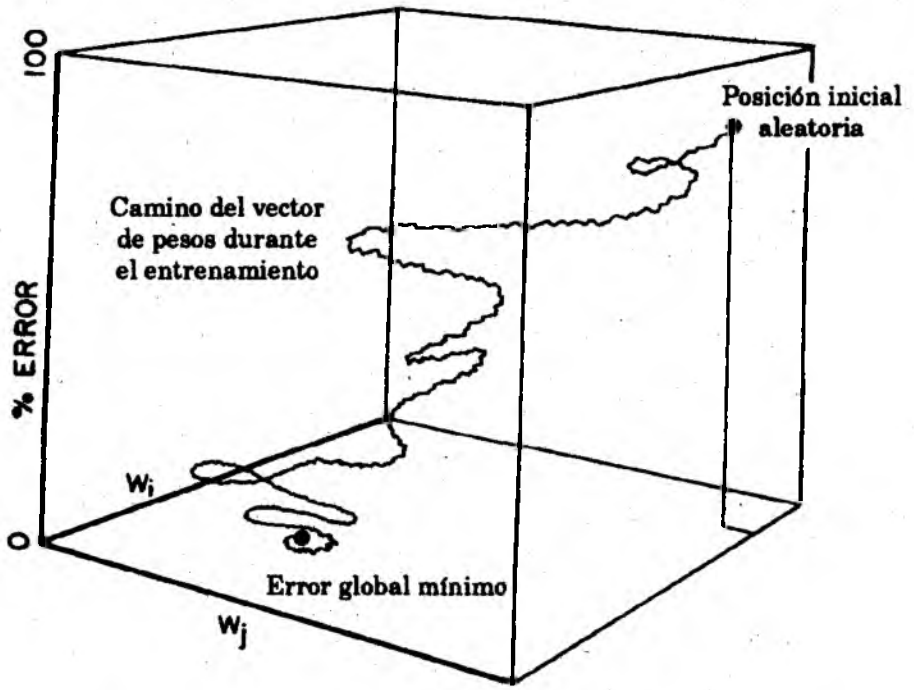
mente para lograr un efecto argumental. Las tecnologías navales existentes ya pueden levantar una lata de cerveza de un fondo arenoso e incluso decir su marca, utilizando principios muy diferentes de análisis. Lo que es interesante es que un sistema tan *simple* pueda realizar la sofisticada tarea de reconocimiento que hemos descripto.

La primera maravilla es que una red adecuadamente adaptada haga este trabajo. La segunda maravilla es que exista una regla que *moldeará* exitosamente la red como una configuración necesaria de pesos, aunque comience con una configuración aleatoria. Esa regla hace que el sistema aprenda de las 100 pruebas que le proporcionamos, más los errores que produce. Este proceso se denomina *aprendizaje automatizado por la propagación regresiva del error,* y es inexorablemente eficiente. Pues con frecuencia encontrará orden y estructura, por sí solo, allí donde inicialmente vemos sólo caos y confusión. Este proceso de aprendizaje es un ejemplo de *descenso de gradiente,* porque la configuración de pesos puede verse como el descenso por una pendiente variable de errores cada vez menores hasta que entra en la estrecha región de un valle muy bajo en el que los mensajes de error se acercan cada vez más a cero. (Véase la figura 7.20 para una representación parcial de este proceso.)

Con errores tan pequeños, la eficiencia de un aprendizaje ulterior naturalmente disminuye, pero en ese punto el sistema ya ha alcanzado un alto nivel de fiabilidad.

El entrenamiento de la red con los ecos de prueba puede llevar un par de horas, pero una vez que el sistema ha sido entrenado, dará su veredicto sobre cualquier prueba en un instante. Al ser un sistema en paralelo, la red transforma todos los elementos del vector de entrada al mismo tiempo. Aquí finalmente tenemos el reconocimiento "perceptual" de estímulos complejos en una escala temporal igual o mejor que la de las criaturas vivas.

APRENDIZAJE: DESCENSO DE GRADIENTE
EN EL ESPACIO DE PESOS




(Los ejes corresponden a los pesos de sólo 2
de las 105 conexiones sinápticas)

Figura 7. 20

## Otros ejemplos y observaciones generales

He analizado en detalle la red roca/mina para proporcionar algunos detalles reales de cómo funciona una red en paralelo. Pero el ejemplo es sólo uno de muchos. Si los ecos de las minas pueden ser reconocidos y diferenciados de otros sonidos, entonces una red adecuadamente entrenada de esta clase general debería poder reconocer los diversos *fonemas* que constituyen el idioma inglés sin preocuparse por las diferencias en el carácter de las voces de las personas, como lo hacen los programas tradicionales de IA. Por lo tanto ahora está a nuestro alcance un reconocimiento del lenguaje por parte de las máquinas.

Tampoco hay nada esencialmente auditivo en las aptitudes de estas redes. Pueden ser "entrenadas" para reconocer complejos estímulos visuales también. Una red nueva puede distinguir la forma tridimensional y la orientación de superficies físicas suavemente curvadas dada sólo una fotografía en la gama del gris de la superficie estudiada. Es decir que resuelve el problema de la "forma a partir de sombreados". Y una vez entrenada, una red así dará su veredicto de salida sobre cualquier muestra casi instantáneamente.

Tampoco hay nada esencialmente perceptual en sus aptitudes. Pueden utilizarse para producir una interesante salida motora de un modo igualmente fácil. Por ejemplo, una red bastante grande ya ha aprendido a resolver el problema de transformar un texto impreso en un discurso audible (el NETtalk de Sejnowski y Rosenberg). El sistema utiliza un esquema de codificación vectorial para las letras de entrada, otro para los fonemas de salida y aprende la transformación vector a vector correspondiente. Es decir que en inglés corriente aprende a pronunciar palabras impresas. Y lo hace sin que se le dé *ninguna* regla a seguir. Esta no es una hazaña cualquiera, especialmente dadas las irregularidades de la ortografía inglesa estándar. El sistema no sólo debe aprender a transformar la letra "a" en un determinado sonido. Debe aprender a transformar "a" en un sonido cuando aparece en "save" ("salvar"), en otro sonido cuando aparece en "have" (tener) y en un tercero cuando aparece en "ball" (balón). Debe aprender que "c" es suave en "city" (ciudad), pero fuerte en "cat" (gato). Y así sucesivamente.

