UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE FILOSOFÍA Y LETRAS

INSTITUTO DE INVESTIGACIONES FILOSÓFICAS

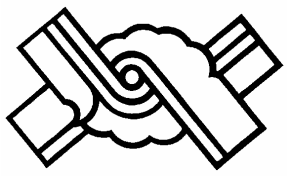
Autoengaño e irracionalidad

T E S I S

QUE PARA OBTENER
EL TÍTULO DE
MAESTRO EN FILOSOFÍA
P R E S E N T A

Roberto Parra Dorantes

MÉXICO, D.F. 2008



Agradecimientos

Doy gracias, en primer lugar, a mi asesor, Mark Platts, quien excedió con mucho el cumplimiento de su deber como director de esta tesis. Me siento muy afortunado de haber contado con su apoyo a lo largo de estos años de formación. Agradezco su infinita paciencia, su incansable rigor y su inagotable sentido del humor. Doy gracias también a los catedráticos de los cursos de filosofía que tomé en la UNAM, especialmente a Gustavo Ortiz Millán, Héctor Islas, Faviola Rivera, Olbeth Hansberg y Héctor Zagal. Todos ellos contribuyeron de manera sustancial a mi desarrollo filosófico.

Agradezco también al Instituto de Investigaciones Filosóficas por haberme dado la oportunidad de participar durante tres años en el programa de Estudiantes Asociados, y al Consejo Nacional de Ciencia y Tecnología (Conacyt) por el apoyo económico que recibí durante mis estudios de maestría.

Es importante para mí reconocer aquí también a aquellas personas que influyeron en mi decisión de perseguir la filosofía como una profesión: Ramón Miranda, Rogelio Larios, Guillermo Lariguet y Pablo Perot. Ellos no sólo despertaron mi interés por la filosofía sino que también me alentaron a emprender este camino.

Por último, quiero agradecer a mi familia y a mis amigos. A mi madre, a mi hermano Ramón, y a la memoria de mi padre y mi hermana Liz, quienes siempre han estado ahí para mí y a quienes esta tesis está dedicada. Y a Azuvia, por el infinito cariño y apoyo que me ha brindado desde el principio. Gracias a todos ustedes.

The life of the mind... There's no roadmap for that territory.

- Joel & Ethan Cohen, *Barton Fink*

Índice

Introducción

Esta tesis, dividida en tres capítulos, intenta desarrollar una explicación original del fenómeno del autoengaño. En el primer capítulo, elaboro un esquema general para explicar la irracionalidad práctica, basándome en gran medida en la teoría de la acción de Donald Davidson. Dedico una parte de este capítulo a la presentación y discusión de los puntos centrales de dicha teoría de la acción. La explicación de la acción irracional, para Davidson, amenaza con conducirnos a dos paradojas a las cuales debe hacer frente la teoría de la acción. A partir de la disolución de tales paradojas que, según Davidson, parece presentar la explicación de la irracionalidad, presento una distinción propia entre razones aléticas (basadas en un interés por la verdad) y razones oréticas (basadas en algún otro interés práctico).[*] Esta distinción, según lo que planteo, es aplicable por igual a las razones para actuar, para creer y para desear. Por último, utilizando dicha distinción, propongo un criterio para caracterizar y evaluar la racionalidad (o irracionalidad) de acciones, creencias y deseos.

En el segundo capítulo, analizo tres de las propuestas contemporáneas más importantes para explicar el autoengaño: la de Donald Davidson, la de Alfred Mele y la de Robert Audi. La explicación de Donald Davidson, que postula barreras o particiones mentales en el agente, se enfrenta a ciertas objeciones que Davidson no resuelve; sin embargo, según digo, tiene la virtud de identificar los casos de autoengaño filosóficamente más problemáticos. La postura de Alfred Mele para explicar el autoengaño, que en los últimos años se ha perfilado como la de mayor popularidad entre

[*] Estando ya muy cerca de concluir esta tesis, encontré que H. P. Grice utiliza en su obra póstuma *Apects of Reason* (Oxford, Clarendon Press, 2001) las frases "alethic reasons" y "practical reasons" para referirse a una distinción similar a la que yo manejo aquí de manera independiente.

los teóricos del autoengaño por resolver, en apariencia, las paradojas relacionadas con el autoengaño y explicar este fenómeno sin necesidad de recurrir a estrategias demasiado sofisticadas, no encara, según argumento, aquellos casos bien identificados por Davidson que plantean los problemas más serios a la explicación filosófica del autoengaño. La explicación de Audi, por otra parte, sí hace frente a dichos casos de autoengaño utilizando una estrategia que consiste en distinguir entre diversos órdenes de creencias del agente; sin embargo, no arroja suficiente luz sobre los problemas de cuáles son los estados mentales que guían al agente a través del proceso del autoengaño y los mecanismos psicológicos que hacen posible este proceso.

Después de analizar y ofrecer críticas a cada una de estas propuestas, finalmente presento una explicación propia del autoengaño que retoma varios elementos de las tres propuestas analizadas, principalmente de la de Robert Audi. Esta explicación se enfoca en los casos más problemáticos de autoengaño, donde la característica principal de estos casos es que una creencia inicial verdadera parece causar y mantener en existencia la creencia falsa resultante del autoengaño. Argumento, siguiendo la distinción jerárquica entre creencias propuesta por Audi, que, en dichos casos, la creencia inicial verdadera puede coexistir con y mantener en existencia a la creencia falsa inducida por el autoengaño sin que éstas sean contradictorias, por la razón de que la creencia inicial (verdadera) es una creencia de primer orden, mientras que la creencia final (falsa) es una creencia de segundo orden. Esta explicación evade exitosamente los peligros planteados por las paradojas del autoengaño y logra al mismo tiempo resolver los problemas planteados por los casos más difíciles de este fenómeno. Sostengo también que,

considerado de esta manera, el autoengaño resulta ser un caso especial del fenómeno, aparentemente más sencillo, de creencias ilusorias [*wishful thinking*].

En el tercero y último capítulo trato el tema del autoengaño en relación con los temas de responsabilidad y autonomía. A pesar de que, según mi explicación del autoengaño, este fenómeno no sucede a partir de una intención específica de engañarse del propio agente, argumento que el agente autoengañado es responsable de su autoengaño y de las consecuencias de éste, pues existen muchas cosas que un agente pudo haber hecho para evitar el autoengaño, o mejor dicho, para disminuir su propia tendencia hacia el autoengaño. Finalmente, ante la sugerencia de que el autoengaño puede en, ciertos casos, ser considerado como una manifestación del ejercicio de la autonomía del agente, argumento que el autoengaño en ningún caso se encuentra bajo el control directo del agente, por lo cual no puede ser considerado como una manifestación de autonomía, y que el autoengaño y la propensión a éste, al disminuir la medida en la que el agente tiene control sobre el mundo, disminuyen considerablemente la autonomía del agente.

Capítulo 1. Irracionalidad práctica

Las acciones intencionales y muchos estados mentales, incluyendo intenciones, deseos, emociones y creencias, son susceptibles de ser explicados por razones. Una explicación por razones es un tipo de explicación causal en la que se proporciona el motivo del agente —una combinación apropiada de deseos o actitudes favorables y creencias— como causa de la acción o el estado mental que es el efecto. Así, para muchas acciones o estados mentales que hayan sido causados directamente por otros estados mentales, es posible dar una explicación que cite una razón. Esto aclara por qué todas las acciones intencionales, y muchos estados mentales, poseen en su núcleo un elemento de racionalidad. Si es posible encontrar acciones o estados mentales que tuvieron en su origen una motivación que sean irracionales, es necesario explicar en qué sentido ulterior son irracionales, sin que por ello dejen de ser racionales en el primer sentido.[1]

Si las explicaciones por razones son aplicables a las creencias que se originan en otros estados mentales a través de la deliberación (excluyendo, por ejemplo, a las que se derivan directamente de la percepción), parece difícil ver cómo esas creencias pueden depender causalmente de otras creencias *y* deseos. Es claro que las creencias basadas en evidencia dependen de otras creencias, pero no es claro cómo podrían depender, conjuntamente, de deseos o actitudes favorables. Parecería necesario postular para todos estos casos un deseo apropiado en el agente a efecto de que éste realice la inferencia y llegue a la conclusión. Esto parece simplemente falso, pero, suponiendo que fuese cierto, ¿cuál podría ser ese deseo? Retomo esta pregunta al final de este capítulo.

[1] Sigo aquí las ideas de Davidson en "Paradoxes of Irrationality", p. 293 y ss., publicado en R.Wollheim and J. Hopkins (eds.): *Philosophical Essays on Freud*: pp. 289-305. Cambridge, Cambridge University Press, 1982.

Donald Davidson menciona dos principios que se han propuesto como explicaciones de la relación entre la acción y la racionalidad[2]. El primero es el que él llama el "principio Platón", según el cual ninguna acción intencional puede ser internamente irracional. Este principio deriva de la tesis socrática según la cual nadie puede actuar en contra de su propio mejor juicio, excluyendo así por completo la posibilidad de irracionalidad práctica. Este principio, por tanto, atribuye a toda acción intencional más que el elemento de racionalidad interno a cualquier acción antes explicado, pues, como ya se dijo, la posibilidad de que una acción sea explicada por razones no excluye la posibilidad de que esa misma acción sea, en algún otro sentido, irracional. Al segundo principio lo llama el "principio Medea", según el cual una persona puede actuar en contra de su propio mejor juicio —permitiendo así la posibilidad de la irracionalidad práctica— pero sólo cuando una fuerza ajena abruma su voluntad.

Según Davidson, los casos definitorios de la debilidad de la voluntad, sin embargo, no se explican por ninguno de estos dos principios. Las acciones que pueden explicarse por el principio Medea no son intencionales, y las que se explican según el análisis de Aristóteles de la *akrasia*, el cual sigue el principio Platón, son casos en los que el agente pierde "contacto activo" con una parte de su conocimiento acerca de los cursos posibles de acción, actuando sólo con base en un conocimiento parcial. En los casos de debilidad de la voluntad, el agente actúa intencionalmente y se da cuenta de que un mejor curso de acción, según sus propios estándares, está disponible. Si las acciones intencionales irracionales son posibles, debemos encontrar otra manera para explicarlas como irracionales. Davidson piensa que estas acciones son, en efecto, posibles. Pero, ¿cómo podemos explicar que alguien vaya en contra de su propio mejor juicio? La

[2] *Ibid.*, p. 294.

explicación, dice Davidson, debe ir más allá que el principio Platón (pues, si no, la acción será racional), pero debe conservar el núcleo de este principio (pues, si no, la acción no será intencional).

Con el fin de aclarar la postura de Davidson acerca de la irracionalidad práctica, me parece conveniente presentar un breve bosquejo de su teoría de la acción.[3] Para Davidson, una acción es un suceso que es causado por las razones del agente para hacer esa acción. Una razón para actuar consiste en una actitud favorable, valor o meta del agente (un deseo, en sentido muy amplio), y una creencia del agente acerca de que, actuando de cierta manera, promoverá la satisfacción de ese deseo. Entre razones y acción pueden encontrarse dos tipos de relaciones: una relación lógica, que muestra por qué cierta acción es deseable a la luz de las creencias y los deseos del agente, así racionalizando la acción, y una relación causal, que nos sirve para entender a las razones del agente como causas de sus acciones. Durante la deliberación práctica, diversos deseos y creencias son considerados por el agente como relevantes para la acción acerca de cuya realización el agente delibera. Estos deseos y creencias pueden constituir razones a favor o en contra de la realización de la acción. Comúnmente pueden encontrarse tanto razones a favor como en contra de cualquier acción acerca de la cual se delibere; esto no involucra al agente en un marco mental contradictorio, pues tales razones son sólo razones *prima facie* a favor o en contra de la realización de la acción. El agente, al deliberar, sopesa estas razones y luego forma un juicio práctico basado en las razones

[3] En lo que sigue, además de apoyarme en los textos relevantes de Davidson (como los influyentes artículos "Actions, Reasons and Causes" y "How is Weakness of the Will Possible?"), me baso en el ensayo de Ariela Lazar, "Akrasia and the Principle of Continence or What the Tortoise Would Say to Achilles", publicado en *The Philosophy of Donald Davidson*, Lewis Edwin Hahn (ed.), The Library of Living Philosophers, Open Court Publishing, Chicago, 1999, pp. 381-400, cuya interpretación de la teoría de la acción de Davidson me parece acertada.

consideradas. Éste es un juicio del tipo "hechas todas las consideraciones", o, para abreviar, juicio HTC, al cual el agente puede llegar de las siguientes dos maneras: idealmente, todas las razones relevantes del agente son consideradas al formar este juicio; éste sería el "juicio óptimo" (*best judgment*). Típicamente, sin embargo, el agente forma su juicio considerando sólo un subconjunto de sus razones relevantes; podemos llamar a éste el "mejor juicio" (*better judgment*). El juicio HTC, según Davidson, es "condicional", es decir, está condicionado por la evidencia (total o parcial) que ha sido considerada por el agente. El juicio HTC, si ha de llevar a la acción, es seguido por un juicio incondicional de la forma "debo hacer *x*" o "es deseable que haga *x*"; éste es llamado por Davidson un juicio entero (*outright* o *all-out judgment*), y corresponde a un compromiso del agente por perseguir un curso de acción especificado. Es idéntico, entonces, a la intención del agente. Debido a que este juicio es incondicional y el juicio HTC es condicional, una persona puede concluir sin contradicción, a través de su deliberación, que cierto curso de acción es el mejor y, sin embargo, actuar de otra manera. Así, mejor juicio e intención son lógicamente independientes. Comúnmente, el agente actúa de acuerdo con su juicio HTC; al actuar incontinentemente, no obstante, el agente ha formado un juicio HTC pero luego actúa de una manera distinta.

