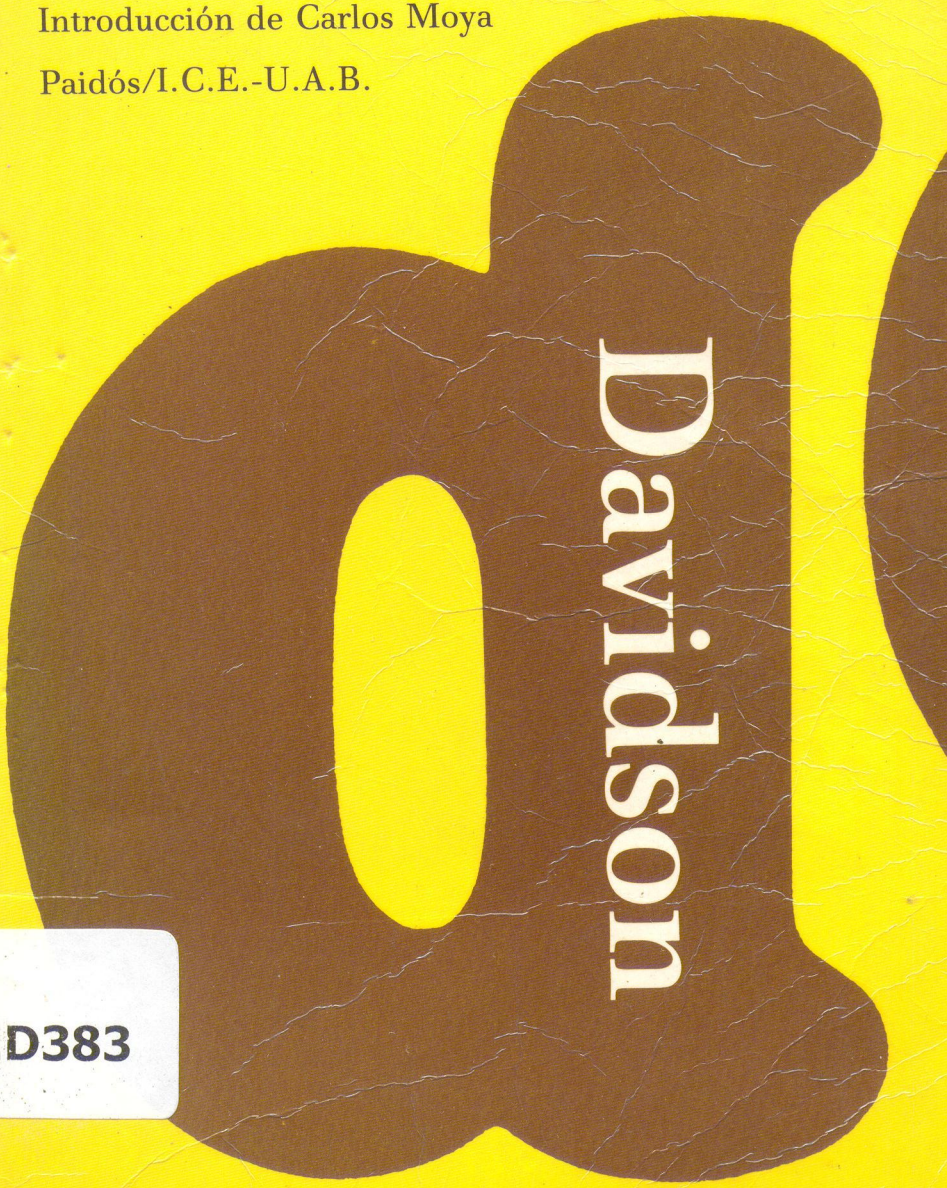
Donald Davidson

Mente, mundo y acción

Introducción de Carlos Moya

Paidós/I.C.E.-U.A.B.



Pensamiento Contemporáneo 20

Mente, mundo y acción
Donald Davidson

La obra de Donald Davidson representa un punto de referencia central en distintos e importantes campos de la filosofía del presente. Sus ideas han influido decisivamente en el desarrollo de la filosofía de la mente y de la acción, la filosofía del lenguaje, la ontología y la teoría del conocimiento. Partiendo de la obra de Quine, Davidson ha avanzado hacia posiciones propias en todos estos ámbitos, renunciando al empirismo de su maestro. Su concepción causal de la mente, el significado y la acción no excluye el énfasis en el carácter racional que singulariza estos fenómenos frente a otros procesos causales y los sitúa más allá del alcance de las leyes científicas. La original combinación de estas dos perspectivas, causal y racional, naturalista y humanista, tradicionalmente enfrentadas, ha suscitado una viva polémica, testimonio claro del interés y profundidad de la posición que le ha dado origen.
El presente volumen incluye una amplia y representativa muestra del pensamiento davidsoniano, seleccionada expresamente por el propio autor para esta edición castellana. Estos trabajos van precedidos por una cuidada introducción a cargo de Carlos Moya, profesor titular de filosofía en la Universidad de Valencia, que ha llevado también a cabo la traducción y la ha completado con notas aclaratorias. El conjunto constituye un instrumento fundamental para el conocimiento de la filosofía davidsoniana.

ISBN 84-7509-790-1

9 788475 097909 46020

Diseño: Mario Eskenazi, Pablo Martín

Donald Davidson

Mente, mundo y acción

Claves para una interpretación

Introducción y traducción de Carlos Moya

Ediciones Paidós
I.C.E. de la Universidad Autónoma de Barcelona
Barcelona - Buenos Aires - México

Título original: *The Myth of the Subjective*
A Coherence Theory of Truth and Knowledge
Deception and Division
Knowing One's Own Mind
The conditions of Thought
Publicado en inglés en *Bewusstsein. Sprache und Kunst; Kant oder Hegel; The Multiple Self; Proceedings and Adresses of the American Philosophical Association* y *Le Cahier du Collège International de Philosophie* respectivamente

Traducción de Carlos Moya Espí

Cubierta de Mario Eskenazi y Pablo Martín

1.ª edición, 1992

ISBN 84-7509-790-1
Depósito legal: B-17.542/1992

Impreso en Nova-Gràfik, S.A.
Puigcerdà, 127 - 08019 Barcelona

Impreso en España - Printed in Spain

SUMARIO

Introducción a la filosofía de Davidson: mente, mundo y acción, por *Carlos Moya* 9

1. El contexto filosófico 9
2. Razones y causas: la acción intencional 14
3. El monismo anómalo 22
4. Significado, verdad e interpretación 26
5. Mente, comunidad y mundo objetivo 33
6. Conclusión: sujeto, causa e intencionalidad 38

MENTE, MUNDO Y ACCION

Prefacio 49
El mito de lo subjetivo 51
Verdad y conocimiento: una teoría de la coherencia 73
Engaño y división 99
El conocimiento de la propia mente 119
Las condiciones del pensamiento 153

INTRODUCCIÓN A LA FILOSOFÍA DE DAVIDSON: MENTE, MUNDO Y ACCIÓN

La obra de Donald Davidson constituye un punto de referencia central en la filosofía del presente. Esta Introducción a dicha obra persigue un doble objetivo. En primer lugar, trata de ofrecer una visión de conjunto del intrincado territorio del pensamiento davidsoniano, disperso insularmente en diversos ensayos y artículos. En segundo lugar, aspira a ser a la vez una invitación y una guía básica para el recorrido directo por los textos davidsonianos, una importante muestra de los cuales, sugerida por el propio autor, se incluye en el presente volumen.

1. El contexto filosófico

Si tuviéramos que indicar la influencia intelectual que gravita con más peso sobre la filosofía de Davidson, sin duda deberíamos mencionar a Willard O. Quine. Esta deuda es reconocida con generosidad y sin ambages por nuestro autor. Con cierto grado de arbitrariedad, podríamos distinguir, en el pensamiento de Quine, los siguientes aspectos fundamentales, cuyo contenido será desarrollado en breve: el empirismo, la concepción naturalizada de la reflexión filosófica, el naturalismo y el materialismo. La actitud de Davidson hacia esos puntos de referencia quinianos nos servirá para iniciar este recorrido introductorio por su pensamiento.

Según interpreta nuestro autor la situación de la filosofía en el presente, estamos asistiendo a un cambio fundamental, de consecuencias enormes, en este campo. En la raíz misma de este cambio se halla, según Davidson, la crítica a que está

siendo sometida la concepción tradicional de las relaciones entre la subjetividad y el mundo objetivo. Desde Descartes hasta nuestros días, esta concepción se ha basado en la postulación de entidades mediadoras entre ambos términos de la relación: las ideas de Descartes y Locke, las impresiones e ideas de Hume, las intuiciones y conceptos de Kant, los datos sensoriales del positivismo lógico. No pretendo pasar por alto las diferencias entre estos diversos tipos de entidades intermedias, que en ocasiones revelan un fuerte racionalismo, incluso platonismo, y en ocasiones manifiestan un decidido y radical empirismo. Más allá de estas diferencias, sin embargo, todas ellas coinciden en su función de mediación entre el sujeto y el mundo objetivo; todas ellas presuponen, además, la posibilidad de distinguir, en el marco del conocimiento y del pensamiento humanos, entre los conceptos y un contenido no contaminado por ellos. Uno de los aspectos fundamentales del pensamiento davidsoniano –que en este libro viene representado sobre todo, aunque no exclusivamente, por el artículo «El mito de lo subjetivo»– está constituido por la crítica a dicha separación entre los conceptos y un material neutro, no conceptualizado. Esta crítica conlleva un ataque frontal a la tradición epistemológica que arranca en Descartes y a la concepción de la mente asociada a ella, así como un cambio de rumbo decisivo en la reflexión filosófica. Con este cambio de rumbo no sólo se pone en cuestión la inteligibilidad de la idea de un dato sensorial absolutamente libre de conceptuación, sino también la comprensión tradicional de los conceptos mismos y de su función, y el hecho de que esta función haya sido definida bajo el supuesto de la separación entre ellos y un elemento no conceptual; definición que ahora es sometida a revisión crítica. La idea de que los conceptos son formas o estructuras de organización de un material conceptualmente neutro pierde contenido desde el momento en que no hay tal material en espera de organización.

Podemos caracterizar ahora el empirismo de modo muy sumario como una determinada forma de entender la relación entre ese material libre de conceptuación y los conceptos mismos. Entre el sujeto y el mundo objetivo se sitúan en-

tidades intermedias que corresponden a ese material no conceptual: las impresiones humeanas o los datos sensoriales del positivismo lógico. Para el empirismo, los conceptos proceden de esas impresiones o datos sensoriales a través de diversos procesos de elaboración mental. (El racionalismo, en cambio, reivindica la autonomía de determinados conceptos frente al material sensible.) La concepción empirista persiste todavía en la obra de Quine: la impresionante riqueza de nuestro aparato conceptual tiene su origen en determinadas excitaciones de las superficies de los órganos sensoriales, y la tarea de la epistemología consiste en explicar el proceso por el que se llega de éstas a los conceptos y teorías. Las excitaciones quinianas son el correlato científicamente ilustrado de las viejas impresiones humeanas. Es obvio que la inteligibilidad de la tesis empirista acerca de la procedencia del elemento conceptual del conocimiento depende de la inteligibilidad previa del dualismo entre ese elemento y el puramente sensorial. Por lo tanto, un golpe asestado a la plausibilidad de dicho dualismo repercute, de modo ampliado, en el empirismo como tal. Para Davidson, el dualismo en cuestión «es él mismo un dogma del empirismo, el tercer dogma. El tercero y tal vez el último, pues si lo abandonamos no resulta claro que quede ya algo distintivo que merezca el nombre de empirismo».[1] Al abandonar ese dualismo, Davidson abandona también el empirismo de sus raíces quinianas.

Otro aspecto esencial del pensamiento de Quine está constituido por su concepción naturalizada de la reflexión epistemológica y filosófica. De acuerdo con esta concepción, la filosofía no es una investigación distinta de la ciencia empírica; no se distingue de ésta por un supuesto carácter de investigación puramente conceptual y *a priori*, sino sólo por un grado mayor de generalidad. Este aspecto del pensamiento de Quine ha contribuido decisivamente –aunque, en mi opinión, en un sentido empobrecedor– a configurar el ca-

1. D. Davidson, «On the Very Idea of a Conceptual Scheme», en *Inquiries into Truth and Interpretation*, Clarendon Press, Oxford, 1984, pág. 189.

rácter de gran parte de la investigación filosófica en el mundo anglosajón. La naturalización quiniana de la filosofía deriva del carácter refinado del empirismo de este autor. Aunque, según la tesis empirista, son las aportaciones sensoriales las que dan origen al esquema conceptual, el contenido sensorial se distribuye vagamente a través de la estructura de conceptos y juicios, con lo que no es posible establecer una distinción clara entre enunciados analíticos, cuya verdad dependería únicamente del significado de los términos empleados en ellos, y enunciados sintéticos, cuya verdad dependería de su confrontación con la experiencia sensorial. Es el sistema conceptual como un todo, y no partes aisladas de éste, el que se confronta con el tribunal de la experiencia. No resulta posible, entonces, llevar a cabo una investigación puramente conceptual en cuanto opuesta a una investigación empírica. De ahí que la filosofía no constituya un modo de conocimiento distinto de la ciencia. La naturalización quiniana de la filosofía se lleva a cabo, sin embargo, en el marco general del empirismo y de la distinción entre contenido empírico y sistema conceptual. El contenido empírico se distribuye difusamente en el esquema conceptual, de modo que no es posible reconocer con nitidez, en un concepto determinado, las aportaciones respectivas de uno y otro. La actitud de Davidson ante la naturalización quiniana de la filosofía es ambivalente. Acepta la carencia de criterios claros para la distinción entre enunciados analíticos y sintéticos, pero, al mismo tiempo, al rechazar el dualismo de contenido y concepto, y con él el empirismo, se concede a sí mismo una mayor libertad para la reflexión puramente *a priori*. De hecho, Davidson utiliza sin reservas este modo de investigación *a priori*, provocando con ello la reacción indignada de los representantes del empirismo y de la filosofía naturalizada.[2] Así, por ejemplo, su desconfianza en las perspectivas de la

2. Véase, por ejemplo, M. Johnston, «Why Having a Mind Matters», en E. LePore y B.P. McLaughlin, comps., *Actions and Events. Perspectives on the Philosophy of Donald Davidson*, Basil Blackwell, Oxford, 1985, págs. 408-426.

psicología como ciencia natural obedece, como veremos, a razones vinculadas a la pura reflexión conceptual. No se trata, ciertamente, de que Davidson menosprecie en algún sentido los métodos o los resultados de la investigación científica; pero es consciente de que, cuando se llega a determinadas cuestiones, este tipo de investigación no puede servir de ayuda. La obra de Davidson no representa, pues, únicamente una ruptura con el empirismo, sino también una marcada recuperación de la autonomía de la reflexión filosófica frente al discurso científico.

Por lo que respecta al naturalismo y al materialismo, la actitud de Quine hacia estas posiciones no se puede desgajar por completo de su compromiso con la ciencia natural, y especialmente con la física, como único modo legítimo de conocimiento, es decir, con el ciencismo como actitud filosófica, del que la naturalización de la filosofía es una manifestación. Así, por ejemplo, dadas las discrepancias, incluso de comportamiento lógico, entre el discurso de la ciencia física y el discurso psicológico cotidiano, en el que hablamos de creencias, intenciones, deseos, etc., el materialismo de Quine, inspirado en el primero, tiende a la eliminación del segundo. En cambio, el materialismo davidsoniano, en consonancia con su desconfianza hacia la concepción naturalizada de la filosofía y hacia el carácter omniabarcante de la ciencia, no presenta tendencia alguna hacia la eliminación del discurso mental. Propongo que entendamos por materialismo o fisicismo la tesis metafísica según la cual el mundo está constituido únicamente por objetos, estados y eventos físicos, es decir, por objetos, estados y eventos que tienen descripciones verdaderas en el lenguaje de las ciencias físicas. La adopción de esta tesis por sí misma no conlleva, a menos que vaya acompañada por un fuerte compromiso cientista, el rechazo de la existencia de, por ejemplo, estados y eventos mentales, la reducción del discurso mental al discurso físico o la pura y simple eliminación de lo mental. No conlleva estas consecuencias si se admite, como David-son hace, que los eventos y estados pueden tener, además de la descripción física, otras descripciones verdaderas

no menos legítimas desde el punto de vista epistemológico.

Finalmente, el naturalismo davidsoniano expresa la convicción general de que los seres humanos no constituyen ún *imperium in imperio* en el seno de la naturaleza física, sino que forman parte de ella. La perspectiva davidsoniana sobre los seres humanos está condicionada por este robusto sentido naturalista, poco dado a ensoñaciones trascendentes: es la perspectiva objetiva del observador, la perspectiva de la tercera persona. De nuevo, sin embargo, hemos de insistir en los matices. El naturalismo davidsoniano no va acompañado, como en Quine, de una naturalización de la reflexión filosófica. Y es incluso compatible con la defensa de tesis tradicionalmente vinculadas al humanismo, como la autonomía de las ciencias sociales y humanas frente a las ciencias naturales y de los fenómenos mentales frente a las leyes científicas.

Sirva lo dicho hasta aquí para situar el pensamiento davidsoniano en un marco filosófico más general. Es el momento de prestar algo de carne y sangre a esta delgada estructura de conceptos y posiciones filosóficas.

2. Razones y causas: la acción intencional

Davidson alcanzó notoriedad en el mundo filosófico con la publicación, en 1963, de un artículo titulado «Acciones, razones y causas».[3] En él defendía la tesis según la cual las explicaciones de una acción mediante razones constituyen una forma de explicación causal, siendo las razones causas de la acción. Por otra parte, e inspirándose a este respecto en Elizabeth Anscombe, Davidson tendía a caracterizar implícitamente una acción intencional como aquella que tiene una explicación verdadera en términos de razones. Tal explicación ha de cumplir dos requisitos: las razones deben justificar racionalmente la acción y, además, deben causarla.

3. D. Davidson, «Actions, Reasons, and Causes», en *Essays on Actions and Events*, Clarendon Press, Oxford, 1982, págs. 3-19.

Con ello, Davidson sentaba las bases de una concepción causal de la acción intencional humana según la cual una acción intencional es un proceso causal de cierto tipo, y se distingue de otros procesos por el tipo de causas que dan lugar a ella. Una acción intencional es un fragmento de conducta cuyas causas son razones, en virtud de las cuales resulta justificado. Esta concepción causal de la acción intencional expresa con claridad la inspiración naturalista del pensamiento davidsoniano que destacábamos en el apartado anterior.

La importancia de estas tesis sólo puede apreciarse si tenemos en cuenta el contexto intelectual en cuyo seno surgen y frente al cual reaccionan. Cuando se publicó el artículo al que nos hemos referido, constituía casi un lugar común en el mundo filosófico anglosajón la idea de que las razones no son causas de la acción. Esta idea había sido defendida profusamente por diversos filósofos de inspiración wittgensteiniana, que desarrollaron algunas intuiciones de Wittgenstein acerca de la conducta regida por reglas. Tras la idea de que las razones no son causas de la acción podemos descubrir una concepción general marcadamente antinaturalista del agente humano y de sus acciones; si las razones no son causas de la acción intencional y si ésta puede entenderse en términos de su relación con dichas razones, la conducta intencional humana no se halla en pie de igualdad con otros procesos de la naturaleza física, sino que constituye un proceso *sui generis*, no sometido al alcance del conocimiento nomológico. Uno de los libros más representativos de esta perspectiva es el titulado *The Idea of a Social Science*, de Peter Winch.[4]

En este contexto, la defensa davidsoniana de la tesis, aparentemente puntual y restringida, según la cual las razones *son* causas de la acción, tenía consecuencias de amplio alcance. Los partidarios del naturalismo y de la unidad de la ciencia bajo el modelo de la física vieron reivindicadas sus posiciones frente a los wittgensteinianos. El trabajo de Da-

4. P. Winch, *The Idea of a Social Science and its Relation to Philosophy*, Routledge & Kegan Paul, Londres, 1958.

vidson parecía constituir el comienzo de un proceso en el que la acción humana acabaría siendo entendida como un proceso causal más, como una parte más de los cambios en la naturaleza física. Veremos, sin embargo, cómo estas esperanzas estaban sólo parcialmente justificadas.

En defensa de la tesis del carácter no causal de las razones, sus partidarios habían desarrollado una considerable diversidad de argumentos. Entre ellos destaca, sin embargo, por su fuerza y profundidad, el llamado «argumento de la conexión lógica» entre razones y acción. Este argumento partía de la concepción humeana de la relación causal según la cual los términos de esta relación, causa y efecto, son eventos distintos e independientes, no habiendo entre ellos otro vínculo de unión que la regularidad y constancia con que se presentan juntos en la experiencia. A partir de esta premisa, el argumento trata de mostrar que la razón y la acción no poseen la independencia recíproca propia de la causa y el efecto, sino que entre ellas hay una relación conceptual o «lógica», por lo cual no están unidas por una simple regularidad o conjunción constante, para concluir entonces que las razones no podían ser causas de la acción.

El argumento que acabamos de describir en su estructura general ha tenido varias concreciones, de las que quisiera destacar dos. De acuerdo con la primera de ellas, debida especialmente a A.I. Melden,[5] si se ofrece como razón para una acción el deseo de llevarla a cabo, no se está ofreciendo una causa de la acción, ya que el concepto mismo del deseo (digamos, el deseo *de ir al cine*) contiene el concepto de la acción que explica (ir al cine, en este caso), y se supone que los conceptos de la causa y del efecto son mutuamente independientes. En el caso del deseo de ir al cine, no podemos entender la naturaleza de ese deseo sin incluir el concepto de la acción deseada, y por ello el deseo en cuestión no puede ser la causa de esa acción. El deseo de ir al cine explica que vayamos al cine, pero esa explicación no es causal. Una se-

5. Véase A.I. Melden, *Free Action*, Routledge & Kegan Paul, Londres, 1961, especialmente caps. 8-10.

gunda forma que ha adoptado el argumento de la conexión lógica podemos encontrarla en autores como Charles Taylor o William Dray.[6] De acuerdo con éstos, causa y efecto se conectan mediante una ley o regularidad empírica que expresa la conjunción constante entre tipos de fenómenos semejantes a la causa y al efecto. Sin embargo, la conexión entre razón y acción no se establece mediante una ley o regularidad empírica. Supongamos que se nos ofrece la siguiente explicación de que *X* levantara el brazo: *X* levantó el brazo porque deseaba indicar un giro y creía o sabía que levantar el brazo es una manera de indicar un giro. La explicación ofrece una razón para la acción de *X*. La cuestión es ahora la siguiente: ¿qué es lo que nos permite conectar el *explanandum* (la acción) con el *explanans* (el deseo más la creencia) y considerar adecuada la explicación? No hay una «conjunción constante» entre la razón y la acción. No todo aquel que desea indicar un giro, y cree que levantar el brazo es un modo de hacerlo, levanta el brazo (ni siquiera señala el giro de otro modo: esto es algo que muchos conductores hemos comprobado). Pero en realidad –proseguiría el argumento– no necesitamos conjunciones constantes entre razones y acción. Nos basta comprender lo que significa «desear indicar un giro» y «creer que levantar el brazo es un modo de indicar un giro» para ver que levantar el brazo es una acción adecuada en esas circunstancias. Es el concepto mismo de desear algo el que establece, a través de la creencia pertinente, la conexión con la acción destinada a satisfacer el deseo. Son, pues, simples conceptos, y no regularidades empíricas entre fenómenos separados, lo que nos permite conectar la razón y la acción. Por otra parte, razón y acción se hallan también vinculadas por una norma de racionalidad o prudencia del siguiente tipo: es razonable hacer A si uno desea lograr B y cree que A es un modo de conseguirlo. La explicación apare-

6. Véase Ch. Taylor, *The Explanation of Behaviour*, Routledge & Kegan Paul, Londres, 1964, especialmente cap. II y W. Dray, *Laws and Explanation in History*, Clarendon Press, Oxford, 1957, especialmente cap. V.

ce como una aplicación de esta norma. Pero la norma *no es una ley o regularidad empírica*, pues una regularidad sólo expresa el modo en que suceden las cosas y no el modo en que sería razonable actuar. Así, pues, si la conexión entre causa y efecto se establece mediante una ley o regularidad empírica y la conexión entre razón y acción se produce mediante simples conceptos y principios normativos, las razones no pueden considerarse como causas de la acción.

En su respuesta al argumento de la conexión lógica, la estrategia de Davidson consiste, básicamente, en aceptar buena parte de las premisas de sus adversarios y en negar, sin embargo, que de ellas se desprenda la conclusión que ellos pretenden. Davidson admite, por ejemplo, que no hay conjunciones constantes o leyes estrictas que conecten las razones con las acciones. Lo que hay, a lo sumo, son generalizaciones vagas y llenas de excepciones. Pero señala que esto mismo sucede en el caso de muchos enunciados singulares de carácter claramente causal. Cuando vemos, por ejemplo, que alguien rompe un cristal de una pedrada, podemos decir que la pedrada causó la rotura del cristal, pero es obvio que no tenemos leyes estrictas que conecten las pedradas con la rotura de cristales. Esto no significa que estos enunciados causales no estén respaldados por leyes. «Estar respaldado por una ley» puede entenderse de dos formas distintas, ambas compatibles con la concepción humeana de la causalidad: o bien puede significar que el enunciado «*A* causó *B*» involucra una ley que contiene los predicados usados en las descripciones «A» y «B», o bien puede significar que hay una ley ejemplificada por *alguna* descripción verdadera de *A* y *B*. Según Davidson, sólo esta segunda versión del principio humeano se adapta a la mayoría de los enunciados causales ordinarios, y puede también aplicarse a las explicaciones de la acción mediante razones. No hay, ciertamente, leyes de la física que hablen de pedradas y roturas de cristales, pero si el enunciado «la pedrada rompió el cristal» es verdadero, hay leyes físicas de las cuales – en unión con una descripción física adecuada del impacto de la piedra, la estructura microscópica del cristal, etc.– , podría deducirse ló-

gicamente la rotura del cristal (descrita asimismo mediante el vocabulario adecuado de la física teórica). Una cosa son, pues, los eventos singulares, unidos por relaciones causales que expresamos mediante enunciados causales singulares, y otra distinta es la explicación causal, que conecta entre sí determinadas *descripciones* de esos eventos a través de leyes generales. Puede, pues, haber relaciones lógicas entre determinadas descripciones de eventos sin que esto impida que tales eventos se relacionen como causa y efecto. La relación causal entre dos eventos es independiente de las relaciones, lógicas o no, que pueda haber entre determinadas descripciones de ambos eventos. Davidson concede a sus adversarios que entre las *descripciones* de las razones y la acción puede haber una conexión «lógica» en algún sentido, pero esto no impide que las razones puedan ser causas de la acción.

Podemos, pues, concebir la acción intencional humana como una conclusión justificada a partir de determinadas razones del agente sin vernos obligados a situarla más allá de la naturaleza física. La conducta intencional es un proceso causal como cualquier otro, aunque la *describimos* e *interpretamos* de tal manera que la singularizamos frente a otros procesos causales, otorgándole precisamente el aspecto racional que le es característico.[7]

En su artículo «Acciones, razones y causas», Davidson concibe la acción intencional como una especie de conclusión a partir de determinadas premisas, correspondientes a ciertas razones (creencias y deseos) del agente, y, además, como efecto de éstas. Más adelante, Davidson revisará esta concepción, cuyas raíces se remontan a la doctrina aristotélica del silogismo práctico, ante el problema que plantean las situaciones de conflicto de deseos y/o creencias.[8] En estas situaciones el agente tiene razones tanto a favor como en con-

7. Para una discusión más detallada y crítica de la teoría causal de la acción de Davidson, puede verse mi libro *The Philosophy of Action. An Introduction*, Polity Press, Cambridge, 1990, caps. 10-13.
8. Ibíd., cap. 13.

tra de un determinado curso de la acción. Si ésta fuese una consecuencia lógica de las razones del agente, esto supondría que en los casos de conflicto el agente extrae conclusiones contradictorias y lleva a cabo acciones incompatibles. Este resultado imposible lleva a Davidson a modificar su concepto inicial de la acción intencional de tal modo que el nuevo análisis pueda dar cuenta tanto de los casos normales como de los casos de conflicto de razones. Se trata de explicar estos casos, que en ocasiones conllevan una actuación irracional, sin atribuir al agente una contradicción flagrante y manifiesta. El artículo «Engaño y división», incluido en el presente volumen, guarda relación con este campo de problemas.

No es extraño, por otra parte, que la explicación de la irracionalidad constituya un problema importante de la filosofía davidsoniana. No es extraño porque, desde el trabajo que acabamos de exponer, Davidson concibe la mente bajo la idea (o «principio constitutivo», como lo denomina más adelante) de la racionalidad. Creencias y deseos, por ejemplo, son estados que atribuimos a los demás en el proceso de interpretación de su conducta, y en este proceso el contenido de estos estados ha de ser tal que la conducta aparezca como racional en relación con ellos, o al menos como inteligible; de otro modo no hemos conseguido entenderla como acción intencional; por ello es esencial la forma en que describimos las creencias, los deseos y las acciones. Una acción puede estar justificada a la luz de ciertos deseos y creencias cuando se describe de cierta forma y no cuando se describe de otra. Y lo mismo sucede con las creencias y deseos: pueden justificar la acción bajo cierta descripción y no bajo otra. El énfasis en la concepción de la mente y de la conducta intencional se sitúa, pues, en la descripción que una persona hace de otra con el fin de entenderla. Dicho de otro modo: la concepción davidsoniana de la mente está dominada por la perspectiva de la tercera persona, por la perspectiva del proceso por el que cada uno de nosotros trata de entender a los demás y ellos a nosotros. La mente, podríamos decir, es lo que atribuimos a los demás para hacernos inteli-

gible su conducta, lingüística y no lingüística. El problema, pues, no afecta sólo a la comprensión de la acción, sino también del lenguaje. El punto de vista de Davidson es el del observador que trata de hallar sentido en ciertos fenómenos. Ello define, por tanto, otro de los problemas que plantean dificultades a la filosofía de la mente davidsoniana: la autoridad cognoscitiva del sujeto acerca de *sus propios* estados mentales, autoridad que no posee sobre los estados mentales de los demás. Sin embargo, lejos de negar lo que se ha dado en llamar la perspectiva de la primera persona, Davidson la reconoce y la defiende, tratando de integrarla en su propia concepción, centrada en torno a la tercera persona. El ensayo «El conocimiento de la propia mente», que forma parte del presente volumen, intenta precisamente mostrar que la adopción de una aproximación objetiva a los estados mentales no tiene por qué llevar a la negación del conocimiento de primera persona. Este ensayo tiene, además, la virtud de ofrecer un amplio panorama de buena parte de la reflexión filosófica contemporánea en torno a la mente.

Junto a la aproximación a la mente y a la conducta basada en las ideas de racionalidad e interpretación, que hemos destacado, tenemos también la imagen que de la mente y de la conducta arroja la categoría de causalidad. Los estados mentales no sólo justifican y hacen inteligible la conducta, sino que, como hemos visto, también la causan. La consideración de los estados mentales como causas, y no sólo como justificaciones, obliga a concederles una realidad ontológica más robusta: los estados mentales son eventos y, como veremos, eventos físicos. Junto a las relaciones de justificación que resultan de un proceso de interpretación guiado por la idea de racionalidad, tenemos las relaciones causales entre la mente y la conducta, relaciones que comprometen a Davidson con una concepción materialista.

Los resultados de ambas aproximaciones, que por abreviar podríamos denominar racional y causal, no conviven siempre sin tensiones. La acción intencional aparece bajo un doble aspecto: como conducta racionalmente justificada y

como proceso causal físico.[9] También la mente presenta una doble faz: el principio constitutivo de la racionalidad la sitúa más allá del alcance explicativo de las leyes físicas, como veremos en el siguiente apartado, mientras que la consideración causal la presenta como una parte más de la naturaleza física. Se trata de un equilibrio de fuerzas contrapuestas, y no es extraño que autores de distintas tendencias hayan intentado romperlo aumentando la tensión, bien por el lado causal, bien por el lado racional. Davidson, sin embargo, se mantiene equidistante entre ambos extremos, pretendiendo así hacer justicia tanto a la concepción cotidiana de los seres humanos como agentes racionales y responsables de sus acciones como a la concepción científica, para la que los seres humanos son tan sólo sistemas físicos de alta complejidad.

Este equilibrio se manifiesta palmariamente en lo que el propio Davidson ha denominado «monismo anómalo», que caracteriza su concepción de lo mental y cuyas raíces se hallan también en el temprano trabajo de 1963 que hemos comentado en este apartado.

3. El monismo anómalo

La defensa davidsoniana de la tesis según la cual las razones son (también) causas de la acción compromete a nuestro autor, como hemos apuntado, con una concepción materialista de la mente, que será desarrollada en trabajos como «Mental Events» y «The Material Mind».[10] Veamos cómo la primera tesis conduce a la segunda.

Supongamos que una determinada explicación de una acción particular en términos de razones es verdadera. En

9. En su reciente libro titulado *Donald Davidson* (Polity Press, Cambridge, 1991), Simons Evnine ha subrayado la existencia de un doble proyecto en la obra de Davidson, doble proyecto relacionado con la doble aproximación de que hablamos aquí.

10. En *Essays on Actions and Events*, págs. 207-225 y 245-259 respectivamente.

ese caso, según la concepción davidsoniana, las razones causan la acción (además de justificarla). Dada la concepción humeana de la causalidad, que Davidson acepta, ello supone que hay una ley general,estricta que respalda dicho enunciado causal. Pero no hay leyes estrictas que conecten razones con acciones. Por lo tanto, las leyes en cuestión serán de carácter físico o neurofisiológico. Pero si estas leyes, en unión con descripciones adecuadas de la causa, han de permitirnos deducir el efecto, ello significa que la causa (creencias y deseos) y el efecto (la acción) han de tener descripciones verdaderas en el lenguaje de la neurofisiología o de la física. Es decir, el deseo de que *P*, por ejemplo, pueda ser descrito también, digamos, como tal configuración de neuronas en tal o cual estado de excitación, y lo mismo sucederá con la creencia y con la acción. Ahora bien, si un determinado evento tiene una descripción física verdadera, es un evento físico. Y, así, los eventos mentales (creencias, deseos, intenciones) que explican una acción son también eventos o estados físicos del cuerpo y del sistema nervioso del agente. De este modo, la defensa de la tesis según la cual las razones son causas de la acción compromete a Davidson con una posición materialista que adopta la forma de una teoría de la identidad entre lo mental y lo físico (monismo), entre la mente y el cuerpo. Los estados mentales son estados físicos bajo otras descripciones. En su conocido artículo «Men-tal Events», Davidson desarrolla un argumento general en favor del monismo según el cual todo evento mental que interactúa causalmente con un evento físico es un evento físico. Las premisas y principios que conducen a esta conclusión son, básicamente, los que acabamos de exponer de modo más informal.

Sin embargo, el monismo no es el único aspecto de la concepción davidsoniana de la mente. El anomalismo de lo mental constituye un aspecto no menos importante. La unión de ambos aspectos forma el monismo anómalo. Se trata de una posición que tiene cierto aire de paradoja; sostiene, en efecto, que los fenómenos físicos están sometidos a leyes estrictas que permiten su predicción en principio; afir-

ma además que los fenómenos mentales son fenómenos físicos; y, sin embargo, niega que los fenómenos mentales estén sometidos a leyes estrictas que permitan su predicción (anomalismo de lo mental). En *este* sentido, la mente se halla más allá del alcance de las leyes físicas. Trataré ahora de disipar ese aire paradójico.

Pensemos, en primer lugar, que la argumentación davidsoniana en favor del monismo no permite deducir que exista una identidad entre *propiedades* mentales y físicas, o entre *tipos* de estados mentales y de estados físicos. Lo único que se deduce de ella es que cada evento mental *particular* es idéntico a un estado físico particular. Es decir, si el deseo que Juan tenía el jueves pasado de ir al cine causó que fuese al cine, ese deseo es un determinado evento físico en el cuerpo de Juan. Pero ello no permite afirmar que, en general, el deseo de ir al cine sea idéntico a una determinada propiedad o tipo físico que hubieran de compartir todos aquellos agentes de los que puede decirse que desean ir al cine. Ni siquiera cabría suponer que un individuo se halla en un estado físico del mismo tipo cada vez que desea ir al cine. En términos algo más técnicos, la identidad psicofísica defendida por Davidson se limita a la identidad de *casos* particulares (*tokens*) y no alcanza la identidad de *tipos* (*types*) o propiedades mentales y físicas. Ahora bien, para poder predecir estados mentales sobre la base de leyes físicas sería necesaria una identidad, no sólo de casos particulares de eventos mentales y físicos, sino también una identidad, o cuando menos una correspondencia sistemática, entre propiedades mentales y físicas, puesto que las leyes expresan relaciones entre propiedades o tipos de fenómenos. Y el argumento davidsoniano en favor del monismo no garantiza esta segunda condición.