Inicialmente, por supuesto, no hace nada de esto. Cuando se le da el texto impreso, su vector de salida produce, a través de un sintetizador de sonidos, balbuceos sin sentido como los de un bebé: "nananunu nunanana". Pero cada uno de los vectores de salida equivocados es analizado por el ordenador estándar que monitorea el proceso. Los pesos de la red se ajustan de acuerdo con la regla delta generalizada. Y la calidad de sus balbuceos mejorará lentamente. Luego de diez horas de entrenamiento sobre una muestra de 1000 palabras, produce sonidos coherentes e inteligibles dado cualquier texto

en inglés. Y lo hace sin que se representen reglas explícitas en ningún lugar del sistema.

¿Hay límites para las transformaciones que puede realizar una red en paralelo de esta clase general? La opinión actual entre los investigadores en este campo se inclina hacia la idea de que no hay límites teóricos, ya que las nuevas redes tienen importantes características que las redes de fines de la década de 1950 no tenían. Lo más importante es que la señal de salida axonal producida por cualquier unidad no es una función directa o "lineal" del nivel de excitación en la unidad misma, sino que sigue una curva en forma de S. Esta simple técnica permite que una red calcule lo que se denominan transformaciones no lineales y esto amplía enormemente la gama de problemas que puede resolver.

También es importante destacar que las nuevas redes tienen una o más capas de unidades "ocultas" que intervienen entre los niveles de entrada y de salida, mientras que las primeras redes solamente tenían una capa de entrada y de salida. La ventaja de la capa intermedia es que, dentro de ella, el sistema puede analizar estímulos posibles que no están explícitamente representados en los vectores de entrada. Así puede encontrarse con simetrías que están detrás o debajo de las simetrías superficiales que conectan los estímulos explícitos en los vectores de entrada. Esto le permite al sistema *teorizar*. Para tomar un ejemplo conocido, resulta ser que lo que las unidades ocultas en la red mina/roca realmente aprenden a codificar es si el pulso sonar ha rebotado contra algo de *metal* o *no metal*.

Tercero, las redes actuales pueden construirse mediante el algoritmo de la propagación regresiva: la regla delta generalizada. Este reciente descubrimiento es una regla de aprendizaje muy poderosa, pues permite que una red analice el espacio vectorial de sus unidades ocultas y encuentre transformaciones efectivas de toda clase posible, tanto lineales como no lineales. Permite que una gran red *encuentre* un complejo conjunto de pesos que nunca podríamos haber identificado como correctos con anticipación. Este es un importante adelanto en la tecnología del "aprendizaje de las máquinas".

Ahora el lector puede apreciar por qué las redes artificiales han captado tanta atención. Su microestructura es similar en muchos aspectos a la del cerebro y por lo menos tienen algunas de las mismas propiedades difíciles de simular.

¿Hasta dónde llega la analogía? ¿Es realmente así como trabajaría el cerebro? Permítaseme terminar este apartado ocupándome de un serio problema. Con las redes artificiales podemos construir sistemas adecuados para calcular el error de salida y para modificar los pesos de acuerdo con ello. (Por cuestiones de simplicidad ninguno de nuestros diagramas intenta mostrar esto.) Pero en un cerebro verdadero, ¿a través de qué caminos se propaga el error regresivamente hasta el conjunto pertinente de conexiones sinápticas, de modo que sus pesos puedan ser modificados y que pueda llevarse a cabo el aprendizaje? La pregunta constituye una medida de lo valioso que es contar con alguna nueva teoría, porque sin ella ni siquiera estaríamos planteando una pregunta tan específica, ni escudriñando las partes específicas del cerebro a la espera de encontrar una respuesta.

Cuando estudiamos el cerebelo, por ejemplo, encontramos que contiene un segundo sistema de entrada importante: las fibras ascendentes. Estas no se dibujaron en la figura 7.16 para evitar el amontonamiento, pero se ven fácilmente. Una fibra ascendente, como lo indica su nombre, es como una delgada enredadera que trepa por la gran célula de Purkinje desde la base y se envuelve alrededor del cuerpo celular y alrededor de las ramificaciones de sus tupidas dendritas. Cada célula de Purkinje termina envuelta en una fibra ascendente, como un roble cubierto de hiedra. Así las fibras ascendentes están en la posición correcta para hacer exactamente el trabajo que se necesita, es decir, modificar los pesos de todas las conexiones sinápticas entre las fibras paralelas y las células de Purkinje.

Desafortunadamente, aún no entendemos cómo pueden hacer esto. Tampoco estamos muy seguros de que hagan algo remotamente similar a esto. Quizá la teoría cognitiva impulsará aquí a la neurociencia a descubrir algo acerca de las actividades de las fibras ascendentes que todavía no conocía. Por otro lado, los datos neurocientíficos pueden mostrar que

una atrayente teoría del aprendizaje en el cerebelo (propagación regresiva de los errores) posiblemente no sea correcta.

Esto sería sólo una efímera decepción para la teoría cognitiva. Hay otros procedimientos para el aprendizaje, de eficiencia comparable, que sacan provecho de las limitaciones locales y no requieren propagación regresiva. Quizás el cerebro utilice alguno de ellos. Evidentemente hay que seguir investigando en el tema. Lo que es alentador acerca de esta situación, más allá de los sorprendentes éxitos ya registrados, es que la IA, las ciencias cognitivas y la neurociencia ahora interactúan enérgicamente. Ahora se enseñan entre sí, proceso del que todos se beneficiarán.

Una observación final. De acuerdo con el estilo de teoría que hemos estado estudiando aquí, son los vectores de actividad los que forman la clase más importante de representación en el cerebro. Y son las transformaciones vector a vector las que constituyen la clase más importante de cálculos. Quizás esto sea correcto o no, pero da verdadera solidez a la idea mencionada anteriormente del materialismo eliminativo (apartado 2.5) de que los conceptos de la psicología popular no necesitan aprehender los estados y actividades de la mente dinámicamente significativos. Los elementos de la cognición, tal como se describieron en las páginas precedentes, tienen un carácter desconocido para el sentido común. Quizá deberíamos esperar que, a medida que aumenta nuestra comprensión teórica, nuestra idea de los fenómenos que estamos tratando de explicar también pase por una revisión significativa. Este es un patrón común a través de la historia de la ciencia y no hay razón por la cual las ciencias cognitivas deban ser una excepción.

## Lecturas complementarias


Rumelhart, D. E., Hinton, G. E. y Williams, R. J., "Learning Representations by Back-propagating Errors", *Nature*, 323, 9 de octubre de 1986, págs. 533-36.

Sejnowski, T. J. y Rosenberg, C. R., "Parallel Networks that Learn to Pronounce English Text", *Complex Systems*, vol. 1, 1987.