Dos paradojas surgen de la noción de irracionalidad práctica, según Davidson.[4] Estas paradojas se hacen más visibles al considerar el tema de la debilidad de la voluntad, pero también están presentes en otros tipos de irracionalidad práctica, como el autengaño. La primera paradoja proviene de la idea de que siempre es posible explicar una acción por razones. Dentro del marco de la psicología del sentido común, las razones prácticas

[4] "Paradoxes of Irrationality", p. 293 y ss.

convierten una acción en inteligible presentando el punto de vista del agente; muestran por qué tiene sentido la acción a la luz de sus deseos y creencias. Pero en los casos de irracionalidad práctica, las razones del agente no hacen que la acción sea inteligible desde el punto de vista del agente. En los casos típicos, la acción refleja el contenido del juicio HTC, que a su vez puede ser explicado por las razones del agente, pero en los casos de acción incontinente, la acción entra en conflicto con el juicio HTC y con el balance total de las razones del agente. Si la acción incontinente se intenta explicar por las razones del agente, la explicación a la cual se llega es la misma que hubiera explicado la acción continente. "Es imposible que una explicación explique dos eventos que constituyen, uno de otro, sus grupos de contraste."[5]

La segunda paradoja proviene de lo que Davidson denomina el "holismo" de lo mental. Según Davidson, la identidad de un deseo o una creencia está constituida por sus relaciones con sucesos u objetos en el mundo, así como por sus relaciones con otras creencias y deseos. Bajo esta perspectiva, la racionalidad es constitutiva de lo mental: establece el marco dentro del cual entendemos el comportamiento que es descrito con términos mentales. La identidad de los estados mentales está determinada entonces por la manera en que éstos funcionan en explicaciones psicológicas. Si el comportamiento del agente es considerado como "desviado" [*deviant*] porque el agente no actuó por las razones que consideraba superiores, la visión holista implica que hay cierta evidencia para afirmar que la atribución inicial de creencias y deseos para este sujeto era incorrecta, y que más bien puede atribuírsele un juicio práctico que corresponda con su acción. De ahí el problema de explicar una acción que fue hecha por una razón y sin embargo no corresponde con los propios valores, creencias y deseos del agente. A la luz de esta

[5] Ariela Lazar: "Akrasia and the Principle of Continence", p. 385.

teoría, la tarea de explicar las acciones incontinentes, según Davidson, consiste en "encontrar el mecanismo [psicológico] que pueda ser aceptado como apropiado para procesos mentales y que, sin embargo, no racionalice lo que explica"[6]. En otras palabras, este proceso debe respetar ciertas exigencias mínimas de racionalidad y, al mismo tiempo, dejar espacio para la irracionalidad.

Una parte de la explicación que da Davidson de la debilidad de la voluntad en "Paradoxes of Irrationality", sostendré aquí, es suficiente para disolver ambas paradojas. Esa parte de la explicación es la siguiente: la acción incontinente es susceptible de ser explicada por razones, pero, al realizarla, el agente ignora el principio de continencia, su principio de segundo orden que le dicta actuar según su propio mejor juicio. El agente tiene una razón (es decir, un motivo) para ignorar tal principio, a saber, el deseo de realizar la acción que en este caso resulta ser incontinente. La irracionalidad se introduce cuando este deseo causa (no razonablemente, sino como mera causalidad) que el agente ignore el principio, pues aunque ese deseo es una razón para ignorar el principio, no es una razón en contra del principio mismo, y así, cuando entra por segunda vez en la deliberación, es irrelevante como razón para la acción.

Con esto, Davidson ha mostrado cómo no es incompatible sostener que cierta acción puede basarse en una razón —como, de hecho, todas las acciones intencionales se basan en una razón— con sostener que esa misma acción puede ser irracional. La primera paradoja queda entonces resuelta, pues para explicar una acción es suficiente citar la razón del agente para realizar esa acción, independientemente de que haya otras razones mejores por las cuales el agente pudo haber hecho otra acción. No es exactamente la misma explicación la que explica el acto incontinente que la que hubiera explicado el

[6] "Paradoxes of Irrationality", p. 297.

hipotético acto continente; la explicación para esta última acción necesitaría citar otras razones del agente.

Sin embargo, Davidson dice que todavía quedan otras fuentes de paradoja acerca de la irracionalidad práctica.[7] Según él, la segunda paradoja emerge de lo siguiente: si explicamos un caso de irracionalidad en términos puramente neurofisiológicos o físicos, la irracionalidad no aparece en la explicación. (Hasta aquí no parece haber nada paradójico: no hay lugar para la irracionalidad en los sucesos descritos en términos puramente físicos, y además se ha estado considerando hasta aquí a la irracionalidad como incoherencia interna, es decir, entre estados proposicionales. No es sorprendente que en descripciones no-proposicionales de acciones irracionales no aparezca el elemento irracional.) Davidson dice que, entonces, para explicar o siquiera describir la irracionalidad del acto, debemos introducir una descripción mental de la causa, lo cual la convierte en un candidato para ser una razón. Pero incluso esto no es suficiente, dice, pues todavía estamos fuera del "único patrón claro de explicación que aplica a lo mental". Para que ese patrón sea aplicable no sólo necesitamos que la causa sea un candidato a ser una razón, dice Davidson, sino que *sea* una razón, "lo cual, en el presente caso, no puede ser", y esto específicamente parece ser, para él, la fuente de la segunda paradoja. Él dice "no puede ser" porque la acción a explicar en este caso es irracional, pero aquí parece olvidar que la causa en este caso sólo necesita el "elemento de racionalidad" que, habíamos dicho, está presente en toda acción intencional sin que ello excluya la posibilidad de que la acción sea irracional, y esa paradoja ya había sido

[7] *Ibid.* p. 299.

disuelta. Él mismo lo afirma justo en el párrafo anterior.[8] No es necesario que, en los casos de acción que se considera desviada, la atribución inicial de deseos y creencias del agente siempre haya sido equivocada; simplemente sucede que al actuar incontinentemente el agente no actúa basándose en los deseos y las creencias que constituyen sus mejores razones. Davidson había explicado ya cómo esto es posible al mostrar que el agente no cae en un marco mental contradictorio al formar un juicio HTC acerca de cómo debe actuar y después actuar de manera distinta. Dice Davidson al llegar a este punto que todavía existe una tensión conceptual entre el principio Platón y la existencia de acciones irracionales. No veo cuál pueda ser esa tensión conceptual: si existen acciones irracionales, el principio Platón es sencillamente falso. Si no existen, es verdadero. Así, esta segunda paradoja resulta ser sólo una reformulación de la primera, y se disuelve de la misma manera.

De esta segunda paradoja Davidson intenta derivar la necesidad teórica de hacer particiones en la mente para poder explicar la irracionalidad práctica. Si lo que he dicho es correcto y esta segunda paradoja sólo es una reformulación de la paradoja que ya se había resuelto antes, no parece existir otra razón por la cual sea conveniente introducir las particiones a la mente. Más aún, aparentemente el único argumento que ofrece Davidson para explorar el camino de las particiones a la mente como posible solución de la segunda paradoja de la irracionalidad es que, tomando en consideración su hipótesis general para explicar los casos de irracionalidad como casos en los cuales se produce una acción o estado mental para el cual existe una causa que no es una razón, hay otro tipo de casos en los que él puede pensar en donde hay estados mentales que causan otros sin ser razones, a

[8] Es sumamente extraño que Davidson no haya visto esto y me parece más probable que yo lo esté interpretando inadecuadamente. Sin embargo, no he podido encontrar otra explicación.

saber, cuando se trata de mentes distintas.[9] Pone como ejemplo el deseo de alguien de cultivar un jardín con el propósito de provocar que alguien más desee entrar en ese jardín. Suponiendo que esta persona ponga en marcha su plan y consiga que la otra persona efectivamente entre en él, el deseo de entrar en el jardín de la segunda persona habrá sido causado por el deseo citado de la primera persona. Pero también pueden encontrarse otros ejemplos donde algo muy similar ocurre dentro de una misma mente, como cuando alguien desea programar un reloj con alarma para que suene una hora después y le recuerde que hay que sacar el pavo del horno; si pone en práctica este plan con éxito, el que la persona se haya acordado específicamente una hora después de sacar el pavo del horno habrá sido causado, entre otras cosas, por su deseo de programar el reloj con alarma. Davidson mismo, en otro lugar, pone otro ejemplo de causas mentales que no son razones para los efectos mentales que producen que no involucra dos mentes distintas: el de alguien que tararea cierta canción dentro de su cabeza, por así decirlo, con el fin de recordar alguna información. (Dentro de la explicación de Davidson, incluso este ejemplo requeriría una partición de la mente; ello me parece completamente contraintuitivo, pues, por el contrario, parece que el tararear en este caso es justamente un recurso mental para conectar contenidos mentales.) Aun suponiendo que la explicación correcta de la irracionalidad efectivamente requiera hacer referencia a estados mentales que causan acciones u otros estados mentales sin ser razones para ello, al mostrarse que esto puede ocurrir dentro de una misma mente en los ejemplos citados sin que haya actitudes mentales que necesariamente deban permanecer separadas para que el proceso funcione, se desvanece la necesidad teórica de hacer particiones en la mente. Lo que los dos primeros ejemplos tienen en común es que los estados mentales que funcionan como

[9] *Ibid.,* p. 300.

causas no actúan de manera inmediata para producir otros estados mentales, sino que primero funcionan como causas de sucesos fuera de la mente, los cuales a su vez causan otros estados mentales. En el ejemplo de la persona que tararea una canción en su mente, la persona realiza una actividad mental que tiene por resultado el facilitar el acceso a información detallada que se encuentra almacenada dentro de su memoria a largo plazo pero cuyo contenido tal como es recordado —la tonada y la letra de la canción, supóngase, sobre el orden de los planetas— fue aprendido de una manera en que el recuerdo de las percepciones auditivas ha tomado preponderancia por encima del significado de las palabras que componen la letra de la canción, y por ello el repaso mental de la tonada ayuda a encontrar otra vez (o incluso por primera vez) los contenidos lingüísticos de los enunciados de la canción y sus conexiones lógicas. La relación causal entre unos estados mentales y otros en todos estos ejemplos, sin embargo, no es inmediata (en el caso de la canción, el agente no recuerda directamente el orden de los planetas, sino que recurre a la tonada y a la entonación de las palabras dentro de la canción para recordar la letra), y es plausible que en esto resida una explicación más completa sobre cómo en estos casos unos estados mentales pueden causar otros para los cuales no son razones.

Creo que puede construirse una explicación plausible de al menos varios tipos de irracionalidad práctica, como la debilidad de la voluntad y el autoengaño, utilizando algunos de los elementos que ha ofrecido Davidson pero rechazando la conveniencia de admitir particiones en la mente. Conviene aclarar aquí que algunos casos de irracionalidad práctica, como, por ejemplo, ciertos tipos de autoengaño, parecen exigir

que el agente no conecte ciertos estados mentales suyos para que su acción o estado mental irracional sea posible; para estos casos especiales podría seguir existiendo una conveniencia teórica de hacer particiones a la mente, como sugiere Davidson. Pero tales divisiones en la mente podrían no ser más que "ayudas conceptuales para la descripción coherente de irracionalidades genuinas"[10], en el sentido de que, como resultado de una estipulación, un agente no puede considerar ciertos estados mentales relevantes al mismo tiempo *mientras actúa irracionalmente*; la restricción impuesta por estas fronteras mentales no describiría una imposibilidad empírica del agente, sino sólo un requisito conceptual para la descripción de ciertos estados mentales irracionales. En muchos otros casos de irracionalidad, sin embargo, no es difícil suponer que el agente se da cuenta de todos sus estados mentales relevantes para la acción o el estado mental irracional.

Antes de presentar un intento paralelo al de Davidson por explicar la irracionalidad, me parece necesario introducir, sin embargo, otra distinción trazada por el mismo Davidson. Se trata de una distinción entre razones para tener una creencia. "Si alguien desea que cierta proposición fuese verdadera, es natural asumir que él o ella disfrutaría más creyéndola verdadera que no creyéndola verdadera. Tal persona tiene, por tanto, una razón para creer la proposición. […] Aquí debemos hacer una distinción obvia entre tener una razón [o motivo] para ser creyente en una proposición, y tener evidencia a la luz de la cual es razonable pensar que la proposición es verdadera." [11] En el primer sentido, un agente tiene una razón para creer que cierta proposición es verdadera en virtud de que el tener esa creencia específica se ajusta a alguno de sus deseos. En cambio, en el segundo sentido, el agente tiene una razón para creer en esa proposición, donde el

[10] "Deception and Division", publicado en *The Multiple Self*, John Elster (ed.), Cambridge, Cambridge University Press, 1986), p. 92.
[11] *Ibid.*, p. 85.

valor de tener esa creencia es independiente de si la creencia se ajusta a los deseos del agente y la creencia tiene el mismo valor que tendría la creencia contraria si la evidencia apuntara a favor de ella.

A continuación sugiero que esta distinción entre razones para tener creencias puede extenderse a las razones para tener deseos y a las razones para actuar. Llamaré "razones aléticas" a *las razones para actuar, creer o desear cuyas afirmaciones correspondientes de la forma "*A *tiene una razón para* x*" están justificadas por la evidencia disponible al agente*[12], y "razones oréticas" a *aquellas que racionalizan*[13] *la acción, creencia o deseo resultante como consecuencia de la combinación de deseos y creencias de tipo medio-fin preexistentes del agente*.

En lo que resta de este capítulo, intentaré aclarar la distinción entre razones aléticas y oréticas. Trataré primero las razones aléticas para tener creencias, pues creo que son las menos problemáticas. Pasaré después, todavía dentro del tema de las razones aléticas, a las razones para desear y actuar. Por último, relacionaré las razones aléticas con las razones oréticas. Las razones aléticas son, según lo dicho, un tipo de exigencias normativas de la racionalidad práctica que dictan qué acciones, creencias o deseos el agente está requerido a hacer o tener en una situación concreta, dada la evidencia disponible, para ser considerado como un agente racional. Vale la pena resaltar que el hecho de que toda la evidencia disponible apunte hacia cierta proposición no garantiza que esa proposición sea verdadera, pues podría surgir nueva evidencia con mayor peso

[12] Cuando utilizo el término "evidencia" en este contexto, me refiero sólo a los estados mentales cognoscitivos (representados principalmente por las creencias) del agente cuyos contenidos proposicionales justifican una inferencia al contenido proposicional del estado mental resultante. Recojo así la intuición de que una creencia sólo puede ser justificada por otras creencias.