¿Por qué no podría haber, sin embargo, correspondencias sistemáticas con fuerza de ley (leyes-puente) entre propiedades mentales y propiedades físicas? Si las hubiera, habríamos dado el primer paso hacia la reducción de la psicología intencional a la neurofisiología o a la física. Según Davidson, sin embargo, tal reducción no es posible, porque

no es posible descubrir tales leyes-puente psicofísicas, y no es posible descubrirlas porque no las hay. El principal argumento davidsoniano en contra de la existencia de leyes psicofísicas se basa en el hecho de que la adscripción de predicados (propiedades) mentales y la de predicados (propiedades) físicos están regidas por principios constitutivos diferentes. Davidson ilustra la noción de principio constitutivo mediante el ejemplo de la medida de longitudes. La medida de longitudes, y con ello la adscripción de longitudes a objetos, es posible en el marco de un principio o postulado que caracteriza la relación «más largo que» como transitiva y asimétrica: es decir, si un objeto es más largo que otro y éste a su vez más largo que un tercero, el primero es más largo que el tercero. Sin aceptar este principio no es posible medir longitudes inteligiblemente. Pues bien, la adscripción de predicados mentales (a diferencia de la de predicados físicos) se lleva a cabo en el marco del principio constitutivo de la coherencia y la racionalidad, en el sentido de que no podemos atribuir inteligiblemente un estado mental (una creencia, un deseo, una intención, etc.) a un agente salvo en el marco de una teoría global sobre sus estados mentales que atribuye al agente un amplio grado de coherencia y racionalidad. El contenido de cada estado mental deriva de su lugar en este contexto global (holista) regido por principios de coherencia racional. Según esto, cualquier conexión entre una propiedad mental y una propiedad física que pudiera llegar a establecerse tendría un carácter meramente accidental y no sería proyectable hacia el futuro con vistas a la predicción de estados mentales en los agentes. Lo mental, aun no siendo distinto de lo físico, es, sin embargo, anómalo, al menos por lo que respecta a las leyes físicas. Por otra parte, tampoco puede haber, según Davidson, leyes puramente psicológicas que conecten tipos de estados mentales entre sí y con acciones intencionales. En favor de esta tesis hemos de mencionar, junto a las razones de globalidad y racionalidad ya indicadas, el hecho de que lo mental no constituye un sistema causalmente cerrado, de modo que hay muchos factores no mentales que inciden sobre los estados mentales de

las personas.[11] Así, cualquier conexión entre propiedades mentales que pueda establecerse tendrá el carácter de una simple generalización no proyectable y llena de excepciones.

Así, pues, el monismo anómalo davidsoniano combina una actitud naturalista hacia la mente y hacia la acción intencional humana con una defensa de la autonomía de las ciencias que se ocupan de ellas (la psicología y las ciencias sociales y humanas en general) frente a las ciencias naturales. Como indica Davidson en «The Material Mind»: «No hay ningún sentido importante en el que la psicología pueda reducirse a las ciencias físicas».[12] La filosofía davidsoniana de la mente se opone, pues, a cualquier tipo de reducción de los conceptos mentales, ya sea de carácter conductista, neurofisiológico o funcional. Davidson no participa, pues, de la fuerte corriente funcionalista presente en la filosofía de la mente y en la psicología actuales.

La concepción de la mente en términos del proceso de interpretación, regido por principios normativos de coherencia y racionalidad, desempeña, como hemos visto, un papel central en la filosofía davidsoniana de la acción y de la mente. Los fundamentos de dicha concepción de la mente y del lugar central que ocupa en ella el principio constitutivo de la racionalidad se hallan, sin embargo, en la teoría davidsoniana del significado y de la comunicación lingüística. En este sentido, la filosofía del lenguaje representa un aspecto crucial de la filosofía de Davidson.

4. Significado, verdad e interpretación

La filosofía davidsoniana del lenguaje se halla imbuida

11. El artículo «El conocimiento de la propia mente», contenido en el presente volumen, desarrolla esta idea, entre otras.

12. *Essays on Actions and Events*, pág. 259.

del naturalismo que impregna otras partes de su obra.[13] El ser humano es una parte de la naturaleza. Sin embargo, frente a otros seres naturales, el ser humano es, por decirlo con Aristóteles, un animal que habla, y al hablar se comunica con otros acerca de diversos asuntos. El habla, un fenómeno que bajo cierto aspecto es puramente físico, una misión de sonidos, posee, sin embargo, frente a otros fenómenos físicos, la propiedad del significado. Al emitir los sonidos «la hierba es verde» estoy diciendo algo acerca de la hierba. Pero los sonidos «la hierba es verde» y el color de la hierba son dos partes del mundo completamente distintas. ¿Qué hace entonces de la emisión de esos sonidos una afirmación acerca de la hierba? ¿Qué hace de dicha emisión un discurso significativo? ¿Qué es, en suma, el significado?

Antes estas preguntas es posible adoptar diversas actitudes. La actitud introspectiva busca la raíz del significado en la conciencia. Lo que confiere significado a ciertos sonidos son ciertos fenómenos psíquicos, como imágenes o representaciones, que acompañan su emisión. Así, lo que confiere a los sonidos «la hierba es verde» el significado de que la hierba es verde es una imagen de la hierba verde en la mente del que los emite. A diferencia de los sonidos, la imagen mental de la hierba verde se parece a la hierba verde. Pero entonces, ¿por qué un trozo de hierba verde, que ciertamente se parece al resto de hierba verde más que la imagen mental, no significa que la hierba es verde? Basar el significado en la relación de semejanza explica demasiado, pues cualquier cosa se parece a otra en algún sentido. Por otra parte, si la relación entre los sonidos «la hierba es verde» y el color de la hierba constituía una problema, no lo es menos la relación entre di-

13. La filosofía del lenguaje de Davidson es el objeto de estudio específico de algunos libros recientes, como el de Bjorn T. Ramberg, *Donald Davidson's Philosophy of Language. An Introduction*, Basil Blackwell, Oxford, 1989, y, en nuestro país, el de Manuel Hernández Iglesias, *La semántica de Davidson*, Visor, Madrid, 1990. Dado el carácter básico y general de esta Introducción, el lector interesado en una exposición más detallada y crítica de la filosofía davidsoniana del lenguaje hará bien en consultar los estudios mencionados.

chos sonidos y la imagen mental en cuestión. En ambos casos se trata de dos cosas distintas y separadas. En realidad, el problema se ha complicado: si lo que da significado a los sonidos «la hierba es verde» es una imagen mental, ¿cómo puedo saber que esos sonidos significan lo mismo en tu boca que en la mía? La comunicación se convierte en un azar incognoscible.

Estas y otras dificultades pueden generar la tentación platónica. Lo que hace de la emisión de ciertos sonidos un discurso significativo es un contenido objetivo, no mudable ni dependiente de la estructura psíquica individual. Junto al habla, los hablantes y el mundo, la ontología se ve enriquecida con entidades ideales: significados, proposiones, intenciones, sentidos. La escasa simpatía de Davidson hacia esta actitud no se debe tan sólo a una tendencia hacia la sobriedad ontológica, sino también, y sobre todo, a su pobre valoración de las virtudes explicativas de las entidades postuladas. Al postular estas entidades no se ha hecho sino reificar el problema con el que tropezábamos, no resolverlo. Si alguien dice que la oración inglesa «it rains» significa lo mismo que la oración castellana «llueve», se ha limitado a constatar un hecho. No ha explicado este hecho. Sin embargo, podemos vernos tentados, ilusoriamente, a pensar que si decimos: «la oración inglesa "it rains" y la oración castellana "llueve" expresan la misma proposición», hemos avanzado, con respecto a la afirmación anterior, hacia una explicación de la identidad de significado de ambas oraciones. En realidad, lo que hemos hecho al postular la existencia de una proposición es hipostasiar la comunidad de significado que percibimos entre ambas oraciones. No la hemos explicado. Si la obviedad inicial era la simple constatación de un hecho semántico, la reformulación posterior no posee tampoco mayores virtudes explicativas. Es cierto, sin embargo, que, si los significados son entidades objetivas, la continuidad del significado entre diversos hablantes no representa un problema, a diferencia de lo que sucede en la aproximación introspectiva. Como veremos, la concepción davidsoniana del significado mantiene esta importante ventaja de la concepción platónica.

A diferencia de las perspectivas introspectiva y platónica, la filosofía davidsoniana del lenguaje no pretende encontrar *algo* (representaciones mentales o entidades objetivas ideales) que haga significativa el habla. La pregunta davidsoniana no es «¿qué es el significado?», ni «¿qué hace significativa la emisión de ciertos sonidos?», sino más bien la siguiente: dado que los seres humanos son animales que hablan, ¿cómo podemos entender lo que dicen? El problema del significado se convierte en el problema de la interpretación y de la comunicación entre los hablantes.

La investigación davidsoniana de la comunicación y la interpretación lingüística es heredera del análisis quiniano de la traducción radical, que en Davidson pasa a denominarse interpretación radical. Este no es un simple cambio terminológico. Podemos saber que una oración traduce otra sin saber qué significa ninguna de las dos. En cambio, la interpretación de una oración ha de proporcionarnos su significado. El intérprete radical pretende construir una teoría del significado de las emisiones aparentemente lingüísticas de un sujeto cuyo lenguaje le es totalmente desconocido. Situar el punto de partida del análisis de la interpretación en esta situación extrema es un artificio metodológico (cuyas virtudes, sin embargo, no me parecen del todo claras) destinado a poner de manifiesto los aspectos implicados en la comunicación normal entre los seres humanos. La ventaja de este punto de partida consiste en que nos permite evitar que nos pasen inadvertidos presupuestos importantes de la comunicación, cosa que puede fácilmente suceder si analizamos la comunicación en el caso de sujetos que comparten un lenguaje y una cultura.

El intérprete radical cuenta sólo con la observación de la conducta del sujeto (los sonidos que emite, los movimientos que lleva a cabo) y del entorno en el cual se desarrolla. El intérprete radical ha de suponer, sin embargo, que es capaz de detectar en el sujeto una actitud básica, a saber, la de tener por verdadera una emisión. Esta actitud básica corresponde a la noción de creencia. Esta noción, junto con la noción de verdad, relacionada con ella, constituyen el bagaje de con-

ceptos semánticos del intérprete. Aunque se trata de conceptos semánticos, no vician el proceso de la interpretación, ya que no presuponen que el intérprete conozca ya las creencias del sujeto ni el significado de sus emisiones. Sin embargo, a diferencia de la conducta de asentimiento, que Quine toma como base de la traducción radical, la admisión de la actitud psicológica de creencia testimonia la convicción davidsoniana de que no es posible proceder a la interpretación de un sujeto sobre bases exclusivamente conductistas, como Quine pretende. En cuanto a la verdad, Davidson la considera como una noción primitiva, una noción, como diría Descartes, trascendentalmente clara, no susceptible de ser definida en términos de otras nociones más claras que ella misma. Entendemos mejor la noción de verdad que cualquier otra noción semántica como la de significado, referencia o traducción. Es posible, en cambio, construir estas otras nociones sobre la noción de verdad. Podemos, por ejemplo, concebir el significado de una oración o emisión lingüística como las condiciones en que esa oración o emisión es verdadera: si sabemos en qué condiciones es verdadera la oración «it rains», sabemos qué significa la oración, sin necesidad de postular significados.

De acuerdo con esto, la tarea del intérprete radical consiste en elaborar una teoría de la verdad acerca de las emisiones que pretende interpretar, es decir, cuyo significado pretende conocer. Esta teoría debe dar como resultado teoremas que expresen, para cada oración que se interpreta, las condiciones en que esa oración es verdadera. Formalmente, los teoremas en cuestión son enunciados bicondicionales, que son verdaderos en el caso de que ambas partes del bicondicional sean verdaderas. Así, por ejemplo, si el sujeto a interpretar habla inglés y el intérprete radical habla castellano, la oración del primero «snow is white» estará interpretada mediante una teoría, uno de cuyos teoremas es un bicondicional como el siguiente:

«Snow is white», emitida por el sujeto, es verdadera si, y sólo si, la nieve es blanca.

Que la nieve sea blanca es la condición de verdad de la oración «snow is white», y el conocimiento de esta condición nos permite entender la oración en cuestión.

Ahora bien, pensemos que el siguiente bicondicional es igualmente verdadero:

«Snow is white», emitida por el sujeto, es verdadera si, y sólo si, la hierba es verde.

Intuitivamente, este bicondicional no constituye una interpretación adecuada de la oración «snow is white». Que la hierba sea verde no es una condición de verdad de «la nieve es blanca». Lo que podría excluir este tipo de bicondicionales es, en primer lugar, el hecho de que la interpretación de una oración se produce en el marco global de la teoría y de las relaciones de coherencia entre sus axiomas y teoremas; es la acumulación progresiva de estas relaciones lo que va aislando ciertos bicondicionales como interpretaciones correctas. Y, en segundo lugar, las condiciones de verdad de una oración como «snow is white», a saber, que la nieve sea blanca, causan en el agente, a diferencia del hecho de que la hierba sea verde, una disposición a asentir o tener por verdadera la oración «snow is white». Es, no obstante, muy dudoso que estas restricciones permitan eliminar completamente bicondicionales como el citado, que no constituyen una interpretación adecuada de una determinada oración. Estos bicondicionales podrían eliminarse exigiendo que la oración que formula las condiciones de verdad sea una traducción de la oración a interpretar, pero esta condición parece presuponer ya el concepto mismo de significado que la teoría davidsoniana pretendía aclarar mediante la noción de verdad.[14]

Estas son dificultades importantes para el proyecto davidsoniano, pero no podemos entrar más a fondo en ellas. Admitiendo que puedan ser resueltas, ¿cuáles son los su-

14. Sobre estos problemas véase M. Hernández Iglesias, *La semántica de Davidson*, cap. 5.

puestos que harían posible construir una teoría de la verdad para la interpretación de las emisiones de un hablante? Estos supuestos deberían arrojar luz sobre aquellos que subyacen a la comunicación normal entre los seres humanos.[15]

El proceso de interpretación constituye un proceso global en el que la asignación de condiciones de verdad a emisiones y la asignación de estados mentales, como creencias y deseos, al agente, se llevan a cabo simultáneamente y se condicionan de manera recíproca.[16] Según Davidson, dicha asignación no puede llevarse a cabo inteligiblemente a menos que el intérprete respete ciertos supuestos acerca del sujeto al que pretende interpretar. En primer lugar, habrá de aceptar que los contenidos de las creencias más básicas del sujeto están constituidos por determinados rasgos objetivos del entorno, los cuales causan dichas creencias en el sujeto. En segundo lugar, y en relación con el primer supuesto, habrá de aceptar que, en los casos más básicos, lo que el sujeto considera verdadero será también verdadero para él mismo. En tercer lugar, habrá de atribuir al sujeto la capacidad de pensar, por lo general, de modo coherente (de acuerdo con lo que el intérprete mismo considera como pensamiento coherente). A menos que acepte estos supuestos acerca del sujeto, el intérprete no será capaz de dar sentido a sus emisiones. Por lo tanto, si a partir de la interpretación radical es posible extraer conclusiones sobre la comunicación entre los seres humanos (condición ésta que no resulta obvia, pero que no podemos discutir en esta Introducción) y si en general es cierto que podemos comunicarnos con nuestros

15. Sobre la interpretación radical véase «Verdad y conocimiento: una teoría de la coherencia», en el presente volumen.

16. Por ello Davidson considera que la teoría de la interpretación debe avanzar hacia una teoría unificada del significado y de la acción. En el marco de esta teoría, sólo bosquejada por Davidson, la evidencia ha de servir simultáneamente para asignar significados a las emisiones del agente y estados mentales que hagan inteligible su conducta, incluidas esas emisiones. Véase «Toward a Unified Theory of Meaning and Action», *Grazer Philosophische Studien*, 2 (1980), págs. 1-12.

semejantes, habrá de ser cierto que la mayor parte de las creencias de los seres humanos sobre el mundo son objetivamente verdaderas y que sus estados mentales están regidos, en general, por normas objetivas de coherencia. Las consecuencias filosóficas de todo esto son de enorme importancia: el escepticismo global sobre la verdad de nuestras creencias es una posición necesariamente errónea, la concepción de la mente como un conjunto de representaciones internas, de raíz cartesiana, es incorrecta y el relativismo cultural extremo es una posición incoherente y, por tanto, necesariamente falsa.

La importancia y rotundidad de estas consecuencias exige que retrocedamos hacia los supuestos que las generan. La justificación de estos supuestos reside, para Davidson, en que sin ellos no sería posible la interpretación. Y si aceptamos que la interpretación es un hecho, es decir, que en muchos casos entendemos las emisiones lingüísticas de los demás, habremos de aceptar que los supuestos de los que depende son verdaderos. La argumentación davidsoniana parece tener, pues, estructura trascendental (en sentido kantiano): se remonta desde un hecho (la interpretación y la comunicación intersubjetiva) hacia sus condiciones de posibilidad. Lo que se trata de mostrar ahora es que los supuestos mencionados constituyen realmente condiciones de posibilidad de la interpretación. Para ello, partiremos de la negación de dichos supuestos y veremos qué resulta de ella.

5. Mente, comunidad y mundo objetivo

Supongamos que el sujeto emite ciertos sonidos al mismo tiempo que un animal pasa frente a él. El intérprete atribuye al sujeto la actitud de considerar algo como verdadero, pero no relaciona el contenido de esa actitud con el evento constituido por el paso de dicho animal (o con algún otro rasgo sobresaliente del entorno objetivo) ni supone que este evento tenga relación causal con la actitud que atribuye al sujeto. En este caso, el intérprete se hallará en completa oscuridad acerca de la creencia del sujeto y del significado de

su emisión. El intérprete pues, no tiene otra opción que considerar ciertos rasgos del entorno objetivo (en este caso, presumiblemente, el paso del animal) como contenido de la creencia que atribuye al sujeto y a la vez como causa de que el sujeto posea dicha creencia. Naturalmente, puede equivocarse, y esto es algo que el proceso ulterior de la interpretación podrá mostrarle. Pero el supuesto, como punto de partida, es indispensable para la interpretación, y si en algunos casos conduce al error, habrá de llevar a la verdad en muchos otros. Estos últimos constituyen la base inicial e indispensable de la interpretación. El supuesto que consideramos, pues, no es optativo para el intérprete. Ha de ser correcto si la interpretación es posible. Por lo tanto, el contenido de las creencias más básicas de los seres humanos acerca del mundo no está formado por representaciones mentales privadas, sino por situaciones y eventos comunes e intersubjetivos. Y dado el papel central de estas creencias básicas en el conjunto de la vida mental, la concepción representativa de la mente, de raíz cartesiana, no es correcta. No hay representaciones mentales intermedias entre el sujeto y el mundo. Y en la medida en que estas representaciones introducen la posibilidad del escepticismo, éste no puede llegar a plantearse inteligiblemente.[17]

Neguemos ahora el segundo supuesto y veamos qué resulta de ello. Supongamos que, ante el paso del animal a que nos referíamos, el intérprete atribuye al sujeto, como no puede menos de hacer, la actitud de tener algo por verdadero, pero no acepta que aquello que el sujeto tiene por verdadero sea verdadero objetivamente. En este caso, lo que el intérprete atribuye al sujeto es una creencia que para él mismo es falsa. Ahora bien, el intérprete no dispondrá en este caso de clave alguna para establecer el contenido de la creencia del sujeto. Si parte de aquello que el sujeto cree verdadero es en realidad falso, ha bloqueado el camino que conduce a la interpretación. En la interpretación, pues, hemos de emplear,

17. Véase «El mito de lo subjetivo» y «Las condiciones del pensamiento», en el presente volumen.

necesariamente, un concepto objetivo de verdad, no relativizado a sujetos o perspectivas subjetivas. Como la opción de considerar falsas la mayoría de las creencias del sujeto no es viable para el intérprete, el supuesto contrario habrá de ser necesariamente correcto. Esto no supone que el intérprete no pueda atribuir al sujeto una creencia falsa, pero sí implica que esta atribución ha de hacerla sobre la base de la atribución de creencias básicas mayoritariamente verdaderas. La creencia falsa no puede constituir la base de la interpretación del significado y de la comunicación humana. Si ésta tiene lugar, como de hecho parece suceder, nuestras creencias más básicas sobre el mundo han de ser verdaderas.

Finalmente, veamos qué resulta de la negación del tercer supuesto, es decir, de la negación de que el sujeto sea fundamentalmente coherente en sus creencias, estados mentales y acciones, de acuerdo con los criterios del intérprete. Esto supone que el intérprete atribuye al sujeto creencias, intenciones y estados mentales masivamente contradictorios. Pero la atribución al sujeto de la creencia de que *P* y de la creencia de que no *P*, o de la intención de *A* y la intención de no *A*, deja totalmente indeterminado el contenido de sus creencias e intenciones, y con ello la interpretación se ve nuevamente bloqueada. Si un sujeto cree que el pelo de cierto animal es marrón y cree al mismo tiempo que no es cierto que el pelo de ese mismo animal sea marrón, no sabemos qué es lo que cree y no podemos tampoco asignar condiciones de verdad a sus emisiones. El intérprete, pues, no tiene tampoco opción en este caso: ha de suponer que el sujeto es fundamentalmente coherente en su vida mental. La negación de esta coherencia al sujeto conlleva negarle la posesión de creencias, intenciones y, en general, de propiedades mentales. Como señalábamos más arriba, la atribución de estados mentales a un sujeto está regida por el principio constitutivo de la racionalidad y la coherencia. Esta idea, central en la concepción davidsoniana de la mente y de la acción intencional, encuentra su fundamento en la teoría del significado y de la interpretación. Naturalmente, esta idea no excluye que un agente pueda cometer errores en el razonamiento o tener al-

gunas creencias contradictorias. Pero estos errores e incoherencias habrán de ser locales, pues un error o una incoherencia masivos destruye las mínimas bases necesarias para detectar errores e incoherencias locales, al hacer imposible la fijación de contenidos para las creencias e intenciones del sujeto.[18]

Los supuestos cuya justificación hemos bosquejado dan lugar a una imagen del significado y de la mente humana de carácter profundamente opuesto a la tradición mentalista y subjetiva que procede de la obra de Descartes y se prolonga en la actualidad en importantes corrientes de la llamada ciencia cognitiva.

Nuestro propio aprendizaje del lenguaje comparte aspectos importantes con la interpretación radical. Del mismo modo que en esta última la perspectiva del intérprete y la del sujeto han de converger en una situación o evento común a ambos en el espacio público para que la interpretación sea posible, también en la situación de aprendizaje participan al menos dos sujetos, aprendiz y maestro, cuyas perspectivas han de converger también en un objeto o evento situado en un espacio común a ambos.[19] El carácter social del lenguaje y del pensamiento es subrayado por Davidson con toda claridad. Sólo en el marco de la relación intersubjetiva en un mundo común a los sujetos puede haber pensamiento, conceptos y significado. La posibilidad de la interpretación no es compatible con el supuesto de que los objetos de nuestros estados mentales son representaciones privadas, conceptualmente neutras, de las que surgirían los conceptos –como sucede en el empirismo– a través de sus relaciones de semejanza o contigüidad, o que una actividad conceptual autónoma ordenaría –como sucede en el kantismo– en la representación de un mundo objetivo. Ambas concepciones parten de una idea que la concepción davidsoniana de la interpretación hace ininteligible: la distinción entre un esquema

18. Véase «Engaño y división», en el presente volumen.
19. Véase «Las condiciones del pensamiento», en el presente volumen.

conceptual y un contenido neutro, llámese éste realidad, experiencia o datos sensoriales. Los supuestos de la interpretación obligan a concebir el contenido de nuestras creencias básicas como un evento u objeto público, y no como una entidad intermedia entre el sujeto y el mundo. Los contenidos de nuestras creencias básicas son parte del mundo público e intersubjetivo, y no objetos intermedios entre éste y nosotros.

Ahora bien, es esa distinción entre esquema conceptual y contenido la que permite formular la idea del relativismo cultural extremo, la idea de que puede haber concepciones de la realidad absolutamente impermeables e inconmensurables entre sí. Esta idea parte del supuesto de que dos sujetos podrían tener esquemas conceptuales que ordenasen u organizasen los contenidos de su experiencia o la realidad de formas tan dispares que sus mundos serían completamente distintos. Pero este supuesto no es inteligible sin la distinción implicada en él, y esta distinción no es compatible, según Davidson, con la posibilidad de la interpretación y del aprendizaje del lenguaje. La posesión de creencias, conceptos y significados no es inteligible sin la relación que vincula a dos sujetos entre sí y a ambos con objetos y eventos públicos y comunes, y esta relación excluye precisamente la posibilidad de mundos y esquemas conceptuales inconmensurables, ya sea en virtud de una distribución absolutamente dispar de valores de verdad a oraciones, ya sea en virtud del empleo de formas de razonar recíprocamente incompatibles.

La concepción representativa de la mente como un espectáculo privado y accesible sólo al «ojo interior» de cada sujeto se ve asimismo, por las mismas razones, desacreditada por el análisis de los supuestos necesarios de la interpretación y la comunicación intersubjetiva, y con ella la legitimidad de los problemas epistemológicos a que ha dado lugar, como el escepticismo, el conocimiento de otras mentes o la existencia del mundo externo. Davidson concibe su propia obra como parte del movimiento que conduce a una transformación profunda de los supuestos en que se ha movido la reflexión filosófica y epistemológica desde los tiem-

pos de Descartes. En este movimiento se incluyen también ciertos aspectos de la filosofía de Hilary Putnam, Tyler Burge y otros.[20]

6. Conclusión: sujeto, causa e intencionalidad

Las importantes consecuencias de la filosofía davidsoniana que hemos ido señalando a lo largo de esta Introducción dependen crucialmente de la concepción de la mente humana desde la perspectiva de la interpretación del lenguaje y de la conducta, es decir, desde la perspectiva del proceso por el que un sujeto trata de hallar sentido en la conducta y las emisiones lingüísticas de otro. Los estados mentales, creencias, intenciones, deseos y significados, son precisamente aquello que se atribuye a un sujeto para hacer inteligible su conducta. De ello resulta la concepción de la mente bajo el principio constitutivo de la racionalidad y bajo el supuesto de la veracidad de las creencias. La mente, por decirlo en una palabra, es un producto de la interpretación y de la comunicación intersubjetiva. Esta concepción de la mente sitúa el estudio de ésta más allá del alcance de un modelo explicativo inspirado en las ciencias naturales. Las ciencias que se ocupan de la acción intencional humana –que, recordémoslo, se concibe en términos de sus relaciones de coherencia y causalidad con creencias, deseos e intenciones– han de proceder de modo holista e interpretativo, ajustando sus resultados a la evidencia en desarrollo bajo la guía del carácter globalmente coherente de la vida mental y la conducta de los agentes. Este modo de proceder las separa de la búsqueda de leyes y de la explicación nomológica que caracteriza las ciencias de la naturaleza.

Lo cierto, sin embargo, es que esta concepción interpretativa de la mente, basada en el principio constitutivo de la racionalidad, no se compadece fácilmente con el

20. Véanse «El conocimiento de la propia mente» y «El mito de lo subjetivo», en el presente volumen.

materialismo de Davidson. En efecto, si un estado mental es un estado físico, ha de tener una realidad ontológica propia, independiente de sus relaciones de coherencia con otros estados mentales. En cambio, en la concepción interpretativa de la mente los estados mentales tienden a convertirse en constructos teóricos, en resultados del proceso de interpretación, y su contenido puede ser reajustado a medida que el proceso se desarrolla. Esta tendencia a concebir los estados mentales como constructos teóricos se ve asimismo contrapesada, en el marco de la teoría de la interpretación, por la idea de que, en los casos más básicos, el contenido de una creencia coincide con su causa externa: de acuerdo con esta idea, el contenido de la creencia parecería estar fijado, con independencia de sus relaciones con otros estados mentales, por la situación o evento que la causa. Finalmente, la concepción interpretativa de la mente se ve también compensada por la concepción davidsoniana de los estados mentales como causas, y no sólo como razones, de la conducta. No está claro, cuando menos, que estas tendencias contrapuestas sean finalmente conciliables, por más que su conciliación sea un rasgo medular del proyecto filosófico davidsoniano, que aúna el naturalismo con la importancia central de la normatividad en la imagen cotidiana de los seres humanos.

En estas circunstancias, pensadores actuales como John McDowell, de orientación wittgensteiniana, han instado a Davidson a abandonar el monismo materialista (sin recalar en el dualismo substancial) y a profundizar en las consecuencias de la concepción interpretativa de la mente, en un sentido más afín a las reflexiones wittgensteinianas sobre el concepto de regla y el carácter social del lenguaje que al holismo de raíz quiniana.

Pensadores como Fodor, en cambio, insistirían más bien en profundizar en la concepción causal de los estados mentales y en desarrollar los aspectos congeniales a la misma.[21]

21. Véase J. Fodor, *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, M.I.T. Press, Cambridge Mass., 1987.

Por lo que respecta a los principios de racionalidad y coherencia, que presiden la concepción interpretativa de la mente, deberían contemplarse en la perspectiva de su posible naturalización, tratando de convertirlos en, o reducirlos a, leyes empíricas contingentes que relacionaran tipos de estados mentales en virtud de sus contenidos. La existencia de estas leyes o generalizaciones empíricas es, según Fodor, un presupuesto de la tesis davidsoniana según la cual los estados mentales causan acciones para las cuales constituyen también razones. La concepción interpretativa de las ciencias sociales, y en especial de la psicología, debería abandonarse, consecuentemente, en favor de una concepción explicativa de la psicología intencional, basada en la idea de ley.

Estas exigencias contrapuestas son síntomas de la existencia de tensiones reales en el seno de la filosofía davidsoniana. La concepción davidsoniana de la mente tiene sin duda dificultades para captar todos los aspectos intuitivamente relevantes de la vida mental. La primacía del punto de vista de la tercera persona, la concepción de la mente como un resultado del proceso de interpretación, no puede dar cuenta fácilmente del hecho de que, por lo que respecta a nosotros mismos, no nos atribuimos (normalmente) estados mentales a través de la interpretación de nuestra propia conducta. En nuestro propio caso, los estados mentales parecen tener una realidad inmediata e independiente de su posible papel en el proceso de explicación e interpretación de la conducta y del lenguaje. En este hecho se ha fundado la concepción cartesiana de la mente, y los cartesianos actuales, como Thomas Nagel, se basan en él para tratar de mostrar la insuficiencia de las concepciones contextuales o relacionales de la vida mental, entre las que se incluye la aproximación interpretativa de Davidson. A fin de cuentas, la concepción representativa de la mente, de raíz cartesiana, surgió, en parte, como un intento de explicar la certeza indudable que acompaña los enunciados en que un sujeto se atribuye a sí mismo, sinceramente, un estado mental. Davidson rechaza, a mi entender por buenas razones, esta concepción representativa, pero no logra proporcionar una explicación alternativa

satisfactoria de ese hecho fundamental acerca de la mente. Yo sé lo que deseo sin esperar a ver cómo actúo y sin tener en cuenta las relaciones de coherencia entre ese deseo y otros estados mentales o acciones. Una teoría adecuada de la vida mental humana debe dar cuenta de ello.

En su artículo «First Person Authority»,[22] Davidson pretende explicar la autoridad de la primera persona como un supuesto esencial de la interpretación, es decir, como un aspecto necesario del conocimiento de tercera persona acerca de la mente de otros y del significado de sus palabras. Lo que debe aceptarse es que, en general, el hablante sabe lo que quiere decir con sus emisiones. Ciertamente, el proceso de interpretación pierde sentido si no se admite tal cosa. Por lo tanto, este supuesto debe ser respetado, y de él se deduce la autoridad de la primera persona. Al mismo tiempo, el hablante, si desea ser entendido, ha de proporcionar a su oyente las claves necesarias, empleando coherentemente los sonidos ante objetos y situaciones que considera presentes tanto para él como para su oyente. Así, aunque el hablante no determina los contenidos de sus propios estados mentales observando las relaciones causales que median entre él y su entorno o las relaciones de coherencia que guardan con otros estados mentales, su autoridad sobre su propia mente deriva de y presupone el contexto público de la interpretación recíproca, por lo que no puede, en general, entrar en conflicto con él.

La explicación davidsoniana es sugerente, pero no acabo de ver que pueda evitar finalmente el conflicto entre la concepción interpretativa de la mente y ciertos aspectos de nuestra vida mental que involucran la autoridad de la primera persona.

De acuerdo con Davidson, el significado de una emisión y el contenido de la creencia que expresa dependen del objeto o situación pública que regularmente las causan, en unión con las restricciones introducidas por las relaciones de coherencia racional entre ese contenido y el conjunto de los esta-

22. *Dialectica*, 38 (1984), págs. 101-111.

dos mentales del sujeto. Causalidad y racionalidad son, pues, categorías básicas en la concepción davidsoniana de la mente y de la intencionalidad. Esta concepción se enfrenta con problemas, en primer lugar, en casos posibles de ilusión o alucinación sistemática, donde el objeto o situación pública que, desde la perspectiva del observador, causan la emisión del hablante, no son aquellos en cuya existencia el hablante, aparentemente, afirma creer. La concepción interpretativa de la mente no admite la existencia de objetos mentales. Así, desde dicha concepción, puesto que en casos como éste no hay un objeto determinado *O* frente al hablante, sus palabras no pueden querer decir que lo hay. Al mismo tiempo, como el supuesto necesario de la interpretación es que, en general, el hablante *sabe* lo que quiere decir, lo que quiere decir no es que hay un objeto *O* frente a él. El problema es que eso es precisamente lo que el hablante parece querer decir y creer. La ilusión perceptiva sistemática es, sin duda, un problema difícil para *cualquier* concepción de la mente, pero la perspectiva interpretativa de Davidson parece tener dificultades especialmente agudas con estos casos.

En segundo lugar, racionalidad y causalidad por sí solas parecen insuficientes para dar cuenta de ciertos rasgos de la relación intencional normal. Si yo creo que hay un conejo tras aquel árbol, el objeto intencional de mi creencia involucra el conejo como tal conejo, como objeto individual unitario, no como un ejemplo de conejidad o como conjunto de partes de conejo no separadas.[23] Pero la causalidad y la racionalidad no pueden discriminar entre esas distintas formas de concebir el conejo. Esas distinciones, que *yo sí puedo hacer* y entre las cuales puedo discriminar, no pueden llevarse a cabo desde la causalidad y la racionalidad. Como señala Lycan, «la causalidad no discrimina entre resecciones. Por usar el ejemplo de Loar, allí donde hay un corazón que bombea sangre, hay también conjuntos de partes de corazón no

23. El origen de este problema se halla en Quine. Véase *Word and Object*, M.I.T. Press, Cambridge Mass., 1960, especialmente págs. 72-79.

separadas, segmentos temporales cardíacos, ejemplos de la propiedad "ser un corazón", y así sucesivamente».[24] Y, en cuanto a la racionalidad, la atribución racional de actitudes desde la tercera persona puede funcionar exactamente igual tomando como contenido de mis creencias acerca de conejos o corazones cualquiera de estas distintas formas de concebirlos, a pesar de que, desde mi punto de vista de primera persona, mis creencias versan acerca de conejos y corazones como objetos unitarios y no como otra cosa. La concepción davidsoniana de la mente, basada en la perspectiva objetiva de la tercera persona, no puede dar cuenta de este importante aspecto de nuestra vida mental.

Por otra parte, un énfasis exclusivo en la perspectiva subjetiva de la primera persona, como el que se produce en el cartesianismo, da lugar a una concepción de la mente no menos insatisfactoria. Este énfasis tiene como consecuencia, por ejemplo, que yo podría tener intenciones incompatibles y creencias sistemáticamente contradictorias con la sola condición de que cada una de esas intenciones y creencias me fuera introspectivamente accesible. En este punto, sin embargo, las exigencias de la perspectiva de la tercera persona parecen insoslayables: el mero acceso introspectivo no puede ser condición suficiente de mi posesión de dichas creencias e intenciones. Si yo tengo intención de *A* y también de no *A*, si creo que *P* y también que no *P*, el contenido de mi intención y mi creencia no está fijado, ni siquiera para mí mismo. La mente se halla sometida a normas objetivas que ella misma no crea. Este aspecto es subrayado con toda razón por Davidson.