Churchland, P. S. y Sejnowski, T. J., "Neural Representation and Neural Computation", *Neural Connections and Mental Computation*, Nadel, L. (comp.) Cambridge, MA, The MIT Press, 1988.

Rumelhart, D. E. y McClelland, J. L., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cambridge, MA, The MIT Press, 1986.

# 8

# Una perspectiva más amplia

## 1. La distribución de la inteligencia en el universo

Existen pruebas, como se estudió en los capítulos precedentes, de que la inteligencia consciente es un fenómeno completamente natural. De acuerdo con un amplio y creciente consenso entre los filósofos y científicos, la inteligencia consciente es la actividad de una materia organizada adecuadamente y la sofisticada organización responsable de ello, en este planeta por lo menos, es el resultado de miles de millones de años de evolución química, biológica y neurofisiológica.

Si la inteligencia se desarrolla naturalmente, a medida que el universo avanza, entonces ¿podría haberse desarrollado, o estar desarrollándose, en muchos lugares de todo el universo? La respuesta es evidentemente afirmativa, a no ser que el planeta Tierra sea absolutamente el único que posee la constitución física requerida o las circunstancias energéticas necesarias ¿Es único en los aspectos pertinentes? Analicemos el proceso evolutivo, así como lo entendemos ahora, y veamos lo que el proceso requiere.

### Flujo de energía y la evolución del orden

Básicamente se requiere un sistema de elementos físicos (como los átomos) capaces de muchas combinaciones diferen-

tes y un flujo de energía (como la luz solar) a través del sistema de elementos. Esto describe la situación en la tierra prebiológica, unos cuatro mil millones de años atrás, durante el período de evolución puramente química. El flujo de energía, que entra al sistema y sale otra vez, es esencial. En un sistema *cerrado* a la entrada y salida de energía externa, las combinaciones ricas en energía gradualmente se separarán y distribuirán su energía entre los elementos con poca energía hasta que el nivel energético sea el mismo en todas las partes del sistema: éste es el estado de *equilibrio.* Se puede decir que, como el agua, la energía busca su propio nivel; tiende a fluir "cuesta abajo" hasta que el nivel sea el mismo en todas partes.

Esta modesta analogía expresa el contenido esencial de una ley física fundamental llamada segundo principio de la termodinámica: En un sistema cerrado que aún no está en equilibrio, cualquier intercambio tiende inexorablemente a llevar el sistema hacia el equilibrio. Y una vez que un sistema ha alcanzado su estado más bajo o de equilibrio, tiende a permanecer allí para siempre, como una oscuridad uniforme e indistinguible. Por lo tanto la formación de estructuras complejas, interesantes y ricas en energía es muy improbable, ya que ello requeriría que parte de la energía interna del sistema fluyera "cuesta arriba" nuevamente. Requeriría que apareciera un *des*equilibrio energético significativo espontáneamente dentro del sistema. Y esto es efectivamente lo que prohíbe el segundo principio. Evidentemente, la evolución de estructuras complejas no se encontrará en un sistema cerrado.

Sin embargo, si un sistema está abierto a un flujo continuo de energía, la situación cambia completamente. Como ejemplo esquemático, consideremos una caja de vidrio llena de agua, con una fuente de calor constante en un extremo y una pileta de calor constante (algo que absorba la energía calórica) en el otro, como en la figura 8.1. En el agua hay un poco de nitrógeno y de anhídrido carbónico disueltos. Un extremo de la caja se calentará mucho pero, a medida que el fuego vierte energía en este extremo del sistema, es conducida hacia el extremo más frío y hacia afuera nuevamente. Por lo tanto,

la temperatura promedio en el interior de la caja es una constante.

Consideremos el efecto que esto tendrá sobre el fluido caldo en el interior de la caja. En el extremo caliente, el extremo de energía elevada, las moléculas y los átomos absorben esta energía extra y se elevan a estados de excitación. A medida que son arrastradas por el sistema, estas partes energizadas son libres de formar uniones químicas de energía elevada entre sí, uniones que habrían sido estadísticamente imposibles con el sistema en equilibrio global. Por lo tanto, es probable la formación de una variedad de compuestos químicos complejos que se agrupen hacia el extremo frío del sistema, compuestos de una mayor variedad y complejidad que los que se podrían haber formado sin el flujo continuo de energía calórica. En conjunto, el carbono, hidrógeno, oxígeno y nitrógeno son capaces de producir literalmente millones de combinaciones químicas diferentes. Con el flujo de calor encendido, este sistema abierto o *semicerrado* comienza vigorosamente a analizar estas posibilidades de combinación.

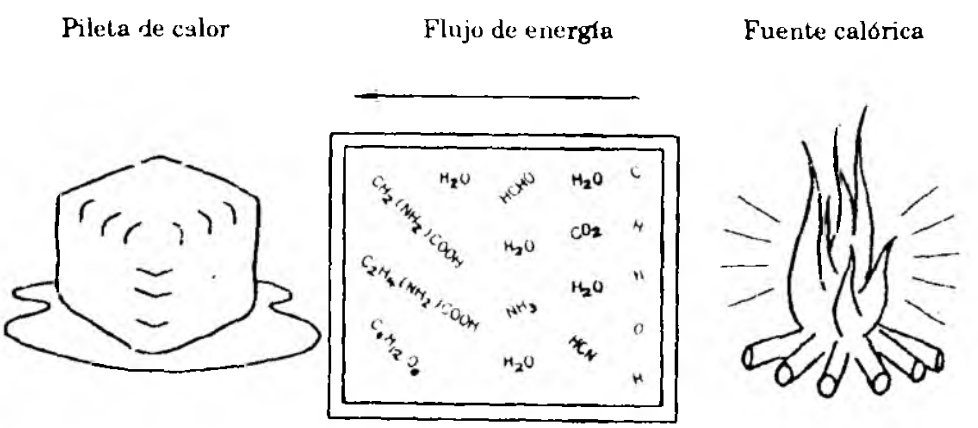



Figura 8. 1.

Es fácil ver que entonces se está llevando a cabo una especie de competencia en el interior de la caja. Algunas clases de moléculas no son muy estables y tenderán a separarse poco después de la formación. Otras clases estarán he-

chas de material más firme y tenderán a aguantar durante un tiempo. Otras clases, aunque muy inestables, pueden formarse muy frecuentemente, y entonces habrá muy pocas de ellas en el sistema en un momento dado. Algunas clases catalizan la formación de sus propios bloques de construcción, intensificando así la formación de más combinaciones. Otras clases inician ciclos catalíticos mutuamente beneficiosos y forman un par simbiótico de clases prósperas. De estas y otras formas, los diversos tipos de moléculas *compiten* por la dominación del medio líquido. Aquellas clases con gran estabilidad y/o porcentajes altos de formación constituirán las poblaciones más grandes.