[13] En el sentido usado por Davidson en "Actions, Reasons and Causes", publicado en *Essays on Actions and Events*, Oxford, Clarendon Press, 1980.

que apunte hacia la proposición contraria, o, aunque no surja nueva evidencia, la creencia puede ser simplemente falsa; esto, sin embargo, no impide que pueda hablarse de una exigencia racional, en cierto momento, de dar crédito a la proposición que en ese momento está mejor apoyada por la evidencia disponible, al grado de volver razonable la adquisición de esa creencia. De esta manera, puede haber una razón alética ahora para creer una proposición que después resulte ser falsa, y sin embargo, en ese caso, la afirmación: "dada la evidencia *e*, el agente *A* tiene una razón alética para creer que *p* en el momento *t*" (donde *t* hace referencia al momento en que el balance de la evidencia apunta hacia *p*) es una proposición "eternamente" verdadera.

Lo que digo no es equivalente a sostener que los agentes tienen la obligación, para ser considerados como sujetos racionales, de extraer todas las consecuencias que se siguen lógicamente del contenido de sus creencias y convertirlas en nuevas creencias suyas. Dado que las consecuencias lógicas de cualquier conjunto de proposiciones son infinitas, un requisito de tal naturaleza volvería imposible que cualquier agente con tiempo y capacidad de razonamientos limitados sea racional. Cuando hablo de adquisición racionalmente requerida de creencias me refiero a los casos en los cuales un agente, durante cierta etapa del proceso deliberativo, llega a darse cuenta de que cierta proposición relevante para sus fines que no forma parte de sus creencias en ese momento es, en vista del resto de sus creencias, apoyada de manera conclusiva por la evidencia con la que cuenta. En tales casos puede hablarse de la adquisición de ciertas creencias como requisito para la racionalidad de un agente. Lo mismo es aplicable a la pérdida o modificación racionalmente requerida de creencias.

Por otra parte, el mero hecho de que cierta proposición verdadera apoye conclusivamente la verdad de *p* no constituye por sí mismo una razón alética para creer que *p*. Decir lo contrario incurriría en un error paralelo al del "teórico de las razones externas" del que habla Williams al tratar el tema de las razones para actuar.[14] Michael Smith, creo, cae en este error[15]; él piensa que pueden existir razones para que A crea que *p* aun cuando esas razones no estén disponibles, incluso si esas razones *no pudiesen nunca* estar disponibles. Es suficiente, considera Smith, que una proposición sea verdadera y apoye la creencia de que *p* para que sea una razón para creer que *p*. Reconstruyo aquí el argumento de Williams en contra de tal postura, aplicándolo a las razones para creer: si algo es una razón para creer que *p*, debe ser posible que al menos alguien alguna vez creyese que *p* por *esa* razón, y entonces esa razón debería poder figurar en alguna explicación de por qué A tiene esa creencia. Éste es un requisito que no cumplen las llamadas razones externas (es decir, las que no dependen de ningún elemento subjetivo del agente), y, por tanto, ésas en realidad no son razones. Si nadie pudiese tener nunca acceso a la creencia en cierta proposición verdadera que apoya la verdad de *p*, esa proposición no podría ser nunca la razón por la cual alguien creyese que *p*, y entonces no podría figurar en ninguna explicación de por qué alguien llegó a creer que *p*.

Smith dice que las razones para creer son los estados de cosas en el mundo que "vuelven racional, para alguien que cree que las cosas son de esa manera, que crea cierta proposición". Extrañamente, la formulación me parece correcta, pero sólo si la interpretamos de una manera sutilmente distinta a aquélla en que Smith la toma. Smith

[14] Williams: "Internal and External Reasons", publicado en *Moral Luck*, Cambridge University Press, Cambridge, 1981.
[15] En "Is There a Nexus Between Reasons and Rationality", publicado en Sergio Tenembaum (ed.): *New Trends in Moral Psychology*, Rodolphi, Amsterdam, 2006.

piensa que "los estados de cosas en el mundo" se refiere a las cosas acerca de las cuales tratan las proposiciones en cuestión. Así, si *p* significa "las ballenas son los mamíferos más grandes", las razones que podrían apoyar la creencia en *p* serían verdades acerca de las ballenas y los mamíferos que harían racional que alguien que crea en esas verdades crea también que *p*. Pero el conocimiento de tales verdades, según el mismo Smith admite, podrían nunca ser accesibles para un ser racional y conservar, sin embargo, su estatus de razones para que ese ser crea que *p*. "Puede haber razones para creer que *p* de las que nadie sabe, y quizá incluso de las que nadie pudiese saber". Yo sostengo que las razones (aléticas) pueden caracterizarse, en efecto, como un tipo de verdades o "estados de cosas en el mundo", pero no verdades acerca de ballenas o cualquier otra cosa que sea el objeto de la proposición *p*, sino que son un tipo de proposiciones verdaderas acerca de cuáles son las creencias de A en ese momento. Creo que esto va de acuerdo con la intuición que tenemos de que ser racional, o comportarse de manera racional, aplicado a un ser humano, debe ser algo que esté, al menos en principio, a su alcance. Por supuesto, está también la intuición de que el hecho de que algo es verdadero nos proporciona buenas razones para creerlo, pero creo que es permisible adicionar esta intuición (para hacerla más precisa) con la idea de que, si ahora es imposible para nosotros saber si cierta proposición es verdadera o falsa, no es racional ahora considerarla (y utilizarla) como una razón para creer otras proposiciones, aun cuando después resulte ser verdadera. Este punto se relaciona con el hecho de que no consideramos que algo sea conocimiento por el mero hecho de que sea verdadero, pedimos además que sea creído con alguna justificación.

Según lo que digo, alguien puede tener buenas razones, en el sentido de exigencias racionales, para creer que *p* aun cuando *p* sea falso. Esto a Smith le parece inapropiado, pues él quiere que la diferencia entre razones "normativas" y razones "motivantes" para tener una creencia[16] (como él las llama) se traduzca en que las primeras lleven siempre a creencias verdaderas, y las segundas puedan contingentemente llevar o no a creencias verdaderas, dependiendo de si son "buenas" o "malas" razones; justamente serán buenas razones si llevan a creencias verdaderas (correspondiendo así con las creencias a las cuales llevan las razones normativas). Creo que no es ése el valor real de la distinción entre estos dos tipos de razones para creer. Las razones motivantes para creer pueden ser buenas o malas sin importar si llevan a creencias verdaderas o no; son buenas, por ejemplo, si están guiadas por un interés por la verdad y por ello siguen los estándares de la racionalidad teórica (como la regla de *modus ponens*). Son malas si llevan a la creencia de una manera distinta, por ejemplo, a través de las "creencias ilusorias"[17] [*wishful thinking*] o el autoengaño. Me parece que Smith no se da cuenta de estos tipos de maneras de llegar a creencias *a través de* razones motivantes, y eso lo hace colocar el valor de la distinción en la verdad o falsedad de las creencias a las cuales se llega en virtud de esas razones. Pero alguien puede llegar a creencias verdaderas por razones motivantes malas (como en algunos casos de creencias ilusorias) y alguien puede llegar a creencias falsas por buenas razones motivantes (cuando la persona está genuinamente interesada en conocer la verdad y sigue los estándares de la racionalidad teórica, pero parte de premisas que son falsas), y la explicación de Smith no da cuenta de esto. Retomaré la distinción entre razones normativas y motivantes más adelante.

[16] La distinción de Smith entre razones motivantes y normativas para actuar se comentará más adelante en este capítulo.
[17] Este fenómeno se explica más adelante en este capítulo. Su definición puede encontrarse en la nota 31.

Lo anterior me lleva a sugerir que las razones aléticas para tener creencias son relativas a los estados mentales de un agente en cierto momento. Esto permite que un agente tenga una razón alética para creer que *q*, por ejemplo, si cree que *p* y cree que "si *p* entonces *q*". Pero puede resultar que a fin de cuentas el agente no tenga ninguna razón alética para creer que *p* o para creer que "si *p* entonces *q*", y, al corregir sus creencias en esta dirección, su razón alética inicial para creer que *q* desaparezca. De manera análoga, puede decirse apropiadamente que *A* tiene una razón alética para creer que *p* si se enfrenta ante la decisión de escoger entre *p* y *q*, y *p* está mejor apoyada por la evidencia (constituida por las propias creencias de *A*) que *q*, pero puede ser que tenga razones aléticas también para creer que *r* en vez de *p*, si resulta que hay evidencia (es decir, otras creencias) de todavía mayor peso que justifique *r*. En otras palabras, basta un "mejor" [*better*] juicio HTC de que hay más evidencia para creer cierta proposición que otra para que haya una razón alética para creer en esa proposición; el "óptimo" [*best*] juicio HTC sobre razones de un agente para tener una creencia será el que tome en cuenta *toda* la evidencia disponible en ese momento.

Así como la adquisición, conservación y modificación de creencias puede estar justificada por la evidencia disponible (es decir, por las creencias ya existentes del agente), de manera análoga, los deseos y las acciones pueden estar justificados por la existencia de ciertos estados mentales del agente que también constituyen evidencia a favor de esos deseos y creencias. Mejor dicho, puesto que los deseos y las acciones no tienen valor de verdad, las *afirmaciones* del tipo "en las circunstancias *C*, *A* está racionalmente autorizado a desear/hacer *x*" pueden estar justificadas por el conjunto de los deseos y creencias de un agente en determinado momento del mismo modo en que las

creencias que constituyen evidencia pueden justificar otra creencia. Así, por ejemplo, si un agente desea hacer ψ, cree que la única manera de lograrlo es haciendo φ y cree que hacer φ no tiene ningún costo considerable para él, el agente está entonces racionalmente autorizado a desear hacer φ y a hacer φ. Debe quedar claro que al menos una parte fundamental (si no es que el todo) de la evidencia que justifica afirmaciones del tipo "A está racionalmente autorizado a desear/hacer *x*" consiste en las creencias tipo medio-fin y los deseos que el agente tiene en ese momento. Las combinaciones de estos tipos de estados mentales, según dije antes, constituyen las razones del otro tipo, las razones oréticas. Una razón orética para hacer φ está compuesta mínimamente por un deseo de hacer ψ y una creencia de que, si hace φ, el agente hace ψ. (En casos límite, φ y ψ representan la misma acción, y entonces la creencia, trivial, que forma parte de la razón orética será simplemente que si hace φ, el agente hace φ. Son los casos donde se dice que el agente hace cierta acción por sí misma, sin ningún propósito ulterior.) El que las razones oréticas, es decir, combinaciones apropiadas entre deseos y creencias, constituyan parte fundamental de la evidencia para establecer la verdad de enunciados sobre razones aléticas, nos ayuda a entender la relación entre ambos tipos de razones: las razones aléticas son proposiciones verdaderas acerca de cuáles son las mejores razones oréticas de un agente, bajo circunstancias específicas, para creer, desear o hacer algo.

Tratando el tema de las razones para actuar, Michael Smith, como ya se mencionó, distingue entre razones motivantes y razones normativas. Esta distinción tiene algunas semejanzas importantes con la distinción entre razones que aquí planteo. Él dice: "hay dos conceptos completamente diferentes de una razón para actuar, dependiendo de si

enfatizamos la dimensión explicativa y aminoramos la justificativa, o viceversa".[18] En el primer caso se trata de razones motivantes, en el segundo, de razones normativas.

> Ambas tienen algo en común en virtud de lo cual ambas cuentan como razones. Pues citar cualquiera de ellas nos permite volver inteligible la acción de un agente. (…) Sin embargo, en virtud de sus diferencias, las razones motivantes y las normativas hacen que las acciones sean inteligibles por razones completamente distintas.
>
> Decir que alguien tiene una razón normativa para hacer φ es decir que hay alguna exigencia normativa para que él haga φ, y es entonces decir que esa acción está justificada desde la perspectiva del sistema normativo que genera tal exigencia. (…) La mejor manera de considerar a las razones normativas es como verdades. Esto es, como proposiciones de la forma general: "Es deseable o exigido que A haga φ." (…)
>
> Las razones motivantes son, sin embargo, diferentes. La característica distintiva de una razón motivante para hacer φ es que, en virtud de tener tal razón, un agente está en un estado que permite explicar su realización de tal acción. (…) Es entonces natural suponer que su razón motivante es en sí misma *psicológicamente real*. Por contraste con las razones normativas, entonces, las cuales parecen ser verdades de la forma "es deseable o exigido que yo haga φ", las razones motivantes parecen ser *estados psicológicos*, estados que juegan un papel explicativo en la producción de la acción.

No estoy de acuerdo con Smith cuando él afirma que las razones normativas "vuelven inteligible" la acción del agente. Bajo la definición de Smith, un agente puede tener razones normativas para hacer cierta acción por el hecho de que cierto sistema de racionalidad así lo exija, independientemente de que el agente tenga deseos o creencias

[18] *The Moral Problem*, Oxford, Blackwell, 1994; p. 95 y ss.

relacionados con esa acción. Una acción, sin embargo, sólo es inteligible cuando podemos entender desde el punto de vista del agente qué fue lo que vio de deseable en realizarla, es decir, cuáles fueron los deseos y las creencias relevantes en la producción de esa acción. El hecho de que sea verdadera una proposición que diga que, según cierto sistema de racionalidad, el agente debe realizar cierta acción, por sí mismo, no puede explicar ninguna acción del agente. Por eso dije antes que las razones aléticas deben basarse en las razones evaluativas (los deseos y creencias reales) del agente.

Las razones normativas para actuar que propone Smith, si no consideran las creencias y los deseos del agente, no son más que, usando otra vez los términos de Bernard Williams, afirmaciones acerca de razones externas, y, si Williams tiene razón, como creo que la tiene, cualquier afirmación de este tipo es falsa.[19] Parte del concepto de razones aléticas que propongo, entonces, es el hecho de que sirvan para determinar qué es lo racional para el agente dados sus deseos y creencias.