La tarea de la filosofía de la mente consiste en integrar en una visión coherente ambas series de exigencias, derivadas, respectivamente, de las perspectivas de la primera y de la tercera persona. La dificultad formidable involucrada en esta tarea, que hasta ahora permanece irresuelta, consiste

24. W.G. Lycan, «Semantics and Methodological Solipsism», en E. LePore, comp., *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson*, Basil Blackwell, Oxford, 1986, pág. 261.

en que uno y otro conjunto de exigencias parecen apuntar en direcciones opuestas e inconciliables. Los esfuerzos davidsonianos por integrar en su concepción interpretativa el conocimiento de primera persona son altamente instructivos: ponen de manifiesto las condiciones que deberían ser cumplidas por una comprensión satisfactoria de la mente humana. Que la tarea *debe* tener una solución es la idea regulativa que mueve a la reflexión filosófica. La concepción davidsoniana de la acción, la mente y el significado descansa en dos categorías fundamentales: causalidad y racionalidad. Las dificultades que hemos detectado podrían apuntar al hecho de que, si bien dichas categorías pueden formar parte de una teoría adecuada de dichos fenómenos, que podríamos subsumir bajo el epígrafe «intencionalidad», tal vez no constituyan categorías realmente básicas, sino derivadas de conceptos más primitivos. Un síntoma de ello sería el hecho de que Davidson se ve llevado, por coherencia con su propia posición, a negar la posesión genuina de estados mentales a aquellos seres que carecen de un lenguaje racionalmente interpretable, como los animales o los niños muy pequeños. En el caso de estos últimos, dicha negación es especialmente injustificable, no sólo por la naturalidad con que les atribuimos deseos e intenciones, sino también porque, de acuerdo con el propio Davidson, el aprendizaje del lenguaje requiere la atribución al niño de actitudes mentales. Por otra parte, en la medida en que Davidson concibe la acción intencional por su relación causal y racional con estados mentales, habrá de negar a tales seres la realización de acciones intencionales. Estas consecuencias, intuitivamente inaceptables, parecen indicar insuficiencias importantes en la aproximación davidsoniana a la intencionalidad. Las relaciones causales entre estados mentales y conducta y entre estados mentales y entorno objetivo, así como las relaciones de coherencia racional en el seno de la vida mental, parecen constituir aspectos derivados de fenómenos intencionales más básicos, en lugar de cimientos sobre los que descansarían las relaciones intencionales entre la mente, el mundo y la acción.

Hay otra razón por la que no considero adecuada la concepción interpretativa de la mente: esta concepción sitúa la mente en el marco de la explicación de la conducta de nuestros semejantes. Pero no siempre, y en realidad bastante raramente, adoptamos actitudes explicativas hacia nuestros semejantes. En realidad lo hacemos sólo cuando el sentido de una conducta no nos es obvio. La consideración de los demás como seres que poseen una mente no deriva sólo de la actividad explicativa, sino que se relaciona también con el tipo de actitudes que adoptamos ante ellos, especialmente con actitudes no explicativas, sino evaluativas. La mente no es parte de una teoría explicativa de sentido común. *Vemos* que ciertos seres tienen mente: no llegamos a ello como resultado de nuestros intentos de explicar su conducta. En relación con lo dicho, la concepción davidsoniana de la racionalidad es exclusivamente instrumental: nuestro autor concibe la racionalidad de una acción como su adecuación al logro de los deseos o fines del agente, dadas sus creencias. Esta concepción puede hacernos ciegos para hechos importantes acerca de otras vidas y otras culturas. Aunque comparto la actitud negativa de Davidson hacia el relativismo cultural, creo que su forma de criticarlo puede llevar a cierta trivialización o descuido de las diferencias culturales, que no son siempre favorables a nuestra época. Nuestra racionalidad predominantemente instrumental no puede constituir un patrón universal de juicio.[25]

CARLOS MOYA
Universidad de Valencia

25. Agradezco a mis colegas y amigos Josep Lluis Prades y Josep Corbí sus valiosas observaciones sobre este trabajo, excusándoles por completo de cualquier error que pudiera contener.

PREFACIO A LA EDICION ESPAÑOLA

La traducción es siempre una empresa tentativa y delicada, pero especialmente en el caso de la filosofía, donde el alma de la cuestión se halla a menudo en la elección precisa de las palabras, resulta esencial un toque de exactitud y compenetración. Por lo tanto, me complace mucho que los cinco ensayos contenidos en este libro estén a disposición de los lectores de habla española en una traducción preparada por las expertas manos del profesor Carlos Moya. Me siento afortunado de llegar a los lectores bajo tan buenos auspicios.

Al ser traducido a otra lengua, un autor se encuentra con ventajas inesperadas. Una de ellas consiste en que su mente se concentra otra vez en problemas y pensamientos que los años habían apartado de su atención. Otra ventaja la constituye la inusual oportunidad que se le brinda de estudiar sus propias teorías como si procediesen de una fuente distinta; expresadas en nuevas palabras, las ideas pueden ser contempladas y revisadas con un grado de objetividad que nunca es posible en su antigua forma de expresión. El trabajo que el traductor ha llevado a cabo para llegar a entender su fuente revela defectos y oscuridades semiocultas; el producto final añade al original la visión creadora del traductor.

El beneficio mayor para el autor traducido tal vez provenga de su presentación ante una audiencia cuyo abanico de intereses, problemas, expectativas y sensibilidades es algo distinto de aquel en el que originalmente pensó. No sé si otros tienen la misma experiencia, pero en mi caso siempre me sorprenden e ilustran las respuestas de mis lectores. Reparan en cuestiones que yo apenas había advertido, se preocupan por ambigüedades a las que no presté atención y a menudo ven aplicaciones e implicaciones que nunca se me

hubieran ocurrido; y, desde luego, descubren dificultades que yo no siquiera imaginé. Los lectores y las lenguas extranjeras amplifican estos efectos y con ellos los beneficios consiguientes en cuanto a penetración y objetividad. Sé por mi experiencia pasada cuánto puedo aprender de los filósofos de habla hispana, y me sorprendería que un resultado del presente libro no fuera el ensanchamiento de mi horizonte.

DONALD DAVIDSON
California, diciembre de 1991

EL MITO DE LO SUBJETIVO

El tema del que se ocupa este ensayo tiene una larga tradición: se trata de la relación entre la mente humana y el resto de la naturaleza, entre lo subjetivo y lo objetivo, según hemos dado en pensarlos. Este dualismo, aun siendo a su modo demasiado obvio para ser cuestionado, arrastra consigo en nuestra tradición una pesada, y no necesariamente apropiada, carga de ideas asociadas. En la actualidad, algunas de dichas ideas están siendo sometidas a un detallado examen crítico, cuyo resultado lleva consigo la promesa de un cambio abismal en el pensamiento filosófico contemporáneo, un cambio de tal hondura que podría llegar a pasarnos inadvertido.

Aunque el presente ensayo es claramente tendencioso, no tiene como objetivo primario la conversión del escéptico; su propósito principal consiste en describir, desde un determinado punto de vista, un episodio reciente, ampliamente reconocido, en el desarrollo de la reflexión sobre los contenidos de la mente, y en sugerir algunas de las consecuencias que, en mi opinión, se siguen de él.

Las mentes son muchas; la naturaleza es una. Cada uno de nosotros ocupa su propia posición en el mundo y tiene, por tanto, su propia perspectiva del mismo. Es fácil dejarse deslizar desde esta verdad obvia hacia una noción confusa de relativismo conceptual. El punto de partida no es más que el relativismo, familiar e inocuo, de la posición que se ocupa en el espacio y el tiempo. Puesto que cada uno de nosotros ocupa con exclusividad un determinado volumen de espacio-tiempo, dos de nosotros no podemos hallarnos exactamente en el mismo lugar al mismo tiempo. Las relaciones entre nuestras posiciones respectivas son inteligibles debido

a que podemos situar a cada persona en un mundo común y único y en un marco temporal compartido.

El relativismo conceptual puede parecer similar a éste, pero es difícil completar la analogía. ¿Cuál es, en efecto, el punto de referencia común, o sistema de coordenadas, al que cada esquema es relativo? Sin una buena respuesta a esta pregunta, la afirmación de que cada uno de nosotros habita, en algún sentido, un mundo propio deja de ser inteligible.

Por esta y otras razones he venido sosteniendo, desde hace tiempo, que la amplitud de las diferencias entre individuos o sistemas sociales de pensamiento tiene límites. Si se entiende por relativismo cultural la idea de que los esquemas conceptuales y los sistemas morales, o los lenguajes asociados a ellos, pueden diferir globalmente entre sí, hasta el punto de ser mutuamente ininteligibles o inconmensurables, o de situarse para siempre más allá del alcance de un dictamen racional, en ese caso rechazo el relativismo conceptual.[1] Entre distintas épocas, culturas y personas hay, desde luego, contrastes que todos reconocemos y con los cuales nos enfrentamos; pero se trata de contrastes que, con una actitud comprensiva y con esfuerzo, podemos explicar y entender. Los problemas se presentan cuando intentamos incluir la idea de que podría haber diferencias más globales, ya que esto parece exigir de nosotros (de manera absurda) que adoptemos una actitud externa a nuestro propio modo de pensar.

En mi opinión, no entendemos la idea de un esquema realmente extraño. Sabemos qué son los estados mentales y cómo se identifican correctamente; son, sencillamente, esos estados cuyo contenido puede llegar a descubrirse por medios bien conocidos. Si otras personas o criaturas se hallan en estados que no es posible descubrir mediante esos métodos, puede que esto no se deba a un fracaso de nuestros mé-

1. He ofrecido argumentos en favor de esta posición en «On the Very Idea of a Conceptual Scheme», reimpreso en *Inquiries into Truth and Interpretation*, The Clarendon Press, 1984.

todos, sino a que dichos estados no merecen propiamente el nombre de estados mentales: no son creencias, deseos, anhelos o intenciones. El sinsentido en la idea de un esquema conceptual situado para siempre más allá de nuestro alcance no responde a nuestra incapacidad de comprender un esquema semejante o a otras de nuestras limitaciones humanas; se debe simplemente a lo que entendemos por un sistema de conceptos.

Muchos filósofos no se sienten satisfechos con argumentos de este tipo, ya que consideran que el relativismo conceptual puede hacerse inteligible de otro modo. Parece, en efecto, que seríamos capaces de entenderlo a condición de que pudiéramos encontrar en la mente un elemento no afectado por la interpretación conceptual. En este caso, sería posible considerar los distintos esquemas como relativos a este elemento común y asignarles la tarea de organizarlo. Este elemento común es, desde luego, alguna versión del «contenido» de Kant, de las impresiones e ideas de Hume, de los datos sensoriales, de las sensaciones no interpretadas o de lo dado a los sentidos. Kant pensaba que tan sólo era posible un esquema; pero una vez que el dualismo de esquema y contenido se hizo explícito, se puso también de manifiesto la posibilidad de esquemas alternativos. Esta idea se expresa con claridad en la obra de C.I. Lewis:

> En nuestra experiencia cognitiva hay dos elementos: los datos inmediatos, como los de los sentidos, que se presentan o se dan a la mente, y una forma, construcción o interpretación, que representa la actividad del pensamiento.[2]

Si pudiésemos concebir de este modo la función de los esquemas conceptuales, el relativismo aparecería como una

2. C.I. Lewis, *Mind and the World Order*, Scribner's, 1929, pág. 38. Lewis declara que es tarea de la filosofía «revelar los criterios categoriales que la mente aplica a lo que le es dado» (pág. 36).

posibilidad abstracta, pese a las dudas acerca de cómo podría descifrarse un esquema extraño: la idea sería que los distintos esquemas o lenguajes constituyen formas distintas en que se puede organizar lo dado en la experiencia. Según esta concepción, no habría punto de vista alguno desde el cual pudiéramos inspeccionar tales esquemas ni, probablemente, modo alguno de compararlos o evaluarlos en general; no obstante, en la medida en que creyésemos haber comprendido la dicotomía esquema-contenido, podríamos imaginar que distintas mentes o culturas reconstruyen de formas diversas el flujo inmaculado de la experiencia. De este modo, cabría sostener, el relativismo conceptual puede disponer del elemento con el que se relacionan los esquemas alternativos: ese elemento es lo dado sin interpretación, los contenidos de la experiencia no sometidos a categorías.

Esta imagen de la mente y de su lugar en la naturaleza ha definido, en gran medida, los problemas que la filosofía moderna se consideró obligada a resolver. Entre ellos se encuentran muchas de las cuestiones básicas referentes al conocimiento: cómo conocemos el «mundo externo», cómo sabemos de otras mentes e incluso cómo llegamos a conocer los contenidos de la propia. Pero también deberíamos incluir el problema de la naturaleza del conocimiento moral, el análisis de la percepción y muchas cuestiones inquietantes en el ámbito de la filosofía de la psicología y en el de la teoría del significado.

En correlación con este catálogo de problemas o de áreas problemáticas tenemos una larga lista de formas en que el supuesto contraste esquema-contenido ha hallado expresión. El esquema puede concebirse como una ideología, un conjunto de conceptos adecuados a la tarea de organizar la experiencia en objetos, eventos, estados y combinaciones de ellos; o bien, el esquema puede ser un lenguaje, tal vez con predicados y otro utillaje asociado, interpretado para estar al servicio de una ideología. Los contenidos del esquema pueden consistir en objetos de un tipo especial, como datos sensoriales, objetos de percepción, impresiones, sensaciones o apariencias; o los objetos pueden disolverse en modifica-

ciones adverbiales de la experiencia: puede «aparecérsenos rojo».* Los filósofos han mostrado cierto ingenio al inventar formas de expresar en palabras los contenidos de lo dado; tenemos, por ejemplo, esas extrañas oraciones, carentes de verbo, como «rojo aquí ahora», y las diversas formulaciones de las oraciones protocolares sobre las que discutían los positivistas lógicos.

Sin embargo, expresar en palabras la materia o contenido no es necesario y, según determinados puntos de vista, ni siquiera es posible. La división entre esquema y contenido puede sobrevivir incluso en un entorno preservado de la distinción analítico-sintética, de los datos sensoriales o del supuesto de que puede haber pensamientos o experiencias libres de teoría. Si estoy en lo cierto, éste es el tipo de entorno que nos ofrece W.V. Quine. De acuerdo con la «epistemología naturalizada» de este autor, no deberíamos pedir a la filosofía del conocimiento más que una explicación de nuestra capacidad de elaborar una teoría satisfactoria del mundo a partir de la evidencia con la que contamos. Dicha explicación se inspira en la mejor teoría de que disponemos: nuestra ciencia actual. La evidencia, de la que dependen en último término los significados de nuestras oraciones y todo nuestro conocimiento, está constituida por las estimulaciones de nuestros órganos sensoriales. Estas estimulaciones representan las únicas claves con las que cuenta una persona acerca de «lo que ocurre a su alrededor». Quine no es, desde luego, un reduccionista: «No podemos quitar los adornos conceptuales oración por oración». Sin embargo, según Quine, hay que trazar una distinción clara entre el contenido invariable y los adornos conceptuales cambiantes, entre «in-

* Traduzco «redly» por «rojo» para evitar el horrísono «rojamente». Sería impropio preguntar *qué* es lo que se nos aparece rojo. En este tipo de (pseudo) enunciados el adjetivo funciona adverbialmente para describir la manera o calidad del aparecer, no para atribuir una propiedad a un objeto. Fue R. Chisholm quien formuló claramente esta «teoría adverbial de la percepción» (en *Perceiving: A Philosophical Study*, Cornell Univ. Press, Ithaca, 1957, cap. 8), aunque hay algunos precedentes. (T.)

forme e invención, substancia y estilo, claves y conceptuación», ya que podemos investigar el mundo, y al ser humano como parte de él, y descubrir así qué claves podría tener acerca de lo que ocurre a su alrededor. Substrayendo entonces dichas claves de su concepción del mundo obtenemos como diferencia la contribución neta del ser humano. Esta diferencia acota la extensión de la soberanía conceptual del hombre, el ámbito en el que puede revisar la teoría salvando los datos.[3]

Concepción del mundo y claves, teoría y datos: éstos son el esquema y el contenido de los que he estado hablando.

Lo importante, pues, no es que podamos o no describir los datos en un lenguaje neutral, libre de teoría; lo importante es que tenga que haber una fuente última de evidencia cuyo carácter pueda ser plenamente especificado sin referencia a aquello de lo que es evidencia. Así, las pautas de estimulación, al igual que los datos sensoriales, pueden ser identificadas y descritas sin referencia a «lo que ocurre a nuestro alrededor». Si nuestro conocimiento del mundo deriva enteramente de una evidencia de este tipo, no sólo puede suceder que nuestros sentidos nos engañen a veces, sino que es también posible que estemos engañados de forma general y sistemática.

No es difícil recordar lo que conduce a esta concepción: se cree necesario aislar del mundo externo las fuentes últimas de la evidencia, con el fin de garantizar la autoridad de ésta para el sujeto. Puesto que no podemos estar seguros de cómo es el mundo fuera de la mente, lo subjetivo sólo puede mantener sus virtudes –su castidad, su certeza para nosotros– si se impide que sea contaminado por el mundo. El problema, bien conocido, consiste, desde luego, en que

3. Este pasaje y las citas que le preceden provienen de W.V. Quine, *Word and Object*, M.I.T. Press, 1960. En justicia debería advertirse que Quine ha declarado a menudo, explícitamente, que él no es un relativista conceptual.

ningún razonamiento o construcción permite salvar plausiblemente el abismo creado por esa desconexión. Una vez elegido el punto de partida cartesiano, no es –o, cuando menos, no parece– posible decir, acerca de la evidencia, de qué es evidencia. Se vislumbran ya, amenazantes, el idealismo, las formas reduccionistas del empirismo y el escepticismo.

La historia es bien conocida, pero permítaseme que pase a narrar, en mi apresurado estilo, un capítulo más. Si la evidencia última de nuestros esquemas y teorías, el material bruto en el que se basan, es subjetivo en el modo en que lo he descrito, también lo es, entonces, lo que descansa directamente en ello: nuestras creencias, deseos, intenciones y lo que queremos decir con nuestras palabras. Aunque constituyan la progenie de nuestra «concepción del mundo» –de hecho, en su conjunto forman nuestra concepción del mundo– conservan también, sin embargo, la misma independencia cartesiana, frente a aquello de que pretenden tratar, que poseía la evidencia en que se basan: como las sensaciones, también nuestras creencias, deseos, intenciones, etc., podrían ser exactamente como son aun cuando el mundo pudiera ser muy diferente. Nuestras creencias pretenden representar algo objetivo, pero su carácter subjetivo nos impide dar el primer paso para determinar si corresponden a aquello que afirman representar.

Así, en lugar de decir que lo que ha dominado y definido los problemas de la filosofía moderna ha sido la dicotomía esquema-contenido, se podría perfectamente afirmar que ha sido la forma en que se ha concebido el dualismo de lo subjetivo y lo objetivo, ya que ambos dualismos tienen un origen común, a saber, un concepto de la mente como algo dotado de sus estados y objetos privados.

Henos ahora en el lugar al que nos dirigíamos, pues me parece que la impugnación de estos dualismos por nuevas vías o su remodelación radical constituye el cambio más prometedor e interesante que está teniendo lugar en la filosofía actual. Es muy probable que dichos dualismos acaben siendo abandonados, al menos en la forma que hoy presen-

tan. El cambio está empezando a hacerse patente, y sus consecuencias apenas han sido advertidas, incluso por parte de aquellos que lo están propiciando. Como era de esperar; se enfrenta, y se enfrentará, con una fuerte resistencia. Estamos a punto de asistir a la emergencia de una concepción radicalmente revisada de la relación entre la mente y el mundo.

Voy a describir ahora algunos de los presagios que, desde mi punto de vista, anuncian esta transformación.

La acción se ha centrado en torno al concepto de subjetividad, de lo que está «en la mente». Comencemos atendiendo a lo que sabemos o captamos cuando conocemos el significado de una palabra u oración. Constituye un lugar común de la tradición empirista la idea de que aprendemos nuestras primeras palabras, como «manzana», «hombre», «perro», «agua», que al principio desempeñan la función de oraciones, mediante un condicionamiento de determinados sonidos o conducta verbal frente a fragmentos apropiados de materia en el ámbito público. El condicionamiento funciona de forma óptima con objetos que despiertan el interés del aprendiz y que difícilmente pueden pasar inadvertidos al maestro y al discípulo. Este no es sólo un relato acerca de cómo aprendemos a usar palabras, sino que ha de ser también, necesariamente, parte esencial de una explicación adecuada de la referencia y el significado de las palabras.

Huelga decir que el relato completo no puede ser tan sencillo, pero, por otra parte, resulta difícil creer que esta especie de interacción directa entre usuarios del lenguaje y eventos y objetos públicos no sea una pieza básica del mismo, la pieza que, directa o indirectamente, determina en gran medida el modo en que las palabras se relacionan con las cosas. Sin embargo, este relato tiene consecuencias que parecen haber sido ignoradas hasta tiempos muy recientes. Una de ellas es que los detalles de los mecanismos constitutivos de las cadenas causales que unen a los hablantes entre sí, así como al hablante con el aprendiz del lenguaje y con el objeto del que se habla, no pueden tener relevancia, por sí misma, para el significado y la referencia. La captación de

significados viene determinada únicamente por los elementos terminales del proceso de condicionamiento y se pone a prueba tan sólo mediante el producto final, a saber, el uso de palabras engranadas con objetos y situaciones apropiadas. La mejor forma de percibir esto tal vez sea advertir que dos hablantes que «entienden lo mismo» ante una expresión no necesitan tener en común más que sus disposiciones para una conducta verbal apropiada; sus estructuras neurológicas pueden ser muy diferentes. Dicho a la inversa: dos hablantes pueden ser semejantes en todos los aspectos físicos relevantes y, sin embargo, entender cosas distintas ante las mismas palabras debido a diferencias en las situaciones externas en que las aprendieron. Así, pues, en la medida en que se concibe lo subjetivo o lo mental como algo que sobreviene* a las características físicas de una persona, y nada más, los significados no pueden ser puramente subjetivos o mentales. Como lo expresó Hilary Putnam, «los significados no están en la cabeza».[4] La cuestión estriba en que la interpretación correcta de lo que un hablante quiere decir no está determinada únicamente por lo que hay en su cabeza, sino que depende también de la historia natural de lo que hay en la cabeza. El argumento de Putnam depende de experimentos conceptuales bastante elaborados que algunos filósofos no encuentran convincentes. Pero, por lo que se me alcanza, la mejor forma de defender su posición es apelar directamente a hechos obvios sobre el aprendizaje del lenguaje y a hechos relativos al modo en que interpretamos palabras y lenguajes

* La «sobrevinencia» de lo mental es un término técnico en la filosofía de la psicología o en la filosofía de la mente. Atribuir a lo mental un carácter «sobreviniente» respecto de lo físico significa comprometerse con la idea de que no puede haber una diferencia mental entre dos organismos sin una correspondiente diferencia física, de modo que no podría haber dos organismos físicamente iguales que difiriesen en alguna propiedad mental. (T.)

4. Hilary Putnam, «The Meaning of "Meaning"», reimpreso en *Philosophical Papers*, vol. 2: *Mind, Language, and Reality*, Cambridge University Press, 1975, pág. 227.

con los que no estamos familiarizados.[5] Los hechos relevantes ya han sido mencionados anteriormente; en los casos más simples y básicos, las palabras y las oraciones derivan su significado de los objetos y circunstancias en las que fueron aprendidas. Si en el proceso de aprendizaje hemos sido condicionados para considerar verdadera una oración en presencia del fuego, esta oración será verdadera cuando el fuego esté presente; si hemos sido condicionados para considerar aplicable una palabra en presencia de serpientes, esta palabra hará referencia a serpientes. Muchas palabras y oraciones no se aprenden de este modo, por supuesto; pero son las que se aprenden así las que sujetan el lenguaje al mundo.

Si los significados de las oraciones son proposiciones, y las proposiciones son los objetos de actitudes como la creencia, la intención y el deseo, lo que hemos dicho acerca de los significados debe aplicarse también a todas las actitudes proposicionales. El punto esencial puede plantearse sin recurrir a las proposiciones o a otros supuestos objetos de dichas actitudes. En efecto, del hecho de que los hablantes son, en general, capaces de expresar sus pensamientos en el lenguaje se deriva que, en la misma medida en que la subjetividad del significado esté sometida a duda, también lo estará la del pensamiento en general.

Las consecuencias de estas consideraciones para la teoría del conocimiento son (o deberían ser) sencillamente revolucionarias. Si, en los casos más básicos, las palabras y los pensamientos tratan necesariamente de los tipos de objetos y eventos que los causan, no hay espacio alguno para dudas cartesianas acerca de la existencia independiente de tales objetos y eventos. Puede haber dudas, desde luego. Pero no es necesario que haya algo sobre lo cual estemos indudablemente en lo cierto para que sea correcto afirmar que estamos generalmente en lo cierto sobre la naturaleza del mundo. En ocasiones el escepticismo parece descansar en una

5. Donald Davidson, «Knowing One's Own Mind», *Proceedings and Addresses of the American Philosophical Association*, 1986. (Traducción incluida en el presente volumen.)

simple falacia, consistente en pasar del hecho de que no hay nada sobre lo que no pudiéramos estar equivocados a la conclusión de que podríamos estar equivocados acerca de todo. Esta segunda posibilidad queda excluida si aceptamos que nuestras oraciones más simples reciben sus significados de las situaciones que generalmente causan que las consideremos verdaderas o falsas, puesto que considerar verdadera o falsa una oración que entendemos equivale a tener una creencia. Siguiendo en esta misma dirección, vemos que el escepticismo general sobre las aportaciones de los sentidos ni siquiera puede ser formulado, ya que, si los contenidos de la mente dependen de las relaciones causales, sean cuales fueren, entre las actitudes y el mundo, los sentidos y sus aportaciones no desempeñan un papel *teórico* central en la explicación de la creencia, el significado y el conocimiento. Con esto no se niega, por supuesto, la importancia del papel causal efectivo de los sentidos en el conocimiento y en la adquisición del lenguaje.

La razón por la que los sentidos no son de importancia teórica primaria en la explicación filosófica del conocimiento consiste en que el hecho de que nuestros oídos, ojos, papilas gustativas y órganos táctiles y olfativos tengan un papel causal en la formación de nuestras creencias acerca del mundo constituye un simple accidente empírico. Las conexiones causales entre el pensamiento y los objetos y eventos del mundo podrían haberse establecido de forma totalmente distinta sin que esto supusiera diferencia alguna en los contenidos o en el carácter verídico de la creencia. La filosofía ha cometido el error de suponer que, puesto que a menudo es natural terminar la defensa de una determinada pretensión de conocimiento con la frase «lo vi con mis propios ojos», toda justificación del conocimiento empírico debe remontarse a la experiencia sensorial. Es cierto que determinadas creencias causadas directamente por la experiencia sensorial son con frecuencia verídicas y, por lo tanto, ofrecen a menudo buenas razones para ulteriores creencias. Pero esto no sitúa dichas creencias, por principio, en un lugar aparte ni les confiere prioridad epistemológica alguna.

Si lo dicho es correcto, la epistemología (segregada, quizá, del estudio de la percepción, cuyo parentesco con la epistemología se nos presenta ahora como lejano) no tiene necesidad básica alguna de «objetos de la mente» subjetivos, puramente privados, ya sea en calidad de experiencia o de datos sensoriales no interpretados, por un lado, o de proposiciones plenamente interpretadas, por otro. Contenido y esquemas, según veíamos en el texto de C.I. Lewis citado más arriba, se presentan en forma de pareja; podemos, pues, dejar que desaparezcan juntos. Una vez dado este paso, no quedarán ya *objetos* con respecto a los cuales pueda plantearse el problema de la representación. Las creencias son verdaderas o falsas, pero no representan nada. Es una buena cosa librarse de las representaciones, y con ellas de la teoría de la verdad como correspondencia, ya que es la idea de que hay representaciones lo que engendra los pensamientos relativistas. Las representaciones *son* relativas a un esquema; un mapa representa, pongamos por caso, México, pero sólo en relación con una proyección de Mercator o con alguna otra.

Hay abundancia de enigmas en torno a la sensación y a la percepción, pero estos enigmas, como ya indiqué, no afectan a los fundamentos de la epistemología. La cuestión de qué es lo experimentado directamente en la sensación y cómo se relaciona con los juicios de percepción sigue siendo hoy tan difícil de responder como lo ha sido siempre, pero ya no se puede dar por supuesto que constituya una cuestión central para la teoría del conocimiento. La razón de ello ya la hemos indicado: aunque la sensación desempeña un papel crucial en el proceso causal que conecta las creencias con el mundo, es un error pensar que desempeña un papel *epistemológico* en la determinación de los contenidos de dichas creencias. Al aceptar esta conclusión, estamos abandonando el dogma crucial del empirismo tradicional, el que yo he denominado tercer dogma del empirismo.* Pero esto es lo que

* Davidson denomina este dogma, a saber, la separación entre contenido no interpretado y esquema conceptual, el «tercero» en relación con los dos primeros, criticados por Quine en su célebre artículo «Dos dogmas del empirismo». (T.)

cabía esperar, pues el empirismo es la doctrina según la cual lo subjetivo constituye el fundamento del conocimiento empírico objetivo. Lo que estoy sugiriendo es que el conocimiento empírico no tiene fundamento epistemológico alguno y tampoco lo necesita.[6]

Hay otro problema, bien conocido, que resulta transformado una vez reconocemos que las creencias, los deseos y el resto de las llamadas actitudes proposicionales no son subjetivas del modo en que pensábamos que lo eran. Me refiero al problema de cómo una persona conoce la mente de otra. Quizá resulte obvio que, si es correcta la explicación que he esbozado de nuestra comprensión del lenguaje y de su conexión con los contenidos del pensamiento, la accesibilidad de las mentes ajenas está asegurada desde el principio. De este modo queda descartado el escepticismo acerca de la *posibilidad* de conocer otras mentes. Pero reconocer esto no equivale a responder a la pregunta sobre las condiciones conceptuales que debe cumplir la estructura intelectual que posibilita a un intérprete el paso desde la conducta observada al conocimiento de las actitudes intencionales de otro sujeto. Sin embargo, el hecho de que el lenguaje y el pensamiento tengan una naturaleza que los hace interpretables garantiza que esa pregunta *tiene* una respuesta.[7]

No debemos suponer que todos los problemas de la epistemología vayan a evaporarse si nos libramos de la tiranía o seducción de las dicotomías esquema-contenido y subjetivo-objetivo. Pero los problemas que parecen más importantes serán distintos. Responder al escéptico global dejará de constituir un reto, la búsqueda de fundamentos epistemológicos en la experiencia aparecerá como una tarea huera y el relativismo conceptual perderá su atractivo. No obstante, un buen número de cuestiones de igual o mayor interés permanecerán o serán generadas por la nueva perspectiva. La ex-

6. Donald Davidson, «A Coherence Theory of Truth and Knowledge», en *Kant oder Hegel*, comp. D. Henrich, Klett-Cotta, 1983. (Traducción incluida en el presente volumen.)

7. Donald Davidson, «First Person Authority», *Dialectica*, 38 (1984).

tinción de los subjetivo, tal como había sido previamente concebido, nos deja sin fundamentos para el conocimiento y nos libera de la necesidad de tenerlos. Surgen, sin embargo, nuevos problemas, que se agrupan en torno a la naturaleza del error, pues resulta difícil identificar y explicar el error si no se restringe de algún modo el holismo que acompaña a una concepción no fundamentalista. La posibilidad del conocimiento del mundo y de otras mentes no resulta problemática; pero la forma en que alcanzamos dicho conocimiento y las condiciones que la creencia ha de satisfacer para que pueda constituirse en conocimiento siguen siendo cuestiones a resolver. No se trata tanto de problemas de epistemología tradicional como de problemas acerca de la naturaleza de la racionalidad, problemas que, como los de tipo epistemológico a los que sustituyen, no tienen una solución definitiva, pero que, a diferencia de éstos, merece la pena tratar de resolver.

Hoy en día, la familiaridad con muchos de los aspectos que he indicado es bastante amplia entre los filósofos. Pero, por lo que yo sé, sólo unos pocos de ellos han advertido el alcance de la revolución que todo esto implica en nuestras formas de pensar acerca de la filosofía. Parte, al menos, de la razón de esta inadvertencia podría residir en ciertos malentendidos sobre la naturaleza de lo que cabría denominar el nuevo antisubjetivismo. He aquí tres de ellos.

1. Han sido los ejemplos, más que los argumentos de tipo general, los que han persuadido a mucha gente de que los significados dependen de factores exteriores a nuestras cabezas. Como consecuencia de ello, hay una fuerte tendencia a suponer que la dependencia se limita a los tipos de expresiones que aparecen una y otra vez en los ejemplos, a saber, los nombres propios, los términos de tipos naturales como «agua» y «oro» y las expresiones indicativas.* Pero, de hecho, el fenómeno es ubicuo, ya que es inseparable del ca-

* Traduzco «indexicals» por «expresiones indicativas». Estas expresiones abarcan los pronombres demostrativos, pero también pronombres personales y adverbios como «aquí» o «ahora». (T.)

rácter social del lenguaje. No se trata de un problema local que haya de resolverse mediante alguna argucia semántica; se trata de un hecho perfectamente general acerca de la naturaleza del pensamiento y del habla.[8]

2. Si los estados mentales como la creencia, el deseo, la intención y el significado no sobrevienen únicamente a los estados físicos del agente, en ese caso, han argüido algunos, las teorías que identifican los estados y eventos mentales con estados y eventos físicos en el cuerpo han de ser erróneas. A esto apunta el *dictum* de Putnam según el cual «los significados no están en la cabeza», que Tyler Burge y Andrew Woodfield defienden explícitamente.[9] El argumento presupone que si un estado o evento es identificado (quizá de modo necesario, si se trata de un estado o evento mental) por referencia a cosas externas al cuerpo, ese mismo estado o evento ha de hallarse fuera del cuerpo, o al menos no puede ser idéntico a un evento que tenga lugar en el cuerpo. Este supuesto constituye sencillamente un error: con el mismo derecho se podría argüir que un eritema solar no se encuentra en el cuerpo de la persona que sufrió la quemadura (puesto que el estado de la piel ha sido identificado por referencia al sol). De forma similar, caracterizamos los estados mentales, en parte, por sus relaciones con eventos y objetos externos a la persona, pero esto no demuestra que los estados mentales sean estados de algo distinto de la persona misma o que no sean idénticos a estados físicos.

3. El tercer malentendido se relaciona estrechamente con el segundo. Piensan algunos que si la determinación correcta de los pensamientos de un agente depende, al menos hasta cierto punto, de la historia causal de los mismos, entonces, como el agente puede ignorar esa historia, puede asimismo no saber lo que piensa (ni, *mutatis mutandis*, lo que quiere decir, lo que pretende, etc.). El «nuevo antisubjetivis-

8. Tyler Burge, «Individualism and the Mental», en *Midwest Studies in Philosophy*, vol. 4, comps. Peter French, Theodore Vehling y Howard Wettstein, University of Minnesota Press, 1979.

9. Ibíd. pág. 111 y Andrew Woodfield, *Thought and Object*, comp. Andrew Woodfield, Clarendon Press, 1982, pág. 8.

mo» se interpreta, pues, como una amenaza a la autoridad de la primera persona: al hecho de que, en general, una persona sabe lo que ella misma piensa, desea y pretende sin recurrir a inferencia alguna a partir de la evidencia y, por tanto, de un modo distinto de como lo saben los demás. Una reacción natural, aunque injustificada, consiste en recurrir a maniobras tendentes a aislar, una vez más, los estados mentales de sus determinantes externos.

Estas maniobras no son necesarias y, además, nada nos apremia a adoptarlas si lo que pretendemos es defender el conocimiento, pues la autoridad de la primera persona no se halla realmente amenazada. Puede que yo no conozca la diferencia entre un equidna y un puercoespín; en consecuencia, puede que llame puercoespín a todo equidna con el que me encuentre. Sin embargo, debido al entorno en el que aprendí la palabra «puercoespín», mi término «puercoespín» se refiere, no a los equidnas, sino a los puercoespines. A ellos pienso que se refiere el término y uno de ellos es lo que creo tener ante mí cuando, con toda sinceridad, digo: «Eso es un puercoespín». Mi ignorancia de las circunstancias que determinan lo que quiero decir y lo que pienso no muestra en modo alguno que yo no sepa lo que quiero decir y lo que pienso. Suponer otra cosa no hace sino poner de manifiesto la fuerza con que nos aferramos a la concepción subjetiva de los estados mentales, según la cual éstos podrían ser exactamente como son con independencia del resto del mundo y de su historia.