El resultado típico de tales procesos es que el sistema pronto muestra una gran cantidad de casos de una variedad bastante pequeña de distintas clases de moléculas complejas y almacenadoras de energía. (Qué clases, de los millones de clases posibles, realmente llegan a dominar el sistema depende de la composición original del caldo y del nivel de flujo y es sumamente sensible a estos factores.) El sistema presenta un orden, una complejidad y una distribución de energía en desequilibrio que serían impensables sin el flujo de energía a través del sistema. El flujo mueve el sistema. Lo obliga a salir de su caos inicial y lo lleva hacia las muchas formas de orden y complejidad de las que es capaz. Lo que era improbable se ha tornado inevitable.

El experimento precedente es esquemático, creado para ilustrar un principio general, pero se han realizado efectivamente casos del mismo. En un ahora famoso experimento, Urey y Miller, en 1953, recrearon la atmósfera de la Tierra antes de la vida (hidrógeno, metano, amoníaco y agua) y la sometieron a una descarga eléctrica constante. Luego de varios días de este flujo energético, el análisis de los contenidos del frasco mostraba que se habían formado muchos compuestos orgánicos complejos, incluyendo una cantidad de aminoácidos, las unidades a partir de las cuales se construyen las moléculas proteicas. Otras versiones del experimento probaron con diferentes fuentes de energía (luz ultravioleta, calor, ondas de choque) y todas presentaron el mismo patrón: un

flujo de energía induce orden y complejización dentro de un sistema semicerrado.

La naturaleza también ha efectuado este experimento, con toda la Tierra y con miles de millones de otros planetas. Pues la Tierra en su totalidad también es un sistema semicerrado, con el sol como fuente de energía y el vacío negro que nos rodea la pileta de energía a baja temperatura (figura 8.2). La energía solar ha estado fluyendo a través de este sistema gigante durante más de cuatro mil millones de años, analizando pacientemente las infinitas posibilidades de orden, estructura y complejidad inherentes a la materia que contiene. No es sorprendente que haya superado a los sistemas artificiales descriptos.

Desde esta perspectiva es evidente que *cualquier* planeta soportará un rico proceso evolutivo, si posee una rica variedad de elementos en alguna solución líquida y si cuenta con un adecuado flujo energético desde una estrella cercana. ¿Aproximadamente cuántos planetas en la Vía Láctea reúnen estas condiciones?

## Distribución de los lugares evolutivos

En nuestra galaxia hay aproximadamente 100 mil millones o $10^{11}$ estrellas. ¿Cuántas de ellas poseen planetas? Las teorías de la formación estelar, los estudios espectrográficos de la rotación de las estrellas y los estudios telescópicos de los efectos dinámicos de los compañeros negros coinciden al señalar que efectivamente todas las estrellas, excepto las gigantes supercalientes, son sólo un pequeño porcentaje de la cantidad total, de modo que su supresión aún nos deja cerca de $10^{11}$ sistemas planetarios en la galaxia.

¿Cuántos de ellos contendrán un planeta adecuadamente constituido y situado? Una constitución adecuada indica que deberíamos considerar sólo sistemas de segunda generación, formados de los desperdicios de explosiones estelares anteriores, ya que éstos son la principal fuente de los elementos, además del helio y el hidrógeno. Esto nos deja un poco menos

de la mitad de los sistemas disponibles, por lo que bajamos a $10^{10}$. En estos sistemas que nos quedan, los planetas con una constitución aceptable prometen ser bastante comunes. En nuestro sistema solamente, la Tierra, Marte y dos de las lunas de Júpiter presentan agua en cantidades significativas, si es que exigimos que la misma sea nuestro solvente evolutivo. Las lunas de Júpiter tienen una importancia extra, ya que el gigante Júpiter y sus doce satélites casi constituyen un sistema solar en miniatura, el único ejemplo adicional de que disponemos para estudiar en detalle. Curiosamente, los satélites segundo y tercero de Júpiter, Europa y Ganimedes, contienen cada uno tanta agua como la Tierra: aunque sus superficies son menores, sus océanos son mucho más profundos que los nuestros. Entonces, si podemos generalizar a partir de estos dos sistemas, los planetas con agua se encontrarán en una amplia gama de sistemas estelares y algunos harán alarde de dos o más.

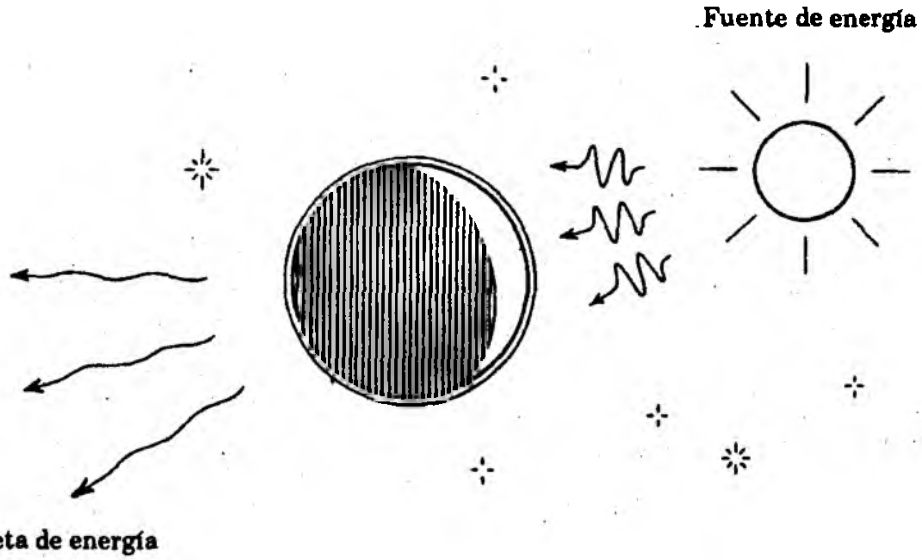



Figura 8. 2

Los planetas con agua tampoco agotan las posibilidades. El amoníaco líquido y el metano líquido también son solventes

comunes, y por lo tanto pueden sostener procesos evolutivos. Tales océanos se encuentran en planetas mucho más fríos y permitirían el análisis de uniones químicas de energía mucho menores que las que caracterizan la bioquímica de la Tierra. Aquellos medios tonificantes constituyen un nicho evolutivo alternativo. En conjunto, la constitución adecuada parece no ser un problema. Quedémonos con un cálculo de por lo menos $10^{10}$ planetas adecuadamente constituidos para una evolución química significativa.