La distinción entre razones aléticas y razones oréticas nos proporciona nuevas bases para explicar, entre otras cosas, la debilidad de la voluntad como un tipo de irracionalidad. Es claro que puede haber casos donde lo que dictan algunas razones aléticas del agente choque con algunas de sus razones oréticas. En estos casos, a través de la deliberación práctica, un agente puede llegar a darse cuenta de que tiene razones aléticas que requieren racionalmente la realización de cierta acción (según su propio mejor juicio), pero de cualquier manera omitir intencionalmente realizar esa acción, haciendo algo más que sólo puede apoyar con razones oréticas; su acción es entonces incontinente. El mismo tratamiento puede aplicarse a una descripción del autoengaño: una persona tiene un

[19] "Internal and External Reasons", p. 109.

conflicto de razones oréticas y aléticas acerca de adquirir o conservar una creencia; la persona se da cuenta de que tiene razones aléticas, por ejemplo, para adquirir cierta creencia, pero se apoya en sus razones oréticas para hacer ciertas acciones u omisiones intencionales que le impidan adquirir esa creencia y que le permitan conservar la creencia contraria. A partir de estos elementos es posible formular una hipótesis general paralela a la de Davidson acerca de cómo explicar la irracionalidad práctica: una acción o estado mental es irracional si en su producción participaron razones oréticas que no estaban *apoyadas* por razones aléticas que el agente conocía o estaba en posibilidad de conocer, cuando, según los estándares de racionalidad, se requería contar con el apoyo de razones aléticas (es decir, excluyendo casos como los de las creencias provenientes de la percepción y la memoria). Una razón alética apoya una razón orética cuando la acción que la primera recomienda es la misma hacia la cual esa razón orética se dirige.

Bernard Williams hace una distinción entre dos aplicaciones de la noción "querer creer que *p*" que tiene alguna relación con la distinción entre razones aléticas y oréticas.[20] "*A* desea creer que p" quiere decir, bajo la primera aplicación, que *A* desea que *p* sea verdadero. No basta, para satisfacer este deseo, que él adquiera la creencia; además se necesita que haya bases suficientes para apoyar esta creencia. En este caso *A*, al querer creer que *p*, tiene un "motivo centrado en la verdad". Bajo la segunda aplicación, "*A* desea creer que *p*" sólo significa que a *A* le gustaría adquirir esa creencia, independientemente de que haya bases suficientes para creerla. En este caso *A* tiene un "motivo no-centrado en la verdad".

[20] "Deciding to Believe", en *Problems of the Self*, Cambridge University Press, New York, 1973, p. 149-150.

En esta distinción, al hablar de "bases suficientes" para una creencia, me refiero exactamente a lo mismo a lo que Davidson se refiere con "razones apoyadas por la evidencia" para tener una creencia. En el caso de los motivos centrados en la verdad, la persona desea tener razones aléticas para creer que cierta proporción es verdadera. En ambos casos, la persona *tiene ya* razones oréticas para creer que *p*.

Puede parecer razonable preguntarse por qué, en presencia de un conflicto entre razones aléticas y oréticas, debemos preferir a las primeras para decidir qué es más racional. ¿Podría haber casos donde las razones "superiores" del agente sean sus razones oréticas y no sus razones aléticas? Davidson plantea una cuestión similar al preguntarse si un agente puede no aceptar como suyos ciertos principios básicos de racionalidad, como el principio de continencia o el principio de evidencia total, y en tales casos sus acciones podrían no ser irracionales desde su propio punto de vista.[21] La respuesta de Davidson, con la que estoy de acuerdo, dice que no es posible que un agente no acepte principios de racionalidad tan fundamentales como ésos y siga pudiendo ser considerado un agente. Aceptar ciertos principios básicos de racionalidad es parte de lo que constituye ser un agente. Complementando la respuesta de Davidson, creo que tenemos bases, en cualquier caso, para establecer que, en general, es irracional no seguir lo que dictan las razones aléticas. Esto es porque podemos asumir que todo agente desea (se dé cuenta o no de ello) tener creencias verdaderas y utilizar razonamientos apropiados en lo que respecta a sus intentos por alcanzar sus fines. Habrá, como en el autoengaño, casos en los cuales el tener cierta creencia falsa es parte esencial de uno de los fines del agente, pero en estos casos excepcionales el deseo de tener una creencia que es falsa coexiste con el deseo del agente por tener creencias verdaderas. Siempre que estemos en presencia de un conflicto

[21] "Deception and Division", p. 83-84.

entre razones aléticas y oréticas del agente, puede decirse que las razones aléticas están basadas en ciertas *otras* razones oréticas del agente: todo agente tiene un interés por estar correctamente informado en lo que respecta a la consecución de sus fines.

Williams establece un punto similar al explicar por qué él considera que alguien que se basa en creencias falsas o argumentos incorrectos para afirmar que tiene una razón para hacer determinada acción está equivocado, como en su famoso ejemplo del hombre que cree que lo que va a beber es ginebra, cuando realmente es gasolina, mientras que alguien puede tener una razón para hacer algo y no saberlo, por no tener cierta creencia verdadera relevante o por no haber efectuado cierto razonamiento válido. Dice Williams:

> Los fundamentos para establecer este punto general acerca de las consideraciones de hecho y razonamiento, a diferencia de las consideraciones prudenciales y morales, son bastante sencillas: cualquier agente deliberativo racional tiene en su [conjunto motivacional subjetivo] un interés general por estar factual y racionalmente bien informado. Podría haber un caso de alguien que tuviese una necesidad abrumadora de ser engañado; y si sus relaciones con la realidad estuviesen tan pobremente negociadas que él de hecho necesitase creer lo que es falso, entonces tal vez él tendría razón para adquirir creencias falsas —en este respecto en particular. El punto básico, sin embargo, es que en la perspectiva internista existe ya una razón para incluir, en general, las exigencias de información y razonamiento correctos en la noción de una ruta deliberativa sólida, pero no una razón similar para incluir las exigencias de la prudencia y la moralidad.[22]

[22] "Internal Reasons and the Obscurity of Blame", en su *Making Sense of Humanity*, Cambridge, Cambridge University Press, 1995, p. 37.

Esto nos proporciona una respuesta a la pregunta que hice al principio de este capítulo: ¿cuál es el supuesto deseo que proporciona la motivación para adquirir creencias, a través de la deliberación, que están basadas en otras creencias? En la mayoría de los casos, sería el deseo correspondiente a este "interés general por estar factual y racionalmente bien informado". Excepcionalmente, como en los motivos no-centrados en la verdad, sería un deseo que se satisface con cierta creencia específica, independientemente de que esté apoyada por la evidencia.

Capítulo 2. Autoengaño

En algunos ejemplos de autoengaño es posible encontrar que se satisfacen, al menos en apariencia, los criterios para afirmar que el agente cree una proposición y también su contradictoria. De otra manera, parece imposible explicar cómo el agente puede sinceramente afirmar tener cierta creencia y al mismo tiempo evadir inteligentemente la evidencia preponderante que se le presenta a cada momento a favor de la creencia opuesta. Al igual que con el problema de la debilidad de la voluntad, a pesar de la gran cantidad de aparentes pruebas a favor de la existencia del autoengaño que la realidad cotidiana nos proporciona, se ha puesto en duda incluso que este fenómeno exista, con base en algunas paradojas que parecen indicar que ciertas características esenciales del estado del autoengaño, o de su proceso de formación, nos llevan a contradicciones en la descripción o a la atribución de estados mentales imposibles. Algunos intentos de explicar el autoengaño han seguido la estrategia de reconstruirlo de maneras que esas paradojas no emerjan, presentándolo como un fenómeno plenamente plausible y libre de extravagancia —en ocasiones dejándonos con la sensación de que en algún momento de la explicación el autor ha permitido que lo que era filosóficamente interesante acerca del problema escapara intacto.

En este capítulo intentaré explicar las dificultades que enfrenta el teórico que busca explicar el autoengaño, así como tres de las posturas más reconocidas que se han propuesto para explicarlo y críticas a esas posturas. Después presentaré un intento de solución alternativa que, en mi opinión, es superior. La explicación del autoengaño que

presento aquí no es original en cuanto a sus rasgos generales, aunque me parece que en las discusiones contemporáneas se le ha prestado menos atención que la que merece.[1]

I

Un problema preliminar a la explicación del autoengaño consiste en cómo identificar este fenómeno. Basándonos en los componentes de la expresión "autoengaño", parecería que una buena táctica para hacer esto es comenzar por explicar el engaño interpersonal común, en donde una persona *A* engaña a una persona *B*, y después definir al autoengaño como una aplicación de ese mismo modelo, en donde *A* y *B* son la misma persona. Esta estrategia ha sido llamada por Alfred Mele la "aproximación léxica"[2]. Según él, las siguientes dos afirmaciones acerca del engaño son típicas de esta aproximación: 1) por definición, una persona *A* engaña a una persona *B* (donde *B* puede o no ser la misma persona que *A*) haciéndole creer que *p*, sólo si *A* sabe, o al menos cree verdaderamente, que no-*p* y causa a *B* que crea que *p*; y 2) por definición, engañar es una actividad intencional: el engaño no-intencional es conceptualmente imposible.[3] Según Mele, ciertos teóricos que se adhieren a la aproximación léxica de manera estricta sostienen que cualquier caso de autoengaño, como condición para que sea posible su existencia, debe satisfacer los requisitos planteados por esas dos afirmaciones.

Si se acepta esta aproximación, sin embargo, una vez que se han elucidado las implicaciones que conlleva el imponer tales requisitos al autoengaño, es difícil encontrar

[1] Esta postura, con algunas variantes, ha sido defendida anteriormente por D.H. Mellor: "Conscious Belief", en *Proceedings of the Aristotelian Society*, vol. 77, 1977-78; Robert Audi: "Self-Deception and Rationality", en Mike W. Martin (ed.): *Self-Deception and Self-Understanding*, University Press of Kansas, 1985, pp. 169-194, entre varios otros artículos suyos; y Eric Funkhouser: "Do the Self-Deceived Get What They Want?", *Pacific Philosophical Quarterly*, 86, no. 3, Sept. 2005; pp. 295-312.
[2] Mele: *Unmasking Self-Deception*, Princeton University Press, 2001; p. 5.
[3] *Idem*.

casos que adecuadamente puedan ser recibir el nombre de autoengaño. Dos paradojas bien conocidas acerca del autoengaño resultan de esa aproximación, la primera acerca del estado del autoengaño y la segunda acerca del proceso que lleva a él. Mele las ha denominado, respectivamente, la paradoja *estática* y la paradoja *dinámica*.[4]

La primera afirmación de la aproximación léxica impone a cualquier caso de autoengaño el requisito de que la misma persona, en tanto que es a la vez perpetradora y víctima del engaño, crea que *p* y crea que no-*p*. Esto aparentemente requiere que la persona crea simultáneamente en ambas partes de una contradicción explícita, lo cual parece ser un estado mental imposible. Ésta es la paradoja estática del autoengaño.

La segunda afirmación de la aproximación léxica, a su vez, impone el requisito de que el proceso del autoengaño sea intencional. Una característica de las acciones intencionales, al menos a primera vista, es que el agente está consciente de su intención al momento de realizarlas. Sin embargo, para que un engaño sea eficaz la víctima no puede conocer la intención de quien intenta engañarla. ¿Cómo sería posible, entonces, que la persona actúe con la intención de engañarse y al mismo tiempo no conozca esa intención? Ésta es la paradoja dinámica del autoengaño.

II

Hay quienes afirman, con base en estas paradojas, que el autoengaño es de hecho imposible y que los casos así llamados que frecuentemente encontramos en la vida cotidiana deben ser explicados de otra manera, por ejemplo, apelando a la insinceridad, o bien a algún proceso de formación de creencias motivadas que no requiera la posesión de creencias contradictorias ni la intención de engañarse en la persona.

[4] *Ibid.*, pp. 7-8.

Para quien desee seguir explorando la posibilidad de que el autoengaño exista, al menos dos tipos de estrategias están abiertas. Por una parte, se puede negar que la aproximación léxica sea la correcta; la analogía entre engaño interpersonal y autoengaño puede no ser necesariamente tan estrecha, de manera que sea posible hablar justificadamente de autoengaño aun cuando los requisitos planteados por las dos afirmaciones de esa aproximación no se satisfagan estrictamente. Por otra parte, se puede también intentar lidiar con las paradojas de manera que describir correctamente un caso de autoengaño no implique necesariamente la atribución de creencias abiertamente contradictorias o una intención de engañarse que sea al mismo tiempo consciente e inconsciente.