Otra reacción ante la supuesta amenaza a la realidad de nuestra vida interior consiste en admitir que las creencias y otros estados mentales que identificamos de manera normal no son verdaderamente subjetivos, pero sin dejar de sostener al mismo tiempo que hay otros estados mentales internos, similares a aquéllos, que sí lo son. Una forma que podría adoptar esta idea sería, por ejemplo, la siguiente: puesto que no hay nada en mi estado interno o en mi conducta que corresponda a la distinción entre puercoespines y equidnas, lo que realmente creo cuando veo un equidna (o un puercoespín) es que tengo ante mi un animal con ciertas característi-

cas generales, compartidas de hecho por puercoespines y equidnas. Dado que mi *palabra* «puercoespín» se refiere sólo a puercoespines, el problema consiste en que, aparentemente, no sé lo que quiero decir cuando afirmo: «Eso es un puercoespín». Esta solución tan poco atractiva es en realidad innecesaria, puesto que se basa en una confusión acerca de lo que es «interno». No hay una diferencia *física* que distinga mi estado actual del estado en que me encontraría si quisiera decir «equidna o puercoespín» o «animal con tales y cuales propiedades» en lugar de «puercoespín» y creyese lo que entonces querría decir, pero de ello no se sigue que no haya una diferencia *psicológica*. (Puede que no haya una diferencia física entre estar bronceado por el sol y estarlo gracias a una lámpara solar, pero hay una diferencia, puesto que un estado fue causado por el sol y el otro no. Los estados psicológicos son, a este respecto, como el bronceado.) Así, pues, nada impide decir que yo puedo saber lo que quiero decir al usar la palabra «puercoespín» y lo que creo al pensar en los puercoespines, aun cuando no pueda distinguir un equidna de un puercoespín. La diferencia psicológica, que es precisamente la diferencia que existe entre querer decir y creer que hay un puercoespín ante mí y querer decir y creer que hay una criatura con ciertos rasgos comunes a los puercoespines y los equidnas, es exactamente la diferencia que se necesita para garantizar que yo sé lo que quiero decir y lo que pienso. Todo lo que Putnam y otros han mostrado es que esta diferencia no tiene por qué reflejarse en el estado físico del cerebro.

Inventar un nuevo conjunto de estados psicológicos realmente «internos» o «estrechos» no es, pues, una forma de restaurar la autoridad de la primera persona en el ámbito de lo mental; más bien lo contrario. Sin embargo, quedaría por considerar la afirmación según la cual una ciencia psicológica sistemática requiere estados del agente susceptibles de ser identificados sin referencia a su historia u otras conexiones con el mundo externo. En caso contrario, se afirma, no habría explicación alguna del hecho de que yo, que con mi palabra «puercoespín» puedo referirme sólo a

puercoespines, no sepa indicar la diferencia entre un puercoespín y un equidna mejor ni peor que si, en lugar de ello, quisiera decir (sin cambio físico alguno) «puercoespín o equidna».

Las esperanzas de lograr una psicología científica no tienen relevancia directa para el tema del presente ensayo, por lo que podemos dejar de lado la cuestión de si hay estados internos de los agentes capaces de explicar su conducta mejor que las creencias y deseos ordinarios. Sin embargo, resulta pertinente considerar si hay estados de la mente que puedan reclamar la denominación de subjetivos con más derecho que las actitudes proposicionales, tal y como éstas se conciben e identifican habitualmente.

A este respecto nos encontramos con dos sugerencias. La más modesta (presente, por ejemplo, en la obra de Jerry Fodor) consiste en decir que los auténticos estados internos o solipsistas podrían ser ciertos estados seleccionados entre las actitudes habituales y modificaciones de éstas. Los pensamientos acerca de puercoespines y equidnas quedarían eliminados, pues sus contenidos se identifican en virtud de relaciones con el mundo externo. Serían admisibles, sin embargo, pensamientos acerca de animales que satisfacen ciertos criterios generales (por ejemplo, los que usamos para decidir si algo es un puercoespín).[10]

Semejantes estados internos, si es que realmente existen, contarían como subjetivos según todos o casi todos los cánones: serían susceptibles de identificación y clasificación sin referencia a objetos y eventos externos, cabría acudir a ellos para que sirvieran como fundamentos del conocimiento empírico y estarían sometidos, con toda verosimilitud, a la autoridad de la primera persona.

Sin embargo, parece claro que no existen estados semejantes, al menos si han de poder expresarse en palabras. Los «rasgos generales» o «criterios» que usamos para identificar

10. Jerry Fodor, «Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology», *The Behavioral and Brain Sciences*, 3 (1980).

puercoespines son, pongamos por caso, tener cuatro patas, hocico, ojos y púas. Pero resulta evidente que los significados de las palabras que se refieren a estos rasgos y los contenidos de los conceptos expresados por ellas dependen de la historia natural de la adquisición de dichas palabras y conceptos en no menor medida de lo que sucedía con «puercoespín» y «equidna». No hay palabras, o conceptos vinculados a palabras, que no hayan de ser entendidas e interpretadas, directa o indirectamente, en términos de relaciones causales entre las personas y el mundo (y, desde luego, de relaciones entre palabras y entre conceptos).

En este punto cabe imaginar una propuesta que apuntaría a la existencia de criterios fenoménicos inexpresables a los cuales pudieran reducirse los criterios expresables de modo público; es de esperar, en este caso, que el recuerdo de pasados fracasos cosechados por fantasías reduccionistas semejantes sirvan para disipar la idea de que dicha propuesta pueda llevarse a cabo. Sin embargo, aun prescindiendo de cavilaciones nostálgicas en torno a la reducción fenomenista, resulta instructivo encontrarse con el esfuerzo tendente a dar rango científico a la psicología transformado en una búsqueda de estados proposicionales que puedan ser detectados e identificados con independencia de sus relaciones con el resto del mundo, a semejanza del afán de filósofos anteriores por hallar algo «dado en la experiencia» que no contuviese clave necesaria alguna de lo que ocurría en el exterior. El motivo es similar en ambos casos: la idea de que un apoyo sólido, para el conocimiento o para la psicología, requiere de algo interno, en el sentido de no relacional.

La segunda, y más revolucionaria, sugerencia consiste en sostener que los estados mentales necesarios para una psicología científica, aun manteniendo cierto carácter proposicional, no guardan relación directa con las creencias, deseos e intenciones cotidianas.[11] Estos estados se conciben, por esti-

11. Esta idea ha sido propuesta por Stephen Stich, *From Folk Psychology to Cognitive Science: The Case Against Belief*, M.I.T. Press, 1983.

pulación, como aquello que explica la conducta, y son, por lo tanto, internos o subjetivos únicamente en el sentido de que caracterizan a una persona o sujeto similar y se hallan debajo de la piel. No hay razón alguna para suponer que las personas puedan decir cuándo se hallan en tales estados.

Recapitulando lo dicho, he hecho cinco consideraciones, conectadas entre sí, acerca de los «contenidos de la mente».

En primer lugar, los estados de la mente, tales como dudas, anhelos, creencias y deseos, se identifican, en parte, por el contexto social e histórico en que se adquieren; en este aspecto son semejantes a otros estados que se identifican por medio de sus causas, como por ejemplo padecer ceguera de nieve o favismo (una dolencia causada por contacto con las habas).

En segundo lugar, lo anterior no demuestra que los estados mentales no sean estados físicos de una persona; la manera en que describimos e identificamos eventos y estados no tiene relación directa con el lugar en que están.

En tercer lugar, el hecho de que los estados de la mente –incluyendo lo que un hablante quiere decir– se identifiquen por relaciones causales con objetos y eventos externos es esencial para la posibilidad de la comunicación y hace que una mente sea, en principio, accesible a las demás; pero este aspecto público e interactivo de la mente no lleva a disminuir la importancia de la autoridad de la primera persona.

En cuarto lugar, la idea de que hay una división básica entre la experiencia no interpretada y un esquema conceptual organizador constituye un profundo error, nacido de una imagen esencialmente incoherente de la mente como un espectador pasivo, pero crítico, de un espectáculo interior. Una explicación naturalista del conocimiento no apela a intermediarios epistemológicos tales como datos sensoriales *qualia* o sensaciones puras. Como resultado de ello, el escepticismo global acerca de los sentidos no es una posición susceptible de ser siquiera formulada.

Finalmente, he argumentado en contra de la posibilidad de «objetos del pensamiento», tanto si se conciben bajo el modelo de los datos sensoriales como si se les concede ca-

rácter proposicional. Hay una gran diversidad de estados mentales, pero su descripción no requiere la existencia de entidades fantasmales contempladas de algún modo por la mente. Prescindir de semejantes entidades equivale a eliminar, más que a resolver, cierto número de fastidiosos problemas, ya que, si no hay tales objetos, no resulta pertinente preguntarse cómo pueden representar el mundo o devanarse los sesos con la cuestión de cómo puede la mente tener conocimiento directo de ellos.

¿Qué queda del concepto de subjetividad? Por lo que se me alcanza, quedan en pie dos rasgos de lo subjetivo en su concepción clásica. Los pensamientos son privados en el sentido, obvio pero importante, en que la propiedad puede ser privada, es decir, pertenecer a una persona. Y el conocimiento de los pensamientos es asimétrico, en cuanto que la persona que tiene un pensamiento sabe, por regla general, que lo tiene de un modo en que los demás no pueden saberlo. Pero esto es todo lo que resta de lo subjetivo. Así, lejos de constituir un coto cerrado, hasta el punto de que el modo en que pueda aportar conocimiento de un mundo externo o ser conocido por otros se convierta en un problema, el pensamiento es, necesariamente, parte de un mundo público común. No sólo pueden otras personas llegar a saber lo que pensamos al advertir las dependencias causales que dan a nuestros pensamientos su contenido, sino que la posibilidad misma del pensamiento exige patrones compartidos de verdad y objetividad.

VERDAD Y CONOCIMIENTO: UNA TEORÍA DE LA COHERENCIA

En este trabajo voy a defender lo que muy bien puede denominarse una teoría de la coherencia acerca de la verdad y el conocimiento. La teoría que sostengo no entra en competencia con una teoría de la correspondencia, pero su defensa depende de un argumento cuyo propósito es mostrar que la coherencia genera correspondencia.

La importancia del tema resulta obvia. Si la coherencia es una prueba de la verdad, la conexión con la epistemología es directa, ya que tenemos razones para pensar que muchas de nuestras creencias son coherentes con muchas otras, lo que a su vez nos proporciona razones para pensar que muchas de nuestras creencias son verdaderas. Y allí donde las creencias son verdaderas, parece que las condiciones primarias del conocimiento han sido satisfechas.

Alguien podría tratar de defender una teoría de la coherencia acerca de la verdad sin ampliarla al conocimiento, basándose tal vez en que el sujeto de un conjunto coherente de creencias podría carecer de razones para creer que sus creencias son coherentes. No es muy probable que esto suceda, pero bien puede ser que alguien, aun teniendo creencias verdaderas y buenas razones para tenerlas, no aprecie la relevancia de las razones para las creencias. La mejor forma de concebir a una persona semejante podría ser considerarla como alguien que tiene conocimiento sin saber que lo tiene: este sujeto piensa de sí mismo que es un escéptico. En una palabra, es un filósofo.

Dejando al margen los casos aberrantes, lo que mantiene unidos la verdad y el conocimiento es el significado. Si los significados vienen dados por las condiciones objetivas de

verdad, el problema es cómo sabemos que tales condiciones han sido satisfechas, ya que esto parecería requerir una confrontación entre lo que creemos y la realidad, y la idea de una confrontación semejante es absurda. Pero si la coherencia es una prueba de la verdad, es entonces también una prueba del juicio de que las condiciones objetivas de verdad han sido satisfechas, de modo que ya no necesitamos explicar el significado sobre la base de una posible confrontación. Mi lema es: correspondencia sin confrontación. Dada una epistemología correcta, podemos ser realistas en todos los campos. Podemos aceptar las condiciones objetivas de verdad como la clave del significado, podemos aceptar una concepción realista de la verdad y podemos también insistir en que el conocimiento lo es de un mundo objetivo, independiente de nuestro pensamiento o lenguaje.

Puesto que no hay, por lo que yo sé, una teoría que merezca llamarse «la» teoría de la coherencia, voy a caracterizar el tipo de concepción que quiero defender. Resulta obvio que no todos los conjuntos coherentes de oraciones interpretadas contienen sólo oraciones verdaderas, puesto que uno de tales conjuntos podría contener únicamente la oración coherente *S* y otro únicamente la negación de *S*. Y no será de ninguna ayuda añadir más oraciones preservando la coherencia. Podemos imaginar innumerables descripciones de estados –descripciones de máxima coherencia– que no describen nuestro mundo.

Mi teoría de la coherencia se aplica a creencias, u oraciones que son verdaderas para alguien que las entiende. No deseo afirmar, en este punto, que todo posible conjunto coherente de creencias es verdadero (o contiene creencias mayoritariamente verdaderas). Rehúyo afirmar esto debido a la escasa claridad acerca de lo que es posible en este campo. En un extremo, se podría sostener que el ámbito de posibles conjuntos de creencias máximamente coherentes es tan amplio como el de posibles conjuntos de oraciones máximamente coherentes, con lo cual no tendría sentido seguir insistiendo en que una teoría defendible de la coherencia se aplica a creencias y no a proposiciones u oraciones. Pero

hay otras formas de concebir lo que es posible creer, las cuales harían justificable afirmar no sólo que todos los sistemas reales de creencias coherentes son ampliamente correctos, sino que lo son también todos los sistemas posibles. La diferencia entre las dos nociones acerca de lo que es posible creer depende de nuestros supuestos en torno a la naturaleza de la creencia, su interpretación, sus causas, sujetos y configuraciones. Para mí las creencias son estados de las personas que tienen intenciones, deseos, órganos sensoriales; son estados causados por –y que causan a su vez– eventos internos y externos al cuerpo de sus poseedores. Pero, aun con todas estas restricciones, hay muchas cosas que las personas creen y muchas más que podrían creer. La teoría de la coherencia se aplica a todos estos casos.

Desde luego, algunas creencias son falsas. Gran parte del interés del concepto de creencia lo constituye la brecha potencial que introduce entre lo que se tiene por verdadero y lo que es verdadero. Así, la mera coherencia, por robusta y plausible que sea la definición que demos de ella, no puede garantizar que aquello que se cree sea efectivamente así. Todo lo que una teoría de la coherencia puede mantener es que en un conjunto coherente de creencias, la mayoría de ellas son verdaderas.

Esta forma de exponer la posición puede tomarse, en el mejor de los casos, como una indicación, ya que probablemente no hay ninguna forma útil de contar creencias, y con ello, a su vez, la idea de que la mayoría de las creencias de una persona son verdaderas no tiene un significado claro. Un modo mejor de indicar la clave del asunto es tal vez decir que hay una presunción en favor de la verdad de una creencia que es coherente con una masa significativa de otras creencias. Toda creencia, en un conjunto total coherente de ellas, está justificada a la luz de esta presunción, no de modo muy distinto de como lo está toda acción intencional emprendida por un agente racional (cuyas elecciones, creencias y deseos son coherentes en el sentido de la teoría bayesiana de la decisión). Así, pues, por decirlo una vez más, si el conocimiento es creencia verdadera justificada, parecería en-

tonces que todas las creencias verdaderas de un sujeto coherente constituyen conocimiento. Por más que esta conclusión sea demasiado vaga y precipitada para ser correcta, contiene, no obstante, un importante núcleo de verdad, como trataré de argüir. Entretanto, me limitaré a indicar los múltiples problemas que aguardan tratamiento: ¿Qué exige exactamente la coherencia? ¿Qué dosis de práctica inductiva habría que incluir? ¿Qué proporción de la teoría verdadera (si la hay) del apoyo evidencial ha de encontrarse en ella? Puesto que nadie posee un cuerpo de convicciones completamente coherente, *¿cuáles* son las creencias con las cuales la coherencia crea una presunción de verdad? Algunos de estos problemas se situarán en una perspectiva mejor a lo largo de este ensayo.

Debería estar claro que no espero definir la verdad en términos de coherencia y creencia. La verdad es bellamente transparente en comparación con la creencia y con la coherencia, de modo que la tomaré como una noción primitiva. La verdad, aplicada a las emisiones de oraciones, muestra el carácter *desentrecomillador* que se encierra en la Convención T de Tarski*, lo cual es suficiente para fijar su ámbito de aplicación; para fijarlo en relación con un lenguaje o un hablante, desde luego, por lo que la verdad no resulta agotada por la Convención T; la verdad contiene también lo que lleva de un lenguaje a otro lenguaje o de un hablante a otro. Lo que revela la Convención T, y las triviales oraciones que declara verdaderas, como «"la hierba es verde", dicha por un hablante hispano, es verdadera si, y sólo si, la hierba es verde», es que la verdad de una emisión depende de dos únicas cosas: lo que significan las palabras dichas y el modo en que

* Según la Convención T de Tarski, cualquier teoría adecuada de la verdad para un lenguaje (formal) L debe tener como consecuencia lógica, cuando la teoría está formulada tomando como metalenguaje el mismo lenguaje que es objeto de ella, teoremas de la forma siguiente: «*P*» es verdadera en L si, y sólo si, *P*, donde *P* es una oración cualquiera de L. De ahí el carácter *desentrecomillador* del que habla Davidson. Véase también el apartado 4 de nuestra Introducción a este volumen. (T.)

está dispuesto el mundo. No hay un relativismo adicional respecto de un esquema conceptual, una forma de ver las cosas, una perspectiva. Dos intérpretes, tan diferentes como queramos en cuanto a cultura, lenguaje y punto de vista, pueden estar en desacuerdo sobre la verdad de una emisión, pero sólo si difieren acerca de cómo son las cosas en el mundo que comparten o acerca del significado de la emisión.

Creo que estas simples reflexiones nos permiten extraer dos conclusiones. En primer lugar, la verdad es correspondencia con el modo en que son las cosas. (No hay una forma sencilla y libre de confusión de formular esto; para poner las cosas en orden es necesario dar un rodeo a través del concepto de satisfacción, en cuyos términos se caracteriza la verdad.)[1] Así, pues, si una teoría de la coherencia acerca de la verdad es aceptable, ha de estar de acuerdo con una teoría de la correspondencia. En segundo lugar, una teoría del conocimiento que admita que podemos conocer la verdad ha de ser una forma de realismo no interno ni relativizado. Por lo tanto, si una teoría del conocimiento en términos de coherencia es aceptable, ha de estar de acuerdo con ese tipo de realismo. Mi propia forma de realismo no parece ser ni el realismo interno de Hilary Putnam ni su realismo metafísico.[2] No es realismo interno porque éste hace de la verdad algo relativo a un esquema, y ésta es una idea que no considero inteligible.[3] De hecho, constituye una importante razón para aceptar una teoría de la coherencia la falta de inteligibilidad del dualismo de un esquema conceptual y un «mundo» en espera de ser capturado. Pero mi realismo no es tampoco, ciertamente, el realismo metafísico de Putnam, ya que *éste* se caracteriza por ser «radicalmente no epistémico», lo que implica que todos nuestros pensamientos y nuestras teorías

1. Véase mi artículo «True to the Facts», *Journal of Philosophy* (1960), págs. 216-234.

2. Hilary Putnam, *Meaning and the Moral Sciences* (Routledge & Kegan Paul, Londres, 1978), pág. 125.

3. Véase mi artículo «On the Very Idea of a Conceptual Scheme», en *Proceedings and Addresses of the American Philosophical Association* (1974), págs. 5-20.

mejor investigadas y establecidas pueden ser falsas. Considero que la independencia de creencia y verdad requiere únicamente que *cada una* de nuestras creencias pueda ser falsa. Pero, por supuesto, una teoría de la coherencia no puede admitir que todas ellas puedan serlo.

Pero, ¿por qué no? Es tal vez obvio que la coherencia de una creencia con un cuerpo importante de creencias incrementa las posibilidades de que sea verdadera, a condición de que haya razones para suponer que el cuerpo de creencias sea verdadero o que lo sea en gran parte. Pero, ¿cómo puede la coherencia por sí sola sentar bases para la creencia? Tal vez lo mejor que podamos hacer para justificar una creencia sea apelar a otras creencias. Pero entonces el resultado sería, en apariencia, que nos veríamos obligados a aceptar el escepticismo filosófico, por muy firmes que sigan siendo en la práctica nuestras creencias.

Este es el escepticismo en uno de sus ropajes tradicionales. La pregunta es: ¿por qué no podrían todas mis creencias ser coherentes entre sí siendo al mismo tiempo falsas acerca del mundo real? El simple reconocimiento de que es absurdo –o algo peor que eso– tratar de *confrontar* nuestras creencias, una a una o en su conjunto, con aquello de que tratan no responde a la pregunta ni muestra que sea ininteligible. En suma: incluso una teoría moderada de la coherencia como la mía ha de proporcionar al escéptico razones para suponer que las creencias coherentes son verdaderas. El partidario de una teoría de la coherencia no puede permitir que la seguridad proceda del exterior del sistema de creencias si nada en su interior puede ofrecer apoyos, a menos que se pueda mostrar que descansa, en último término o de modo inmediato, en algo independientemente fidedigno.

Es natural distinguir las teorías de la coherencia de teorías de otro tipo por referencia a la cuestión de si la justificación puede o tiene que llegar a un fin o no. Esto, sin embargo, no define las posiciones, sino que se limita a sugerir una forma que puede adoptar la argumentación, pues aunque hay teóricos de la coherencia que sostienen que algunas creencias pueden servir de base a las demás, seguiría siendo posible

mantener que la coherencia no es suficiente aun cuando la aportación de razones nunca llegue a un término. Lo que distingue una teoría de la coherencia es simplemente la idea de que nada puede contar como una razón para sostener una creencia excepto otra creencia. El defensor de esta idea rechaza por ininteligible la demanda de fundamentos o fuentes de justificación de una especie distinta. En palabras de Rorty, «nada cuenta como justificación salvo por referencia a lo que ya aceptamos, y no hay forma de salir de nuestras creencias y lenguaje para hallar alguna otra prueba que no sea la coherencia».[4] En esto estoy de acuerdo con Rorty, como puede verse. Nuestras discrepancias, si las hay, conciernen a la permanencia de la pregunta acerca de si, dado que no nos es posible «salir de nuestras creencias y lenguaje para hallar alguna otra prueba que no sea la coherencia», podemos no obstante conocer y hablar sobre un mundo público objetivo que no hemos producido. Yo pienso que esta pregunta subsiste, pero sospecho que Rorty no lo cree así. Si éste es su punto de vista, habrá de pensar que estoy en un error al intentar responderla. En cualquier caso, voy a tratar de hacerlo.

Será de ayuda en este punto dar un apresurado repaso a algunas de las razones para abandonar la búsqueda de una base para el conocimiento más allá del alcance de nuestras creencias. Por «base» entiendo específicamente una base epistemológica, una fuente de justificación.

Los intentos dignos de ser tomados en serio tratan de fundamentar la creencia, de un modo u otro, en el testimonio de los sentidos: sensación, percepción, lo dado, la experiencia, los datos sensoriales, el espectáculo cambiante.* Todas estas teorías han de explicar al menos dos cosas: ¿cuál es exactamente la relación entre sensación y creencia que permite a la primera justificar la segunda?, y ¿por qué

4. Richard Rorty, *Philosophy and the Mirror of Nature* (Princeton University Press, Princeton, 1979), pág. 178.

* Con la expresión «el espectáculo cambiante» alude Davidson a la concepción empirista, y más específicamente humana, de la mente como un escenario privado e interno a cada sujeto de conciencia. (T.)

habríamos de creer que nuestras sensaciones son confiables, esto es, por qué deberíamos confiar en nuestros sentidos?

La idea más simple consiste en identificar ciertas creencias con sensaciones. Así, no parece que Hume haya distinguido entre percibir una mancha verde y percibir que una mancha es verde. (Una ambigüedad en la palabra «idea» fue aquí de gran ayuda.) Otros filósofos advirtieron la confusión de Hume, pero trataron de alcanzar los mismos resultados reduciendo a cero el hiato entre percepción y juicio mediante el intento de formular juicios que no rebasan el enunciado de que la percepción o sensación o presentación existen (cualquiera que sea el significado de esto). Dichas teorías no justifican las creencias sobre la base de sensaciones, sino que tratan de justificar ciertas creencias sosteniendo que tienen exactamente el mismo contenido epistémico que una sensación. Esta concepción se enfrenta con dos dificultades: en primer lugar, si las creencias básicas no exceden en contenido a la sensación correspondiente, no pueden servir de apoyo a inferencia alguna acerca de un mundo objetivo; y, en segundo lugar, no hay tales creencias.

Una aproximación más plausible consiste en sostener que no podemos estar equivocados acerca de cómo nos parece que son las cosas. Si creemos que tenemos una sensación, la tenemos; ésta, sostienen, es una verdad analítica, o un hecho acerca de cómo se usa el lenguaje.

Es difícil explicar esta supuesta conexión entre las sensaciones y algunas creencias de un modo que no invite al escepticismo acerca de otras mentes y, en ausencia de una explicación adecuada; deberían ponerse en duda las implicaciones de dicha conexión para la cuestión de la justificación. En cualquier caso, sin embargo, no resulta claro cómo, según esta concepción, las sensaciones justifican la creencia en ellas mismas. El punto central es, más bien, que dichas creencias no requieren justificación, pues la existencia de la creencia implica la existencia de la sensación, de modo que la existencia de la creencia implica su propia verdad. A menos que se añada algo más, nos vemos llevados a la teoría de la coherencia en otra de sus formas.

El énfasis en la sensación o percepción en cuestiones epistemológicas surge de la idea obvia según la cual las sensaciones son lo que conecta el mundo con nuestras creencias y aspiran a desempeñar el papel de justificaciones porque a menudo somos conscientes de ellas. La dificultad con la que hemos tropezado consiste en que la justificación parece depender de la conciencia, que no es sino otra creencia.

Tomemos ahora un rumbo más atrevido. Supongamos que decimos que las sensaciones mismas, verbalizadas o no, justifican ciertas creencias que sobrepasan lo dado en la sensación. Así, bajo ciertas condiciones, tener la sensación de ver una luz verde relampagueante puede justificar la creencia de que una luz verde está relampagueando. El problema es ver cómo la sensación justifica la creencia. Desde luego, si alguien tiene la sensación de ver una luz verde relampagueante, es probable, bajo ciertas circunstancias, que una luz verde esté relampagueando. *Nosotros* podemos decir esto, puesto que sabemos de su sensación, pero *él* no puede decirlo, ya que estamos suponiendo que está justificado sin que tenga que depender de la creencia de que tiene la sensación. Supongamos que creyese que no tuvo la sensación. ¿Justificaría aún la sensación su creencia en una luz verde relampagueante objetiva?

La relación entre una sensación y una creencia no puede ser de carácter lógico, pues las sensaciones no son creencias ni otras actitudes proposicionales. ¿Cuál es, entonces, la relación? Creo que la respuesta es obvia: la relación tiene carácter causal. Las sensaciones causan algunas creencias, y en *este* sentido constituyen la base o sustento de esas creencias. Pero una explicación causal de una creencia no muestra cómo o por qué está justificada la creencia.

La dificultad de transmutar una causa en una razón acosa una vez más al adversario de la coherencia si trata de responder a nuestra segunda pregunta: ¿qué justifica la creencia de que nuestros sentidos no nos engañan sistemáticamente? Pues aun si las sensaciones justifican la creencia en la sensación, todavía no vemos cómo justifican la creencia en eventos y objetos externos.

Quine afirma que la ciencia nos dice que «nuestra única

fuente de información sobre el mundo externo viene a través del impacto de rayos de luz y moléculas en nuestras superficies sensoriales».[5] Lo que me preocupa es cómo leer las palabras «fuente» e «información». Sin duda es verdad que eventos y objetos en el mundo externo causan que creamos cosas sobre el mundo externo y que buena parte de la causalidad, si no toda, se orienta a través de los órganos sensoriales. Sin embargo, la noción de información sólo se aplica de modo no metafórico a las creencias generadas. Así, «fuente» ha de leerse simplemente como «causa» e «información» como «creencia verdadera» o «conocimiento». La justificación de las creencias causadas por nuestros sentidos no se vislumbra todavía.[6]

La aproximación al problema de la justificación que he-

5. W.V. Quine, «The Nature of Natural Knowledge», en *Mind and Language*, S. Guttenplan S., comp. (Clarendon Press, Oxford, 1975), pág. 68.

6. Muchos otros pasajes en Quine sugieren que tiene la esperanza de asimilar las causas sensoriales a la evidencia. En *Word and Object* (M.I.T. Press, Massachussets, 1960), pág. 22, escribe que «las irritaciones de la superficie... agotan nuestras claves del mundo externo». En *Ontological Relativity* (Columbia University Press, Nueva York, 1969), pág. 75, encontramos que «la estimulación de los receptores sensoriales es toda la evidencia con que cualquiera ha podido contar, en último término, para seguir adelante en la construcción de su imagen del mundo». En la misma página leemos: «Dos principios cardinales del empirismo siguen siendo inatacables... El primero es que cualquier evidencia que *haya* para la ciencia *es* evidencia sensorial. El segundo... es que toda inculcación de significados de palabras ha de basarse en último término en la evidencia sensorial». En *The Roots of Reference* (Open Court Publishing Company, Illinois, 1974), págs. 37-38, dice Quine que las «observaciones» son básicas «tanto para el apoyo a la teoría como para el aprendizaje del lenguaje». Y prosigue: «¿Qué son las observaciones? Son visuales, auditivas, táctiles, olfativas. Son, evidentemente, sensoriales y con ello subjetivas... ¿Tendríamos que decir entonces que la observación no es la sensación...? No...». «A continuación Quine deja de hablar de observaciones para pasar a hablar de oraciones observacionales. Pero, por supuesto, las oraciones observacionales, a diferencia de las observaciones, no pueden desempeñar el papel de evidencia a menos que tengamos razones para creer que son verdaderas.

mos estado rastreando tiene que ser errónea. Hemos tratado de verlo del siguiente modo: una persona tiene todas sus creencias sobre el mundo, esto es, todas sus creencias. ¿Cómo puede decir si son verdaderas o aptas para serlo? Unícamente, lo hemos supuesto, conectando sus creencias con el mundo, confrontando algunas de sus creencias, una por una, con las aportaciones de los sentidos, o tal vez confrontando la totalidad de sus creencias con el tribunal de la experiencia. Ninguna confrontación semejante tiene sentido, pues, desde luego, no podemos salir de nuestra piel para descubrir lo que causa los aconteceres internos de los que tenemos conciencia. Introducir en la cadena causal pasos intermedios o entidades, como sensaciones u observaciones, sólo sirve para hacer más obvio el problema epistemológico, pues si los intermediarios son simplemente causas, no justifican las creencias que causan, y si suministran información, pueden estar engañándonos. La moraleja es obvia: puesto que no podemos tomar juramento de veracidad a los intermediarios, no debemos permitir intermediarios entre nuestras creencias y sus objetos en el mundo. Desde luego, hay intermediarios causales. De lo que debemos guardarnos es de los intermediarios epistémicos.

Hay puntos de vista comunes sobre el lenguaje que fomentan la mala epistemología. Esto no es un accidente, desde luego, ya que las teorías del significado están conectadas con la epistemología mediante los intentos de responder a la pregunta sobre el modo en que se determina que una oración es verdadera. Si conocer el significado de una oración (saber cómo darle una interpretación correcta) implica o equivale a saber cómo se podría reconocer su verdad, la teoría del significado plantea el mismo problema con el que hemos estado bregando, pues dar el significado de una oración exigirá especificar aquello que justificaría su aserción. En este punto, el defensor de la coherencia mantendrá que es inútil buscar una fuente de justificación más allá de otras oraciones que se tienen por verdaderas, mientras que el fundamentalista tratará de anclar al menos algunas palabras u oraciones a las rocas no verbales. Esta

es la posición que mantienen, creo, tanto Quine como Michael Dummett.

Dummett y Quine difieren, sin duda. En particular, discrepan en lo referente al holismo, la tesis según la cual la verdad de nuestras oraciones ha de ponerse a prueba en conjunto, y no una por una, y discrepan también, en consecuencia, acerca de la existencia de una distinción útil entre oraciones analíticas y sintéticas, así como sobre la posibilidad de que una teoría satisfactoria del significado permita el tipo de indeterminación por la que Quine aboga. (En todos estos puntos, yo soy un fiel discípulo de Quine.)

Sin embargo, lo que me importa aquí es el hecho de que Quine y Dummett concuerdan en un principio básico, según el cual todo lo relativo al significado ha de remontarse de algún modo a la experiencia, a lo dado o a pautas de estimulación sensorial, a alguna cosa intermedia entre la creencia y los objetos usuales sobre los que versan nuestras creencias. Una vez dado este paso, abrimos la puerta al escepticismo, porque entonces hemos de conceder que un gran número –tal vez la mayoría– de las oraciones que tenemos por verdaderas pueden de hecho ser falsas. Hay algo de ironía en esto. El intento de hacer accesible el significado ha hecho inaccesible la verdad. Cuando el significado discurre de este modo por la senda epistemológica, se produce necesariamente un divorcio entre verdad y significado. Siempre se puede, desde luego, arreglar un casamiento a punta de pistola, redefiniendo la verdad como aquello en cuya aserción estamos justificados. Pero esto no casa a los novios originales.

Consideremos la propuesta de Quine según la cual todo lo referente al significado (valor informativo) de una oración observacional se halla determinado por las pautas de estimulación sensorial que causarían en un hablante el asentimiento o disentimiento con respecto a la oración. Este es un modo maravillosamente ingenioso de retener lo que resulta atrayente en las teorías verificacionistas sin tener que hablar de significados, datos sensoriales o sensaciones; esto hizo plausible, por primera vez, la idea de que se podría, e incluso se debería, hacer lo que yo llamo teoría del significado sin

necesidad de lo que Quine llama significados. Pero la propuesta de Quine, como otras formas de verificacionismo, conduce al escepticismo, ya que, claramente, las estimulaciones sensoriales de una persona podrían ser precisamente como son y en cambio el mundo exterior podría ser muy diferente. (Recordemos el cerebro en la cubeta.)*

El modo en que Quine prescinde de los significados es sutil y complicado. Vincula los significados de algunas oraciones directamente a patrones de estimulación (que, en su opinión, constituyen también la evidencia para asentir a la oración), pero los significados de las demás oraciones están determinados por el modo en que las oraciones originales, u oraciones de observación, los condicionan. Los hechos relativos a ese condicionamiento no permiten una distinción tajante entre oraciones que se consideran verdaderas en virtud del significado y oraciones que se consideran verdaderas sobre la base de la observación. Quine formuló esta tesis mostrando que si una forma de interpretar las emisiones de un hablante era satisfactoria, también lo eran muchas otras. Esta doctrina de la indeterminación de la traducción, como Quine la denominó, no debería considerarse ni misteriosa ni amenazante. No es más misteriosa que el hecho de que la temperatura pueda medirse en grados centígrados o Fahrenheit (o cualquier transformación lineal de estos números). Y no es amenazante porque el mismo procedimiento que demuestra el grado de indeterminación demuestra al mismo tiempo que lo que está determinado es todo lo que necesitamos.

En mi opinión, la supresión de la línea divisoria entre lo analítico y lo sintético salvó la filosofía del lenguaje como un campo de estudio serio al mostrar cómo podría cultivarse sin aquello que no puede haber: significados determinados. Lo que ahora sugiero es que abandonemos también la dis-

* Alude Davidson con esto a un famoso artículo de Hilary Putnam en el que se plantea, mediante un ejemplo de ciencia ficción, el viejo reto cartesiano sobre la posibilidad de que todas nuestras creencias basadas en los sentidos fuesen sistemáticamente erróneas. (T.)

tinción entre oraciones de observación y el resto, pues la distinción entre oraciones en cuya verdad está justificada la creencia por sensaciones, y oraciones en cuya verdad está justificada la creencia solamente por mediación de otras oraciones es tan anatema para el partidario de la coherencia como la distinción entre creencias justificadas por sensaciones y creencias justificadas solamente por apelación a otras creencias. En consecuencia, sugiero que abandonemos la idea de que el significado o el conocimiento se fundamenten en algo que valga como fuente última de evidencia. Sin duda, el significado y el conocimiento dependen de la experiencia y ésta a su vez depende en último término de la sensación. Pero este «depende» es el de la causalidad, no el de la evidencia o la justificación.