¿Cuántos de ellos estarán correctamente situados en relación con la estrella proveedora de energía? La órbita del planeta debe estar dentro de la "zona de vida" de su estrella: lo suficientemente lejos de ella como para evitar que el solvente esté en ebullición y se evapore, y a la vez lo suficientemente cerca como para evitar que se solidifique por congelamiento. Para el agua esa zona es bastante amplia, y hay grandes probabilidades de que alguna órbita planetaria esté dentro de ella. Sin embargo, necesitamos un planeta con *agua* y quizá sólo haya uno de cada diez así. Consideremos, de un modo tradicional, que sólo uno de cien de nuestros sistemas restantes contiene un planeta con agua adecuadamente situado. También esperamos que haya planetas con amoníaco y metano adecuadamente situados, pero las mismas consideraciones dan un cálculo similar para ellos y nos queda un resultado de aproximadamente $10^8$ planetas que estén tanto adecuadamente situados como constituidos.

Este cálculo supuso una estrella como nuestro sol. Pero mientras el sol ya es una estrella pequeña y común, la mayoría de las estrellas son más pequeñas aún y más frías y por lo tanto tendrán zonas de vida más pequeñas. Esto podría reducir las probabilidades de una posición adecuada en uno o dos factores más de diez. Aun así, las estrellas del tamaño del sol constituyen aproximadamente el 10 por ciento de la población pertinente y su sola consideración nos dejaría con por lo menos $10^7$ planetas en condiciones de lograrlo.

Entonces nuestro cálculo tradicional es que el proceso evolutivo se está alejando ruidosa y velozmente, en una u otra etapa, en por lo menos 10.000.000 de planetas de esta galaxia solamente.

## Vida e inteligencia

Lo importante de este número es que es grande. La clase de proceso que nos dio origen es aparentemente común en todo el universo. La conclusión es emocionante, pero la pregunta real sigue sin ser respondida: ¿En cuántos de estos casos el proceso evolutivo ha dado expresión a la materia en el nivel de *vida* verdadera y en cuántos de ellos ha producido *inteligencia consciente*?

Estas fracciones son imposibles de calcular con seguridad, ya que ello requeriría una comprensión de las *velocidades* a las que se lleva a cabo el desarrollo evolutivo y de los caminos alternativos que puede tomar. Hasta ahora tenemos una comprensión insuficiente de la volátil dinámica de la evolución para descifrar estos asuntos. Estamos reducidos a analizar las consideraciones pertinentes, pero éstas aún pueden ser informativas. Comencemos con el concepto corriente, insinuado en el párrafo precedente, de que la evolución tiene dos grandes brechas discontinuas que cubrir: la brecha entre la no vida y la vida, y la brecha entre la conciencia y la no conciencia. Ambas distinciones, tan arraigadas en el sentido común, contienen un grado de *error* de concepto. De hecho, ninguna corresponde a ninguna discontinuidad bien definida o imposible de llenar en la naturaleza.

Consideremos la noción de vida. Si tomamos la capacidad de autoduplicación como rasgo esencial, entonces su aparición no tiene que representar una discontinuidad. Las moléculas que catalizan la formación de sus propios bloques de construcción representan una posición inferior en el mismo espectro. Sólo tenemos que imaginar una serie de moléculas progresivamente más eficientes y de rápida acción de esta clase y podemos culminar llanamente en una molécula que catalice sus bloques de construcción en una secuencia tal que puedan acoplarse tan rápidamente como son producidas: una molécula que se autoduplica. Aquí no hay discontinuidad ni brecha que cubrir. El *medio* puede presentar discontinuidades, cuando la eficiencia de un determinado duplicador pasa un punto crítico relativo a su competencia, pero ésta es una dis-

continuidad en las consecuencias de la autoduplicación, no en los mecanismos que la producen.

Por otro lado, la mera autoduplicación puede ser un concepto de vida demasiado simple. Hay algunas razones para rechazarlo. Podemos imaginar algunas moléculas muy simples que, en un medio químico adecuadamente artificial y fabricado, se autoduplicarían. Pero esto solo no debería tentarnos a considerarlas como vivas. En todo caso, hay una caracterización de la vida más aguda a mano, que podemos ilustrar con la célula, la unidad de vida más pequeña, según algunas descripciones. Una célula es un diminuto sistema semicerrado y autoorganizado, dentro del sistema semicerrado más grande de la biosfera de la Tierra. La energía que fluye a través de una célula sirve para mantener e incrementar el orden interno de la célula. En la mayoría de las células el flujo energético es químico —ingieren moléculas ricas en energía y se apropian de la energía que liberan—, pero las células capaces de fotosíntesis hacen uso directo del flujo solar para accionar sus procesos metabólicos. Todo esto indica que definimos un objeto vivo como cualquier sistema físico semicerrado que saca provecho del orden que ya posee y del flujo energético que lo atraviesa de modo de mantener y/o incrementar su orden interno.

Esta caracterización efectivamente tiene en cuenta algo muy importante en relación con las cosas que corrientemente consideramos vivas. Y abarca sin ningún problema a los organismos multicelulares, pues una planta o un animal también es un sistema semicerrado, compuesto por diminutos sistemas semicerrados: una vasta conspiración de células más que (sólo) una vasta conspiración de moléculas. Aun así, la definición tiene algunas consecuencias levemente sorprendentes. Si lo aceptamos, un panal de abejas es un objeto vivo. Como también lo es una colonia de termitas. Y también una ciudad. De hecho, toda la biosfera es un objeto vivo. Pues todas estas cosas satisfacen la definición propuesta.

En el otro extremo del espectro —y esto nos conduce nuevamente al tema de la discontinuidad—, algunos sistemas muy simples pueden reivindicarse como vivos. Consideremos

la gota incandescente de la llama de una vela: es también un sistema semicerrado y, aunque su orden interno es pequeño y su automantenimiento débil, puede satisfacer apenas las condiciones de la definición propuesta. Otros sistemas fronterizos presentarán problemas similares. ¿Entonces debemos rechazar la definición? No. La lección más sabia es que los sistemas vivos se diferencian de los no vivos sólo por grados. No hay una brecha metafísica que cubrir: sólo una suave pendiente que escalar, una pendiente medida en grados de orden y de autorregulación.