Donald Davidson ha defendido una postura que explica al autoengaño siguiendo la segunda de las dos estrategias recién mencionadas.[5] Según Davidson, una persona puede poseer creencias explícitamente contradictorias, incluso siendo ambas independientemente, pero no conjuntamente, accesibles a su conciencia. Frente a la enorme evidencia empírica que, según Davidson, encontramos según la cual la gente de hecho mantiene separadamente creencias relacionadas pero con contenidos opuestos —donde además una de esas creencias mantiene en existencia a la otra, característica que se explicará más adelante—, Davidson postula, como una ayuda conceptual, una frontera mental entre esas creencias de quien se autoengaña, a efecto de que, mientras el autoengaño dura, la persona no es capaz de conectar ambas creencias.[6] Además, para Davidson, el autoengaño debe ser intencional, pues de otra manera, argumenta, cualquier formación accidental de creencias falsas debida a un acto del propio agente, como leer

[5] Donald Davidson: "Deception and Division", en Jon Elster (ed.): *The Multiple Self*, Cambridge University Press, 1985; pp. 79-92.
[6] *Ibid.*, pp. 90-92.

equivocadamente una nota en el periódico, tendría que ser considerada autoengaño.[7] Esta intención, dice Davidson, no es la intención específica de engañarse; para que haya autoengaño respecto de *p* basta con que el motivo de la acción intencional que lleva a la persona al autoengaño se origine en la creencia en la negación de *p*, y que la acción sea realizada con la intención de producir una creencia de que *p*.[8]

Para explicar el fenómeno del autoengaño, Davidson lo compara con dos fenómenos relacionados: 1) la debilidad de la justificación [*weakness of the warrant*] y 2) las creencias ilusorias [*wishful thinking*]. La debilidad de la justificación es una violación a un principio de racionalidad, la "exigencia de evidencia total para el razonamiento inductivo" (llamado así por Carnap y Hempel), según el cual un agente epistémico, al enfrentarse a la situación de tener que decidir entre un conjunto de hipótesis que se excluyen mutuamente, debe dar más crédito a la hipótesis mejor apoyada por el balance de toda la evidencia disponible al agente, donde el agente (insiste Davidson) toma sus creencias que constituyen esa evidencia *como* evidencia. Los casos de autoengaño, según Davidson, siempre involucran debilidad de la justificación, por la simple razón de que quien se autoengaña acepta cierta proposición que no es la que está mejor apoyada por toda la evidencia a su alcance. Tiene mejores razones, basándose en la evidencia que reconoce, para aceptar la negación de esa proposición. La causa de que el agente se desvíe de este principio en los casos de autoengaño, sin embargo, es lo que nos sirve para distinguir al autoengaño de otros casos de debilidad de la justificación. Siempre que hay autoengaño, asegura Davidson, existe un elemento motivacional en funcionamiento. La motivación para creer en contra del principio de evidencia total, en el autoengaño, está

[7] *Ibid*., p. 87.
[8] *Ibid*., p. 88.

constituida por un deseo del agente. Por ejemplo, el deseo de una persona de creer que ella es inteligente le da a esa persona una razón (una motivación) para creer que es inteligente. Esta razón, obviamente, no es una razón cognoscitiva para creer que es inteligente; sin embargo, le da a esa persona un motivo para "actuar de tal manera que promueva tener esa creencia". En virtud de este elemento motivacional, el autoengaño se ha comparado frecuentemente con el fenómeno de las creencias ilusorias [*wishful thinking*]: el agente cree una proposición porque desea que esa proposición sea verdadera.[9]

Davidson piensa, sin embargo, que es posible distinguir entre autoengaño y creencias ilusorias. El autoengaño, dice, exige la intervención del agente; debe haber acciones intencionales cuyo fin sea producir una creencia que el agente no posee al momento de instituir ese comportamiento. En cambio, según él, un "análisis mínimo" de las creencias ilusorias muestra que la creencia resultante de este fenómeno no es el resultado de una actividad del agente. Más bien, el deseo del agente de que *p* "puede engendrar fácilmente" un deseo de creer que *p*, y este deseo puede provocar pensamientos y acciones que enfaticen, o resulten en obtener, razones cognoscitivas para tener esa creencia. Autoengaño y creencias ilusorias se parecen en que, en ambos, un elemento motivacional debe estar en funcionamiento; sin embargo, dice Davidson, el autoengaño "requiere la intervención del agente" y las creencias ilusorias no.

[9] El *Oxford English Dictionary* define el término que yo traduzco por "creencias ilusorias" [*wishful thinking*] como: "creencia o expectativa que está influenciada por los deseos de uno al grado que los datos relevantes conocidos (conscientemente) son ignorados o distorsionados (subconscientemente) [belief or expectation, that is influenced by one's wishes to the extent that relevant (consciously) known facts are (subconsciously) ignored or distorted]". Este término en inglés ha sido traducido de muchas otras maneras al español, sin que ninguna de ellas haya tomado preeminencia hasta ahora. Algunas de estas traducciones son: "creencias por deseos", "creencias esperanzadas" e "ilusiones vanas".

Hay otro aspecto en el que autoengaño y creencias ilusorias parecen diferir. Podría pensarse que aunque no todos los casos de creencias ilusorias, según lo que ya se dijo, son casos de autoengaño, las creencias ilusorias siempre son un ingrediente en los casos de autoengaño. Davidson dice que esto es incorrecto, pues hay excepciones. En las creencias ilusorias, dice, la creencia siempre toma la dirección del afecto positivo, nunca del negativo. En cambio, existen casos de autoengaño (los que Alfred Mele denomina "autoengaño invertido" [*twisted self-deception*]) donde la creencia inducida es dolorosa en vez de placentera: el agente se autoengaña creyendo precisamente aquello que desea que no suceda. Por ejemplo, una persona conducida por los celos puede encontrar "evidencia" por doquier que confirme sus peores sospechas.

La comparación que Davidson hace entre autoengaño y creencias ilusorias me parece insuficiente. Nunca explica él cómo en las creencias ilusorias el deseo de creer que *p* puede provocar pensamientos y acciones que enfaticen, o resulten en obtener, razones cognoscitivas para creer que *p*, y tampoco explica cómo esto es posible sin la intervención del agente, que, después dice, es necesaria para el autoengaño. En cuanto a la segunda diferencia, si bien es cierto que el nombre de "creencias ilusorias" sólo se aplica a creencias placenteras basadas directamente en los deseos que el agente tiene, esto no significa que el proceso psicológico que lleva a las creencias ilusorias no tenga una contraparte "invertida". Si aceptamos la explicación de Davidson de las creencias ilusorias, es razonable, al imaginar más ejemplos, pensar que existe un fenómeno completamente análogo por medio del cual el agente llega, en virtud de sus deseos y sin que medie su intervención, a creencias dolorosas basadas en lo que el agente desea que

no suceda, por ejemplo, en sus miedos. (Tal vez "*aversive thinking*" sería por tanto un nombre apropiado.)

Habiendo hecho esta comparación entre autoengaño, por una parte, y debilidad de la justificación y creencias ilusorias, por la otra, Davidson expone las condiciones que él considera suficientes para un caso de autoengaño. Para él, un agente *A* está autoengañado con respecto a una proposición *p* cuando:

1) "posee evidencia con base en la cual él cree que *p* es más probablemente verdadera que su negación",
2) "el pensamiento de que *p*, o el pensamiento de que él racionalmente debe creer que *p*, lo motiva a actuar de tal manera que se cause a sí mismo creer la negación de *p*. La acción involucrada puede no ser más que un desviamiento intencional de la atención lejos de la evidencia a favor de *p*; o puede involucrar una búsqueda activa de evidencia en contra de *p*. Todo lo que el autoengaño requiere de la acción es que su motivo se origine en la creencia de que *p* es verdadera (o el reconocimiento de que la evidencia la vuelve más probablemente verdadera que falsa), y que la acción sea hecha con la intención de producir una creencia en la negación de *p*." Y, por último,
3) "el estado que motiva el autoengaño y el estado que éste produce coexisten; en el caso más fuerte [cuando el agente originalmente, así como durante y después del proceso de autoengaño, posee la creencia de que *p*], la creencia de que *p* no sólo causa una creencia en la negación de *p*, sino que también la sostiene."[10]

[10] "Deception and Division", p. 91.

Tomando en consideración las dos paradojas del autoengaño mencionadas al principio de este capítulo, parece difícil que un caso real satisfaga estas tres condiciones. La solución que Davidson ofrece para este problema es introducir una división en la mente del agente. De acuerdo con la tercera condición, el estado que produce el autoengaño y el estado producido por el autoengaño coexisten. En los casos más fuertes, incluso, la creencia de que *p* (es decir, el estado que produce el autoengaño) no sólo causa la creencia en la negación de *p* (es decir, el estado producido por el autoengaño), sino que la sostiene, o mantiene en existencia. Por ejemplo, cuando alguien, después de algún tiempo con problemas de caída del cabello, llega a darse cuenta de que se está quedando calvo y ese reconocimiento le lleva a autoengañarse creyendo que no se está quedando calvo; en este caso, algunas acciones que el agente realiza, una vez autoengañado, para evitar confrontarse con la evidencia de que se está quedando calvo (por ejemplo, peinarse de cierta manera o utilizar siempre cierta postura cuando lo fotografían) sólo se explican si la creencia original, de que se está quedando calvo, coexiste todavía con la creencia inducida por el autoengaño. Más aún, si la creencia original desapareciera, el autoengaño se desvanecería, pues el agente quedaría indefenso, por así decirlo, ante la evidencia. En este sentido es que Davidson dice que la creencia original "sostiene" o mantiene en existencia, a la creencia inducida por el autoengaño.

Los casos más débiles, aunque Davidson no se detiene mucho en explicarlos, son para él aquellos en donde el estado que produce el autoengaño no es la creencia de que *p*, sino solamente la creencia de que la evidencia apunta preponderantemente hacia la verdad de *p*; esta creencia causa, a través del autoengaño, la creencia en la negación de *p*, y después coexiste con ella, pero no la mantiene en existencia.

Para entender los propósitos de Davidson al tratar el tema del autoengaño, es importante ver por qué él enfatiza el punto de que, en los casos más fuertes, la creencia original sostiene o mantiene en existencia a la creencia inducida. Los casos que Davidson tiene en mente aquí son aquéllos en los que el agente autoengañado, para mantenerse en ese estado, debe en algún sentido estar consciente de [*be aware of*] la realidad para poder evitarla. Si olvida o pierde de vista su conocimiento de esa realidad, corre el riesgo de quedar indefensamente expuesto a la evidencia y el autoengaño habrá sido inútil. Esto refuerza la idea de Davidson de que el agente debe intervenir para que el autoengaño funcione. Esta intervención debe durar no sólo mientras se instituye el autoengaño, sino durante todo el tiempo que el agente dure autoengañado.

Así, en los casos más fuertes, el autoengaño requiere la existencia simultánea de dos creencias mutuamente incompatibles. Normalmente, una persona no puede mantener creencias explícitamente contradictorias de manera simultánea conscientemente, pues si así fuera, estas creencias chocarían y al menos una de las dos perdería toda su fuerza. La única manera en que un agente puede poseer creencias explícitamente contradictorias es manteniéndolas separadas. Davidson propone entonces, para explicar cómo estas dos creencias contradictorias no chocan entre sí, la introducción de una barrera o división mental en el agente. No es exactamente la idea de múltiples agentes dentro de una misma mente. Lo único que sugiere Davidson de esta partición es que es "un muro metafórico que separe las creencias que, de permitirse que entrasen a la conciencia juntas, al menos una se destruiría."[11] La imagen que Davidson nos invita a ver con esta idea es la de "una sola mente no completamente integrada". El "muro metafórico" que mantiene separadas a ambas creencias no es la conciencia, pues el agente puede estar consciente de cada una de

[11] *Ibid.*, p. 90.

las dos creencias contradictorias por separado; lo que es imposible, mientras el autoengaño dura, es que "esté consciente de ambas en un mismo vistazo".

Davidson completa su cuadro del autoengaño señalando en qué punto, durante el proceso del autoengaño, el agente comete un paso irracional. La irracionalidad del autoengaño, dice, consiste en el hecho de que exige creencias inconsistentes. El paso irracional, entonces, consiste en el paso que hace esto posible: el trazo de la barrera que mantiene a las creencias inconsistentes separadas. En los casos más débiles, lo que se debe aislar del resto de la mente es el principio de la exigencia de evidencia total (pues este principio es el que normalmente haría ver al agente que su creencia inducida es inconsistente con la evidencia preponderante que posee en contra de esa creencia). En los casos más fuertes, además de la exigencia de evidencia total, la creencia original que causa el autoengaño y que directamente contradice a la creencia resultante del autoengaño debe ser también separada.

Davidson dice que, aunque la causa de la división involucrada en el autoengaño es un motivo, no puede ser una razón para invalidar una exigencia racional, pues "nada puede ser una razón para rechazar tal exigencia. Nada puede ser visto como una buena razón para razonar de manera distinta a la que va de acuerdo con los propios mejores estándares de racionalidad."

La explicación de Davidson acerca del autoengaño es controversial y provocadora. En efecto, él da una explicación que al final evade las paradojas que al principio parecía enfrentar; sin embargo, en ciertos aspectos parece que sólo ha llevado los problemas a un nivel de abstracción superior. ¿Cómo es que se trazan las barreras mentales que permiten el autoengaño? ¿Por qué en algunos casos un motivo se alza por

encima de las mejores razones del agente y las anula, y en otros casos no? Davidson, en un artículo publicado muchos años más tarde, reconoce que su tratamiento del autoengaño no explica la etiología de este estado mental, pero dice también que nunca tuvo la intención de hacerlo[12]. No podemos culparlo por su selección de temas. Sin embargo, es posible que haya buenas razones para explorar otro tipo de teorías, por ejemplo, si encontramos que, por una parte, hay razones para rechazar las afirmaciones de la aproximación léxica acerca del autoengaño (de la cual proviene la explicación de Davidson) y que, por otra parte, hay alguna explicación alternativa que puede dar cuenta de los mismos fenómenos sin necesidad de postular divisiones en la mente.

Alfred Mele defiende una postura de este tipo. Su postura es, como él mismo la llama, deflacionaria[13], ya que, al no considerar al autoengaño como un fenómeno estrechamente paralelo al engaño interpersonal, su postura no pretende satisfacer los requisitos planteados por las afirmaciones de la aproximación léxica. Su argumento para apoyar esta estrategia es que en el habla cotidiana pueden identificarse también usos de la palabra "engaño" que no refieren a situaciones donde alguien que posee cierta creencia verdadera actúa con la intención de producir la creencia opuesta en otra persona, causándole así esa creencia falsa. Decimos en ocasiones, por ejemplo, que alguien "vive engañado" en algún respecto por el mero hecho de que está equivocado sobre ese tema, sin que sea posible atribuir a ninguna acción específica de alguien más (o suya), con la intención de producir una creencia falsa, la causa de su creencia equivocada relevante; o decimos también que "las apariencias engañan", sin que ello implique que en tales casos

[12] "Who is Fooled" [1997], en *Problems of Rationality*, Oxford, Clarendon Press, 2004; p. 221
[13] Mele, *op. cit*., p. 2

haya algo o alguien que tenga la intención de engañar. En este contexto, "engaño" significa simplemente la acción o el estado de creer falsa o equivocadamente.