He planteado mi problema lo mejor que he podido. La búsqueda de un fundamento empírico para el significado o para el conocimiento conduce al escepticismo, mientras que una teoría de la coherencia parece estar en aprietos cuando se trata de proporcionar a un sujeto de creencias alguna razón para creer que sus creencias, si son coherentes, son verdaderas. Estamos atrapados entre una respuesta errónea al escéptico y la falta de respuesta.

Este no es un dilema genuino. Lo que se necesita para responder al escéptico es mostrar que alguien que posea un conjunto de creencias (más o menos) coherente tiene una razón para suponer que sus creencias no son en su mayor parte erróneas. Lo que hemos puesto de manifiesto es que resulta absurdo buscar un fundamento que justifique la totalidad de las creencias, algo situado fuera de dicha totalidad que podamos usar para poner a prueba nuestras creencias o compararlas con ello. La respuesta a nuestro problema, pues, ha de ser el hallazgo de una *razón* para suponer que la mayoría de nuestras creencias son verdaderas que no sea sin embargo una forma de *evidencia*.

Mi argumento tiene dos partes. En primer lugar, insistiré en el hecho de que una comprensión correcta del habla, creencias, deseos, intenciones y otras actitudes proposicionales de una persona lleva a la conclusión de que la mayoría de las

creencias de una persona han de ser verdaderas, de modo que hay una presunción legítima en favor de la verdad de cualquiera de ellas si es coherente con la mayoría de las demás. A continuación pasaré a sostener que cualquiera que tenga pensamientos, y en particular cualquiera que se pregunte si tiene alguna razón para suponer que está generalmente en lo cierto acerca de la naturaleza de su entorno, ha de saber qué es una creencia y cómo han de detectarse e interpretarse las creencias en general. Puesto que éstos son hechos perfectamente generales que no podemos dejar de usar cuando nos comunicamos con otros, o cuando tratamos de hacerlo, o incluso cuando simplemente pensamos que lo estamos haciendo, hay un sentido muy fuerte en el que se puede decir de nosotros que sabemos que hay una presunción en favor de la veracidad general de las creencias de cualquiera, incluyendo las nuestras. Por lo tanto, resulta vano que alguien exija una seguridad *adicional*, pues ello no haría sino incrementar el conjunto de sus creencias. Todo lo que se requiere es que reconozca que la creencia es verídica por su propia naturaleza.

Se puede apreciar el carácter verídico de la creencia considerando qué es lo que determina la existencia y los contenidos de una creencia. La creencia, como el resto de las llamadas actitudes proposicionales, sobreviene* a hechos de diverso tipo, conductuales, neurofisiológicos, biológicos y físicos. La razón para indicar esto no es el fomento de la reducción definicional o nomológica de los fenómenos psicológicos a algo más básico, y tampoco la sugerencia de prioridades epistemológicas. Se trata más bien de una cuestión de comprensión. Obtenemos cierta clase de penetración en la naturaleza de las actitudes proposicionales cuando las relacionamos sistemáticamente entre sí y con fenómenos de otros niveles. Puesto que las actitudes proposicionales se hallan profundamente entrelazadas, no podemos conocer la naturaleza de una de ellas obteniendo previamente la comprensión de otra. En cuanto intérpretes, nos abrimos camino

* Véase N. del T. en la pág. 59.

en el sistema entero, dependiendo en amplia medida de la pauta de relaciones recíprocas.

Tomemos como ejemplo la interdependencia de creencia y significado. Lo que significa una oración depende en parte de las circunstancias externas que causan que alcance cierto poder de convicción y en parte de las relaciones gramaticales, lógicas o algo menos que eso, que la oración guarda con otras oraciones que se tienen por verdaderas con distintos grados de convicción. Puesto que estas relaciones se traducen directamente en creencias, es fácil ver cómo el significado depende de la creencia. La creencia, sin embargo, depende igualmente del significado, pues el único acceso a la fina estructura e individuación de las creencias lo constituyen las oraciones que los hablantes y sus intérpretes usan para expresar y describir creencias. Por lo tanto, si queremos iluminar la naturaleza del significado y de la creencia, tenemos que partir de algo que no presupone ni el uno ni la otra. La sugerencia de Quine, que en lo esencial voy a seguir, consiste en tomar como básico el *asentimiento inducido*, la relación causal entre asentir a una oración y la causa de dicho asentimiento. Este es un buen lugar para iniciar el proyecto de identificar creencias y significados, puesto que el asentimiento de un hablante a una oración depende tanto de lo que quiere decir con la oración como de lo que cree acerca del mundo. Y, sin embargo, es posible saber que un hablante asiente a una oración sin saber qué significa la oración en sus labios o cuál es la creencia expresada por ella. Igualmente obvio es el hecho de que, una vez que se ha dado una interpretación a una oración a la que se asiente, se ha atribuido con ello una creencia. Si las teorías correctas de la interpretación no son exclusivas (no llevan a interpretaciones correctas exclusivas), lo mismo valdría, desde luego, para las atribuciones de creencias, en cuanto que están ligadas a la aquiescencia con respecto a oraciones particulares.

Un hablante que desea que sus palabras se entiendan no puede engañar sistemáticamente a sus supuestos intérpretes acerca de cuándo asiente a oraciones; esto es, las tiene por verdaderas. Como cuestión de principio, pues, el significado

y, por su conexión con él, también la creencia, están abiertos a la determinación pública. En lo que sigue me beneficiaré de este hecho y adoptaré la posición de un intérprete radical al preguntar por la naturaleza de la creencia. Lo que un intérprete plenamente informado podría aprender acerca de lo que un hablante quiere decir es todo lo que se puede aprender, y lo mismo puede decirse de lo que el hablante cree.[7]

El problema del intérprete es que aquello que se supone que conoce –las causas del asentimiento de un hablante a diversas oraciones– es, según hemos visto, el producto de dos cosas que se supone que no conoce: el significado y la creencia. Si conociese los significados conocería las creencias y si conociese las creencias expresadas por las oraciones asentidas tendría conocimiento de los significados. Pero, ¿cómo puede llegar a conocer ambas cosas a la vez, si cada una depende de la otra?

Las líneas generales de la solución, así como el problema mismo, se deben a Quine. Sin embargo, introduciré algunos cambios en la solución quiniana, al igual que hice en el planteamiento del problema. Los cambios son directamente relevantes para la cuestión del escepticismo epistemológico.

Considero que el objetivo de la interpretación radical (que se asemeja mucho, pero no del todo, a la traducción radical de Quine) consiste en producir una caracterización de la verdad, en el estilo de Tarski, para el lenguaje del hablante, así como una teoría de sus creencias. (La segunda deriva de la primera junto con el conocimiento presupuesto de las oraciones tenidas por verdaderas.) Esto no añade mucho al programa quiniano de traducción, puesto que la traducción del lenguaje del hablante al propio, más una teoría de la verdad para este último, equivalen a una teoría de la verdad para el hablante. Pero el tránsito de la noción sintáctica de

7. Pienso ahora que es esencial, al practicar la interpretación radical, incluir desde el principio los deseos del hablante, de forma que los resortes de la acción y la intención, a saber, la creencia y el deseo, se relacionen con el significado. Pero en el presente discurso no es necesario introducir este factor adicional.

traducción a la noción semántica de verdad pone en primer plano las restricciones formales de una teoría de la verdad y subraya un aspecto de la estrecha relación entre verdad y significado.

El principio de caridad desempeña un papel crucial en el método de Quine y un papel aún más importante en mi propia variante. En uno u otro caso, el principio ordena al intérprete traducir o interpretar de modo tal que algunos de sus propios criterios de verdad se lean en la estructura de oraciones que el hablante considera verdaderas. El propósito del principio es hacer inteligible al hablante, puesto que las desviaciones excesivas respecto de la coherencia y de la corrección no dejan un terreno común desde el cual juzgar el acuerdo o la diferencia. Desde un punto de vista formal, el principio de caridad ayuda a resolver el problema de la interacción del significado y la creencia al restringir los grados de libertad concedidos a la creencia mientras se determina el modo de interpretar las palabras.

Quine ha insistido en que no tenemos más opción que leer nuestra propia lógica en los pensamientos de un hablante; Quine señala esto con respecto al cálculo de enunciados, y yo añadiría otro tanto por lo que respecta a la teoría de cuantificadores de primer orden. Esto conduce directamente a la identificación de las constantes lógicas, así como a la asignación de una forma lógica a todas las oraciones.

Algo semejante a la caridad opera en la interpretación de aquellas oraciones a las que se asiente por causas presentes o ausentes en distintos tiempos y lugares: cuando el intérprete encuentra una oración del hablante a la que éste asiente regularmente bajo condiciones que él reconoce, considera éstas como condiciones de verdad de la oración del hablante. Esto sólo es correcto a grandes rasgos, como veremos dentro de un momento. Las oraciones y predicados que engranan de forma menos directa con acontecimientos fácilmente detectables pueden, según el canon de Quine, interpretarse a voluntad, dadas únicamente las restricciones relativas a las conexiones con oraciones condicionadas directamente al mundo. En este punto yo extendería el princi-

pio de caridad con vistas a favorecer interpretaciones que preserven la verdad todo cuanto sea posible: creo que contribuye a la comprensión mutua, y por tanto a una mejor interpretación, interpretar como verdadero, cuando podamos, lo que el intérprete acepta como tal. En esta cuestión tengo menos elección que Quine, porque yo no veo cómo trazar de salida la línea entre oraciones observacionales y teóricas. Hay distintas razones para ello, pero la más relevante para el asunto que nos ocupa es que la distinción se basa en último término en una consideración epistemológica de un tipo al que he renunciado: las oraciones observacionales se basan directamente en algo semejante a la sensación –pautas de estimulación sensorial– y ésta es una idea que, como he subrayado, conduce al escepticismo. Sin el vínculo directo con la observación o la estimulación, no cabe trazar la distinción entre las oraciones observacionales y las demás sobre fundamentos epistemológicamente significativos. Permanece, sin embargo, la distinción entre oraciones a las que se asiente por causas que varían según circunstancias observables y aquellas a las que un hablante se aferra a través del cambio, y esta distinción ofrece la posibilidad de interpretar las palabras y las oraciones que rebasan las puramente lógicas.

Los detalles no vienen ahora al caso. Lo que debería quedar claro es que, si es correcta la explicación que he dado de las relaciones entre creencia y significado y de su comprensión por parte de un intérprete, entonces la mayoría de las oraciones que un hablante tiene por verdaderas –especialmente aquellas que sostiene con más tenacidad, las más centrales en el sistema de sus creencias– *son* verdaderas, al menos en opinión del intérprete. En efecto, el único, y por tanto irrecusable, método a disposición del intérprete pone automáticamente de acuerdo las creencias del hablante con los criterios de la lógica del intérprete, y con ello acredita al hablante como poseedor de las verdades llanas de la lógica. No hace falta decir que hay grados de coherencia lógica y de otros tipos, y que no es de esperar la coherencia perfecta. Lo que hay que acentuar es únicamente la necesidad metodológica de encontrar la coherencia suficiente.

Tampoco hay, desde el punto de vista del intérprete, ninguna forma en que pueda descubrir que el hablante está ampliamente equivocado acerca del mundo, ya que él interpreta las oraciones tenidas por verdaderas (lo que no debe distinguirse de atribuir creencias) según los eventos y objetos del mundo externo que causan que la oración se tenga por verdadera.

Es fácil pasar por alto lo que considero como el aspecto importante del presente planteamiento, porque éste invierte nuestra forma natural de pensar sobre la comunicación, que se deriva de situaciones en que la comprensión ya ha sido asegurada. Una vez asegurada la comunicación, somos a menudo capaces de saber lo que cree una persona con total independencia de lo que causó que lo creyera. Esto puede llevarnos a la crucial, y sin duda fatal, conclusión de que podemos en general fijar lo que alguien quiere decir con independencia de sus creencias y de lo que las causó. Pero si estoy en lo cierto, no podemos en general identificar primero creencias y significados y luego preguntar por sus causas. La causalidad desempeña un papel indispensable en la determinación del contenido de lo que decimos y creemos. Este es un hecho que podemos vernos llevados a reconocer al adoptar, como lo hemos hecho, el punto de vista del intérprete.

La existencia de un amplio grado de verdad y coherencia en el pensamiento y el habla de un agente constituye un artefacto de la interpretación correcta del habla de una persona por parte del intérprete. Pero se trata de verdad y coherencia según los criterios del intérprete. ¿Por qué no podría suceder que hablante e intérprete se entendieran entre sí sobre la base de creencias compartidas pero erróneas? Esto puede ocurrir y sin duda ocurre a menudo, pero no puede constituir la regla. En efecto, imaginemos por un momento a un intérprete omnisciente acerca del mundo y de lo que causa y causaría el asentimiento de un hablante a cualquier oración de su (potencialmente ilimitado) repertorio. El intérprete omnisciente, utilizando el mismo método que el intérprete falible, hallará al falible hablante ampliamente co-

herente y correcto, coherente y correcto según sus propios criterios, desde luego, pero, puesto que éstos son objetivamente correctos, el hablante falible resulta ser ampliamente correcto y coherente de acuerdo con criterios objetivos. Podemos también, si queremos, dejar que el intérprete omnisciente dirija su atención al intérprete falible del hablante falible. Resulta entonces que el intérprete falible puede estar equivocado respecto de algunas cosas, pero no en general, de modo que no puede compartir el error universal con el agente al cual está interpretando. Una vez que aceptamos el método general de interpretación que he esbozado, se hace imposible sostener correctamente que cualquiera podría estar equivocado en general acerca de cómo son las cosas.

Hay, como advertí más arriba, una diferencia crucial entre el método de interpretación radical que aquí recomiendo y el método quiniano de la traducción radical. La diferencia reside en la naturaleza de la elección de las causas que gobiernan la interpretación. Quine hace depender la interpretación de patrones de estimulación sensorial, mientras que yo la hago depender de los eventos y objetos externos acerca de los cuales versa la oración de acuerdo con la interpretación que recibe. Así, la noción quiniana de significado se encuentra ligada a criterios sensoriales, que en su opinión pueden ser tratados también como evidencia. Esto lleva a Quine a conceder significación epistémica a la distinción entre las oraciones de observación y las demás, puesto que se supone que las primeras, al estar directamente condicionadas por los sentidos, poseen una especie de justificación extralingüística. Esta es la concepción contra la que argüí en la primera parte de este artículo, subrayando que las estimulaciones sensoriales son de hecho parte de la cadena causal que lleva a la creencia, pero no pueden, sin confusión, considerarse como evidencia o fuente de justificación de las creencias que ellas estimulan.

Lo que se opone al escepticismo global de los sentidos es, en mi opinión, el hecho de que, en los casos más simples y metodológicamente más básicos, hemos de considerar los objetos de una creencia como las causas de esa creencia. Y

lo que nosotros, en cuanto intérpretes, hemos de considerar que son es lo que de hecho son. La comunicación empieza allí donde convergen las causas: tu emisión significa lo mismo que la mía si la creencia en su verdad es causada sistemáticamente por los mismos eventos y objetos.[8]

Las dificultades a las que se enfrenta esta concepción son obvias, pero creo que pueden ser superadas. El método se aplica directamente, en el mejor de los casos, sólo a oraciones ocasionales, el asentimiento a las cuales está causado sistemáticamente por cambios ordinarios en el mundo. Otras oraciones se interpretan por su relación de condicionamiento con las oraciones ocasionales y por la aparición en ellas de palabras que aparecen también en las oraciones ocasionales. Entre estas últimas, algunas variarán en el grado de creencia que exigen, no sólo ante el cambio en el entorno, sino también ante el cambio en el grado de creencia concedido a oraciones relacionadas. Sobre esta base, es posible desarrollar criterios para distinguir grados en el carácter observacional con fundamentos internos, sin apelar al concepto de una base para la creencia exterior al círculo de éstas.

Relacionado con estos problemas, y más fácil aún de comprender, hallamos el problema del error, pues incluso en los casos más simples es claro que la misma causa (la rápida carrera de un conejo) puede engendrar creencias diferentes en el hablante y en el observador, propiciando así el asentimiento a oraciones que no pueden tener la misma interpretación. Es este hecho, sin duda, el que llevó a Quine a pasar de los conejos a los patrones de estimulación como clave de la interpretación. En cuanto simple cuestión de estadística, no estoy seguro del grado en que una aproximación es mejor que la otra. ¿Es la frecuencia relativa con que patrones de

8. Es claro que la teoría causal del significado tiene poco en común con las teorías causales de la referencia de Kripke y Putnam. Estas últimas atienden a relaciones causales entre nombre y objetos, relaciones que los hablantes pueden muy bien ignorar. Con ello, la posibilidad del error sistemático aumenta. Mi teoría causal procede a la inversa al conectar la causa de una creencia con su objeto.

estimulación idénticos suscitarán el asentimiento a «gavagai» y «conejo» mayor que la frecuencia relativa con que un conejo provocará esas mismas dos respuestas en el hablante y en el intérprete? No es ésta una cuestión que pueda someterse a prueba de modo fácil y convincente. Pero supongamos que los resultados imaginados hablan en favor del método de Quine. Entonces tendré que decir, como tendría que hacerlo en cualquier caso, que el problema del error no puede afrontarse oración por oración, ni siquiera en el nivel más simple. Lo mejor que podemos hacer es dar cuenta del error de forma holista, es decir, practicando la interpretación de modo que el agente resulte tan inteligible como sea posible dadas sus acciones, sus emisiones y su lugar en el mundo. Acerca de algunas cosas le hallaremos equivocado, siendo éste el precio necesario de descubrir que está en lo cierto en otros aspectos. A modo de vaga aproximación, descubrir que está en lo cierto significa identificar las causas de sus creencias con los objetos de las mismas, concediendo un peso especial a los casos más simples y admitiendo el error donde pueda explicarse mejor.

Supongamos que tengo razón al decir que un intérprete ha de ejercer su tarea de modo tal que el hablante o agente resulte estar fundamentalmente en lo cierto acerca del mundo. ¿Cómo puede esto ayudar a la persona misma que se pregunta qué razones tiene para pensar que sus creencias son mayoritariamente verdaderas? ¿Cómo puede llegar a conocer esas relaciones causales entre el mundo real y sus propias creencias que llevan al intérprete a comprenderle como alguien que pisa suelo firme?

La respuesta está contenida en la pregunta. Para poder dudar o preguntarse sobre el origen de sus creencias, un agente debe saber qué es la creencia. Esto lleva consigo el concepto de verdad objetiva, pues la noción de creencia es la de un estado que puede o no concordar con la realidad. Pero las creencias se identifican también, directa o indirectamente, por sus causas. Lo que un intérprete omnisciente sabe, un intérprete falible lo vislumbra de modo suficiente si entiende a un hablante, y ésta es precisamente la complicada verdad

causal que hace de nosotros los sujetos de creencias que somos y fija el contenido de las mismas. El agente no tiene más que reflexionar sobre la naturaleza de la creencia para darse cuenta de que la mayoría de sus creencias básicas son verdaderas, y entre sus creencias, las más propensas a la verdad son aquellas que sostiene con mayor firmeza y que guardan cohesión con el cuerpo principal de sus otras creencias. La pregunta: ¿cómo sé que mis creencias son en general verdaderas? encierra en sí misma la respuesta, a saber: sencillamente porque las creencias son en general verdaderas por naturaleza. Parafraseada o expandida, la pregunta se convierte en la siguiente: ¿cómo puedo determinar si mis creencias, que por naturaleza son en general verdaderas, son en general verdaderas?

Todas las creencias están justificadas en el siguiente sentido: están apoyadas por muchas otras creencias (pues en otro caso no serían las creencias que son) y gozan de una presunción de verdad. La presunción se incrementa cuanto más amplio e importante sea el cuerpo de creencias con el que la creencia en cuestión es coherente, y al no haber cosa tal como una creencia aislada, no hay creencia alguna sin una presunción en su favor. A este respecto, intérprete y sujeto interpretado difieren. Desde el punto de vista del intérprete, la metodología impone una presunción general de verdad para el cuerpo de creencias como un todo, pero el intérprete no necesita suponer que cada creencia particular de otra persona es verdadera. La presunción general aplicada a los otros no hace que estén globalmente en lo cierto, como he subrayado, pero proporciona el fondo sobre el cual tiene lugar la acusación de error. Pero desde el punto de vista aventajado de cada persona, ha de haber una presunción gradual en favor de cada una de sus propias creencias.

No podemos, ¡ay!, extraer la pintoresca y placentera conclusión según la cual todas nuestras creencias verdaderas constituyen conocimiento, pues aunque todas las creencias de un sujeto estén hasta cierto punto justificadas para él, algunas pueden no estarlo lo suficiente, o del modo apropiado, para constituir conocimiento. La presunción general en fa-

vor de la verdad de la creencia sirve para rescatarnos de una forma común de escepticismo al mostrar por qué es imposible que todas nuestras creencias sean falsas en su conjunto. Esto deja casi intacta la tarea de especificar las condiciones del conocimiento. No me he ocupado de los cánones del apoyo evidencial (si hay cosa tal), sino de mostrar que todo lo que cuenta como evidencia o justificación de una creencia ha de proceder de la totalidad misma de creencias a la que aquélla pertenece.

ENGAÑO Y DIVISIÓN

Normalmente, el autoengaño no constituye un gran problema para el que lo practica, sino que, por el contrario, tiende a aligerarle, en parte, de la pesada carga de pensamientos dolorosos cuyas causas se hallan más allá de su control. Pero el autoengaño *es* un problema para la psicología filosófica, pues al reflexionar sobre él, como también sobre otras formas de irracionalidad, nos sentimos tentados por ideas opuestas. Por una parte, no es clara la existencia de un caso genuino de irracionalidad a menos que sea posible identificar una inconsecuencia en el pensamiento del agente, algo que sea inconsecuente con las propias normas de éste. Por otra parte, cuando tratamos de explicar con cierto detalle cómo puede el agente haber llegado a ese estado, nos descubrimos inventando algún tipo de racionalización que podamos atribuir al autoengañador y disolviendo con ello la inconsecuencia que le imputábamos. El autoengaño resulta notoriamente embarazoso, puesto que, en algunas de sus manifestaciones, parece exigir de nosotros que sostengamos, no sólo que alguien cree a la vez cierta proposición y su negación, sino también que una de esas creencias sirve de apoyo a la otra.

Consideremos estos cuatro enunciados:

1. D cree que es calvo.
2. D cree que no es calvo.
3. D cree que (es calvo y no es calvo).
4. D no cree que sea calvo.

En el tipo de autoengaño del que voy a ocuparme, una creencia como la expresada en (1) es una condición causal de una creencia que la contradice, como sucede con (2). Resulta tentador, desde luego, suponer que (2) implica (4),

pero si lo hacemos entraremos en contradicción con nosotros mismos. Al intentar ofrecer una descripción coherente de la incoherente estructura mental de D, podríamos decir que, puesto que D cree que no es calvo y cree que es calvo (lo cual es la razón de que (4) sea falsa), ha de creer también que es calvo y que no lo es, como (3) indica. Pero también deberíamos resistirnos a dar este paso, pues nada de lo que una persona pudiera decir o hacer constituiría un fundamento lo bastante sólido para atribuirle una creencia limpia y obviamente contradictoria, del mismo modo que, dada una oración sincera y literalmente aseverada, nada podría sustentar una interpretación de la misma según la cual dicha oración sería verdadera si, y sólo si, D fuese a la vez calvo y no calvo, por más que las palabras emitidas pudieran haber sido «D es calvo y no lo es». Es posible creer cada miembro de un par de enunciados sin creer la conjunción de ambos.

La tarea que se nos presenta consiste en explicar cómo puede alguien tener creencias como (1) y (2) sin combinarlas en un todo, aun cuando crea (2) *porque* cree (1).

El problema puede generalizarse en los términos que siguen.

Probablemente ocurre muy raras veces que una persona tenga la *certeza* de que cierta proposición es verdadera y tenga también la certeza de que su negación lo es. Más común sería la situación en que la suma de la evidencia de que dispone el agente apunta hacia la verdad de una proposición, lo que inclina a éste a creerla (le lleva a considerar su verdad más probable que su falsedad). Esta inclinación (alta probabilidad subjetiva) actúa casualmente sobre él, en formas de las que aún hemos de tratar, llevándole a buscar, apoyar o acentuar la evidencia en favor de la falsedad de dicha proposición, o bien a desatender la evidencia en favor de su verdad. El agente se halla entonces más inclinado a creer la negación de la proposición original que a lo contrario, a pesar de que la totalidad de la evidencia a la que tiene acceso no sustenta esa actitud. (La frase «inclinado a creer» resulta demasiado anodina para caracterizar algunos de los estados mentales que pretendo que describa; tal vez se pueda decir

que el agente cree que la proposición es falsa, pero no está completamente seguro de ello.)

Esta caracterización del autoengaño lo asimila considerablemente a la debilidad de la voluntad. La debilidad de la voluntad consiste en actuar intencionalmente (o en formarse la intención de hacerlo) sin tomar como base todas las razones cuya relevancia se reconoce. Una acción de este tipo se produce en un contexto de conflicto; el agente acrático tiene razones, que él considera tales, tanto a favor como en contra de cierto curso de acción; sobre la base de todas esas razones, juzga que un determinado curso de acción es el mejor y, sin embargo, opta por otro distinto; con ello, ha actuado «en contra de su mejor juicio».[1] En cierto sentido, es fácil decir por qué actuó como lo hizo, ya que tenía razones en favor de esa acción. Pero esta explicación deja de lado el elemento de irracionalidad; no explica por qué el agente fue contra su mejor juicio.

Un acto que revele debilidad de la voluntad peca contra el principio normativo según el cual no deberíamos llevar a cabo intencionalmente una acción cuando juzgamos, sobre la base de todas las consideraciones que creemos tener a nuestro alcance, que un curso de acción alternativo y accesible sería mejor.[2] Este principio, que yo denomino Principio de Continencia, prescribe un tipo fundamental de coherencia en el pensamiento, la intención, la evaluación y la acción. Un agente que actúa de acuerdo con este principio posee la virtud de la continencia. No es claro que una persona pueda dejar de reconocer la norma de continencia; volveré en breve sobre esta cuestión. En cualquier caso, es claro que hay muchas personas que aceptan la norma pero de vez en cuando dejan de actuar de acuerdo con ella. En tales casos, los agen-

1. El problema de la debilidad de la voluntad lo he discutido en «How is weakness of the will possible?», en *Essays on Actions and Events* (Clarendon Press, Oxford, 1980).

2. ¿Qué consideraciones son «accesibles» al agente? ¿Incluyen sólo la información que posee o también la que podría (¿si lo supiera?) conseguir? En este ensayo tendrá que dejar abiertas la mayoría de estas cuestiones.

tes no sólo no acomodan sus acciones a sus propios principios, sino que tampoco razonan como creen que debieran hacerlo, pues su actuación intencional muestra que han concedido al acto que llevan a cabo un valor superior al que deberían concederle a tenor de sus propios principios y razones.

A menudo, el autoengaño y la debilidad de la voluntad se refuerzan recíprocamente, pero no son lo mismo, como lo muestra ya el hecho de que el resultado de la debilidad de la voluntad es una intención o una acción intencional, mientras que el resultado del autoengaño es una creencia. La primera consiste en una actitud evaluativa a la que se ha llegado de forma defectuosa; la segunda consiste en una actitud cognitiva, asimismo alcanzada defectuosamente.

La debilidad de la voluntad es análoga a cierto error cognitivo que voy a denominar *debilidad de la justificación*. La debilidad de la justificación sólo puede darse cuando una persona tiene evidencia tanto a favor como en contra de una hipótesis. La persona en cuestión juzga que, en relación con toda la evidencia de que dispone, la hipótesis es más probable que su negación, y, sin embargo, no acepta la hipótesis (o la fuerza de su creencia en la hipótesis es menor que la de su creencia en la negación de la misma). El principio normativo contra el que dicha persona ha pecado es aquel que Hempel y Carnap denominaron *el requisito de evidencia global en el razonamiento inductivo*: cuando estamos en el trance de decidir entre una serie de hipótesis mutuamente excluyentes, dicha exigencia nos ordena dar crédito a la hipótesis que se halle mejor sustentada por toda la evidencia relevante disponible.[3] La debilidad de la justificación tiene, obviamente, la misma estructura lógica (o, mejor, ilógica) que la debilidad de la voluntad. La primera involucra una creencia irracional ante una evidencia conflictiva; la segunda, una intención (y quizá también acción) irracional en presencia de valores enfrentados. La existencia de un conflicto es una condición necesaria de

3. Véase Carl Hempel, *Aspects of Scientific Explanation* (The Free Press, Nueva York, 1965), págs. 397-403.

ambas formas de irracionalidad y puede ser en ciertos casos una causa del lapsus; pero no hay nada en los conflictos de ese tipo que exija o revele necesariamente un fracaso de la razón.

La debilidad de la voluntad no consiste simplemente en pasar por alto cierta evidencia que se tiene (aunque el descuido «significativo» pueda ser ya otra cuestión, que además es relevante para el autoengaño), ni tampoco es no advertir el hecho de que ciertas cosas que se saben o se creen constituyen evidencia en favor o en contra de una hipótesis. Tomada en su tenor literal, la historia que sigue no muestra que yo haya sufrido un autoengaño. Un compañero y yo estábamos en el Parque Nacional de Amboseli, en Kenia, acechando a los animales. No habiendo encontrado un guepardo por nuestra cuenta, contratamos los servicios de un guía oficial durante una mañana. Una vez que el guía hubo regresado a las dependencias centrales del parque, hice a mi compañero el siguiente comentario: «Lástima que no encontrásemos un guepardo; es el único animal de gran tamaño que nos hemos perdido. Por cierto, ¿no tenía ese guía una voz extraña, muy aguda? ¿Y crees que llamarse "Helen" será algo normal para un hombre en estos lugares? Supongo que será el uniforme oficial, pero parece raro que un guía lleve falda». Mi compañero dijo: «No era un, sino una guía». El supuesto del que yo partía era estereotipado y estúpido, pero a menos que tuviese en cuenta la hipótesis de que el guía era una mujer y la rechazase a pesar de la evidencia, no se trataría de un simple caso de autoengaño. Tal vez otros piensen en explicaciones más profundas de mi terca suposición de que nuestra guía era un hombre.

Supongamos (sea cual fuere la verdad) que consideré la posibilidad de que nuestro guía fuese una mujer y que rechacé dicha hipótesis a pesar de la abrumadora evidencia que tenía para aceptarla. ¿Mostraría esto necesariamente que fui irracional? Es difícil responder a menos que seamos capaces de establecer, con respecto a ciertas formas de razonamiento, una distinción estricta entre carecer de ellas y no aplicarlas. ¿No podemos suponer, por ejemplo, que, aun te-

niendo evidencia, no advertí de qué era evidencia? Esto *puede* suceder, sin duda. La verosimilitud de una determinada explicación depende de las circunstancias exactas del caso. Hemos de insistir, pues, en que no hay error de razonamiento inductivo a menos que la evidencia se *tenga* por tal. ¿Y no podría suceder que, aunque la evidencia fuese tenida por tal, no se advirtiese el hecho de que la totalidad de la misma hacía abrumadoramente probable una determinada hipótesis? Esto podría también ocurrir, por muy improbable que pueda resultar en ciertos casos particulares. Hay un sinnúmero de preguntas adicionales que la tortuga puede hacer a Aquiles siguiendo esta misma línea (puesto que las brechas que el razonamiento desafortunado puede dejar abiertas son tantas como las que el razonamiento feliz debe cerrar). Así, pues, sin pretensiones de especificar todas las condiciones que nos sitúan ante un caso absolutamente claro de debilidad de la justificación, quisiera formular una nueva pregunta: ¿es preciso que alguien acepte el requisito de evidencia global en el razonamiento inductivo para que el hecho de que no actúe de acuerdo con él constituya una prueba de irracionalidad? En esta pregunta se hallan involucradas varias cuestiones.

El hecho de que una persona acepte el requisito no nos autoriza a exigirle que siempre razone o piense de acuerdo con él; de otro modo sería imposible que se diese una auténtica incoherencia, es decir, una incoherencia *interna*, de este tipo. Por otra parte, no tendría sentido suponer que una persona pueda aceptar el principio en cuestión y no pensar nunca, o muy raras veces, de acuerdo con él; aceptar un principio semejante consiste, al menos en parte, en manifestar dicho principio en el pensamiento y el razonamiento. Así, pues, si admitimos, como creo que hemos de hacerlo, que el hecho de que una persona «acepte» o tenga un principio como el requisito de evidencia global consiste en que su forma de pensar se acomode a éste, tiene sentido entonces imaginar que una persona posea dicho principio sin tener conciencia de él o sin ser capaz de formularlo. Pero tal vez queramos añadir a este obvio enunciado condicional («una persona acepta el requisito de evidencia global en el razona-

miento inductivo sólo si tiene una disposición a ajustarse a él en las circunstancias apropiadas») alguna otra condición o condiciones, como por ejemplo que la conformidad es más probable cuando hay más tiempo para pensar, menor carga emocional asociada a la conclusión o cuando se disfruta de una asistencia socrática explícita.

En una persona que acepta el requisito de evidencia global, la debilidad de la justificación es, como vemos, una cuestión de desviación respecto de una costumbre o hábito. En un caso semejante, la debilidad de la justificación revela una incoherencia y resulta claramente irracional. ¿Qué sucede, sin embargo, si alguien no acepta el requisito? En este punto parece surgir un interrogante muy general acerca de la racionalidad: ¿Qué patrones hemos de considerar como aquellos que fijan la norma? ¿Tendríamos que decir de alguien cuyo pensamiento no satisface el requisito de evidencia global que tal vez sea irracional según los patrones de otra persona, pero no según los suyos propios? ¿O quizá deberíamos hacer de la incoherencia interna una condición necesaria de la irracionalidad? No es fácil ver cómo podrían separarse ambas cuestiones, ya que la coherencia interna es en sí misma una norma fundamental.

Cuando se trata de normas fundamentales, no es posible establecer una clara separación entre una y otra cuestión, puesto que, en general, cuanto más llamativo parezca ser un caso de incoherencia interna para un observador ajeno, tanto más inútil le resultará a éste, en su intento de explicar la aparente aberración, la supuesta distinción entre sus propias normas y las de la persona observada. Las diferencias relativamente pequeñas toman forma y son explicadas sobre un fondo de normas compartidas, pero por lo que se refiere a desviaciones importantes respecto de patrones de racionalidad fundamentales, es más verosímil que se encuentren en el ojo del intérprete que en la mente del sujeto interpretado. La razón de ello no hay que buscarla muy lejos. Una persona entiende las creencias, etc., de otra sólo en la medida en que pueda asignar sus propias proposiciones (u oraciones) a las diversas actitudes de aquélla. Puesto que una creencia no

niendo evidencia, no advertí de qué era evidencia? Esto *puede* suceder, sin duda. La verosimilitud de una determinada explicación depende de las circunstancias exactas del caso. Hemos de insistir, pues, en que no hay error de razonamiento inductivo a menos que la evidencia se *tenga* por tal. ¿Y no podría suceder que, aunque la evidencia fuese tenida por tal, no se advirtiese el hecho de que la totalidad de la misma hacía abrumadoramente probable una determinada hipótesis? Esto podría también ocurrir, por muy improbable que pueda resultar en ciertos casos particulares. Hay un sinnúmero de preguntas adicionales que la tortuga puede hacer a Aquiles siguiendo esta misma línea (puesto que las brechas que el razonamiento desafortunado puede dejar abiertas son tantas como las que el razonamiento feliz debe cerrar). Así, pues, sin pretensiones de especificar todas las condiciones que nos sitúan ante un caso absolutamente claro de debilidad de la justificación, quisiera formular una nueva pregunta: ¿es preciso que alguien acepte el requisito de evidencia global en el razonamiento inductivo para que el hecho de que no actúe de acuerdo con él constituya una prueba de irracionalidad? En esta pregunta se hallan involucradas varias cuestiones.