Surge la misma lección cuando consideramos la inteligencia consciente. Ya hemos visto cómo la conciencia y la inteligencia existen en diferentes grados, se extienden a lo largo de un amplio espectro. Sin duda la inteligencia no es exclusiva de los seres humanos: millones de otras especies la presentan en alguna medida. Si definimos la inteligencia crudamente como la posesión de un conjunto complejo de respuestas apropiadas al medio cambiante, entonces hasta la humilde patata presenta una cierta astucia inferior. Aquí no surgen discontinuidades metafísicas.

Pero esa definición es demasiado cruda. Deja afuera el aspecto del desarrollo o el aspecto creativo de la inteligencia. Entonces consideremos la siguiente y más aguda definición. Un sistema tiene inteligencia sólo en el caso de que aproveche la *información* que ya tiene y el flujo energético que lo atraviesa (esto incluye el flujo de energía a través de sus órganos de los sentidos) de modo tal que *incrementa* la información que contiene. Dicho sistema puede *aprender* y ése parece ser el elemento central de la inteligencia.

Esta definición mejorada sin duda incluye algo muy importante acerca de las cosas que comúnmente consideramos inteligentes. Y espero que el lector ya se haya sorprendido por el cercano paralelismo entre esta definición de la inteligencia y nuestra anterior definición de vida como el aprovechamiento del orden contenido y del flujo energético para lograr un mayor orden. Estos paralelos son importantes por la siguiente razón. Si la posesión de información puede entenderse como la posesión de algún orden físico interno que tiene alguna relación

sistemática con el medio, entonces las operaciones de inteligencia, abstractamente concebidas, resultan ser sólo una versión de alto grado de las operaciones características de la vida, excepto que están más intrincadamente ligadas al medio.

Esta hipótesis es consistente con el uso que el cerebro hace de la energía. La producción de grandes cantidades de clases específicas de orden requiere un flujo energético muy grande. Y mientras que el cerebro constituye sólo el 2 por ciento de la masa corporal, consume, cuando está muy activo, más del 20 por ciento del resto del presupuesto energético del cuerpo. El cerebro también es un sistema semicerrado, curiosamente de alta intensidad, cuyo orden microscópico siempre cambiante refleja el mundo con un impresionante detallismo. Aquí nuevamente la inteligencia no representa una discontinuidad. La vida inteligente es simplemente vida, con una alta intensidad termodinámica y un acoplamiento especialmente estrecho entre el orden interno y las circunstancias externas.

Todo esto significa que, dada energía suficiente, y tiempo, deben esperarse tanto la vida *como* la inteligencia entre los productos naturales de la evolución del planeta. Hay energía suficiente, y planetas. ¿Ha habido tiempo? En la Tierra ha habido tiempo suficiente, pero ¿qué hay acerca de los otros $10^7$ candidatos? Nuestra incertidumbre aquí es muy grande. A priori, la probabilidad de que seamos el primer planeta en haber desarrollado vida inteligente es cada vez más *pequeña*: no más que una probabilidad de uno en $10^7$. Y ésta disminuye aun más cuando consideramos que las estrellas ya habían estado suministrando energía a los planetas durante por lo menos 10 mil millones de años cuando el sistema Sol/Tierra se originó, unos 4,5 mil millones de años atrás. En todo caso ingresamos en la carrera de la evolución con un gran handicap. Por otro lado, los ritmos evolutivos pueden ser muy volátiles, variando en órdenes de magnitud como una función de variables planetarias sutiles. Esto tornaría insignificante nuestro handicap temporal, y hasta podríamos ser el primer planeta de nuestra galaxia en haber desarrollado inteligencia.

Ninguna decisión tomada aquí puede lograr la confianza, pero una decisión obligada, tomada bajo las inciertas conjetu-

ras precedentes, tendría que suponer que aproximadamente la mitad de los candidatos pertinentes están detrás de nosotros, y la mitad están adelante. Esta "mejor suposición" significa que unos $10^6$ planetas en esta galaxia solamente *ya* han producido vida altamente inteligente.

¿Significa esto que deberíamos esperar que aparezcan hombrecitos verdes en platos voladores en nuestra atmósfera? No. Ni siquiera si aceptamos la "mejor suposición". Las razones son importantes y hay tres. La primera razón es la distribución espacial de los $10^6$ planetas. Nuestra galaxia tiene un volumen de más de $10^{14}$ años luz cúbicos (es decir, la distancia recorrida en un año a la velocidad de la luz = 300.000 kilómetros por segundo × un año, que es aproximadamente 9,6 billones de kilómetros), y $10^6$ planetas distribuidos en todo este volumen tendrían una distancia promedio entre sí de más de 500 años luz. Es una distancia muy poco práctica para visitas ocasionales.

La segunda razón, y quizá más importante, es la dispersión temporal. No podemos suponer que los $10^6$ planetas desarrollarán todos vida inteligente simultáneamente. Tampoco podemos estar seguros de que, una vez desarrollada, la vida inteligente dure demasiado. Suceden accidentes, aparece la degeneración, hay autodestrucción. A modo de ejemplo supongamos que el tiempo de vida promedio de la inteligencia en cualquier planeta es de 100 millones de años (es el intervalo entre la aparición de los primeros mamíferos y la destrucción nuclear que acabaría con nosotros en este siglo). Si estos intervalos de inteligencia se distribuyen uniformemente en el tiempo, entonces cualquier planeta con inteligencia probablemente tenga sólo $10^4$ planetas inteligentes simultáneamente como compañeros, con una distancia promedio entre ellos de 2.500 años luz. Más aún, nada garantiza que aquellas otras cunas de inteligencia simultáneas exhiban nada más inteligente que ratones de campo u ovejas. Nuestro propio planeta ha superado ese nivel sólo recientemente. Y las civilizaciones muy inteligentes y con alta tecnología pueden durar en promedio sólo 1000 años, debido a algunas inestabilidades intrínsecas. En tal caso estarán casi siempre absoluta y dramá-

ticamente solas en la galaxia. La posibilidad de compañía altamente inteligente comienza a parecer bastante remota.

Y así es si asignamos tendencias suicidas a toda compañía potencial. Si no lo hacemos, entonces podemos volver a un cálculo más optimista de la compañía actual. Si suponemos una duración promedio de la vida inteligente de entre mil y cinco mil millones de años, entonces la dispersión temporal nos dejará aún con $10^5$ planetas simultáneamente por delante de nosotros en el desarrollo evolutivo. Esto parece finalmente prometer la aparición de hombrecitos verdes y alguna comunicación edificante, aunque sea por radiotelescopio. Pero no lo hace debido a la tercera razón y la más importante de todas: la variedad potencialmente interminable de las diferentes *formas* que pueden tomar la vida y la inteligencia.