El autoengaño, sin embargo, según Mele, no puede consistir simplemente en una creencia equivocada accidental; siempre que se habla de autoengaño, es posible encontrar algún elemento en el contenido de la creencia falsa resultante que hace referencia a un tema acerca del cual la persona, dados sus intereses, no es imparcial. Mele argumenta que los casos más comunes de autoengaño son causados por la influencia tendenciosa [*biasing influence*] que los deseos de una persona tienen en su recolección e interpretación de la evidencia.[14] A diferencia de Davidson, Mele sugiere que hay una explicación relativamente sencilla del fenómeno del autoengaño que no apela a intenciones de creer proposiciones que el agente sabe o al menos tiene la sospecha de que son falsas, o a divisiones en la mente. Podemos explicar esos casos, dice Mele, apelando a la bien conocida influencia tendenciosa que nuestros deseos en ocasiones tienen sobre nuestras creencias. Por estas razones, Mele afirma que su explicación del autoengaño está más estrechamente ligada con lo que él llama la "postura antiagencial" de la creencia motivacionalmente influenciada que con la "postura agencial". Según la postura agencial, "todas las creencias motivacionalmente influenciadas son intencionalmente producidas o protegidas. En cualquier instancia de una creencia motivacionalmente influenciada en *p*, intentamos llevar a cabo que nosotros adquiramos o conservemos la creencia de que *p*." Según la postura antiagencial, "ninguna creencia motivacionalmente influenciada es producida o protegida intencionalmente. En ninguna instancia de creencias

[14] *Ibid.*, p. 11.

motivacionalmente influenciadas en *p* uno intenta llevar a cabo que uno adquiera o conserve la creencia de que *p*."[15]

Sin embargo, ¿cómo podríamos explicar, desde la postura antiagencial, qué es lo que causa, y mantiene, el estado mental de quien se autoengaña, especialmente cuando esta persona se enfrenta a un mundo que presenta a cada momento evidencia cuyo balance apoya preponderantemente a la proposición que es contradictoria con el contenido de su creencia? Mele intenta hacer esto identificando y explicando una serie de procesos que pueden contribuir a que alguien llegue a estar autoengañado, y permanezca así frente a un mundo que le presenta más evidencia en contra que a favor de su estado mental.[16]

Mele identifica cuatro "maneras en que el deseo de *A* de que *p* puede contribuir a que *A* crea que *p*" [17]. Estas cuatro maneras son:

> 1) *interpretación deficiente negativa*: "desear que *p* puede llevarnos a interpretar datos como no contando (o no contando fuertemente) en contra de la creencia de que *p*, datos que, en la ausencia del deseo, reconoceríamos fácilmente como contando (o contando fuertemente) en contra de la creencia en *p*.

[15] Mele: *Self-Deception Unmasked*, Princeton University Press, Princeton, 2001; p. 9.
[16] No trataré todavía la distinción, prevista ya por Davidson, que Mele maneja entre autoengaño directo [*straight-self deception*], donde "la gente se autoengaña al creer algo que ellos quieren que sea verdadero" y autoengaño invertido [*twisted self-deception*], donde "la gente se autoengaña al creer algo que ellos quieren que sea falso (y no quieren además que sea verdadero)".
[17] *Ibid.*, p. 26.

2) *interpretación deficiente positiva*: desear que *p* puede llevarnos a interpretar datos como apoyando a *p*, datos que, en la ausencia del deseo, reconoceríamos fácilmente que cuentan en contra de la creencia de que *p*.

3) *atención y enfoque selectivos*: desear que *p* puede llevarnos tanto a no enfocar la atención en evidencia que cuenta en contra de *p* como a enfocar la atención en evidencia que sugiere que *p*; y

4) *recolección selectiva de evidencia*: desear que *p* puede llevarnos tanto a soslayar evidencia fácilmente accesible a favor de no-*p*, y a encontrar evidencia a favor de *p* que es mucho menos accesible.

Mele da varios ejemplos que muestran cómo estos procesos funcionan y contribuyen a instancias comunes de autoengaño, y enfatiza cómo, según él, ninguno de ellos requiere que la persona primero sostenga la creencia en no-*p* y después se provoque a sí misma sostener la creencia falsa en *p* a través de una estrategia para engañarse. Podemos ver cómo funcionan estos procesos desarrollando un ejemplo que da Mele: un académico se propone averiguar si su filósofo favorito sostenía cierta tesis específica que este académico ha defendido. A través de la interpretación deficiente negativa, este académico descarta pasajes, por ser demasiado "oscuros" o "vagos" donde el filósofo en cuestión parece más bien estar argumentando en contra de dicha tesis. A través de la interpretación deficiente positiva, en cambio, este académico selecciona pasajes realmente vagos y donde la conexión con la tesis defendida por el académico no es de ninguna manera

evidente, y se esfuerza por encontrar una interpretación según la cual ahí es donde el filósofo en cuestión defiende la tesis. La atención y el enfoque selectivos y la recolección selectiva de evidencia para entonces ya han jugado también un papel importante en el ejemplo, pues el académico ha desviado su atención de los pasajes donde el filósofo parece estar en contra de la tesis (él considera una pérdida de tiempo detenerse demasiado tiempo en esos puntos), y concentra su atención sólo en aquellos donde la conexión con su tesis no es nada clara. El académico concluye que su filósofo favorito de hecho defendía la misma tesis que él.

Así, Mele explica cómo una persona puede adquirir la creencia de que *p* aun cuando la verdad de la proposición no-*p* es mejor apoyada por la evidencia, apelando a los mecanismos psicológicos recién descritos. Según Mele, nuestros deseos tienen una influencia tendenciosa en la recolección e interpretación de la evidencia relevante para establecer la verdad de una hipótesis. No hay, para él, una necesidad explicativa de suponer que tales instancias de creencia motivacionalmente tendenciosa requieran que el agente actúe intencionalmente (con la intención de engañarse) o sostenga dos creencias inconsistentes.

Mele propone que las siguientes cuatro condiciones son conjuntamente suficientes para "entrar al estado de autoengaño al adquirir la creencia de que *p*", y nos dan un modelo adecuado para caracterizar correctamente la mayoría de los casos de autoengaño[18]:

1. La creencia en *p* que *S* adquiere es falsa.

[18] *Ibid.*, pp. 50-51.

2. *S* maneja los datos relevantes, o al menos aparentemente relevantes, para establecer la verdad de *p* de una manera motivacionalmente tendenciosa.
3. Este manejo tendencioso es una causa no-desviada [*nondeviant cause*] de que *S* adquiera la creencia en *p*.
4. El conjunto de datos que posee *S* en ese momento proporciona mayor justificación para no-*p* que para *p*.

Acerca de la primera condición, me parece importante hacer un comentario aquí. La mayoría de los autores afirma, junto con Mele, que es un requisito del autoengaño el hecho de que la creencia que el agente adquiere en virtud de este fenómeno sea falsa. De esta manera, es conceptualmente imposible que alguien esté autoengañado al creer una proposición que es verdadera. Estoy dispuesto a admitir que, por razones verbales, comúnmente no llamamos engaño (ni autoengaño) al estado de una persona si esa persona no cree algo que es falso en virtud de ese estado. No obstante, creo que hay que resaltar que, cualquiera que sea el problema filosóficamente interesante acerca del autoengaño, ese problema no depende crucialmente de que el resultado del autoengaño sea una creencia meramente falsa, sino más bien de las razones (o falta de razones) que el agente tiene para poseer cierta creencia. Alguien puede tener muy buenas razones para tener determinada creencia, y sin embargo esa creencia bien puede ser falsa; a la inversa, alguien puede sostener una creencia por muy malas razones y la creencia en cuestión puede resultar ser verdadera. Una explicación adecuada del fenómeno del autoengaño, entonces, será capaz de explicar de la misma manera aquellos casos en los que, a través

de un fenómeno análogo al autoengaño, el agente adquiere (o conserva) cierta creencia verdadera.

A través de estas condiciones conjuntamente suficientes para el autoengaño y los procesos psicológicos de formación de creencias motivacionalmente influenciadas, Mele pretende haber explicado las instancias más comunes [*garden variety cases*] del autoengaño, sin necesidad de postular divisiones en la mente del agente o intenciones de engañarse, y apelando sólo a la influencia que nuestros deseos pueden tener sobre nuestras creencias.

III

La postura de Mele, aunque ha alcanzado considerable aceptación en los últimos años entre muchos teóricos del autoengaño, enfrenta algunas serias objeciones. En primer lugar, su explicación no sirve para distinguir adecuadamente el autoengaño de las creencias ilusorias [*wishful thinking*]. Aunque existen dificultades para identificar satisfactoriamente en qué consiste este último fenómeno, existe cierto acuerdo acerca de que las creencias ilusorias son creencias producidas directamente por deseos, donde el deseo de que *p* produce la creencia de que *p* sin la intervención del agente, como se dijo antes en este capítulo al explicar la comparación de Davidson entre autoengaño y creencias ilusorias. Si el autoengaño no requiere una intención por parte del agente de engañarse y puede ser explicado solamente en virtud de sus deseos, no parece haber diferencia sustancial entre un caso de autoengaño y uno de creencias ilusorias. La única y poco sugestiva diferencia entre las creencias ilusorias y el autoengaño podría consistir, según la explicación de Mele, en que en el caso de las primeras no se hace uso de las

estrategias de manipulación tendenciosa de los datos (interpretación deficiente negativa y positiva, enfoque selectivo de la atención y recolección selectiva de evidencia). Pero si esto es correcto, podemos válidamente preguntarnos entonces qué es lo que causa que en los casos de creencias ilusorias el deseo de que *p* engendre la creencia de que *p*, si ninguna de esas estrategias está en juego.

Por otra parte, y esta crítica es más importante, Mele no da cuenta de aquellos casos en donde pareciera que el autoengaño es mantenido en existencia por la creencia (verdadera) opuesta que lo motivó, es decir, aquellos casos en los que la guía que sólo podría proporcionar la creencia verdadera relevante parece ser necesaria para que la persona autoengañada esquive con precisión la evidencia a la que está constantemente expuesta —sobre todo ciertos tipos de evidencia abrumadora que harían imposible que el autoengaño persistiera. En estos casos de autoengaño, el agente realiza ciertas acciones que son muy difíciles de explicar a menos que le atribuyamos esa creencia verdadera.

Daré a continuación un ejemplo detallado para ilustrar lo anterior. Andrea afirma con sinceridad que cuenta con buena salud. Hace ejercicio regularmente y se esfuerza por llevar una dieta saludable. Hace seis meses, sin embargo, se enteró que un antiguo novio, con quien vivió durante dos años, se encontraba muy enfermo de cáncer. Para entonces, hacía un año que Andrea había roto toda comunicación con él, y no volvió a contactarlo. Poco tiempo después, supo que esta persona había muerto, y escuchó rumores de que su enfermedad no había sido realmente cáncer, sino sida, y que sus familiares y amigos sólo habían dicho que tenía cáncer para "guardar las apariencias". Andrea, desde la muerte de su antiguo novio, ha evitado siempre hablar o tener cualquier contacto con familiares y amigos de él, aunque eso le haya significado perder varias amistades, y no ha hablado

sobre este tema con nadie más. Ella, desde hace tres meses, ha empezado a notar ciertas anomalías en su cuerpo: le han aparecido algunas manchas en la piel, y a menudo se siente débil y con fiebre. No ha ido al médico en todo ese tiempo; piensa que las manchas en su piel se deben a que la última vez que fue a la playa no usó protector solar, y atribuye la debilidad y la fiebre a su intensa carga de trabajo. En una ocasión, hace unas semanas, un familiar de Andrea que fue sometido a cirugía necesitaba una transfusión del mismo tipo de sangre que Andrea tiene, un tipo de sangre poco común. Aunque Andrea ya había donado sangre en dos ocasiones hacía varios años, esa vez se rehusó a hacerlo, argumentando que las personas que tienen tatuajes no pueden donar sangre, algo que escuchó alguna vez, y ella tiene un tatuaje. Su trabajo la agobia y ha estado buscando uno nuevo. Muy recientemente, después de varias largas entrevistas y exámenes en una empresa, ella tuvo la oportunidad de obtener un trabajo en donde le hubieran pagado el doble de lo que gana actualmente y acerca del cual estaba muy entusiasmada, pero cuando le pidieron que se hiciera un examen de sangre como último requisito para comenzar a trabajar ahí, ella rechazó el trabajo, diciendo que esas prácticas atentan contra los derechos de los empleados, especialmente de las mujeres, y que a ella no le gustaría trabajar en un lugar como ése. Además, aunque Andrea toma pastillas anticonceptivas desde muy joven, en estos meses ha accedido a tener relaciones con su pareja actual sólo si usan condón. Cuando llena solicitudes de empleo, Andrea contesta sinceramente "excelente" a la pregunta sobre cómo describiría su estado de salud actual.

Añadiendo la circunstancia de que ella realmente está enferma de sida, creo que la mayoría de los teóricos aceptaría que, en este caso, Andrea está autoengañada acerca de su salud. El aspecto particular que me interesa de este ejemplo (y que, me parece, es

común a muchos otros casos de autoengaño) es la manera en que el comportamiento de Andrea evade sistemáticamente —incluso de manera inteligente y sutil— cualquier situación en la que ella podría tener que enfrentarse con evidencia sustancial de que está enferma de sida, a través de acciones que ella puede explicar con base en otros deseos.

Eric Funkhouser denomina a estas acciones *comportamiento evasivo* [*avoidance behavior*].[19] Aun cuando para cada una de estas acciones el agente autoengañado tiene una explicación con base en otros deseos y creencias genuinas, a medida que estas acciones se van multiplicando, cada vez es más difícil decir que el agente no sabe (o al menos no cree) aquella proposición cuya evidencia está evitando. En el ejemplo, Andrea ha evitado no solamente hacerse la prueba del sida después de tener claros indicios para pensar que su antiguo novio murió de esa enfermedad, que es justo lo que Andrea recomendaría a cualquiera (otra persona) que estuviese en esas circunstancias, sino también ir con el médico para revisar las manchas de su piel, donar sangre para el familiar que la necesitaba, tener contacto con cualquier familiar o amigo de su antiguo novio y hacerse exámenes de sangre para su posible nuevo trabajo. Todo esto, junto con las razones que ella ha ofrecido para su comportamiento, y dados sus intereses, revela que ella está evitando sistemáticamente ponerse en una posición donde potencialmente tuviera que enfrentarse con evidencia sustancial para establecer la verdad de la proposición que dice que ella está enferma de sida. Andrea no es tonta, y fácilmente podría darse cuenta hacia dónde apuntan las pruebas en un caso de este tipo si se tratase de otra persona. ¿Cómo logra, entonces, afirmar sinceramente que cuenta con buena salud?

[19] Funkhouser, "Do the Self-Deceived Get What They Want?", p. 297.