El hecho de que una persona acepte el requisito no nos autoriza a exigirle que siempre razone o piense de acuerdo con él; de otro modo sería imposible que se diese una auténtica incoherencia, es decir, una incoherencia *interna*, de este tipo. Por otra parte, no tendría sentido suponer que una persona pueda aceptar el principio en cuestión y no pensar nunca, o muy raras veces, de acuerdo con él; aceptar un principio semejante consiste, al menos en parte, en manifestar dicho principio en el pensamiento y el razonamiento. Así, pues, si admitimos, como creo que hemos de hacerlo, que el hecho de que una persona «acepte» o tenga un principio como el requisito de evidencia global consiste en que su forma de pensar se acomode a éste, tiene sentido entonces imaginar que una persona posea dicho principio sin tener conciencia de él o sin ser capaz de formularlo. Pero tal vez queramos añadir a este obvio enunciado condicional («una persona acepta el requisito de evidencia global en el razona-

miento inductivo sólo si tiene una disposición a ajustarse a él en las circunstancias apropiadas») alguna otra condición o condiciones, como por ejemplo que la conformidad es más probable cuando hay más tiempo para pensar, menor carga emocional asociada a la conclusión o cuando se disfruta de una asistencia socrática explícita.

En una persona que acepta el requisito de evidencia global, la debilidad de la justificación es, como vemos, una cuestión de desviación respecto de una costumbre o hábito. En un caso semejante, la debilidad de la justificación revela una incoherencia y resulta claramente irracional. ¿Qué sucede, sin embargo, si alguien no acepta el requisito? En este punto parece surgir un interrogante muy general acerca de la racionalidad: ¿Qué patrones hemos de considerar como aquellos que fijan la norma? ¿Tendríamos que decir de alguien cuyo pensamiento no satisface el requisito de evidencia global que tal vez sea irracional según los patrones de otra persona, pero no según los suyos propios? ¿O quizá deberíamos hacer de la incoherencia interna una condición necesaria de la irracionalidad? No es fácil ver cómo podrían separarse ambas cuestiones, ya que la coherencia interna es en sí misma una norma fundamental.

Cuando se trata de normas fundamentales, no es posible establecer una clara separación entre una y otra cuestión, puesto que, en general, cuanto más llamativo parezca ser un caso de incoherencia interna para un observador ajeno, tanto más inútil le resultará a éste, en su intento de explicar la aparente aberración, la supuesta distinción entre sus propias normas y las de la persona observada. Las diferencias relativamente pequeñas toman forma y son explicadas sobre un fondo de normas compartidas, pero por lo que se refiere a desviaciones importantes respecto de patrones de racionalidad fundamentales, es más verosímil que se encuentren en el ojo del intérprete que en la mente del sujeto interpretado. La razón de ello no hay que buscarla muy lejos. Una persona entiende las creencias, etc., de otra sólo en la medida en que pueda asignar sus propias proposiciones (u oraciones) a las diversas actitudes de aquélla. Puesto que una creencia no

puede mantener su identidad al perder sus relaciones con otras creencias, no es posible que la misma proposición sirva para interpretar actitudes particulares de dos personas distintas y guarde al mismo tiempo, con las demás actitudes de una de ellas, relaciones muy diferentes de las que guarda con las de la otra. De ello se sigue que, a menos que un intérprete pueda reproducir en otra persona los contornos principales de su propia pauta de actitudes, no le será posible identificar inteligiblemente ninguna de las actitudes de aquélla. La comprensión de algunas desviaciones respecto de las propias normas en otros sujetos es posible únicamente debido a la gran multitud y complejidad de las formas en que se ramifican las relaciones de una actitud con las demás.

Podemos ver ahora lo engañoso de la cuestión que nos planteábamos un poco más arriba, a saber, si la irracionabilidad en un agente requiere una incoherencia *interna*, una desviación con respecto a las propias normas de esa persona, ya que, cuando se trata de normas básicas, éstas representan elementos constitutivos en la identificación de actitudes, con lo cual la cuestión de si alguien las «acepta» no puede siquiera llegar a plantearse. Esto no se aplica sólo a las incoherencias claramente lógicas, sino también a la debilidad de la voluntad (como ya señaló Aristóteles), a la debilidad de la justificación y al autoengaño.

Todavía he de decir en qué consiste el autoengaño, pero ahora ya estoy en disposición de indicar algunas cosas acerca de él. Es claro que el autoengaño incluye la debilidad de la justificación, pues el sujeto no aceptaría la proposición con respecto a la cual se autoengaña si fuese liberado de su error; tiene, en efecto, mejores razones para aceptar la negación de dicha proposición. Asimismo, como sucede en la debilidad de la justificación, la víctima del autoengaño sabe que tiene mejores razones para aceptar la negación de la proposición que de hecho acepta, al menos en el sentido siguiente: se da cuenta de que, de conformidad con ciertas otras cosas que sabe o acepta como evidencia, es más probable que sea verdadera la negación que la proposición aceptada por él; sin embargo, sobre la base de sólo una parte de lo

que considera como la evidencia relevante, acepta la proposición.

Es precisamente en este punto donde el autoengaño llega más lejos que la debilidad de la justificación, pues la persona que se autoengaña ha de poseer una *razón* para su debilidad de justificación y, además, tiene que haber participado en la generación de esta última. La debilidad de la justificación tiene siempre una *causa* (todas las cosas la tienen), pero en el caso del autoengaño la debilidad de la justificación resulta ser autoinducida (uno mismo la *produjo*). No forma parte del análisis de la debilidad de la justificación o de la debilidad de la voluntad el hecho de que la desviación con respecto a los patrones del agente tenga un motivo (aunque, sin duda, a menudo lo tiene); en cambio, la existencia de un motivo es parte esencial del análisis del autoengaño. Por esta razón resulta instructivo examinar otro fenómeno que guarda ciertas semejanzas con el autoengaño: me refiero al pensamiento desiderativo.

En una elucidación inicial, el pensamiento desiderativo consiste en creer algo debido al deseo que uno tiene de que sea verdad. Esto no es irracional en sí mismo, ya que en general no somos responsables de las causas de nuestros pensamientos. Pero el pensamiento desiderativo es a menudo irracional, por ejemplo cuando sabemos por qué tenemos la creencia y sabemos también que no la tendríamos si no fuera por el deseo.

Es frecuente la opinión de que el pensamiento desiderativo involucra algo más que lo indicado en esa elucidación inicial. Si alguien desea que cierta proposición sea verdadera, es natural suponer que gozará más creyendo que lo es que no creyéndolo. Por lo tanto, una persona tal tiene una razón (en cierto sentido) para creer la proposición. Si esta persona actuase intencionalmente con vistas a promover esa creencia, ¿sería esto irracional? En este punto hemos de hacer una distinción obvia entre tener una razón para creer cierta proposición y tener una evidencia a cuya luz es razonable considerar verdadera la proposición. (Oraciones de la forma «Carlos tiene una razón para creer que *P*» son ambiguas con respecto a esta distinción.) Una razón del primer tipo es eva-

luativa: ofrece un motivo para actuar de modo favorable a la posesión de la creencia. Una razón del segundo tipo es cognitiva: consiste en tener evidencia de la verdad de una proposición. El pensamiento desiderativo no exige una razón de uno u otro tipo, pero, como ya subrayamos, el deseo de que *P* sea el caso (por ejemplo, que alguien me ame) puede engendrar fácilmente el deseo de creer en *P*, y este deseo a su vez puede inspirar pensamientos y acciones que acentúen o que tengan como resultado la obtención de razones del segundo tipo. ¿Hay algo necesariamente irracional en esta secuencia? Una acción intencional que tienda a hacernos felices o a aliviar nuestras penas no es irracional en sí misma ni se convierte en tal si los medios empleados incluyen el intento de disponer las cosas con vistas a tener cierta creencia. En algunos casos, puede ser inmoral hacer esto con otra persona, especialmente si tenemos razones para pensar que la creencia que va a inculcar es falsa, pero no es necesariamente erróneo ni es, ciertamente, irracional. Esto mismo se aplica, en mi opinión, a las creencias autoinducidas; aquello que no es necesariamente irracional cuando se le hace a otra persona sigue sin serlo cuando el objeto es el propio yo futuro.

¿Es necesariamente irracional una creencia generada deliberadamente del modo descrito? Claramente lo es si el sujeto sigue pensando que la evidencia en contra de la creencia es mejor que la evidencia en favor de la misma, pues entonces estamos ante un caso de debilidad de la justificación. Pero si el sujeto ha olvidado la evidencia que originariamente le llevó a rechazar la creencia que ahora abriga, o si la nueva evidencia parece ahora lo bastante buena como para compensar la antigua, el nuevo estado mental no es irracional. Cuando el pensamiento desiderativo tiene éxito, podríamos decir, no hay ningún momento en que el sujeto *tenga que* ser irracional.[4]

4. En «Paradoxes of irrationality», incluido en *Philosophical Essays on Freud*, Richard Wollheim y James Hopkins, comps. (Cambridge University Press, Cambridge, 1982), supuse que, en el pensamiento desiderativo, el deseo producía la creencia sin aportar evidencia en favor de ésta. En semejante caso, la creencia es irracional, desde luego.

Tal vez merezca la pena indicar ahora que tanto el autoengaño como el pensamiento desiderativo pueden ser benignos algunas veces. No resulta sorprendente, ni es tampoco malo, en conjunto, que las personas tengan de sus amigos y familiares una opinión mejor que la que quedaría justificada por un examen lúcido de la evidencia. El aprendizaje se ve más a menudo favorecido que perjudicado por aquellos padres y maestros que sobrevaloran la inteligencia de sus educandos. Con frecuencia, las esposas mantienen la estabilidad familiar ignorando o pasando por alto la mancha de carmín en el cuello de la camisa. Todos estos pueden ser casos de autoengaño caritativo ayudado por el pensamiento desiderativo.

No todo pensamiento desiderativo es autoengaño, pues este último, a diferencia del primero, exige la intervención del agente. Uno y otro se parecen, sin embargo, en que en ambos ha de actuar un elemento motivacional o evaluativo, y en este aspecto difieren de la debilidad de la justificación, en la cual el defecto determinante es cognitivo, sea cual fuere su causa. Esto sugiere que, aun cuando el pensamiento desiderativo pueda ser más simple que el autoengaño, es siempre un ingrediente de éste. Sin duda lo es con frecuencia, pero parece haber excepciones. En el pensamiento desiderativo la creencia toma la dirección del afecto positivo, nunca del negativo; la creencia causada siempre resulta bienvenida. En el autoengaño no sucede lo mismo. El pensamiento alimentado por el autoengaño puede ser doloroso. Una persona movida por los celos puede hallar por doquier «evidencia» que confirma sus peores sospechas; el que ansía la vida privada puede creer que ve un espía detrás de cada cortina. Si un pesimista es alguien que adopta una visión más sombría de las cosas que la justificada por la evidencia de que dispone, todo pesimista cae en cierta medida en el autoengaño de creer lo que desearía que no fuese cierto.

Estas observaciones se limitan a aludir a la naturaleza de la distancia que puede separar el autoengaño y el pensamiento desiderativo. No se trata sólo de que el autoengaño, a diferencia del pensamiento desiderativo, requiera que el

agente *haga* algo con vistas a modificar sus propias opiniones, sino que hay también una diferencia en el modo en que el contenido del elemento afectivo se relaciona con la creencia que produce. En el caso del pensamiento desiderativo, lo que el sujeto llega a creer ha de ser precisamente lo que a él le gustaría que fuese verdad. En cambio, aunque el sujeto del autoengaño pueda estar motivado por un deseo de creer lo que a él le gustaría que fuese cierto, hay sin embargo muchas otras posibilidades. De hecho, es difícil indicar cuál ha de ser la relación entre el motivo de alguien que se engaña a sí mismo y la alteración específica en sus creencias que produce en sí mismo. La relación no es accidental, desde luego; hacer algo intencionalmente con la consecuencia de que el sujeto de la acción resulte engañado no constituye, sin más, autoengaño, pues de otro modo una persona se autoengañaría si leyera y creyese una información falsa en un periódico. El «engaño» ha de ser objeto de la intención del que se autoengaña.

Hasta este punto, al menos, el autoengaño es semejante a la mentira; hay una conducta intencional que tiene como objeto generar una creencia no compartida por el agente en el momento de iniciar esa conducta. La sugerencia implícita en esta comparación es que, mientras que el mentiroso trata de engañar a otro, el que se autoengaña trata de engañarse a sí mismo. La sugerencia no está totalmente desencaminada. Yo me engaño a mí mismo sobre mi grado de calvicie eligiendo aquellas perspectivas e iluminación que favorecen una apariencia hirsuta; un adulador mentiroso podría tratar de obtener el mismo efecto diciéndome que en realidad no soy tan calvo. Pero hay importantes diferencias entre ambos casos. Aunque el mentiroso pueda pretender que su oyente crea lo que dice, esta intención no es esencial al concepto de mentira; un mentiroso que tiene a su oyente por una persona retorcida puede decir lo contrario de lo que intenta que el otro crea. Ni siquiera es necesario que el mentiroso pretenda hacer creer a su víctima que él mismo cree lo que dice. Las únicas intenciones que un mentiroso ha de tener, en mi opinión, son las siguientes: 1) ha de tener la intención de ofre-

cer una imagen falsa de sus auténticas creencias (por ejemplo, en el caso más típico, aseverando aquello que no cree), y 2) ha de tener la intención de ocultar a sus oyentes esa primera intención (aunque no necesariamente lo que de hecho cree). Así, pues, la mentira involucra un tipo muy especial de engaño que afecta a la sinceridad en la representación de las propias creencias. No parece posible que esta forma precisa de engaño pueda practicarse con uno mismo, ya que exigiría hacer algo con la intención de que esa misma intención no sea reconocida por el propio sujeto que la concibe.[5]

En determinado aspecto, pues, el autoengaño no es tan difícil de explicar como lo sería el mentirse a sí mismo, ya que esto último implicaría la existencia de una intención de autoanulación, mientras que el autoengaño se limita a enfrentar intención y deseo a creencia, y creencia a creencia. Aun así, tampoco resulta fácil de entender. Antes de tratar de describir con mayor detalle y plausibilidad el estado mental del agente autoengañado, voy a resumir lo expuesto hasta ahora en cuanto atañe a la naturaleza del autoengaño.

Un agente *A* se autoengaña con respecto a una proposición *P* bajo las siguientes condiciones: *A* posee evidencia sobre la base de la cual cree que *P* es más verosímil que su negación; el pensamiento de *P*, o de que debería creer racionalmente que *P*, ofrece a *A* motivos para actuar con vistas a causar en sí mismo la creencia en la negación de *P*. La acción involucrada puede consistir simplemente en apartar intencionalmente su atención de la evidencia en favor de *P* o puede implicar la búsqueda activa de evidencia en contra de *P*. Todo lo que el autoengaño exige de la acción es que el motivo tenga su origen en una creencia en la verdad de *P* (o en

5. Se puede pretender ocultar una intención presente al yo futuro. Así, yo podría tratar de perderme una reunión desagradable, fijada con un año de antelación, escribiendo deliberadamente una fecha equivocada en mi agenda y contando con mi mala memoria para haber olvidado ya lo que hice cuando llegue el momento. Este no es un caso puro de autoengaño, puesto que la creencia que pretendo tener no está *sustentada* por la intención que la produjo, y no hay necesariamente nada irracional en todo ello.

el reconocimiento de que la evidencia hace más probable la verdad de *P* que su falsedad) y que se lleve a cabo con la intención de producir una creencia en la negación de *P*. Finalmente –y esto es lo que hace del autoengaño un problema– el estado que motiva el autoengaño y el estado que éste produce *coexisten*; en el caso más grave, la creencia de que *P* no sólo causa, sino que incluso sustenta la creencia en la negación de *P*. El autoengaño es, pues, una forma autoinducida de debilidad de la justificación, donde el motivo para inducir una creencia es una creencia que la contradice (o lo que se considera como evidencia suficiente de esta última). En algunos casos, pero no en todos, el motivo nace del hecho de que el agente desea que la proposición, la creencia en la cual él mismo se induce, sea verdadera, o de su temor de que pudiera no serlo. Así, pues, a menudo el autoengaño implica también pensamiento desiderativo.

Lo que resulta difícil de explicar es cómo una creencia, o la constatación de que se tienen razones suficientes para sostener una creencia, puede sustentar una creencia contraria. Desde luego, no puede sustentarla en el sentido de proporcionarle un fundamento racional; en este contexto, «sustentar» únicamente ha de significar «causar». Nuestra tarea consiste en hallar, en la secuencia de estados mentales, un punto en el que haya una causa que no sea una razón; buscamos, pues, una irracionalidad específica en relación con los patrones de racionalidad del propio agente.[6]

Veamos, pues, a grandes rasgos, el modo en que, en mi opinión, puede ocurrir un caso típico de autoengaño. En el ejemplo que vamos a presentar, la debilidad de la justificación resulta autoinducida a través del pensamiento desiderativo. Carlos tiene buenas razones para creer que no superará las pruebas para la obtención del permiso de conducir. Ya

6. La idea de que la irracionalidad implica siempre la existencia de una causa mental de un estado mental para el que no hay ninguna razón se halla ampliamente tratada en «Paradoxes of irrationality».

ha suspendido esas pruebas dos veces y su instructor ha dicho cosas bastante desalentadoras. Por otra parte, sin embargo, conoce personalmente al examinador y tiene fe en su propio encanto personal. Se da cuenta de que la totalidad de la evidencia apunta hacia el fracaso. Como el resto de nosotros, Carlos razona normalmente de acuerdo con el requisito de evidencia global. Sin embargo, la idea de volver a fracasar en estas pruebas le resulta dolorosa (de hecho, Carlos encuentra particularmente mortificante la idea de fracasar en cualquier cosa). Así, pues, tiene un motivo perfectamente natural para creer que no suspenderá las pruebas, es decir, tiene un motivo para hacer que se dé el caso de que él sea una persona que cree que (probablemente) superará las pruebas. Su razonamiento práctico es simple y directo. En igualdad de circunstancias, es mejor evitar el dolor; creer que suspenderá las pruebas es doloroso; por lo tanto (en igualdad de circunstancias) es mejor evitar creer que suspenderá el examen. Puesto que presentarse al examen es una condición del problema con el que se enfrenta, esto significa que será mejor creer que aprobará. Actúa entonces con vistas a favorecer esta creencia, obteniendo quizá nueva evidencia en favor de la creencia de que aprobará. La cuestión puede reducirse simplemente a desplazar hacia el fondo la evidencia negativa o a destacar la positiva. Sin embargo, sean cuales fueren los mecanismos (y hay, desde luego, un buen número de ellos), en los casos centrales de autoengaño se requiere que Carlos siga siendo consciente de que su evidencia apoya la creencia de que fracasará, pues es la conciencia de este hecho lo que motiva sus esfuerzos por librarse del temor al fracaso.

Supongamos que Carlos consigue inducir en sí mismo la creencia de que aprobará el examen. En ese caso, es culpable de debilidad de la justificación, pues, aunque posee evidencia en favor de su creencia, sabe, o en todo caso piensa, que tiene mejores razones para creer que suspenderá. Este estado es irracional, pero ¿en qué punto hizo su entrada la irracionalidad?

Ya he rechazado, explícita o implícitamente, algunas res-

puestas a esta pregunta. Una de ellas es la sugerencia de David Pears según la cual el sujeto del autoengaño ha de «olvidar», o en todo caso ocultarse a sí mismo, el modo en que llegó a creer lo que ahora cree.[7] Convengo en que al autoengañador le *gustaría* hacer esto, y si lo hace ha conseguido, en un claro sentido, engañarse a sí mismo. Pero este grado y este tipo de logro hace del autoengaño un proceso, no un estado, y no es evidente que en algún momento el autoengañador se halle en un estado irracional. Por mi parte, creo que el autoengaño ha de alcanzarse a través de un proceso, pero luego puede ser un estado continuo y claramente irracional. El agente de Pears acaba hallándose en un estado anímico gratamente coherente. Por suerte, esto sucede a menudo. Pero el placer puede ser inestable, como ocurre probablemente en el caso de Carlos, pues el pensamiento placentero se ve amenazado por la realidad, o incluso simplemente por el recuerdo. Si la realidad (o el recuerdo) sigue amenazando la creencia que el sujeto autoengañado ha inducido en sí mismo, resulta necesaria una constante motivación para mantener en vigor el pensamiento feliz. Si estoy en lo cierto, el autoengañador no puede permitirse olvidar el factor que inspiró en primer término su conducta de autoengaño: la preponderancia de la evidencia contraria a la creencia inducida.

Implícitamente, también he rechazado la solución de Kent Bach, ya que, según este autor, el autoengañador no puede realmente creer en el peso de la evidencia contraria. Como Pears, Bach concibe también el autoengaño como una secuencia cuyo producto final se halla, con la motivación original, en un conflicto demasiado agudo para que pueda

7. Véase David Pears, «Motivated irrationality», en *Philosophical Essays on Freud*, así como su contribución a este volumen. [Davidson se refiere al libro *Actions and Events. Perspectives on the Philosophy of Donald Davidson*, Basil Blackwell, Oxford, 1985. T.]. Las diferencias entre mi concepción y la de Pears son pequeñas comparadas con las similitudes. Esto no es accidental, ya que mi exposición es deudora de su primer artículo y del que se contiene en este volumen.

coexistir con la percepción consciente de ésta.[8] Cabría pensar, tal vez, que estas diferencias entre mis puntos de vista y los de Pears y Bach se deben, al menos en parte, a la elección de formas distintas de describir el autoengaño más que a discrepancias sustantivas. A mí me parece importante identificar una incoherencia o inconsecuencia en el pensamiento del que se autoengaña; a Pears y Bach les preocupa más examinar las condiciones del éxito en la tarea de autoengañarse.[9] La dificultad reside en mantener el equilibrio entre ambas consideraciones: recalcar el primer elemento pone en claro la irracionalidad pero la hace difícil de explicar psicológicamente; recalcar el segundo elemento facilita la explicación del fenómeno a costa de menospreciar la irracionalidad.

¿En qué punto de la secuencia que conduce a un estado de autoengaño hay una causa mental que no es una razón del estado mental causado por ella? La respuesta depende, en parte, de la contestación a otra pregunta. Al principio di por supuesto que, aunque es posible, en relación con un conjunto de proposiciones contradictorias entre sí, creer simultáneamente cada una de ellas, no es posible, en cambio, tener una creencia cuyo objeto sea la conjunción de aquéllas cuando su contradicción resulta obvia. El agente autoengañado cree proposiciones contradictorias si cree que es calvo y cree que no lo es; Carlos cree proposiciones contradictorias si cree que superará las pruebas y cree que no las superará. La dificultad no es tan extraordinaria si el conflicto en la creencia es un caso normal de debilidad de la justificación, pero aun así sigue siendo muy considerable, dado el supuesto (para el que ya ofrecí argumentos) de que tener actitudes proposicionales implica aceptar el requisito de evi-

8. Véase Kent Bach, «An analysis of self-deception», *Philosophy and Phenomenological Review*, 41 (1981), págs. 351-370.

9. Así, pues, estoy de acuerdo con Jon Elster cuando dice que el autoengaño requiere «la consideración simultánea de creencias incompatibles»: *Ulysses and the Sirens* (Cambridge University Press, Cambridge, 1979), pág. 174.

dencia global. ¿Cómo puede una persona dejar de reunir las creencias contradictorias o incompatibles?

Sería un error por mi parte tratar de responder a esta pregunta mediante una detallada exposición psicológica. Lo importante es que las personas pueden, y a veces consiguen, mantener separadas creencias estrechamente relacionadas pero opuestas. En esa medida, hemos de aceptar la idea de que puede haber límites entre partes de la mente; allí donde hay creencias (obviamente) antagónicas, postulo la existencia de un límite entre ellas. Tales límites no son descubiertos por la introspección, sino que constituyen apoyos conceptuales para la descripción coherente de irracionalidades genuinas.[10]

No debemos concebir estos límites como barreras permanentes que demarcan territorios separados. Las creencias contradictorias sobre la superación de unas pruebas han de pertenecer a un vasto e idéntico nexo de creencias sobre pruebas y otros temas relacionados si han de ser realmente contradictorias. Aunque han de pertenecer a territorios fuertemente imbricados, dos creencias contradictorias no pertenecen al mismo territorio; borrar la línea existente entre ellas conllevaría la destrucción de una de las dos. No veo ninguna razón obvia para suponer que uno de los territorios haya de estar cerrado a la conciencia, sea cual fuere el significado de esto, pero es claro, en todo caso, que el agente no puede inspeccionar el todo sin borrar los límites.

Es posible ahora sugerir una respuesta a la pregunta que nos planteábamos, a saber, dónde hay un *paso* irracional en la secuencia que acaba en el autoengaño. La irracionalidad del estado resultante reside en el hecho de que contiene creencias contradictorias; el paso irracional es, por lo tanto, el que hace posible tal cosa, a saber, el que consiste en trazar el límite que mantiene separadas las creencias contradictorias. Cuando el autoengaño está constituido por una debilidad de la justificación inducida por el propio agente, lo que

10. Discuto la necesidad de «parcelar» la mente en «Paradoxes of irrationality».

ha de mantenerse apartado del resto de la mente es el requisito de evidencia global. La causa de ese exilio o aislamiento temporal se halla, desde luego, en el deseo de evitar la aceptación de aquello que ese requisito recomienda. Pero ésta no puede ser una *razón* para desatender el requisito. Nada puede considerarse como una buena razón para que una persona no razone según sus mejores normas de racionalidad.

En el caso extremo, cuando el motivo de autoengaño nace de una creencia que contradice directamente la creencia inducida, la creencia motivadora original ha de ser llevada fuera de los límites, junto con el requisito de evidencia global. Pero el hecho de que el pensamiento exiliado se halle fuera de los límites no le priva de poder, sino todo lo contrario, porque la razón no tiene jurisdicción alguna más allá de aquéllos.

EL CONOCIMIENTO DE LA PROPIA MENTE*

No hay ningún secreto acerca de la naturaleza de la evidencia que usamos para decidir lo que piensan otras personas: observamos sus actos, leemos sus cartas, estudiamos sus expresiones, escuchamos sus palabras, nos familiarizamos con sus biografías y atendemos a sus relaciones con la sociedad. El modo en que somos capaces de reunir todo ese material en una imagen convincente de una mente es ya otra cuestión; sabemos cómo hacerlo sin saber necesariamente cómo lo hacemos. Algunas veces averiguo lo que yo creo de forma muy similar a como lo averigua otra persona: reparando en lo que digo y hago. Puede haber ocasiones en que éste es mi único acceso a mis propios pensamientos. Según Graham Wallas:

> Talento poético tenía la pequeña que, ante la sugerencia de que se asegurase de lo que quería decir antes de hablar, respondió: «¿Cómo puedo saber lo que pienso hasta ver lo que digo?».[1]

Una idea similar fue expresada por Robert Motherwell: «Yo diría que la mayoría de los buenos pintores no saben lo que piensan hasta que lo pintan».

Gilbert Ryle coincidía por completo en este asunto con la pequeña poetisa y con el pintor; sostuvo heroicamente que conocemos nuestra propia mente exactamente del mismo modo en que conocemos la mente de los demás, a saber, observando lo que decimos, hacemos y pintamos. Ryle estaba

* Discurso presidencial pronunciado ante la Sexta Reunión Anual de la Sección del Pacífico de la Asociación Filosófica Americana en Los Angeles, California, 28 de marzo de 1986.

1. Graham Wallas, *The Art of Thought*.

equivocado. En muy raras ocasiones necesito recurrir a la evidencia o a la observación para descubrir lo que creo; normalmente sé lo que pienso antes de hablar o de actuar. Incluso cuando tengo evidencia, raras veces hago uso de ella. Puedo estar equivocado acerca de mis propios pensamientos, y por ello el recurso a lo que se puede determinar públicamente no resulta irrelevante. Pero la posibilidad de estar equivocado acerca de los pensamientos propios no puede quebrantar la presunción predominante de que una persona sabe lo que cree; en general, la creencia de que se tiene un pensamiento basta para justificar dicha creencia. Pero, aunque esto es verdad, e incluso obvio para la mayoría de nosotros, el hecho no tiene, hasta donde yo puedo saber, una explicación sencilla. Mientras que los recursos de que disponemos al tratar de penetrar en los pensamientos de los demás resultan bastante claros, al menos a grandes rasgos, es en cambio oscuro por qué, en nuestro propio caso, sabemos tan a menudo lo que pensamos sin recurrir a la evidencia o a la observación.

A causa de que normalmente sabemos lo que creemos (y deseamos y dudamos y pretendemos) sin necesidad de usar la evidencia (incluso cuando disponemos de ella), nuestro testimonio sincero acerca de nuestros estados mentales presentes no se halla sometido a las deficiencias de las conclusiones basadas en la evidencia. Así, pues, las aserciones sinceras en primera persona del presente acerca de pensamientos, aun no siendo infalibles ni corregibles, poseen no obstante una autoridad que no puede tener una aserción en segunda o tercera persona o en un tiempo distinto del presente. Reconocer este hecho, sin embargo, no equivale a explicarlo.

A partir de Wittgenstein se ha convertido en una rutina el intento de aliviar las preocupaciones sobre «nuestro conocimiento de otras mentes» señalando que constituye un aspecto esencial de nuestro uso de ciertos predicados mentales el hecho de que los aplicamos a otros sobre la base de la evidencia conductual, mientras que no disponemos de esa ayuda al aplicarlos a nosotros mismos. La observación es verda-

dera y, adecuadamente elaborada, debería servir como respuesta a quien se pregunte a sí mismo cómo podemos conocer las mentes ajenas. Pero, como respuesta al escéptico, la intuición de Wittgenstein (si realmente es de Wittgenstein) sería escasamente satisfactoria. En primer lugar, es una idea extraña que se deban favorecer las aserciones llevadas a cabo sin el apoyo de la evidencia o la observación frente a las que cuentan con dicho apoyo. Desde luego, si no se aporta evidencia en sostén de una afirmación, ésta no puede impugnarse poniendo en cuestión la verdad o relevancia de la evidencia. Pero difícilmente bastan estas observaciones para sugerir que, en general, las aserciones carentes de apoyo evidencial son más fidedignas que las dotadas de él. La segunda dificultad, y la más importante, es la que sigue. Normalmente diríamos que aquello que cuenta como evidencia para la aplicación de un concepto ayuda a definir el concepto, o al menos impone condiciones a su identificación. Si dos conceptos dependen regularmente en su aplicación de criterios o escalas de apoyo evidencial diferentes, han de ser conceptos diferentes. Así, pues, si lo que aparentemente es la misma expresión se emplea correctamente a veces sobre la base de cierta escala de apoyo evidencial y a veces sobre la base de una escala distinta (o de ninguna), la conclusión obvia parece ser que la expresión es ambigua. ¿Por qué, entonces, habríamos de suponer que un predicado como «*x* cree que Ras Dashan es la montaña más alta de Etiopía», que se aplica unas veces sobre la base de la evidencia conductual y otras no, carece de ambigüedad? Si es ambiguo, no hay razón para suponer que tiene el mismo significado cuando se aplica a uno mismo que cuando se aplica a otros. Si admitimos (como deberíamos hacerlo) que el carácter necesariamente público e interpersonal del lenguaje garantiza que a menudo nuestra aplicación de estos predicados a otros es correcta y que, por lo tanto, sabemos con frecuencia lo que *otros* piensan, ha de plantearse entonces la cuestión de los fundamentos que tiene cada uno de nosotros para creer que sabe lo que *él* (en el mismo sentido) piensa. La respuesta de estilo wittgensteiniano puede resol-

ver tal vez el problema de las otras mentes, pero crea un problema correspondiente en torno al conocimiento de la mente propia. La correspondencia no es, sin embargo, completa. El problema original de las otras mentes invitaba a preguntarse cómo sabe uno que los demás tienen siquiera una mente. El problema al que nos enfrentamos ahora puede formularse del modo que sigue. Yo sé a qué he de atender al atribuir pensamientos a otros. Esos mismos predicados me los aplico a mí mismo usando criterios muy diferentes (o ninguno). Surge, así, la pregunta escéptica: ¿por qué habría de pensar que son *pensamientos* lo que me atribuyo a mí mismo? Sin embargo, puesto que la evidencia que uso en el caso de otros es de carácter público, no hay razón para no atribuirme pensamientos a mí mismo del mismo modo en que los atribuyo a otros, a la manera de Graham Wallace, Robert Motherwell y Gilbert Ryle. En otras palabras, no trato mis propios estados mentales, aunque podría hacerlo, del mismo modo que los de los otros. Semejante estrategia no está a disposición de quien ambicione, con respecto a los pensamientos de otros, el mismo tipo de autoridad que parece tener en el trato con los suyos propios. Así, pues, la asimetría entre los casos sigue siendo un problema, y es la autoridad de la primera persona la que crea ese problema.

He sugerido una respuesta a este problema en otro artículo.[2] En él argüía que el examen del modo en que atribuimos pensamientos y significados a otros explicaría la autoridad de la primera persona sin invitar a la duda escéptica. En años recientes, sin embargo, algunos de los hechos sobre la atribución de actitudes en los que confiaba para defender la autoridad de la primera persona han sido empleados para atacar esa misma autoridad: se ha argüido, sobre bases consideradas nuevas, que, aun cuando los métodos del intérprete en tercera persona determinan lo que consideramos usualmente como los contenidos mentales

2. Donald Davidson, «First Person Authority», *Dialectica*, 38 (1984), págs. 101-111.

de un agente, los contenidos así determinados pueden ser desconocidos para el agente mismo. En el presente artículo examinaré algunos de estos argumentos e insistiré en que no constituyen una amenaza genuina para la autoridad de la primera persona. La explicación que ofrecí en mi artículo anterior de la asimetría entre las atribuciones de actitudes en primera persona y las que se llevan a cabo en segunda y tercera personas me parece en todo caso fortalecida por las nuevas consideraciones, o al menos por aquellas que parecen válidas.

Hay que destacar una vez más que el problema que me ocupa no exige la infalibilidad o incorregibilidad de nuestras creencias acerca de nuestros estados mentales presentes. Podemos cometer, y de hecho cometemos, errores acerca de lo que creemos, deseamos, aprobamos y pretendemos; existe también la posibilidad del autoengaño. Pero tales casos, aunque no infrecuentes, no son ni podrían ser la norma; no argüiré en favor de esto ahora, sino que lo consideraré como uno de los hechos a explicar.

Dejando, pues, de lado el autoengaño y otros fenómenos anómalos o inciertos, la cuestión es si podemos pensar lisa y llanamente, sin irracionalidad, incoherencia o confusión, que tenemos una creencia que no tenemos, o que no tenemos una creencia que de hecho tenemos. Cierto número de filósofos y de psicólogos de mentalidad filosófica han acariciado recientemente puntos de vista que implican o sugieren que tal cosa podría ocurrir fácilmente; más aún, que ha de ocurrir constantemente.

La amenaza se hallaba ya en el concepto russelliano de ciertas proposiciones que contenían «ingredientes» con los que la mente del sujeto cognoscente no estaba familiarizada; y el peligro se tornó más agudo con la evolución del estudio de las actitudes *de re*.*

* Las actitudes *de re* son aquellas que versan sobre una cosa o persona particular, de la que el sujeto cree, desea, etc., algo. De ellas se distinguen las actitudes *de dicto*, que no necesitan involucrar una cosa o persona particular, sino que se adoptan ante cierto *dictum* o proposición, indicando que se cree en su verdad, o se desea que sea verdadera, etc. (T.)

Fue, sin embargo, Hilary Putnam quien descorrió el velo. Atendamos al argumento que Putnam expuso en 1975 a fin de mostrar que los significados, como él lo expresó, «simplemente no están en la cabeza».[3] Putnam arguye persuasivamente que el significado de las palabras depende de algo más que de «lo que está en la cabeza». Narra cierto número de historias cuya moraleja es que ciertos aspectos de la historia natural del modo en que alguien aprendió el uso de una palabra introducen necesariamente una diferencia en el significado de esa palabra. De ello parece seguirse que dos personas podrían hallarse en estados físicamente idénticos y querer decir sin embargo cosas distintas con las mismas palabras.

Las consecuencias tienen un amplio alcance, pues si las personas pueden (normalmente) expresar correctamente sus pensamientos en palabras, entonces sus pensamientos –sus creencias, deseos, intenciones, esperanzas, expectativas– han de identificarse también, en parte, por eventos y objetos exteriores a la persona. Si los significados no están en la cabeza, tampoco lo están entonces, al parecer, las creencias, deseos y demás.

Puesto que algunos lectores podrían estar algo cansados del *doppelgänger* de Putnam en la Tierra Gemela, me permitiré narrar mi propia historia de ciencia ficción –si eso es lo que es–. Mi historia evita algunas dificultades irrelevantes de la de Putnam, aunque presenta algunos problemas nuevos.[4] (Un poco más abajo volveré a la cuestión de la Tierra y la

3. Hilary Putnam, «The Meaning of "Meaning"», reimpreso en *Philosophical Papers, Vol. II: Mind, Language and Reality*, Cambridge University Press, 1975, pág. 227.