Nuestra biosfera ha sido organizada en unidades individuales de vida independiente: células y organismos multicelulares. Nada de esto es estrictamente necesario. Algunas biosferas pueden haber evolucionado en una única "célula" unificada, masivamente compleja y muy inteligente que circunde el planeta. Otras pueden haber sintetizado sus células o elementos multicelulares en un individuo planetario singular y análogamente unificado. Si uno de nosotros tratara de comunicarse con dicho ente sería como una bacteria del pantano local intentando comunicarse con un ser humano, emitiendo algunas sustancias químicas. El ente mayor simplemente no está "interesado".

Incluso con criaturas más conocidas, un medio diferente puede exigir diferentes órganos de los sentidos, y éstos pueden significar cerebros muy diferentes (hablando en general, los cerebros deben evolucionar desde la periferia sensorial hacia adentro, desarrollándose en formas que sirvan a las modalidades disponibles). Las criaturas que navegan por campos eléctricos percibidos, que cazan por buscadores de dirección en el infrarrojo extremo, que guían la manipulación cercana por la estereoaudición en el rango de los 50 kilohertz, y que se comunican mediante fugas de hidrocarburos aromáticos, no tienen probabilidades de pensar como un ser humano.

Dejando de lado los extraños órganos de los sentidos, el conglomerado concreto de aptitudes cognitivas que se encuen-

tra en nosotros no tiene por qué caracterizar a otras especies. Por ejemplo, es posible ser muy inteligente y sin embargo carecer de toda capacidad para manejar números, incluso la capacidad de contar hasta cinco. También es posible ser muy inteligente y sin embargo carecer de toda capacidad para entender o manejar el lenguaje. Estas deficiencias aisladas ocurren ocasionalmente en seres humanos con aptitudes mentales ejemplares de otra clase. El primero es un síndrome raro pero conocido llamado *acalculia*. El segundo, más común, es una afección llamada *afasia global*. Por lo tanto, no debemos esperar que una especie diferente muy inteligente deba saber inevitablemente las leyes aritméticas o que pueda aprender un sistema como el lenguaje, o que tenga siquiera una sospecha de que estas cosas existen. Estas reflexiones indican también que puede haber capacidades cognitivas fundamentales cuya existencia ¡*nosotros* desconocemos por completo!

Finalmente, no debemos esperar que las metas o los intereses de una especie inteligente diferente se parezcan a los nuestros, o ni siquiera que nos resulten inteligibles. La meta final de toda una especie puede ser terminar de componer la indefinidamente larga sinfonía magnética iniciada por sus ancestros prehistóricos, una sinfonía en la que los jóvenes se socializan aprendiendo a cantar los primeros movimientos. Otra especie podría tener una singular devoción por el estudio de las matemáticas superiores, y sus actividades significarían tanto para nosotros como las actividades del departamento de matemáticas de una universidad para un Neanderthal. Igualmente importante es que las metas raciales mismas sufren un cambio evolutivo genético o cultural. Las metas dominantes de nuestra propia especie desde hace 5000 años pueden no tener relación con nuestros intereses actuales. Todo esto significa que no podemos esperar que una especie inteligente diferente comparta los entusiasmos e intereses que caracterizan a nuestra propia efímera cultura.

El punto del análisis precedente ha sido plantear preguntas acerca de la naturaleza de la inteligencia en una perspectiva más amplia de la que generalmente se utiliza, y enfatizar la naturaleza extremadamente general o abstracta de este

fenómeno natural. La inteligencia humana actual es sólo una variación de un tema muy general. Aunque, como parece probable, la inteligencia *esté* bastante expandida dentro de nuestra galaxia, no podemos decir casi nada acerca de lo que estarán haciendo aquellas otras especies inteligentes o acerca de qué forma toma su inteligencia. Si la definición teórica de inteligencia dada anteriormente es correcta, entonces podemos deducir que deben estar utilizando *energía* (quizás en abundantes cantidades) y creando *orden*, y que por lo menos parte del orden creado tiene algo que ver con mantener interacciones productivas con su medio. Más allá de eso, todo es posible. Tanto para nosotros como para ellos.

### Lecturas complementarias


Schrödinger, E., *What is Life?*, Cambridge, Cambridge University Press, 1945.
Shklovskii, I. S. y Sagan C., *Intelligent Life in the Universe*, Nueva York, Dell, 1966.
Cameron, A. G. W., *Interstellar Communication: The Search for Extraterrestrial Life*, Nueva York, Benjamin, 1963.
Sagan, C. y Drake, F., "Search for Extraterrestrial Intelligence", *Scientific American*, vol. 232, mayo de 1975.
Morowitz, H., *Energy Flow in Biology*, Nueva York, Academic Press, 1968.
Feinberg, G. y Shapiro, R., *Life beyond Earth*, Nueva York, William Morrow and Company, 1980.


## 2. La expansión de la conciencia introspectiva

Para dar un cierre a este libro, volvamos del universo en su totalidad y orientemos nuestra atención sobre el fenómeno de la conciencia introspectiva o autoconciencia. He estado utilizando un concepto muy general y neutro de la introspección en todo el libro, que puede describirse del siguiente modo.

Tenemos una gran variedad de estados y procesos inter-

nos. También tenemos ciertos mecanismos innatos para discriminar la aparición de algunos de esos estados y procesos y su no presencia, y para diferenciarlos unos de otros. Y cuando apelamos a dicha actividad discriminatoria y la observamos podemos responder a ella con movimientos explícitamente conceptuales, es decir, con *juicios* más o menos apropiados acerca de aquellos estados y procesos internos, juicios enmarcados en los conceptos conocidos del sentido común: "Tengo una sensación del rosa", "Me siento mareado", "Tengo un dolor", y así sucesivamente. Así tenemos cierto acceso, aunque incompleto, a nuestras propias actividades internas.

Se supone que el autoconocimiento es algo bueno, de acuerdo con la ideología de casi todo el mundo. ¿Entonces cómo podríamos mejorar o intensificar este acceso introspectivo? La modificación quirúrgica o genética de esos mecanismos introspectivos innatos es una posibilidad, pero no es realista en el corto plazo. A falta de esto, quizá podamos aprender a hacer un uso más refinado y agudo de los mecanismos discriminatorios que ya poseemos.