Este tipo de ejemplos, de ser plausibles, demuestran que el análisis de Mele es insuficiente para caracterizar al menos una clase importante de casos de autoengaño, aquélla en la cual el agente presenta comportamiento evasivo sistemático, o, en los términos de Davidson, aquellos casos en los que la creencia original coexiste con, y mantiene en existencia a, la creencia inducida. Cabría mencionar aquí una distinción trazada por Robert Audi relevante al considerar estos casos. Según Audi, podemos distinguir entre autoengaño [*self-deception*] y algo que él nombra "autoilusión" [*self-delusion*]. Él reconoce que "ilusión" [*delusion*, que también puede traducirse como "delirio" o "alucinación"] es un término bastante fuerte para lo que él quiere nombrar; lo único que necesita es una palabra distinta a "autoengaño" para llamar a ciertos casos que tienen características ligeramente distintas al autoengaño, y, a falta de una mejor palabra, utiliza "autoilusión". Después de mencionar un ejemplo en donde una persona se autoengaña acerca de su propia valentía, él dice:

> Hay una cierta tensión que es característica del autoengaño y que explica parcialmente su típica inestabilidad: el sentido de evidencia en contra de *p* jala [a la persona autoengañada] en dirección contraria al engaño y amenaza levantar el velo que esconde de su conciencia su conocimiento de que no es valiente; los deseos o necesidades que fundamentan el autoengaño jalan en dirección contraria de su comprensión de la evidencia y amenazan bloquear su percepción de la verdad. Si la primera fuerza prevalece, uno puede ver la verdad llanamente y deja de estar engañado; si la segunda prevalece, uno pasa del autoengaño a la

> autoilusión y no ve la verdad en absoluto. Pienso que el autoengaño existe sólo cuando hay un balance entre estas dos fuerzas.[20]

Los casos que no presentan lo que aquí se ha llamado comportamiento evasivo (o una disposición a ese tipo de comportamiento) carecen de la tensión que Audi menciona. En tales casos, la víctima del engaño "no ve la verdad en absoluto" acerca del tema relevante, y su comportamiento se ajusta por completo al de alguien que tiene la creencia equivocada. Desde esta perspectiva, Mele sólo ha logrado ofrecer una explicación de la autoilusión, sin siquiera rozar el tema del autoengaño tal como Audi lo considera. En efecto, Mele utiliza consistentemente ejemplos en donde es relativamente claro que el agente autoengañado ha perdido, antes que el autoengaño se consume, la creencia verdadera relevante, si es que alguna vez la tuvo. Por otra parte, de la cita se colige que Audi atribuye al agente autoengañado *conocimiento* —aunque escondido de su conciencia— de esa proposición. ¿Es esta atribución correcta? ¿No nos llevaría esto a postular otra vez creencias abiertamente contradictorias en la mente del agente? De ser así, tendríamos que volver a una explicación del tipo de la de Davidson, que lidie con las paradojas del autoengaño. Pero Audi no piensa que esto sea necesario.

En el ejemplo de Andrea, puede afirmarse con seguridad que ella satisface a primera vista al menos algunos de los requisitos para atribuirle al menos la sospecha de que tiene sida (como su comportamiento evasivo sistemático, que sólo puede ser explicado de esa manera), así como también algunos de los requisitos para atribuirle la creencia de que no tiene sida (como el afirmar sinceramente que cuenta con buena salud).

[20] Robert Audi: "Self-Deception, Rationalization and the Ethics of Belief", en su libro *Moral Knowledge and Ethical Character*, Oxford University Press, 2002; p. 144.

Audi no piensa que debamos concluir entonces que ella tiene creencias abiertamente contradictorias. Según él, las circunstancias del caso pueden justificar que atribuyamos a Andrea, sin caer en paradojas, por una parte, la creencia de que tiene sida, y, por otra, la creencia *de segundo orden* de que *cree* que no tiene sida. A la hora de atribuir creencias, dice Audi, "las acciones hablan más alto que las palabras".[21] Es decir, hay que otorgar preferencia al comportamiento no-lingüístico de la persona por encima de su comportamiento lingüístico al atribuirle creencias de primer orden a esa persona. La expresión más natural de una creencia de primer orden, según este curso de ideas, no es afirmar que *p*, sino comportarse (no-lingüísticamente) como si *p*. Sus afirmaciones sinceras son más relevantes, en cambio, al atribuirle creencias *de segundo orden*, es decir, creencias acerca de sus propias creencias. La creencia de segundo orden puede no reflejar la creencia de primer orden, si esta última es inconsciente. Eric Funkhouser, apoyando este punto, dice:

> Cuando alguien afirma "*p*", generalmente tomamos esto como evidencia de que la persona cree que *p*. Esto es porque generalmente atribuimos a la gente un deseo de decir la verdad *y* asumimos que ellos tienen algún acceso privilegiado con respecto a lo que ellos creen. Así, dados el motivo y la habilidad apropiados, las afirmaciones de los demás pueden proporcionarnos ventanas a su psicología. Pero si se carece ya sea del motivo o de la habilidad, entonces tenemos una condición derrotante. Sugiero que cuando falta el motivo, ellos nos están engañando; pero cuando falta la habilidad, ellos a menudo están autoengañándose.[22]

[21] Audi: "Self-Deception and Rationality", p. 173.
[22] Funkhouser: "Do the Self-Deceived Get What They Want?", p. 309.

Los análisis contemporáneos más reconocidos sobre este tema generalmente dan por supuesto que en cualquier caso de autoengaño el sujeto posee la creencia falsa relevante. Mele y Davidson, por ejemplo, asumen explícitamente que una persona que está autoengañada acerca de la verdad de *p* posee la creencia falsa en no-*p*, y entonces, para evitar la paradoja estática, proceden ya sea a tratar de explicar cómo esto no implica en ningún caso que la persona crea al mismo tiempo que *p* (Mele), o bien que su creencia falsa puede coexistir con la creencia verdadera en *p* si existe una división mental que las mantenga separadas (Davidson). Ni siquiera exploran la posibilidad de que la persona autoengañada no llegue nunca a la creencia de primer orden falsa. La interpretación de Audi, sin embargo, puede ser una manera fiel de representar los datos sin atribuir creencias contradictorias, evitando así caer en la paradoja estática, pues la creencia verdadera de primer orden en *p* de esa persona —en el caso de Andrea, creer que tiene sida— no es contradictoria con su creencia falsa de segundo orden de que cree que no-*p* —en el caso de Andrea, creer que cree que no tiene sida.[23]

En situaciones normales, una creencia de segundo orden refleja acertadamente el contenido de la creencia de primer orden de la cual trata. Es preciso encontrar entonces una explicación para los casos de creencias de segundo orden falsas que coexisten con creencias (de primer orden) opuestas a las que servirían para justificar tales creencias de segundo orden. Audi afirma que en los casos de autoengaño esas creencias de primer orden son inconscientes, y que en tales casos existe al menos un deseo de la persona que explica tanto por qué su creencia de primer orden es inconsciente como por qué la

[23] Un antecedente importante de esta postura puede encontrarse en D.H. Mellor: "Unconscious Belief", en *Proceedings of the Aristotelian Society*, 77, 1977-78; pp. 87-101.

persona está dispuesta a afirmar sinceramente que tiene una creencia de segundo orden en el sentido opuesto.

Audi presenta entonces condiciones que él considera necesarias y suficientes para los casos paradigmáticos de lo que él denomina autoengaño.[24] Según él, una persona *S* está autoengañada acerca de una proposición *p* si, y sólo si:

> 1) *S* inconscientemente sabe que no-*p* (o tiene razones para creer, e inconscientemente cree, que no-*p*).
>
> 2) *S* sinceramente "afirma"[25] [*avows*], o está dispuesto a afirmar sinceramente, que *p*; y
>
> 3) *S* tiene al menos un deseo que explica, en parte, tanto por qué *S* está dispuesto a afirmar que no cree que no-*p*, y a afirmar que *p*, aun cuando se le presenta lo que él ve que es evidencia en contra de *p*.

Tal como las entiende Audi, las creencias inconscientes requeridas por el autoengaño "no necesitan estar profundamente enterradas en una Mente Inconsciente Freudiana. Ellas simplemente no son accesibles a *S* sin ayuda exterior o al menos un auto-escrutinio cuidadoso. Sólo necesitan estar lo suficientemente veladas de *S* como para hacer que sus afirmaciones de las proposiciones contradictorias sean sinceras".[26] Esta característica de los estados mentales participantes en el autoengaño conserva, en mi opinión, un fuerte parecido a las fronteras mentales que postulaba Davidson; de hecho, aunque una creencia

[24] Audi: "Self-Deception and Rationality", p. 173.
[25] Ésta no es una traducción exacta del término "*avows*"; afirmar, en este sentido especial, requiere que el agente acepte la proposición conscientemente dentro de su mente, y sólo eso.
[26] *Ibid.*, p. 174.

de primer orden no pueda ser contradicha por una creencia de segundo orden, una división mental del tipo de las que menciona Davidson seguramente será necesaria para evitar ya sea que la creencia de primer orden sea accesible a la conciencia de la persona y ésta forme una creencia de segundo orden correspondiente acertada (lo cual terminaría con el autoengaño), o bien que la creencia falsa de segundo orden propicie la eliminación de la creencia de primer orden y deje el camino libre para una creencia de primer orden correspondiente (convirtiéndose entonces en un caso de autoilusión).

Acerca del deseo mencionado en la tercera condición, Audi no da mayores especificaciones sobre el contenido de este deseo, excepto el hecho de que surge de una necesidad de la persona de proteger su ego, y por ello en ocasiones llama a estos deseos "necesidades del ego" [*ego needs*].[27] Audi no explica cómo es que este deseo (o cualquiera otro de los elementos que conforman sus condiciones para el autoengaño) puede proporcionar una guía al agente para evadir inteligentemente la evidencia que desvanecería el autoengaño.

Eric Funkhouser aporta, a mi parecer, el elemento de especificidad al contenido de ese deseo que hace más completa la explicación del autoengaño de Audi. Según Funkhouser, quien presenta una explicación del autoengaño que sigue las mismas líneas que la de Audi, la persona que se autoengaña acerca de *p* está motivada, al realizar las acciones que desembocan en su autoengaño, por un deseo de creer que *p*.[28] La guía que este deseo proporciona no es intencional ni consciente, evitando así caer en la paradoja dinámica. El comportamiento evasivo que es característico del autoengaño según la perspectiva de Audi refleja, al menos, que la persona autoengañada cree que no-*p* y desea

[27] *Ibid*., p. 149.
[28] Funkhouser, *op. cit.*, p. 299.

creer que *p* independientemente de que *p* sea verdadera, aunque típicamente refleja también que la persona desea que *p*.

Una objeción que es aplicable a la explicación conjunta de Audi y Funkhouser es que no han explicado de dónde surgen las acciones intencionales que son necesarias para el comportamiento evasivo característico del autoengaño según su misma explicación. Audi sólo menciona un deseo que explica por qué la creencia de segundo orden se mantiene a nivel inconsciente, y Eric Funkhouser especifica el contenido de esta creencia diciendo que es un deseo de creer cierta proposición.

Podría pensarse que también los casos de autoengaño invertido [*twisted self-deception*] presentan un obstáculo para esta postura. En tales casos, quien se autoengaña acerca de *p* tiene el deseo de que no-*p*. (Un ejemplo clásico de este tipo de autoengaño sería cuando un hombre se autoengaña al creer que su esposa le es infiel, mientras que él desea fuertemente que su esposa sea fiel.) Aunque el sujeto del autoengaño desee que no-*p*, esto no es incompatible con que al mismo tiempo él desee *creer* que *p*. En efecto, puede ser que él tenga preparadas algunas estrategias de emergencia para hacer frente a la situación en caso de que *p* suceda, y por tanto él querría enterarse lo más pronto posible de esa situación con el fin de poner en marcha sus estrategias (el marido del ejemplo podría tener el plan de intentar ser más cariñoso con su esposa para conquistarla otra vez, o de reprenderla fuertemente para que ella deje de engañarlo). En tales circunstancias, no es difícil suponer que las personas en ocasiones desean creer, como medida de precaución, algo que no desean que suceda.[29]

[29] Algo análogo puede decirse del pesimista. ¿Qué es un pesimista si no alguien que tiene una tendencia a creer demasiado pronto que algunas de las cosas que desea, o que al menos considera deseables, no se han cumplido ni se cumplirán? Postular en él un deseo genérico de creer, en tanto que pesimista, que no podrán satisfacerse otros deseos suyos no parece demasiado arriesgado.

IV

Recapitulando acerca de las perspectivas sobre el autoengaño que se han presentado aquí, hemos visto que la postura de Davidson complica innecesariamente los casos del autoengaño al seguir una analogía demasiado estrecha entre éste y el engaño interpersonal. No es necesario, para hablar de un caso de autoengaño, hacer referencia a una intención específica de alguien que provoca el engaño, ni a creencias abiertamente inconsistentes de la persona autoengañada. (Recuérdese que Davidson requiere que en el autoengaño la persona posea o bien creencias contradictorias, en los casos más fuertes, o bien una creencia inducida que sea inconsistente con la preponderancia de la evidencia en contra de esa proposición y el principio de evidencia total, en los casos más débiles.) La postura de Mele, por su parte, no es capaz de distinguir entre el autoengaño y otros tipos de creencias motivadas, como las creencias ilusorias, ni da cuenta de los casos filosóficamente más problemáticos de autoengaño, en los cuales la persona guía las acciones que le permiten alcanzar y conservar el estado del autoengaño por medio de la creencia verdadera acerca de la cual está autoengañada. Audi y Funkhouser, sin necesidad de negar que existan los casos que Mele describe, pero otorgando a esos casos el nombre distinto de "autoilusión", presentan una postura que logra rescatar y poner en primer plano los casos más problemáticos de autoengaño, apelando a una creencia verdadera de primer orden inconsciente y a una creencia falsa de segundo orden, junto con un deseo de creer en la proposición sobre la cual el autoengaño versa, siendo su descripción fiel a los detalles de este tipo de autoengaño (una vez que se ha diferenciado éste de la autoilusión) y escapando tanto de la paradoja estática como de la dinámica.