4. No pretendo ser original en este punto; Steven Stich ha usado un ejemplo muy similar en «Autonomus Psychology and the Belief-Desire Thesis»; *The Monist*, 61 (1978), págs. 573 y sigs. Debo subrayar que no estoy sugiriendo que un objeto creado accidental o artificialmente no podría pensar; El Hombre de los Pantanos necesita simplemente tiempo para adquirir una historia causal que dé sentido a la afirmación de que está hablando, recordando, identificando o pensando en cosas del mundo. (Vuelvo sobre esto más adelante.)

Tierra Gemela.) Supongamos que un rayo cae sobre un árbol muerto en una zona pantanosa; yo estoy de pie junto a él. Mi cuerpo es reducido a sus elementos, mientras que, por pura coincidencia (y a partir de moléculas diferentes), el árbol se convierte en una réplica física de mí mismo. Mi réplica, El Hombre de los Pantanos, se mueve exactamente como yo lo hacía; de acuerdo con su naturaleza, abandona los pantanos, se encuentra con mis amigos y parece reconocerlos, y en apariencia responde a sus saludos en inglés. Se traslada a vivir a mi casa y parece escribir artículos sobre la interpretación radical. Nadie puede notar la diferencia.

Sin embargo, *hay* una diferencia. Mi réplica no puede reconocer a mis amigos; no puede *re*conocer cosa alguna, puesto que nunca la conoció anteriormente. No puede saber los nombres de mis amigos (aunque desde luego parece saberlos), no puede recordar mi casa. No puede querer decir lo mismo que yo con la palabra «casa», por ejemplo, puesto que el sonido «casa» emitido por él no fue aprendido en un contexto que le diese su significado correcto, o algún significado siquiera. En realidad, no veo cómo se podría decir que mi réplica quiere decir algo con los sonidos que emite o que tiene pensamientos.

Puede que Putnam no aceptase esta última afirmación, pues dice que si dos personas (u objetos) se hallan en similares estados físicos relevantes, es «absurdo» pensar que sus estados psicológicos «difieren en lo más mínimo».[5] Sin embargo, sería un error asegurar que Putnam y yo discrepamos en este punto, porque todavía no resulta claro cómo se está usando la frase «estado psicológico».

Según Putnam, muchos filósofos han supuesto erróneamente que estados psicológicos como la creencia y el conocimiento del significado de una palabra son a la vez (I) «internos», en el sentido de que no presuponen la existencia de ningún individuo distinto del sujeto al que se adscribe el estado, y (II) son los estados que normalmente identificamos e

5. Hilary Putnam, «The Meaning of "Meaning"», pág. 144.

individuamos como lo hacemos con las creencias y las demás actitudes proposicionales. Puesto que normalmente identificamos e individuamos estados mentales y significados, en parte, en términos de relaciones con objetos y eventos distintos del sujeto, Putnam cree que (I) y (II) se quiebran en pedazos: en su opinión, no hay estados que puedan satisfacer ambas condiciones.

Putnam llama «estrechos» a los estados psicológicos que satisfacen la condición (I). Concibe tales estados como solipsistas y los asocia con la concepción cartesiana de lo mental. Puede que Putnam los considere como los únicos estados psicológicos «auténticos»; en gran parte de su artículo omite el calificativo «estrechos», a pesar de que los (llamados) estados psicológicos estrechos no corresponden a las actitudes proposicionales tal como normalmente se las identifica. No todo el mundo ha quedado convencido de que pueda trazarse una distinción inteligible entre estados psicológicos estrechos (o internos, o cartesianos, o individualistas; todos estos términos son hoy corrientes) y estados psicológicos identificados (si es que algunos lo son) en términos de hechos externos (sociales o de otro tipo). Así, John Searle ha defendido que nuestras actitudes proposicionales cotidianas satisfacen la condición (I), de modo que no hay necesidad de estados que satisfagan la condición (II), mientras que Tyler Burge ha negado que existan, en algún sentido interesante, actitudes proposicionales que satisfagan la condición (I).[6] Pero parece haber un consenso universal en que no hay estados que satisfagan ambas.

La tesis de este artículo es que no hay razón para suponer que los estados mentales cotidianos no satisfacen ambas condiciones, (I) y (II): creo que tales estados son «internos», en el sentido de que son idénticos a estados del cuerpo, y con ello identificables sin referencia a objetos o eventos exteriores al cuerpo; al mismo tiempo son «no-individualistas»,

6. Véase John Searle, *Intentionality*, Cambridge University Press, 1983, y Tyler Burge, «Individualism and Psychology», *The Philosophical Review*, 95 (1986), págs. 3-45.

en el sentido de que pueden ser, y normalmente son, identificados en parte por sus relaciones causales con eventos y objetos exteriores al sujeto del que son estados. Un corolario de esta tesis resulta ser que, en contra de lo que a menudo se supone, la autoridad de la primera persona puede aplicarse sin contradicción a estados que, por lo regular, son identificados en términos de sus relaciones con eventos y objetos externos a la persona.

Comenzaré con el corolario. ¿Por qué es natural suponer que los estados que satisfacen la condición (II) pueden no ser conocidos por la persona que se halla en ellos?

He de referirme ahora a la Tierra Gemela de Putnam. Nos pide este autor que imaginemos a dos personas exactamente iguales desde el punto de vista físico y (por lo tanto) iguales con respecto a todos los estados psicológicos «estrechos». Una de ellas, habitante de la Tierra, ha aprendido a usar la palabra «agua» atendiendo al agua que se le mostraba, oyendo y leyendo sobre ella, etc. La otra, habitante de la Tierra Gemela, ha aprendido a usar la palabra «agua» en condiciones a primera vista iguales, pero la sustancia que le ha sido mostrada no es agua, sino una sustancia de aspecto similar que podemos llamar «gagua». Bajo tales circunstancias, sostiene Putnam, la primera hablante* se refiere al agua cuando usa la palabra «agua»; su gemela se refiere al «gagua» cuando *ella* usa la palabra «agua». Parece, pues, que tenemos un caso en que los estados psicológicos «estrechos» son idénticos y sin embargo los hablantes quieren decir cosas distintas con la misma palabra.

¿Qué sucede con los pensamientos de estas dos hablantes? La primera se dice a sí misma, cuando se halla ante un vaso de agua, «aquí hay un vaso de agua»; la segunda masculla para sí los mismos sonidos cuando se halla ante un

* Mantengo en la traducción el uso davidsoniano del género femenino en la formulación de ideas aplicables a los dos sexos. Esta práctica, de clara inspiración feminista, resulta cada vez más frecuente en la bibliografía filosófica anglosajona. (T.)

vaso de gagua. Cada una de ellas dice la verdad, puesto que sus palabras significan cosas distintas. Y, siendo ambas sinceras, es natural suponer que creen cosas diferentes: la primera, que hay un vaso de agua frente a ella, y la segunda que hay un vaso de gagua frente a *ella*. Pero, ¿saben qué es lo que creen? Si los significados de sus palabras, y con ello las creencias expresadas mediante el uso de las mismas, se hallan en parte determinados por factores externos que las hablantes ignoran, sus creencias y significados no son estrechos en el sentido de Putnam. Por lo tanto, no hay nada sobre cuya base cada hablante pueda decir en qué estado se halla, ya que no dispone de indicios, internos o externos, que le permitan conocer la diferencia. Al parecer, pues, debemos concluir que ninguna de las hablantes sabe lo que quiere decir o lo que piensa. Esta conclusión ha sido explícitamente extraída por algunos filósofos, entre los que se cuenta Putnam, quien declara que «... abandona la idea de que si hay una diferencia en el significado... *tiene* que haber alguna diferencia en nuestros conceptos (o en nuestro estado psicológico)». Lo que determina el significado y la extensión «... no es, en general, plenamente conocido para el hablante».[7] Aquí, «estado psicológico» significa estado psicológico *estrecho*, y se supone que únicamente tales estados son «plenamente conocidos». Jerry Fodor cree que las actitudes proposicionales cotidianas están (prácticamente) «en la cabeza», pero conviene con Putnam en que *sí* las actitudes proposicionales se identificasen en parte mediante factores externos al agente, no estarían en la cabeza y no serían necesariamente conocidas para el agente.[8] También John Searle sostiene, aunque sus razones no son las de Fodor, que los significados están en la cabeza («no hay otro sitio en donde pudieran estar»), pero parece aceptar la inferencia según la cual, si ello

7. Hilary Putnam, «The Meaning of "Meaning"», págs. 164-165.

8. Jerry Fodor, «Cognitive Science and the Twin Earth Problem», *Notre Dame Journal of Formal Logic*, 23 (1982), pág. 103. Véase también su «Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology», *The Behavioral and Brain Sciences*, 3 (1980).

no fuera así, la autoridad de la primera persona se perdería.[9] Tal vez la más clara formulación de esta posición sea la que aparece en la introducción de Andrew Woodfield a un libro de ensayos sobre los objetos del pensamiento. Refiriéndose a la afirmación según la cual los contenidos de la mente se hallan a menudo determinados por factores externos a la persona de cuya mente se trata y son quizá desconocidos para ella, dice Woodfield:

> Puesto que la relación externa no está subjetivamente determinada, el sujeto no tiene autoridad sobre ella. Muy bien podría suceder que una tercera persona se hallase en mejor posición que el sujeto para conocer el objeto en el que dicho sujeto está pensando, y que por tanto estuviera mejor situada para saber cuál era el pensamiento en cuestión.[10]

Aquellos que aceptan la tesis según la cual los contenidos de las actitudes proposicionales se identifican, en parte, en términos de factores externos parecen tener un problema similar al del escéptico que descubre que podríamos estar completamente equivocados sobre el mundo «externo». En el presente caso, el escepticismo común en torno a los sentidos se evita suponiendo que el mundo mismo determina más o menos correctamente los contenidos de los pensamientos acerca de él. (El hablante que piensa que se trata de agua tiene probablemente razón, ya que aprendió el uso de la palabra «agua» en un entorno acuoso; el hablante que piensa que se trata de gagua tiene probablemente razón, pues aprendió la palabra «agua» en un entorno gacuoso.) Pero el escepticismo no queda derrotado, sino que simplemente se traslada al conocimiento de nuestra propia mente. Nuestras creencias cotidianas sobre el mundo externo están (en esta perspectiva) guiadas por el mundo, pero nosotros no sabemos qué es lo que creemos.

9. John Searle, *Intentionality*, capítulo 8.

10. *Thought and Object*, Andrew Woodfield (comp.), Clarendon Press, 1982, pág. 8.

Hay, desde luego, una diferencia entre agua y gagua, y esa diferencia puede ser descubierta por medios normales, sea o no descubierta de hecho. Así, una persona podría averiguar lo que cree descubriendo la diferencia entre agua y gagua e indagando lo suficiente sobre sus propias relaciones con ambas para determinar de cuál de ellas tratan sus palabras y creencias. La conclusión escéptica a la que parece que hemos llegado afecta al alcance de la autoridad de la primera persona: ese alcance es mucho más limitado de lo que suponíamos. Nuestras creencias sobre el mundo son en su mayoría verdaderas, pero podemos fácilmente estar equivocados sobre lo que pensamos. Se trata de una imagen invertida del escepticismo cartesiano.

Aquellos que sostienen que los contenidos de nuestros pensamientos y los significados de nuestras palabras se hallan fijados a menudo por factores que ignoramos no han mostrado mucha preocupación por esa consecuencia aparente de sus opiniones que acabo de subrayar. Han advertido, desde luego, que, si tuviesen razón, la idea cartesiana según la cual lo único de que podemos tener certeza son los contenidos de nuestra propia mente y la noción de Frege de significados plenamente «aprehendidos» han de ser erróneas. Pero, por lo que yo sé, no han hecho un gran esfuerzo para resolver el aparente conflicto entre sus concepciones y la potente intuición de la existencia de la autoridad de la primera persona.

Una razón de esta falta de preocupación podría ser que, en opinión de algunos de ellos, el problema parece limitarse a una serie bastanté restringida de casos, en los que ciertos conceptos o palabras se adhieren a objetos identificados o aludidos mediante el uso de nombres propios, expresiones indicativas y términos de géneros naturales. Otros, en cambio, arguyen que los lazos entre el lenguaje y el pensamiento, por un lado, y los asuntos externos, por otro, son tan profundos que ningún aspecto del pensamiento, en su concepción usual, puede quedar intacto. En esta misma línea señala Daniel Dennett que «... se debe contar con una rica información sobre el mundo en general, sus ocupantes y

propiedades, y estar íntimamente conectado con ellos para poder decir con propiedad que se tienen creencias».[11] Y continúa diciendo que la identificación de *todas* las creencias está infectada por los factores externos, no subjetivos, cuya operación se reconoce en el tipo de caso que hemos estado discutiendo. Burge destaca también la amplitud de la influencia de los factores externos sobre nuestras creencias, aunque, por razones que no explica, no parece considerar esto como una amenaza a la autoridad de la primera persona.[12]

El asunto ha tomado un rumbo inquietante. En una época se invocaba el conductismo para mostrar cómo era posible que una persona supiera lo que había en la mente de otra; el conductismo fue luego rechazado, en parte porque no era capaz de explicar uno de los aspectos más obvios de los estados mentales: el hecho de que en general son conocidos por la persona que los tiene sin recurrir a la evidencia conductista. La moda más reciente, aun no siendo estrictamente conductista, identifica en parte los estados mentales, una vez más, en términos de factores sociales y de otros factores externos, haciéndolos así, en esa medida, objeto de averiguación pública. Pero, al mismo tiempo, reinstaura el problema de explicar la autoridad de la primera persona.

Aquellos que están convencidos de la dimensión externa del contenido de los pensamientos, tal como éstos se identifican e individuan ordinariamente, han reaccionado de distintas formas. Una de las respuestas ha consistido en establecer una distinción entre los contenidos de la mente en cuanto determinados subjetiva e internamente, por un lado, y las creencias, deseos e intenciones cotidianas, que normal-

11. Daniel Dennett, «Beyond Belief», en *Thought and Object*, pág. 76.

12. Tyler Burge, «Other Bodies», en *Thought and Object*; «Individualism and the Mental», en *Midwest Studies in Philosophy, Volume 4*, Peter French, Theodore Vehling, Howard Wettstein (comps.), University of Minnesota Press, 1979; «Two Thought Experiments Reviewed», *Notre Dame Journal of Formal Logic*, 23 (1982), págs. 284-293; «Individualism and Psychology».

mente atribuimos sobre la base de conexiones sociales y otras conexiones externas, por otro. Esta es claramente la tendencia del argumento de Putnam (aunque la palabra «agua» tiene significados diferentes y se usa para expresar creencias distintas al ser empleada para referirse al agua y al gagua, las personas que utilizan la palabra para estos diferentes propósitos pueden estar en «el mismo estado psicológico»). Jerry Fodor acepta la distinción para determinados fines, pero arguye que la psicología debe adoptar la perspectiva del «solipsismo metodológico» (la expresión es de Putnam), esto es, debe tratar exclusivamente con estados internos, estados psicológicos verdaderamente subjetivos que no deban nada a sus relaciones con el mundo externo.[13]

Steven Stich establece esencialmente la misma distinción, pero extrae una moraleja más severa: allí donde Fodor piensa que simplemente hemos de componer un poco las actitudes proposicionales en su concepción usual para separar el elemento puramente subjetivo, Stich sostiene que los estados psicológicos como ahora los concebimos pertenecen a una tosca y confusa «psicología popular» que ha de reemplazarse por una «ciencia cognitiva» aún por inventar. El subtítulo de su reciente libro es «El proceso contra la creencia».[14]

Aquellos que trazan una distinción semejante han logrado sin duda que el problema de la autoridad de la primera persona, al menos tal como yo lo he planteado, no pueda resolverse. En efecto, el problema que he propuesto consiste en cómo explicar la asimetría entre el modo en que una persona conoce sus estados mentales presentes y el modo en que los conocen otros. Los estados mentales en cuestión son creencias, deseos, intenciones, etcétera, según se conciben ordinariamente. Aquellos que aceptan algo parecido a la distinción de Putnam ni siquiera intentan explicar la autoridad de la primera persona con respecto a esos estados; si hay al-

13. Jerry Fodor, «Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology».

14. Steven Stich, *From Folk Psychology to Cognitive Science*, M.I.T. Press, 1983.

guna autoridad de ese tipo, rige en todo caso sobre estados muy diferentes. (En el caso de Stich, no es obvio que pueda regir sobre cosa alguna.)

Creo que Putnam, Burge, Dennett, Fodor, Stich y otros tienen razón en llamar la atención sobre el hecho de que los estados mentales ordinarios, o al menos las actitudes proposicionales, se identifican en parte por sus relaciones con la sociedad y con el resto del entorno, relaciones que en algunos aspectos pueden no ser conocidas para la persona que se halla en esos estados. También tienen razón, en mi opinión, cuando sostienen que por ese motivo (si no por otros) los conceptos de la «psicología popular» no pueden integrarse en un sistema de leyes coherente y comprehensivo del tipo del que la física se afana por conseguir. Estos conceptos son parte de una teoría de sentido común para la descripción, interpretación y explicación de la conducta humana, una teoría de estilo un tanto libre, pero (en mi opinión) indispensable. Puedo imaginar una ciencia que se ocupe de las personas y se halle expurgada de «psicología popular», pero no puedo imaginar qué interés podría tener. Este no es, sin embargo, el tema del presente trabajo.

Lo que aquí me preocupa es el enigmático descubrimiento de que, aparentemente, no sabemos lo que pensamos; al menos del modo en que creemos saberlo. Este es un auténtico enigma para quien, como yo, crea que es cierto que los factores externos determinan en parte los contenidos de los pensamientos y crea también que, en general, sabemos lo que pensamos, y además de un modo en que los demás no lo saben. El problema surge porque admitir el papel de identificación e individuación desempeñado por los factores externos parece llevar a la conclusión de que nuestros pensamientos pueden no ser conocidos para nosotros.

Pero, ¿se sigue de hecho esta conclusión? La respuesta depende, creo, de cómo se conciba la dependencia que guarda la identificación de los contenidos mentales con respecto a los factores externos.

La conclusión se sigue, por ejemplo, para cualquier teoría que sostenga que las actitudes proposicionales se iden-

tifican mediante objetos (tales como proposiciones, casos de proposiciones o representaciones) que se hallan en, o «ante», la mente y que contienen o incorporan (como «ingredientes») objetos o eventos externos al agente, pues resulta obvio que todos ignoramos innumerables rasgos de los objetos externos. Que la conclusión se sigue de estos supuestos es algo generalmente aceptado.[15] Sin embargo, por razones que mencionaré más adelante, yo rechazo los supuestos sobre los cuales se basa la conclusión en este caso.

Tyler Burge ha sugerido que hay otra forma en que los factores externos intervienen en la determinación de los contenidos del habla y del pensamiento. Uno de sus «experimentos mentales» se ajusta de hecho bastante bien a mi propio caso. Hasta hace poco yo creía que la artritis era una inflamación de las articulaciones causada por depósitos de calcio; no sabía que cualquier inflamación de las articulaciones, como la gota, por ejemplo, era también una forma de artritis. Así, cuando un médico me dijo que tenía gota (lo que resultó ser falso) yo creí que tenía gota, pero no artritis. En este punto Burge nos pide que imaginemos un mundo en el que yo fuese físicamente el mismo, pero en el que la palabra «artritis» se aplicase de hecho sólo a la inflamación de las articulaciones causada por depósitos de calcio. En ese caso la oración «la gota no es una forma de artritis» sería verdadera, no falsa, y la creencia que yo hubiera expresado con esa oración no sería la creencia falsa de que la gota no es una forma de artritis, sino una creencia verdadera sobre una enfermedad distinta de la artritis. Y, sin embargo, en este mundo imaginado todos mis estados físicos, mis «experiencias cualitativas internas», mi conducta y mis disposiciones para ella serían los mismos que en este mundo. Mi *creencia* habría cambiado, pero yo no tendría razones para suponer tal cosa, de modo que no cabría decir que yo sabía qué es lo que creía.

15. Véase, por ejemplo, Gareth Evans, *The Varieties of Reference*, Oxford University Press, 1982, págs. 45, 199, 201.

Burge destaca el hecho de que su argumento depende de ...la posibilidad de que alguien tenga una actitud proposicional a pesar de su incompleto ...dominio de alguna noción incluida en el contenido de aquélla... *si* el experimento mental ha de funcionar, hemos de encontrarnos con que el sujeto cree (o tiene una actitud caracterizada por) cierto contenido, a pesar de una comprensión incompleta o una aplicación errónea.[16]

Parece seguirse de esto que, si Burge tiene razón, cada vez que una persona está equivocada, confundida o parcialmente mal informada sobre el significado de una palabra, está asimismo equivocada, confundida o parcialmente mal informada acerca de cualquier creencia suya que se exprese (¿o se expresaría?) mediante el uso de dicha palabra. Puesto que, según Burge, semejante «comprensión parcial» es «común o incluso normal en el caso de un gran número de expresiones en nuestros vocabularios», ha de ser igualmente común o normal en nosotros estar equivocados acerca de lo que creemos (y, desde luego, acerca de lo que tememos, esperamos, deseamos que ocurra, dudamos, y así sucesivamente).

Aparentemente, Burge acepta esta conclusión; así interpreto yo al menos su negación de que «... la plena comprensión de un contenido sea en general una condición necesaria para creer dicho contenido». Burge rechaza explícitamente «... el viejo modelo según el cual una persona ha de estar directamente familiarizada con los contenidos de sus pensamientos o aprehenderlos de manera inmediata... el *contenido* mental de una persona no está fijado por lo que sucede en ella o por lo que le es accesible simplemente a través de una cuidadosa reflexión».[17]

No sé con certeza cómo entender estas afirmaciones,

16. Tyler Burge, «Individualism and the Mental», pág. 83.
17. Ibíd., págs. 90, 102, 104.

pues no estoy seguro de la seriedad con que hemos de tomar las declaraciones sobre la «familiaridad directa» con un contenido o la «aprehensión inmediata» del mismo. Pero, en cualquier caso, estoy convencido de que, si lo que pensamos y queremos decir está determinado por los hábitos lingüísticos de los que nos rodean, del modo en que Burge cree que lo está, la autoridad de la primera persona se halla entonces en serio peligro. Puesto que el grado y carácter del peligro me parecen incompatibles con lo que sabemos sobre el tipo de conocimiento que tenemos de nuestras propias mentes, me veo obligado a rechazar alguna de las premisas de Burge. Convengo, sin embargo, en que aquello que pensamos y queremos decir no está «fijado» (exclusivamente) por lo que sucede en mí. Por lo tanto, lo que he de rechazar es la explicación de Burge del modo en que los factores sociales y otros factores externos controlan los contenidos de la mente de una persona.

Por cierto número de razones, me inclino a rebajar la importancia de los rasgos que Burge destaca como característicos de nuestra atribución de actitudes. Supongamos que yo, que pienso que la palabra «artritis» se aplica a la inflamación de las articulaciones sólo si está causada por depósitos de calcio, y mi amigo Arthur, que posee mejor información, proferimos sinceramente, dirigiéndonos a Smith, las palabras «Carl tiene artritis». Según Burge, si en otros aspectos somos más o menos iguales (Arthur y yo somos en general hablantes competentes del inglés, los dos hemos aplicado a ménudo la palabra «artritis» a casos genuinos de artritis, etc.), entonces nuestras palabras significan lo mismo en esta ocasión, los dos queremos decir lo mismo con nuestras palabras y ambos expresamos la misma creencia. Mi error en torno al significado léxico de la palabra (o acerca de qué es la artritis) no introduce diferencia alguna en lo que pensé y quise decir en esa ocasión. La evidencia en la que Burge se apoya para afirmar tal cosa parece basarse en su convicción de que esto es lo que cualquiera (que no esté infectado por la filosofía) diría so-

bre Arthur y yo. Dudo que Burge tenga razón en esto, pero, aun cuando la tuviera, no creo que ello demuestre su afirmación. Las atribuciones ordinarias de significados y actitudes descansan en vastos y vagos supuestos acerca de lo que es y no es compartido (en el aspecto lingüístico y en otros) por el que lleva a cabo la atribución, la persona objeto de la misma y la audiencia en la que piensa el primero. Cuando algunos de estos supuestos demuestran ser falsos, podemos alterar las palabras que usamos para emitir nuestro informe, a menudo de forma sustancial. Cuando nada importante depende de ello, tendemos a escoger el camino más fácil: aceptamos literalmente las palabras del hablante, aun cuando esto no refleje del todo algún aspecto de su pensamiento o de lo que quiere decir. Pero tal proceder no se debe a que estemos obligados (al menos fuera de una sala de justicia) a ser legalistas en este asunto. Y a menudo no lo somos. Si Smith (que no está infectado por la filosofía) informa a un nuevo interlocutor (tal vez un médico de un lugar lejano que trata de elaborar un diagnóstico sobre la base de un informe telefónico) de que Arthur y yo hemos dicho, y creemos, que Carl tiene artritis, puede inducir activamente a engaño a *su* oyente. Si este peligro se presentara, Smith, atento a los hechos, no diría simplemente «Arthur y Davidson creen que Carl tiene artritis», sino que añadiría algo semejante a esto: «Pero Davidson piensa que la artritis ha de estar causada por depósitos de calcio». Considero la necesidad de esta adición como una muestra de que la atribución simple no era totalmente correcta; había una diferencia relevante entre los pensamientos que Arthur y yo expresábamos al decir «Carl tiene artritis». Burge no está obligado a abandonar su posición por este argumento, desde luego, ya que puede insistir en que el informe es literalmente correcto, aunque pueda, como cualquier informe, inducir a error. Creo, por otra parte, que esta réplica pasaría por alto el alcance de la dependencia necesaria de los contenidos de una creencia con respecto a las otras. Los pensamientos no son átomos independientes, de modo que no puede haber ninguna regla

simple o rígida para la atribución correcta de un único pensamiento.[18]

Aunque rechazo la insistencia de Burge en que no podemos sino dar a las palabras de una persona el significado que tienen en su comunidad lingüística e interpretar sus actitudes proposicionales sobre esta misma base, creo, no obstante, que hay un sentido algo diferente, pero muy importante, en que los factores sociales controlan lo que un hablante puede querer decir con sus palabras. Si un hablante desea ser entendido, ha de pretender que sus palabras se interpreten de cierta manera, y por ello ha de pretender proporcionar a su

18. Burge sugiere que la razón por la que normalmente juzgamos que una persona quiere decir con sus palabras lo que quieren decir con ellas otros miembros de su comunidad lingüística, sepa o no el hablante lo que estos otros quieren decir, es que «frecuentemente hacemos que las personas se atengan, y ellas mismas se atienen, a los criterios de la comunidad cuando está en entredicho el uso incorrecto o el malentendido». Dice también que tales casos «... dependen de cierta responsabilidad hacia la práctica comunitaria» («Individualism and the Mental», pág. 90). No pongo en duda el fenómeno, sino su relevancia para lo que supone que muestra. (a) A menudo es razonable considerar responsables a las personas del conocimiento de lo que significan sus palabras; en tales casos podemos atribuirles un compromiso con posiciones que eran desconocidas para ellas o con las que no creían tener ese compromiso. Esto no tiene nada que ver (directamente) con lo que querían decir con sus palabras ni con lo que creían. (b) Como buenos padres y ciudadanos, deseamos alentar prácticas que incrementen las posibilidades de comunicación; usar las palabras como creemos que otros lo hacen puede incrementar la comunicación. Esta idea (esté o no justificada) puede ayudar a explicar por qué algunas personas tienden a atribuir significados y creencias de forma legalista; con ello esperan alentar la conformidad. (c) Un hablante que desee ser entendido ha de pretender que sus palabras se interpreten (y sean por tanto interpretables) en cierta dirección; esta intención puede verse satisfecha usando las palabras como lo hacen los demás (aunque a menudo esto no es así). De modo similar, un oyente que desea entender a un hablante ha de tener la intención de interpretar las palabras del hablante como éste pretendía emplearlas (tanto si la interpretación es «normal» como si no). Estas intenciones recíprocas se tornan moralmente importantes en un sinfín de situaciones que no tienen conexión necesaria con la determinación de lo que alguien tenía en la mente.

audiencia las claves necesarias para que llegue a la interpretación pretendida por él. Esto es válido tanto si el oyente usa a la perfección un lenguaje conocido por el hablante como si está aprendiendo su primer lenguaje. El lenguaje ha de poder ser aprendido e interpretado, y éste es el requisito que da lugar al irreductible factor social y muestra por qué una persona no puede querer decir con sus palabras algo que no pueda ser correctamente descifrado por otra. (El propio Burge parece apuntar esto mismo en un artículo posterior.) [19]

Quisiera ahora volver a Putnam y a su ejemplo de la Tierra Gemela, que no depende de la idea según la cual el uso lingüístico social dicta (bajo condiciones más o menos normales) lo que los hablantes quieren decir con sus palabras y menos aún cuáles son sus estados psicológicos (estrechos). Como dije, estoy persuadido de que Putnam tiene razón; lo que significan nuestras palabras viene fijado en parte por las circunstancias en que las aprendimos y usamos. Tal vez el ejemplo particular de Putnam (el agua) no es suficiente para asentar este punto, ya que es posible insistir en que «agua» no se aplica simplemente a la materia que tiene la misma estructura molecular que el agua, sino también a aquella materia que se asemeja lo bastante al agua en su estructura para ser inodora, potable, para permitir la natación y la navegación, etc. (Me doy cuenta de que esta observación, como muchas otras en este trabajo, puede mostrar que no soy capaz de reconocer un designador rígido* cuando lo

19. Véase, por ejemplo, «Two Thought Experiments Reviewed», pág. 289.

* El concepto de designador rígido se debe a Kripke. Un designador rígido es aquel término o expresión que denota la misma entidad en todos los mundos posibles. Según ciertos autores, los nombres propios y los términos de clases naturales como «agua», «oro», etc. son designadores rígidos. Así, mientras que podemos imaginar que «el monte más alto de la Tierra» podría no ser el Everest en el supuesto de que nuestro planeta hubiese tenido una historia geológica distinta, lo que nosotros llamamos «agua» (esto es, H_2O) no podría ser una substancia distinta en otro mundo posible, pues si fuera una substancia distinta ya no sería agua. La relación de designación entre «agua» y el agua es rígida (T.).

veo.) La cuestión no depende de casos especiales como ésos ni de cómo los resolvemos o deberíamos hacerlo. La cuestión depende simplemente del modo en que se establece la conexión básica entre palabras y cosas, o entre pensamientos y cosas. Yo mantengo, junto con Burge y Putnam, si les entiendo bien, que dicha conexión se establece mediante interacciones causales entre las personas y determinadas partes o aspectos del mundo. Las disposiciones para reaccionar de forma diferenciada ante objetos y eventos situados en ese marco son de central importancia para la interpretación correcta de los pensamientos y el habla de una persona. Si no fuera así, no habría modo de descubrir lo que otros piensan o lo que quieren decir con sus palabras. El principio es tan simple y obvio como esto: una oración que alguien tiene por verdadera movido por (a causa de) y sólo por la visión de la luna tenderá a significar algo semejante a «ahí está la luna»; el pensamiento expresado tenderá a ser que la luna está ahí; el pensamiento inspirado por y sólo por la visión de la luna tenderá a ser el pensamiento de que la luna está ahí. Tenderá a ser tal, contando con el error inteligible, los informes de terceras personas, y así sucesivamente. No se trata de que todas las palabras y oraciones se hallen tan directamente condicionadas a aquello de que versan; podemos perfectamente aprender a usar la palabra «luna» sin ver nunca la luna. Lo que defiendo es que todo pensamiento y lenguaje ha de tener un fundamento en tales conexiones históricas directas, y que estas conexiones constriñen la interpretación de aquéllos. Tal vez debería subrayar que los argumentos en favor de esta tesis no descansan en intuiciones acerca de lo que diríamos si determinados enunciados contrafácticos fuesen verdaderos. La ciencia ficción o los experimentos mentales no son necesarios.[20]

Así, pues, coincido con Putnam y Burge en que el conte-

20. Burge ha descrito «experimentos mentales» que no involucran el lenguaje en absoluto; uno de estos experimentos le lleva a decir que una persona criada en un entorno en el que no hubiera aluminio no podría tener «pensamientos sobre el aluminio» («Individualism and Psychology», pág. 5). Burge no dice por qué

nido intencional de las actitudes proposicionales ordinarias... no puede explicarse en términos de estados o procesos físicos, fenoménicos, causal-funcionales, computacionales o sintácticos, que sean especificados de modo no intencional y se definan puramente en el marco de un individuo aislado de su entorno físico y social.[21]

Queda en pie la cuestión de si este hecho constituye una amenaza a la autoridad de la primera persona, como Burge parece pensar y Putnam y otros ciertamente piensan. He rechazado uno de los argumentos de Putnam que, si fuera correcto, implicaría esa amenaza. Pero queda la posición descrita en el párrafo anterior, posición que yo sostengo, tanto si otros lo hacen como si no, pues pienso que es necesario al menos ese grado de «externalismo» para explicar cómo se puede aprender el lenguaje y cómo puede un intérprete identificar palabras y actitudes.

¿Por qué piensa Putnam que, si la referencia de una palabra está fijada (a veces) por la historia natural de su adquisición, un usuario de dicha palabra puede perder la autoridad de la primera persona? Putnam defiende (correctamente, en mi opinión) que dos personas pueden ser idénticas en todos los aspectos físicos (químicos, fisiológicos, etc.) relevantes y sin embargo querer decir cosas diferentes con sus palabras y tener actitudes proposicionales diferentes (de acuerdo con la identificación normal de las mismas). La disparidad se debe a diferencias ambientales que ambos agen-

---

cree que ello es así, pero no es en absoluto obvio que se necesiten supuestos contrafácticos para expresar esta idea. En todo caso, los nuevos experimentos mentales parecen descansar en intuiciones bastante distintas de las invocadas en «Individualism and the Mental»; no se ve con claridad el papel que desempeñan las normas sociales en los nuevos experimentos, y los hábitos lingüísticos de la comunidad son aparentemente irrelevantes. Puede que en este punto la posición de Burge sea cercana a la mía.

21. «Two Thought Experiments Reviewed», pág. 288.

tes pueden, en algunos aspectos, ignorar. ¿Por qué, bajo esas circunstancias, habríamos de suponer que estos agentes pueden no saber lo que piensan y quieren decir? Hablar con ellos no lo pondrá de manifiesto fácilmente. Como hemos indicado, cada uno de ellos, situado frente a un vaso de agua o de gagua, dice sinceramente: «Aquí hay un vaso de agua». Si se hallan en sus entornos originales respectivos, uno y otro tienen razón; si han intercambiado los planetas, uno y otro están equivocados. Si preguntamos a cada uno de ellos qué entiende por la palabra «agua», nos da la respuesta correcta, usando las mismas palabras que el otro, desde luego. Si preguntamos a cada uno de ellos qué es lo que cree, nos da la respuesta correcta. Estas dos respuestas son correctas porque, aunque son verbalmente idénticas, han de interpretarse de forma diferente. Y, ¿qué es lo que no conocen (con la autoridad usual) acerca de sus propios estados? Como hemos visto, Putnam distingue los estados a los que nos hemos estado refiriendo de los estados psicológicos «estrechos», que no presuponen la existencia de ningún individuo aparte del que se halla en esos estados. Podemos ahora empezar a preguntarnos por qué Putnam está interesado en los estados psicológicos estrechos. Parte de la respuesta es, desde luego, que son esos estados los que él cree que tienen la propiedad «cartesiana» de ser conocidos de una forma especial por la persona que se encuentra en ellos. (La otra parte de la respuesta tiene que ver con la construcción de una «psicología científica»; esto no nos concierne aquí.)

El razonamiento depende, creo, de dos supuestos en gran medida no cuestionados. Se trata de éstos:

1. Si un pensamiento es identificado por una relación con algo exterior a la cabeza, no está totalmente en la cabeza. (It ain't in the head).

2. Si un pensamiento no está totalmente en la cabeza, no puede ser «captado» por la mente del modo que requiere la autoridad de la primera persona.