Las modalidades del sentido externo brindan muchos precedentes para esta idea. Consideremos el enorme aumento en la habilidad discriminatoria (y el conocimiento teórico) que abarca la brecha entre la comprensión auditiva de un niño no entrenado de la Quinta Sinfonía de Beethoven y la comprensión auditiva de la misma persona de la misma sinfonía cuarenta años más tarde, oída en su calidad de director de la orquesta que la interpreta. Lo que antes era una única voz ahora es un mosaico de elementos diferenciables. Lo que antes era una tonada levemente reconocida es ahora una secuencia racionalmente estructurada de acordes diferenciables que sostienen una línea melódica apropiadamente relacionada. El director oye mucho más de lo que oía el niño y probablemente mucho más que la mayoría de nosotros.

Otras modalidades constituyen ejemplos similares. Consideremos al químicamente sofisticado catador de vinos, para quien la amplia categoría de "vino tinto" utilizada por la mayoría de nosotros se divide en una red de quince o veinte elementos distinguibles: etanol, glucol, fructosa, sacarosa, tanino, ácido, dióxido de carbono y otros, cuyas concentracio-

nes relativas puede calcular con exactitud. Saborea más que nosotros. O consideremos al astrónomo, para quien la negra cúpula moteada de su juventud se ha convertido en un abismo visible, distribuyendo planetas cercanos, estrellas enanas amarillas, gigantes azules y rojas e incluso una o dos remotas galaxias, todos diferenciables como tales y localizables en el espacio tridimensional con su ojo desnudo *(desnudo)*. Ve mucho más que nosotros. Cuánto más difícil es de determinar antes de adquirir realmente la habilidad correspondiente.

Lo que se domina finalmente en cada uno de estos casos es un marco conceptual —ya sea musical, químico o astronómico— un marco que incluye mucho más conocimiento sobre el dominio sensorial pertinente que lo *inmediatamente* evidente para la discriminación no guiada. Estos marcos son generalmente una herencia cultural, reunidos a través de muchas generaciones, y su dominio proporciona una riqueza y penetración en nuestras vidas sensoriales que serían imposibles en su ausencia.

Volviendo ahora a la introspección, es evidente que nuestra vida introspectiva ya es la amplia beneficiaria de este fenómeno. Las discriminaciones introspectivas que hacemos son en su mayor parte aprendidas; se adquieren con práctica y experiencia, a menudo bastante lentamente. Y las discriminaciones específicas que aprendemos a hacer son las que nos resultan útiles de realizar. Generalmente, son las discriminaciones que otros ya están haciendo, las discriminaciones que se encuentran dentro del vocabulario psicológico del lenguaje que aprendemos. El marco conceptual para los estados psicológicos que está arraigado en el lenguaje corriente es, como vimos en los capítulos 3 y 4, un logro teórico modestamente sofisticado por derecho propio, y moldea profundamente nuestra introspección desarrollada. Si incluyera mucho *menos* conocimiento en sus categorías y generalizaciones conectadas, nuestra comprensión introspectiva de los estados y actividades internas sería mucho menor, aunque nuestros mecanismos discriminatorios innatos siguieran siendo los mismos. Correlativamente, si incluyera mucho *más* conocimiento sobre la naturaleza interna de lo que tiene ahora,

nuestra discriminación introspectiva y nuestro reconocimiento podrían ser mucho *mayores* que ahora, aunque los mecanismos discriminatorios innatos siguieran siendo los mismos.

Esto me lleva a la última propuesta positiva del capítulo. Si, finalmente, el materialismo está en lo cierto, entonces es el marco conceptual de una neurociencia madura que encarnará la sabiduría esencial sobre nuestra naturaleza interna. (Por ahora ignoro aquí las sutilezas que dividen a las diversas formas de materialismo.) Entonces consideremos la posibilidad de aprender a describir, concebir y comprender introspectivamente las abundantes complicaciones de la vida interior dentro del marco conceptual de una neurociencia "madura", o de una que haya avanzado más allá de su estado actual. Supongamos que hemos entrenado nuestros mecanismos innatos para realizar un nuevo conjunto más detallado de discriminaciones, un conjunto que correspondiera no a la taxonomía psicológica primitiva del lenguaje corriente, sino a alguna taxonomía más aguda de estados obtenidos de una neurociencia "madura". Y supongamos que nos entrenamos para responder a dicha actividad reconfigurada con juicios que fueron enmarcados, como hábito, en los conceptos apropiados de la neurociencia.

Si los ejemplos del director de orquesta, del experto en vinos y del astrónomo proporcionan un paralelo justo, entonces la intensificación en nuestra visión introspectiva podría aproximarse a una revelación. El consumo de glucosa en el telencéfalo, los niveles de dopamina en el tálamo, los vectores de codificación en vías nerviosas específicas, las resonancias en la eneava capa de la corteza periestriada e innumerables otras sutilezas neurofisiológicas y neurofuncionales podrían llevarse al foco objetivo de nuestra discriminación introspectiva y nuestro reconocimiento conceptual, así como los acordes de séptima en sol menor y de la dominante mayor se llevan al foco objetivo de la discriminación auditiva y el reconocimiento conceptual de un músico entrenado. Por supuesto tendremos que *aprender* el marco conceptual de la neurociencia proyectada para poder salir adelante. Y tendremos que practicar para llegar a tener la habilidad de aplicar

estos conceptos a nuestros juicios no deductivos. Pero éste parece un bajo precio a pagar dado el resultado proyectado.

Esta propuesta se lanzó inicialmente en nuestro análisis del materialismo eliminativo, pero la posibilidad también está abierta a las demás posiciones materialistas. Si el materialista reduccionista está en lo correcto, entonces la taxonomía de la psicología popular encajará más o menos fácilmente en alguna subestructura de la taxonomía de una neurociencia "madura". Pero esa nueva taxonomía aun incluirá por lejos el conocimiento más profundo de nuestra naturaleza. Y si el funcionalista está en lo correcto, entonces la teoría "madura" será más abstracta y computacional en su visión de nuestras actividades internas. Pero esa visión superará los simples conceptos cinemáticos y explicativos del sentido común. En los tres casos, el paso al nuevo marco promete un avance comparable, tanto en conocimiento general como en la autocomprensión.

Entonces digo que la genuina llegada de una cinemática y de una dinámica materialistas para estados psicológicos y procesos cognitivos, constituirá, no una penumbra en la que nuestra vida interior será eclipsada o suprimida, sino un amanecer en el que sus maravillosas complejidades serán finalmente reveladas —tranquilamente, si nos concentramos, en la introspección autoconsciente.