Que éste es el razonamiento de Putnam lo sugiere su afirmación según la cual si dos cabezas son iguales, sus estados psicológicos estrechos han de ser iguales. Así, si supone-

mos que dos personas son idénticas «molécula a molécula» («en el sentido en que dos corbatas pueden ser "idénticas"»; podemos añadir, si queremos, que cada una de las dos «piensa los mismos pensamientos verbalizados..., tiene los mismos datos sensoriales, las mismas disposiciones, etc.» que la otra), entonces «es absurdo pensar que [un] estado psicológico difiere un ápice» del otro. Se trata, por supuesto, de estados psicológicos estrechos, no de aquellos que atribuimos normalmente, que no están en la cabeza.[22]

No es fácil decir exactamente de qué forma pueden ser idénticos los pensamientos verbalizados, datos sensoriales y disposiciones sin remitirnos a las corbatas, así que remitámonos a ellas. La idea es entonces la siguiente: los estados psicológicos estrechos de dos personas son idénticos cuando no es posible distinguir sus respectivos estados físicos. No tendría caso discutir esto, ya que Putnam puede definir los estados psicológicos estrechos como le plazca; lo que me gustaría poner en tela de juicio es el supuesto (1), enunciado más arriba, que llevaba a la conclusión de que las actitudes proposicionales ordinarias no están en la cabeza y, por lo tanto, la autoridad de la primera persona no se aplica a ellas.

Debería quedar claro que del mero hecho de que los significados se identifiquen en parte por relaciones con objetos exteriores a la cabeza no se sigue que los significados no estén en la cabeza. Suponer otra cosa sería tan poco feliz como argüir que, puesto que estar quemado por el sol presupone la existencia del sol, mi eritema no es una condición de mi piel. Mi piel quemada por el sol puede ser indistinguible de la piel de otra persona que se quemó de otra forma (su piel y la mía pueden ser idénticas en «el sentido de las corbatas»); sin embargo, uno de nosotros está realmente quemado por el sol y el otro no. Esto basta para mostrar que la consideración de los factores externos que intervienen en nuestras formas comunes de identificar estados mentales no descalifica una teoría de la identidad entre lo mental y lo fí-

22. «The Meaning of "Meaning"», pág. 227.

sico. Andrew Woodfield parece pensar lo contrario. Así, escribe: ningún estado *de re** acerca de un objeto externo al cerebro de una persona puede ser idéntico a un estado de ese cerebro, puesto que ningún estado cerebral presupone la existencia de un objeto externo.[23]

Los estados y eventos particulares no presuponen cosa alguna en sí mismos desde el punto de vista *conceptual*; sin embargo, algunas de sus *descripciones* pueden hacerlo. Mi abuelo paterno no me presuponía a mí, pero si se puede describir a alguien como mi abuelo paterno, han de existir varias personas además de mi abuelo paterno, incluyéndome a mí.

Burge podría haber caído en un error similar en el siguiente pasaje:

> ... Ningún pensamiento que acontezca... podría tener un contenido diferente y ser el mismo evento particular... Entonces... un evento mental de una persona no es *idéntico* a ningún evento que sea descrito por la fisiología, la biología, la química o la física. Pues suponiendo que *B* es cualquier evento dado descrito en términos de una de las ciencias físicas que ocurra en el sujeto mientras tiene el pensamiento relevante. Y suponiendo que «*B*» sea aquello que denota el mismo evento físico que ocurre en el sujeto en nuestra situación contrafáctica... *B* no tiene por qué verse afectado por diferencias contrafácticas [que no cambien los contenidos del evento mental]. Así, pues,... *B* [el evento físico] no es idéntico al pensamiento que tiene lugar en el sujeto.[24]

* Véase N. del T., pág. 123.

23. Andrew Woodfield, en *Thought and Object*, pág. 8.
24. «Individualism and the Mental», pág. 111.

Burge no pretende haber fundamentado la premisa de su argumento, y, en consecuencia, tampoco la conclusión. Sin embargo, mantiene que la negación de la premisa es «muy poco plausible intuitivamente». Y continúa diciendo que «... las teorías materialistas de la identidad han disciplinado la imaginación para que se figure el contenido de un evento mental como algo que varía mientras el evento permanece fijo. Pero si esas figuraciones son episodios posibles o simples quimeras filosóficas es ya otra cuestión». Debido a que considera muy improbable la negación de la premisa, sostiene que las «teorías materialistas de la identidad» han sido «desprovistas de plausibilidad por los experimentos mentales no individualistas».[25]

Acepto la premisa de Burge; considero su negación no meramente implausible, sino absurda. Si dos eventos mentales tienen contenidos diferentes son sin duda eventos diferentes. Lo que creo que muestran los casos imaginados de Burge y Putnam (y lo que creo que muestra más directamente el ejemplo del Hombre de los Pantanos) es que las personas que son similares (o «idénticas» en el sentido de las corbatas) en todos los aspectos físicos relevantes pueden diferir en lo que piensan o quieren decir, del mismo modo que pueden diferir en ser abuelos o en tener quemaduras del sol. Pero desde luego hay *algo* que las distingue, incluso en el mundo físico; sus historias causales son diferentes.

Mi conclusión es que el mero hecho de que los estados y eventos mentales cotidianos se individúen en términos de relaciones con el mundo externo no conlleva el descrédito de las teorías de la identidad psicofísica como tales. En conjunción con cierto número de supuestos (plausibles) adicionales, el «externalismo» de ciertos estados y eventos mentales puede usarse, creo, para desacreditar las teorías de la identidad tipo-tipo; pero sirve de apoyo, si acaso, a las teorías de la identidad caso-caso. (No veo ninguna buena razón para llamar «materialistas» a todas las teorías de la identidad; si

25. «Individualism and Psychology», pág. 15, nota 7. Véase «Individualism and the Mental», pág. 111.

algunos eventos mentales son eventos físicos, esto no los hace más físicos que mentales. La identidad es una relación simétrica.)

Putnam y Woodfield están equivocados, pues, al afirmar que es «absurdo» pensar que dos personas podrían ser físicamente idénticas (en el sentido de las «corbatas») y diferir sin embargo en sus estados psicológicos ordinarios. Burge, por su parte, a menos que desee jugar más fuerte con supuestos esencialistas, yerra al pensar que ha mostrado la implausibilidad de todas las teorías de la identidad. Por lo tanto, tenemos libertad para sostener que las personas pueden ser idénticas en todos los aspectos físicos relevantes al tiempo que difieren desde el punto de vista psicológico: ésta es de hecho la posición del «monismo anómalo», en favor del cual ha argüido en otro lugar.[26]

Ya ha sido apartado un obstáculo para el conocimiento no evidencial de nuestras actitudes proposicionales cotidianas, porque si las creencias y las demás actitudes pueden estar «en la cabeza» aun cuando en parte sean identificadas como tales en términos de lo que no está en la cabeza, la amenaza a la autoridad de la primera persona no puede entonces venir simplemente del hecho de que los factores externos sean relevantes para la identificación de las actitudes.

Sin embargo, sigue en pie una aparente dificultad. Es verdad que mi quemadura del sol, aunque sólo pueda describirse como tal en relación con el sol, es idéntica a una condición de mi piel que (supongo) puede describirse sin referencia a tales factores «externos». Sin embargo, si, como científico experto en todas las ciencias físicas, tengo únicamente acceso a mi piel, y se me niega el conocimiento de la historia de su condición, entonces, por hipótesis, no tengo forma de saber que estoy quemado por el sol. Quizá, pues, un sujeto tiene autoridad de primera persona con respecto a los contenidos de su mente sólo en cuanto que esos contenidos pueden describirse o ser descubiertos sin referencia a

26. «Mental Events», en Donald Davidson, *Essays on Actions and Events*, Oxford University Press, 1982.

factores externos. En la medida en que los contenidos se identifican en términos de factores externos, la autoridad de la primera persona desaparece necesariamente. Examinando mi piel, puedo decir cuál es mi condición privada o «estrecha», pero nada de lo que pueda aprender en este marco restringido me dirá que estoy quemado por el sol. La diferencia entre referirse al agua y pensar en ella y referirse a la gagua y pensar en ella es semejante a la diferencia entre estar quemado por el sol y el hecho de que mi piel se halle exactamente en la misma condición debido a una causa distinta. La diferencia semántica se halla en el mundo exterior, lejos del alcance del conocimiento subjetivo o sublunar. Así es como podría discurrir el argumento.

Esta analogía entre la visión limitada del dermatólogo y la visión tubular del ojo de la mente tiene un defecto fundamental. Su atractivo depende de una imagen incorrecta de la mente, una imagen que comparten los que han atacado el carácter subjetivo de los estados psicológicos ordinarios y los que han sufrido este ataque. Si logramos ser capaces de abandonar esta imagen, la autoridad de la primera persona dejará de aparecer como un problema; en realidad, resultará que dicha autoridad depende de –y es explicada por– los factores sociales y públicos que supuestamente la socavaban.

Hay una imagen de la mente que ha llegado a estar tan arraigada en nuestra tradición filosófica que resulta casi imposible escapar a su influencia, aun cuando se reconozcan y repudien sus peores defectos. En una versión tosca, pero familiar, la imagen en cuestión es la siguiente: la mente es un teatro en el que el yo consciente contempla un espectáculo cambiante (las sombras en el muro). El espectáculo consiste en «apariencias», datos sensoriales, *qualia*, lo dado en la experiencia. Lo que aparece en el escenario no son los objetos ordinarios del mundo que el ojo exterior registra y el corazón ama, sino sus pretendidos representantes. Todo lo que sabemos del mundo exterior depende de lo que podemos espigar a partir de las claves interiores.

La dificultad manifiesta que desde el principio ha presentado esta descripción de lo mental consiste en ver cómo

es posible abrir un camino desde el interior hasta el exterior. Otra dificultad conspicua, aunque quizá menos reconocida, es la de localizar el yo en el cuadro. En efecto, el yo, por un lado, parece incluir el teatro, el escenario, los actores y la audiencia; por otro lado, lo que es conocido y registrado tiene que ver únicamente con la audiencia. Este segundo problema podría también formularse en términos de la localización de los objetos de la mente: ¿están *en* la mente o simplemente son contemplados *por* ella?

No me ocuparé ahora de esos objetos de la mente (ampliamente repudiados hoy) conocidos como datos sensoriales, sino de sus parientes en el plano del juicio, ya se conciban como proposiciones, casos de proposiciones, representaciones o fragmentos de «mentalés».* El punto central que quisiera atacar es la idea según la cual hay entidades que la mente puede «acoger», «aprehender», «tener ante sí» o «conocer directamente». (Estas metáforas podrían ser instructivas: los *voyeurs* simplemente quieren tener representaciones ante el ojo de la mente, mientras que los más agresivos las aprehenden; los ingleses pueden simplemente conocer directamente los contenidos de la mente, mientras que tipos más cordiales realmente los acogerán.)

Es fácil ver cómo resulta perturbada la imagen de la mente que acabo de describir por el descubrimiento de que los hechos externos intervienen en la individuación de los estados mentales. En efecto, si hallarse en cierto estado mental consiste en que la mente guarda cierta relación con un objeto, como por ejemplo la de aprehenderlo, entonces todo lo que ayude a determinar de qué objeto se trata tiene que ser igualmente aprehendido si es que la mente ha de saber en qué estado se halla. Esto resulta particularmente evidente cuando un objeto externo es un «ingrediente» del objeto que

* El «mentalés» («mentalese» en inglés) es, en la jerga de la ciencia cognitiva, el nombre con el que se suele designar el supuesto «lenguaje del pensamiento» en el que, según la hipótesis de Jerry Fodor, inspirada en Chomsky, se desarrolla realmente la vida mental y que subyace al lenguaje público como un código universal. (T.)

se encuentra ante la mente. Pero, en cualquier caso, la persona que se halla en el correspondiente estado mental puede no saber cuál es ese estado mental en el que se halla.

Es en este punto donde el concepto de lo subjetivo –de un estado de la mente– puede disgregarse. Tenemos, por un lado, los auténticos estados internos, sobre los que la mente mantiene su autoridad, y, por el otro, los estados ordinarios de creencia, deseo, intención y significado, que se encuentran contaminados por sus conexiones necesarias con el mundo social y público.

Por analogía, citemos el problema del experto en quemaduras del sol que no es capaz de decir, mediante una inspección de la piel, si se trata de un caso de eritema solar o simplemente de una condición idéntica con una causa distinta. Podemos resolver este problema distinguiendo entre eritema solar y eritema soleado; el eritema soleado es exactamente como el eritema solar con la excepción de que el sol no tiene por qué intervenir. El experto puede descubrir un caso de eritema soleado con una simple mirada, pero no un caso de eritema solar. Esta solución funciona porque, con respecto a las cualidades de la piel, a diferencia de los objetos de la mente, se requiere que haya alguien especial, capaz de decir, con una simple mirada, si cierta cualidad se da o no.

En el caso de los estados mentales, la solución es diferente y más simple; consiste en deshacerse de la metáfora de los objetos ante la mente. Hace tiempo que la mayoría de nosotros abandonamos la idea de las percepciones, los datos sensoriales, el flujo de la experiencia, como cosas «dadas» a la mente; deberíamos tratar los objetos proposicionales del mismo modo. Las personas, desde luego, tienen creencias, deseos, dudas, etcétera; pero admitir esto no equivale a sugerir que las creencias, deseos y dudas son *entidades* en o ante la mente, o que hallarse en tales estados requiera la existencia de los objetos mentales correspondientes.

Esto ya ha sido dicho anteriormente, en distintos tonos de voz, pero por razones diferentes. Los escrúpulos ontológicos, por ejemplo, no son parte de mi interés. Siempre necesi-

taremos un surtido infinito de objetos que nos ayuden a describir e identificar actitudes como la creencia; no estoy sugiriendo ni por un momento que las oraciones de creencia y las que atribuyen las demás actitudes no sean de naturaleza relacional. Lo que sugiero es que los objetos con los que relacionamos a las personas para describir sus actitudes no tienen por qué ser, en ningún sentido, objetos *psicológicos*, objetos que sean aprehendidos, conocidos o considerados por la persona cuyas actitudes se describen.

También esta observación es conocida; Quine la hace cuando sugiere que podemos usar nuestras propias oraciones para estar al corriente de los pensamientos de personas que no conocen nuestro lenguaje. El interés de Quine es semántico, y no dice nada en este contexto acerca de los aspectos epistemológicos y psicológicos de las actitudes. Es necesario reunir estas diversas perspectivas. Las oraciones acerca de las actitudes son relacionales; por razones *semánticas* ha de haber, por tanto, objetos con los cuales poner en relación a aquellos que tienen actitudes. Pero tener una actitud no es tener una entidad ante la mente; por imperiosas razones *psicológicas* y *epistemológicas* debemos negar que haya objetos de la mente.

El origen de la dificultad es el dogma según el cual tener un pensamiento es tener un objeto ante la mente. Putnam y Fodor (y muchos otros) han distinguido dos clases de objetos: aquellos que son verdaderamente internos y están así «ante la mente» o son «aprehendidos» por ella, y aquellos que identifican el pensamiento de la forma usual. Convengo en que no hay objetos que puedan satisfacer ambos propósitos. Putnam (y algunos de entre los demás filósofos que he mencionado) piensa que la dificultad nace del hecho de que no cabe contar con un objeto identificado en parte en términos de relaciones externas para hacer que coincida con un objeto ante la mente, porque la mente puede ignorar la relación externa. Puede que sea así. Pero de ello no se sigue que podamos encontrar *otros* objetos con los cuales asegurar la coincidencia deseada, pues si el objeto *no está* conectado con el mundo, nunca podremos aprender algo del mundo te-

niendo ese objeto ante la mente; y, por razones recíprocas, sería imposible detectar semejantes pensamientos en otra persona. Parece, pues, que lo que se halla ante la mente no puede incluir sus conexiones externas –su semántica–. Por otra parte, si el objeto *está* conectado con el mundo, no puede estar plenamente «ante la mente» en el sentido relevante. Ahora bien, a menos que un objeto *semántico* pueda estar ante la mente *en su aspecto semántico*, el pensamiento, concebido en términos de tales objetos, no puede escapar a la fatalidad de los datos sensoriales.

La dificultad básica es simple: si tener un pensamiento es tener un objeto «ante la mente» y la identidad del objeto determina de qué pensamiento se trata, siempre habrá de ser posible estar equivocado acerca de lo que uno piensa, pues, a menos que lo sepa *todo* sobre el objeto, siempre habrá algún sentido en que no sepa qué objeto es. Ha habido muchos intentos de encontrar, entre una persona y un objeto, una relación que se mantenga en todos los contextos si, y sólo si, se puede decir intuitivamente que la persona sabe què objeto es. Pero ninguno de estos intentos ha tenido éxito, y creo que la razón es clara. El único objeto que cumpliría los requisitos gemelos de estar «ante la mente» y de determinar también cuál es el contenido de un pensamiento tendría que, como las ideas e impresiones de Hume, «ser lo que parece y parecer lo que es». No hay objetos semejantes, ni públicos ni privados, ni abstractos ni concretos.

Los argumentos que Burge, Putnam, Dennett, Fodor, Stich, Kaplan, Evans y muchos otros han desarrollado con vistas a mostrar que las proposiciones no pueden *a la vez* determinar los contenidos de nuestros pensamientos *y* estar subjetivamente garantizadas son, en mi opinión, otras tantas variantes del argumento simple y general que acabo de esbozar. No son solamente las proposiciones las que no pueden desempeñar ese trabajo; ningún objeto podría hacerlo.

Una vez que nos hemos liberado del supuesto de que los pensamientos han de tener misteriosos objetos, podemos ya apreciar que el hecho de que los estados mentales, tal como los concebimos comùnmente, se identifiquen en parte por su

historia natural no sólo no afecta al carácter interno de tales estados ni amenaza la autoridad de la primera persona, sino que abre también el camino a una explicación de dicha autoridad. La explicación se vincula a la comprensión de que el significado de las palabras de una persona depende, en los casos más básicos, de los tipos de objetos y eventos que han causado que la persona considere aplicables esas palabras; y algo similar sucede con aquello de que versan sus pensamientos. Un intérprete de las palabras y pensamientos de otra persona tiene que depender, en su tarea de comprender a ésta, de una información dispersa, un entrenamiento afortunado y un conjeturar imaginativo. La agente misma,* sin embargo, no está en posición de preguntarse si en general usa sus propias palabras aplicándolas a los objetos y eventos correctos, ya que aquello a lo que regularmente las aplica, sea lo que sea, da a sus palabras el significado que poseen y a sus pensamientos el contenido que tienen. Desde luego puede estar equivocada, en un caso particular, en lo que cree del mundo; pero lo que resulta imposible es que esté equivocada la mayor parte del tiempo. La razón es manifiesta: a menos que haya una presunción según la cual la hablante sabe lo que quiere decir, esto es, entiende con claridad su propio lenguaje, no habría nada que un intérprete pudiera interpretar. Por decirlo de otro modo, nada podría valer como el hecho de que una hablante aplicase de forma sistemáticamente errónea sus propias palabras. La autoridad de la primera persona, el carácter social del lenguaje y los determinantes externos del pensamiento y el significado se combinan entre sí de modo natural una vez que abandonamos el mito de lo subjetivo, la idea de que los pensamientos requieren objetos mentales.

* Véase N. del T., pág. 123.

Nota: Estoy en deuda con Akeel Bilgrami y Ernie LePore por sus críticas y consejos. Tyler Burge trató generosamente de corregir mi comprensión de su obra.

LAS CONDICIONES DEL PENSAMIENTO*

¿Cuáles son las condiciones necesarias para la existencia del pensamiento y con ello, en particular, para la existencia de personas que tengan pensamientos? Creo que no podría haber pensamientos en una mente si no hubiese otras criaturas pensantes con las que dicha mente compartiese un mundo natural. Por pensamiento entiendo un estado mental con un contenido especificable. He aquí algunos ejemplos: la creencia de que esto es un trozo de papel; la intención de hablar despacio y con claridad; la duda sobre si mañana será un día soleado. Es natural suponer que pensamientos como éstos no dependen de nada exterior a la mente; que podrían ser exactamente como son aunque el mundo fuese muy diferente. Es natural pensar tal cosa porque parece obvio que cualquier pensamiento particular acerca de la naturaleza del mundo puede ser erróneo, y de esto parece seguirse que todos los pensamientos de ese tipo podrían serlo también. Los únicos pensamientos que escapan a ese escepticismo preliminar y primitivo son los que versan sobre nuestros propios pensamientos; esos pensamientos son privilegiados porque la fuente de la duda –la posibilidad de que algo exterior a la mente pueda no existir– ha sido eliminada.

Algo semejante a esta línea de razonamiento explica, como todos sabemos, por qué una parte tan amplia de la filosofía occidental se ha sentido obligada a partir de un punto de vista solipsista o de primera persona. También explica

* Este trabajo fue presentado el 23 de agosto de 1988 en la sesión plenaria sobre «Los seres humanos: naturaleza, mente y comunidad» del Congreso Mundial de Filosofía celebrado en Brighton, y fue escrito originalmente para dicha sesión.

el hecho, que de otro modo podría ser misterioso, de que el conocimiento de otras mentes se haya presentado como un problema añadido al del conocimiento empírico. En efecto, si los contenidos de una mente son lógicamente independientes de cualquier otra cosa, esto crea dos problemas distinguibles: cómo puede la mente conocer lo que está separado de ella y cómo puede esto último conocerla a ella. Si yo no puedo ver el exterior, tampoco puede otro (si es que hay algún otro) ver el interior.

Cierto número de filósofos han creído saber, posiblemente bajo la influencia de Wittgenstein, cómo dar respuesta al segundo problema, es decir, el conocimiento que una persona tiene de la mente de otra. La solución, a grandes rasgos, discurría del modo que sigue. Hemos de admitir que hay una diferencia en el modo en que conocemos lo que está en nuestra propia mente y lo que está en la mente de otros; en el primer caso, normalmente no necesitamos evidencia o no la empleamos, mientras que en el segundo caso hemos de observar la conducta, incluida la conducta verbal. Pero esto en sí mismo no plantea problemas. Si comprendemos qué son los estados mentales, sabemos entonces que encierran esta anomalía: a diferencia de casi todo otro tipo de conocimiento, el conocimiento de los estados mentales se caracteriza por el hecho de que su base adecuada es la observación de la conducta cuando dichos estados no son los nuestros, mientras que, cuando se trata de nuestros propios estados, no tiene (normalmente) como base la observación o la evidencia.

Como descripción del modo en que empleamos los conceptos y términos mentales, esto es (en mi opinión) correcto. Pero lo que estos filósofos no advirtieron es que una descripción de nuestra práctica no es una solución al problema original, sino una redescripción de aquello que crea dicho problema. Nuestra práctica nunca se puso en duda; la duda se refería a su legitimidad y ésta se enfrenta a dos cuestiones. La primera es que resulta difícil comprender, a falta de una explicación, por qué el conocimiento que no se basa en la evidencia habría de ser más cierto que aquel que se basa en

ella. La segunda consiste en que, si aquello que parece ser un solo concepto o predicado se aplica correctamente usando dos tipos de criterios muy diferentes (o, en uno de los casos, sin usar ningún criterio en absoluto), no tenemos en tal caso razones para suponer que se trata realmente de un solo concepto. Aparentemente deberíamos concluir que el predicado es ambiguo y que en realidad hay dos conceptos. Este es una vez más, después de todo, el mismo viejo problema: ¿por qué habría de creer una persona que alguien más tiene estados mentales como los suyos? O, por plantear el problema en sentido inverso, ¿por qué habría yo de pensar que tengo estados mentales como aquellos que detecto en otros?

Dejemos a un lado estos problemas por un momento, especialmente porque no voy a poder darles un tratamiento adecuado en este trabajo. Habiendo aceptado, por el momento, lo que podemos llamar la actitud del observador, o de la tercera persona, hacia otras mentes, la siguiente pregunta es cómo puede una persona determinar lo que está en otra mente. La respuesta completa es, sin duda, muy complicada, pero una parte básica de la misma tendría que depender, creo, del hecho de que, en los casos más simples, los eventos y objetos que causan una creencia determinan también los contenidos de la misma. Así, la creencia que es causada, distintivamente y en condiciones normales, por la presencia evidente de algo amarillo, de la propia madre o de un tomate, es la creencia de que algo amarillo, la propia madre o un tomate están presentes. La idea no es, desde luego, que la naturaleza garantiza que nuestros juicios más llanos sean siempre correctos, sino que la historia causal de tales juicios representa un importante rasgo constitutivo de sus contenidos.

En años recientes se ha argüido largamente en favor de esta tesis, que puede denominarse «externalismo», recurriendo usualmente a elaborados experimentos mentales que requieren que se evalúe la verdad de supuestos contrafácticos bastante extremos. Creo que el principio subyacente a esta tesis es a la vez más simple y más universal en su explicación que lo que esos argumentos revelan; no tenemos más

que reflexionar sobre el modo en que se aprenden los significados de las primeras y más básicas palabras y oraciones y sobre la relación obvia entre lo que significan nuestras oraciones y los pensamientos que expresamos al usarlas.

El externalismo pone en claro cómo puede una persona llegar a saber lo que otra piensa, al menos en un nivel básico, pues un intérprete, al descubrir lo que normalmente causa las creencias de otro sujeto, ha dado un paso esencial en la determinación del contenido de esas creencias. No es fácil concebir de qué otro modo sería posible descubrir lo que otro piensa. (No deseo dar la impresión de que este proceso es simple y ciertamente no deseo sugerir que el esbozo altamente esquemático que he ofrecido contenga ya una primera respuesta a las preguntas que requieren contestación para que el cuadro pueda ser completado. Mi meta es aquí solamente sugerir la naturaleza de un externalismo aceptable e indicar algunas de sus consecuencias.) Al reflexionar sobre la forma en que el externalismo opera en la interpretación, podemos explicar en parte la asimetría entre el conocimiento de los pensamientos en primera y en tercera persona. Así, mientras que el intérprete ha de conocer, o conjeturar correctamente, los eventos y situaciones que causan una reacción verbal o de otro tipo en otra persona con vistas a penetrar en sus pensamientos, el sujeto de estos pensamientos no necesita de un conocimiento nómico semejante para decidir lo que él mismo piensa. La historia causal determina en parte lo que piensa, pero esta determinación es independiente de cualquier conocimiento que él pueda tener de dicha historia causal.

Si el externalismo es verdadero, no puede haber una pregunta general adicional acerca de cómo es posible el conocimiento del mundo externo. Si es constitutivo de algunos pensamientos que su contenido venga dado por su causa normal, entonces el conocimiento de los eventos y situaciones causantes no puede requerir que un pensador, de forma independiente, establezca –o halle evidencia en favor de– la hipótesis de que hay un mundo externo que está causando estos pensamientos. Desde luego, el externalismo no mues-

tra que un juicio perceptivo particular, incluso del tipo más simple, no pueda ser erróneo. Lo que muestra es por qué no puede ocurrir que la mayoría de tales juicios sean erróneos, ya que el contenido de los juicios erróneos ha de descansar sobre el de los juicios correctos.

He estado suponiendo que el externalismo es verdadero: que podemos adoptar legítimamente el punto de vista de la tercera persona al considerar la naturaleza del pensamiento. Pero imaginemos que fuéramos a asumir, en lugar de ello, el punto de vista solipsista o de primera persona. ¿Habría entonces alguna razón para aceptar el externalismo? Creo que la habría. Abordaré la cuestión suscitando una dificultad para el externalismo.

Consideremos, en primer lugar, una situación primitiva de aprendizaje. Cierta criatura es educada, o aprende de algún modo, a responder de forma específica a un estímulo o a cierta clase de ellos. El perro oye un timbre y es alimentado; muy pronto saliva cuando oye el timbre. El niño balbucea y cuando emite un sonido como «mesa» en presencia de mesas es recompensado distintivamente; muy pronto el niño dice «mesa» en presencia de mesas. El fenómeno de la generalización, de la similaridad percibida, desempeña un papel esencial en este proceso. Un toque del timbre se asemeja lo bastante a otro, desde el punto de vista del perro, para provocar una conducta similar, al igual que una presentación de alimento se asemeja lo bastante a otra para generar la salivación. Si algunos de estos mecanismos selectivos no fuesen congénitos, ninguno podría ser aprendido. Esto parece claro y sencillo, pero, como los filósofos han advertido, existe un problema concerniente a la localización del estímulo. En el caso del perro, ¿por qué decir que el estímulo es el sonido del timbre? ¿Por qué no el movimiento del aire cercano a las orejas del perro, o incluso la estimulación de sus terminaciones nerviosas? Ciertamente, si se hiciera vibrar el aire del mismo modo en que lo hace vibrar el timbre, ello no supondría diferencia alguna en la conducta del perro. Y si las terminaciones nerviosas adecuadas fuesen activadas del modo apropiado, tampoco habría diferencia. Y de hecho, si hemos

de elegir, parece que la causa próxima de la conducta posee los mejores títulos para recibir la denominación de estímulo, pues cuanto más distante se halle un evento desde el punto de vista causal, tanta más probabilidad hay de que se rompa la cadena causal. Tal vez deberíamos decir lo mismo acerca del niño: su respuesta no obedece a las mesas, sino a pautas estimulativas en la superficie de su piel, puesto que estas pautas siempre producen la conducta, mientras que las mesas sólo la producen en condiciones favorables.

¿Por qué, sin embargo, parece natural decir que el perro responde al timbre y el niño a las mesas? Nos parece natural porque nos es natural. Así como el perro y el niño responden de formas similares a estímulos de cierta clase, así también lo hacemos nosotros. Somos nosotros quienes encontramos natural agrupar conjuntamente las distintas salivaciones del perro; y los eventos del mundo en los que reparamos y que agrupamos conjuntamente, vinculados causalmente a la conducta del perro, son los sonidos del timbre. Encontramos similares las emisiones de la palabra «mesa» que el niño lleva a cabo, y las cosas del mundo que acompañan a esas emisiones y que clasificamos juntas de forma natural son precisamente mesas. No podemos observar fácilmente las pautas acústicas y visuales que fluyen rápidamente, a distinta velocidad, entre el timbre y los oídos del perro, entre las mesas y los ojos del niño, y si pudiéramos observarlas nos sería muy difícil decir qué es lo que las hacía similares. (Excepto si hacemos trampa, desde luego: son las pautas características del sonido de los timbres o de las mesas vistas.) Igualmente, tampoco observamos la estimulación de las terminaciones nerviosas de otras personas y animales y, si lo hiciéramos, probablemente encontraríamos imposible describir de forma no circular qué es lo que hacía esas pautas relevantemente similares de un caso a otro. El problema sería en gran medida el mismo, y no menos imposible de resolver, que el de definir mesas y sonidos de timbres en términos de datos sensoriales.

En nuestra descripción están involucradas no dos, sino tres clases de eventos u objetos entre cuyos miembros tanto

nosotros como el niño hallamos una similitud natural. El niño encuentra las mesas relevantemente similares; nosotros también hallamos las mesas similares; y encontramos también similares las respuestas del niño a las mesas. Dadas estas tres pautas de respuesta, resulta posible localizar los estímulos relevantes que promueven las respuestas del niño. Se trata de objetos o eventos que encontramos naturalmente similares (mesas) y que guardan correlación con respuestas del niño que encontramos similares. Es una forma de triangulación: una línea parte del niño en dirección a la mesa, otra línea parte de nosotros en dirección a la mesa y la tercera va de nosotros al niño. El estímulo relevante se halla allí donde convergen las líneas del niño a la mesa y de nosotros a la mesa.

Tenemos ya ante nosotros los rasgos suficientes para dar un significado a la idea de que el estímulo tiene una localización objetiva en un espacio común; la cuestión consiste en dos perspectivas privadas que convergen para marcar una posición en el espacio intersubjetivo. Hasta ahora, sin embargo, nada en esta descripción muestra que nosotros, los observadores, o nuestros sujetos, el perro y el niño, tengamos el concepto de lo objetivo.

No obstante, hemos hecho progresos, pues, si estoy en lo cierto, el tipo de triangulación que he descrito, aunque obviamente no es suficiente para establecer que una criatura tiene un concepto de un objeto particular o de un tipo de objeto, es sin embargo necesaria si ha de haber siquiera una respuesta a la pregunta de qué es aquello de lo que sus conceptos son tales. Si consideramos una criatura aislada por sí misma, sus respuestas, por complejas que sean, no pueden mostrar que está reaccionando a, o pensando en, eventos situados a cierta distancia en lugar de, digamos, sobre su piel. El mundo del solipsista puede tener cualquier dimensión, lo que equivale a decir que no tiene dimensión alguna, que no es un mundo.

Debo insistir en que el problema no consiste en verificar a qué objetos o eventos responde una criatura; el problema reside más bien en que, sin una segunda criatura que inte-

ractúe con la primera, no puede haber respuesta a esa pregunta. Y si no puede haber respuesta a la pregunta sobre aquello que una criatura quiere decir, desea, cree o pretende, no tiene sentido sostener que esa criatura tiene pensamientos. Podemos, pues, decir, como preámbulo de la respuesta a la pregunta con la que comenzamos, que, antes de que cualquiera pueda tener pensamientos, ha de haber otras criaturas (una o más) que interactúen con el hablante. Pero, desde luego, esto no puede ser suficiente, puesto que la simple interacción no muestra de qué modo importa esa interacción a las criaturas involucradas. A menos que pueda decirse que las criaturas afectadas reaccionan a la interacción, no hay forma de obtener ventajas cognitivas de la triple relación que da contenido a la idea de que reaccionan mejor a una que a otra.

He aquí, pues, parte de lo requerido. La interacción ha de hacerse accesible a las criaturas involucradas en ella. Así, el niño, al aprender la palabra «mesa», ya ha advertido efectivamente que las respuestas del educador son similares (remuneradoras) cuando sus propias respuestas (emisiones de la palabra «mesa») son similares. El educador, por su parte, está adiestrando al niño para que responda de forma similar a lo que él (el educador) percibe como estímulos similares. Para que esto funcione, es claro que las respuestas innatas del niño y el educador a la similitud –aquello que de forma natural agrupan conjuntamente– han de ser muy parecidas, pues en otro caso el niño responderá a lo que el educador considera como estímulos similares de maneras que el educador no encuentra similares. Una condición para ser un hablante o un intérprete es que ha de haber otros que se parezcan lo suficiente a uno mismo.

Vamos ahora a reunir las dos observaciones. En primer lugar, si alguien tiene pensamientos, ha de haber otro ser sensitivo cuyas respuestas innatas a la similitud se parezcan lo bastante a las suyas para proporcionarle una respuesta a la siguiente pregunta: ¿cuál es el estímulo al que está respondiendo? Y, en segundo lugar, si las respuestas de alguien han de valer como pensamientos, han de tener el concepto

de un objeto; el concepto del estímulo: del timbre o de mesa. Puesto que el timbre o una mesa se identifican sólo por la intersección de dos (o más) conjuntos de respuestas a la similitud (líneas de pensamiento, podríamos casi decir), tener el concepto de una mesa o de un timbre es reconocer la existencia de un triángulo uno de cuyos vértices es uno mismo, otro una criatura similar a uno mismo y el tercero un objeto o evento (mesa o timbre) localizado en un espacio que se convierte así en común.

La única forma de saber que el segundo vértice, la segunda criatura o persona, reacciona al mismo objeto que uno mismo es saber que esa otra persona tiene en mente el mismo objeto. Pero entonces la segunda persona ha de saber también que la primera constituye un vértice del mismo triángulo otro de cuyos vértices es ocupado por él mismo. Para que dos personas sepan la una de la otra que se hallan en esa relación, que sus pensamientos se relacionan de ese modo, es necesario que estén en comunicación. Cada una de ellas ha de hablar a la otra y ser entendida por ella.

Si estoy en lo cierto, la creencia, la intención y las demás actitudes proposicionales son de carácter social en cuanto que dependen de la posesión del concepto de verdad objetiva. Este es un concepto que no se puede tener sin compartirlo con alguien más, y saber que se comparte con él, un mundo y una forma de pensar sobre el mismo.

Nota: Este conciso artículo se inspira en trabajos recientes del autor en los que puede hallarse un tratamiento más extenso de las tesis principales. Véase: «Rational Animals», *Dialectica*, 36 (1982), págs. 317-327; «First Person Authority», *Dialectica*, 38 (1984), págs. 101-111; «A Coherence Theory of Truth and Knowledge», en *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, E. LePore, comp., Blackwell, 1986; «Knowing One's Own Mind», *Proceedings and Addresses of the American Philosophical Association*, 1987, págs. 441-458. [Los dos últimos artículos están traducidos en el presente volumen bajo los títulos respectivos de «Verdad y conocimiento: una teoría de la coherencia» y «El conocimiento de la propia mente». T.]