Por último, quisiera considerar la relación entre el autoengaño, desde la concepción desarrollada por Audi y Funkhouser, y las creencias ilusorias. Como ya se mencionó, en un caso de creencias ilusorias el deseo de una persona provoca en ella la creencia correspondiente a ese deseo. Audi argumenta que su explicación sirve para distinguir entre el autoengaño y las creencias ilusorias en el sentido de que "la persona que se autoengaña no llega a creer la proposición que quiere creer"[30]. Aunque es estrictamente cierto que la persona que se autoengaña, desde la perspectiva de Audi y Funkhouser, no llega a creer la proposición que quiere creer (es decir, *p*), ése no es el resultado que podría esperarse de un deseo de creer que *p* en un caso de creencias ilusorias. Si lo que motiva el autoengaño es un deseo de creer que *p*, la creencia correspondiente que podría ser provocada en el agente por este deseo a través del fenómeno de las creencias ilusorias no es la creencia (de primer orden) de que *p*, sino la creencia (de segundo orden) de que cree que *p*, es decir, precisamente la creencia de segundo orden (falsa) que, en la explicación de Audi y Funkhouser, posee quien está autoengañado. En otras palabras, el autoengaño, desde esta perspectiva, es un tipo especial del fenómeno de creencias ilusorias.[31]

Esto no significa que no podamos distinguir entre las creencias ilusorias y el autoengaño; solamente necesitamos decir que hay casos simples de creencias ilusorias, donde un deseo de primer orden provoca la creencia de primer orden correspondiente, y

[30] Audi, "Self-Deception and Rationality", p. 174.

[31] Davidson argumenta que las creencias ilusorias no siempre pueden ser un ingrediente del autoengaño, puesto que en las creencias ilusorias la creencia toma la dirección del afecto positivo, mientras que en el caso del autoengaño el pensamiento puede ser doloroso. Me parece que esto es equivocado, pues las creencias ilusorias requieren solamente que el agente tenga cierto deseo que provoque la creencia correspondiente, sin importar, en mi opinión, si el agente tiene otros deseos contrarios o si la creencia resultante es o no placentera. La definición de creencias ilusorias dada por el *OED* apoya esta sugerencia (ver nota 31).

casos de creencias ilusorias más complejos, donde un deseo de tener una creencia provoca la creencia de segundo orden correspondiente.

Dije con anterioridad que la explicación del autoengaño de Mele no sirve para distinguir entre autoengaño y creencias ilusorias. Siguiendo ahora la postura desarrollada por Audi y Funkhouser, podemos ahora decir que esa explicación no sirve siquiera para identificar el autoengaño, sino un fenómeno distinto, la autoilusión. Sin embargo, las cuatro estrategias que Mele menciona para explicar cómo se produce la manipulación de datos que lleva a lo que él considera que es el autoengaño (interpretación deficiente positiva y negativa, atención y enfoque selectivos y recopilación selectiva de evidencia) muy bien pueden servir para explicar las creencias ilusorias. Después de todo, él mismo presenta esas cuatro estrategias como "maneras en que el deseo de *A* de que *p* puede contribuir a que *A* crea que *p*", y eso es exactamente lo que se necesita para explicar las creencias ilusorias. Y éste sería el elemento que faltaba a la explicación del autoengaño de Audi y Funkhouser: a través de una interpretación deficiente positiva y negativa, atención y enfoque selectivos y una recopilación selectiva de la evidencia (nótese que todas estas actividades están compuestas por acciones intencionales), es posible explicar cómo el deseo de creer cierta proposición, en la explicación de Funkhouser, causa el comportamiento evasivo del agente. Si no me equivoco, Mele habrá explicado entonces al menos parcialmente los mecanismos psicológicos que hacen posible las creencias ilusorias, y esta explicación será útil para explicar parcialmente, a su vez, cómo se da el fenómeno del autoengaño, al no ser éste más que una clase especial y más compleja del fenómeno de las creencias ilusorias.

Capítulo 3. Autoengaño, responsabilidad y autonomía

En este último capítulo intentaré extraer algunas consecuencias, a partir de lo dicho hasta ahora, acerca de la relación entre autoengaño, por una parte, y los temas de responsabilidad y autonomía, por otra. Según dije en el capítulo segundo, una estrategia común para intentar capturar el significado del término "autoengaño" es utilizar lo que Alfred Mele ha llamado "la aproximación léxica", según la cual el autoengaño es sencillamente la acción y el efecto de engañar (en el sentido de engañar intencionalmente a alguien más) cuando el perpetrador y la víctima del engaño son la misma persona. La aproximación léxica hacia el fenómeno del autoengaño conduce a un paralelismo entre autoengaño y engaño a otros. Este paralelismo, a su vez, generalmente lleva a sus proponentes a otras dos tesis acerca del autoengaño. La primera es que el autoengaño es siempre intencional. La segunda es que el autoengaño es siempre, en algún grado, reprochable. He argumentado en contra de la aproximación léxica y también en contra de la tesis de que el autoengaño es intencional. Queda entonces por examinar la cuestión sobre si el autoengaño es en algún sentido reprochable.

Según la explicación del autoengaño que he ofrecido, este fenómeno puede entenderse como un caso especial del fenómeno de las creencias ilusorias [*wishful thinking*]. Para los autores que consideran que las creencias ilusorias no involucran la intervención del agente (por ejemplo, Davidson y Pears), el agente no es responsable de las creencias resultantes. Según ellos, el deseo de que *p* tiene una tendencia a generar la creencia de que *p*, y la persona no es responsable de los efectos que sus deseos puedan

tener por sí mismos. Bajo esta concepción de las creencias ilusorias, y si es correcta mi afirmación de que el autoengaño es sólo un caso especial de este fenómeno, parecería que el agente autoengañado nunca es responsable del autoengaño. El problema con esta concepción de las creencias ilusorias es que no explica cómo es que un deseo puede generar la creencia correspondiente. ¿Por qué generaría exactamente esa creencia, y no cualquier otra? Y, ¿por qué algunos deseos tienen esa capacidad y otros no? Mi opinión es que el fenómeno de las creencias ilusorias involucra un proceso en el cual juegan un papel importante los mecanismos que Mele propone para explicar la formación de creencias motivacionalmente influenciadas (interpretación deficiente negativa y positiva, enfoque y atención selectivos y recolección selectiva de evidencia), y tal vez algunos otros similares. De ser así, las creencias ilusorias sí requieren alguna intervención del agente, de hecho, una intervención en la forma de acciones intencionales.

Antes de comenzar a tratar los temas de responsabilidad y autonomía en relación con el autoengaño, creo que será útil intentar aclarar un poco estos conceptos. Conviene recordar algunas distinciones en el uso del término "responsabilidad", que estarán en juego a lo largo de la discusión. En algunas aplicaciones de este término, ser responsable de algo únicamente quiere decir que es correcto atribuir a alguien (o a algo) el papel de *causa* para cierta situación o fenómeno. Otro uso de esta palabra se hace cuando al afirmar que alguien es responsable de algo se quiere decir con ello que esa persona es *digna* de algún tipo de aprobación o reprobación en virtud de esa relación. Uno más es cuando ser responsable de algo significa tener cierta *obligación* específica, por ejemplo, de tener algún cuidado especial acerca de cierto objeto o actividad. Por

último, cuando se dice que alguien es responsable puede también quererse significar con ello que esa persona posee cierta *virtud* (el vicio correspondiente consistiría en ser irresponsable), la virtud de cumplir normalmente y de manera satisfactoria con sus deberes.

Entre las cuestiones que son relevantes para atribuir responsabilidad a un agente por sus acciones u omisiones, se encuentran las siguientes:[1] a) ¿Son sus acciones u omisiones intencionales? b) ¿Son sus acciones u omisiones el resultado de una decisión deliberada de su parte? c) ¿Son las consecuencias de sus acciones u omisiones previstas por el agente, o previsibles? d) ¿Sabe el agente realmente lo que está haciendo al realizar sus acciones u omisiones? e) ¿Se identifica el agente con los deseos y motivos que lo llevan a realizar sus acciones u omisiones? f) ¿Se aprovecha o beneficia el agente de las consecuencias de sus acciones u omisiones? g) ¿Tiene el agente la posibilidad realista de evitar realizar sus acciones u omisiones? Todas estas cuestiones, me parece, deben ser tomadas en cuenta al examinar si un agente puede ser considerado responsable de caer en el autoengaño y de las consecuencias resultantes de su autoengaño.

En cuanto al concepto de autonomía, esta palabra, que originalmente hacía referencia sólo a la capacidad de una ciudad o estado para regirse según sus propias leyes y sin interferencias externas, en este contexto refiere al "autogobierno" o capacidad de autodeterminación que una persona puede ejercer sobre sí misma, cuando ella misma es la autora, en un sentido relevante, de sus decisiones y actos. Aclarar más específicamente cuál es este sentido relevante es una tarea difícil, pues implica tomar

[1] Sigo aquí las ideas de Mark Platts en la introducción a *Responsabilidad y libertad*, IIF-UNAM, México, 2002.

en cuenta todos los tipos de obstáculos que puede haber para que una persona ejerza tal autogobierno eficazmente. Por otra parte, la discusión contemporánea sobre la idea de autonomía surge, en gran medida, a partir de la introducción del concepto de identificación hecha por Harry Frankfurt en su artículo de 1971 "Freedom of the Will and the Concept of a Person". A partir de ese artículo, y en trabajos posteriores, Frankfurt desarrolla una explicación de la autonomía según la cual ésta debe ser entendida como la identificación de un agente con los motivos que lo llevan a actuar. Se explicará este concepto a continuación.

Un agente puede tener a cada momento muchos deseos diferentes, los cuales dan lugar a diversos motivos del agente para actuar y a diferentes cursos de acción. Cualquier acción intencional del agente estará necesariamente basada en alguno de sus motivos; sin embargo, el agente puede identificarse en mayor o menor grado con los motivos que lo impulsan a la acción. Por ejemplo, una persona que está tratando de dejar de fumar tiene un deseo de fumar (por el placer o la satisfacción que esto le provocaría), pero al mismo tiempo tiene un deseo de no fumar (por ejemplo, para conservar su salud). Esta persona, a pesar de sus intentos por dejar de fumar, puede ser llevada a la acción de fumar por el deseo correspondiente. En tal caso, la persona es movida por un deseo por el cual no desea ser movida, un deseo con respecto al cual la persona se siente ajena; usando la terminología de Frankfurt, en estos casos, la persona no se *identifica* con su motivo que la lleva a actuar, y por tanto, según Frankfurt, la persona carece de autonomía. Identificarse con un deseo significa entonces tener el deseo (de orden superior) de tener cierto deseo (de orden inferior) que el agente ya tiene; significa que un agente aprueba la posesión de cierto deseo suyo. El concepto de

autonomía de Frankfurt, por tanto, descansa sobre una distinción aparentemente sencilla: de los deseos que nos impulsan a actuar, sólo algunos de ellos son deseos que nos gustaría tener o que aprobamos tener. Cuando alguien actúa con base en un deseo que no aprueba (o que no le gustaría) tener, esa persona no es autónoma.

Según Frankfurt, en tal distinción descansa una clave para entender no sólo el concepto de autonomía, sino también el de responsabilidad. Dice: "en la medida en que una persona se identifica con los motivos de sus acciones, ella toma responsabilidad por tales acciones y adquiere responsabilidad moral por ellas; más aún, las cuestiones acerca de cómo las acciones y sus identificaciones con sus motivos son causadas son irrelevantes para las cuestiones sobre si ella realiza las acciones libremente o si es responsable moralmente por realizarlas."[2] El problema al cual se refiere Frankfurt principalmente con esta afirmación es el de si las acciones de una persona pueden ser consideradas libres o autónomas (y si puede atribuírsele responsabilidad a la persona por ellas), frente a la posibilidad de un marco determinista donde tanto las acciones como los estados mentales de la persona son productos de otras causas más remotas. Acerca del tema de la responsabilidad, no es nada claro por qué Frankfurt piensa que una persona es responsable de sus acciones sólo en la medida en que ella se identifica con sus motivos, como puede notarse en el pasaje citado. Comúnmente no eximimos a las personas de la responsabilidad por sus acciones por el simple hecho de que ellos mismos no aprueben los motivos que los condujeron a ellas.

Por otra parte, y sin necesidad de adentrarse en el problema del marco determinista, es posible señalar cómo la identificación de una persona con sus motivos para actuar en muchos casos no es suficiente para que su acción pueda ser considerada

[2] Frankfurt: *The Importance of What We Care About*, Cambridge University Press, 1988; p. 54

libre o autónoma. Sobre esto, Gerald Dworkin afirma: "las reflexiones de segundo orden no pueden ser el todo de la historia de la autonomía, pues tales reflexiones (…) pueden haber sido influenciadas por otras personas o circunstancias de tal manera que no vemos a esas evaluaciones como propias de la persona."[3]

En los casos de manipulación, la víctima usualmente termina identificándose con los motivos por los cuales actúa; sin embargo, normalmente no diríamos que en esos casos la víctima de la manipulación actúa de manera autónoma. La manipulación puede tomar varias formas. En algunos casos (por ejemplo, en el adoctrinamiento o en el "lavado de cerebro"), la interferencia se ejerce en buena parte sobre los deseos, preferencias y valores del agente, haciéndole adquirir, sin su consentimiento y probablemente sin que se dé cuenta de ello, materiales que podrán influir considerablemente en sus futuras decisiones y acciones. Aunque el agente, llegado el momento, se identifique con sus motivos para actuar, la idea de autogobierno a la cual la noción de autonomía parece hacer referencia no está presente en estos casos.

Otras formas distintas de manipulación (y éstas serán especialmente importantes para el tema del autoengaño) son las que se dan a través del engaño y la omisión de dar información relevante que el agente tiene derecho a conocer. En el primer caso, el agente, debido a alguna interferencia con su proceso normal de formación de creencias, adquiere una o más creencias falsas, que, en combinación con los deseos del agente, pueden dar lugar a motivos para actuar. En tales situaciones, si el engaño puede ser correctamente calificado como una influencia externa, el agente que actúa influido por engaño ha sido obstaculizado en su capacidad para decidir por cuál de sus diversos motivos actuar. En el segundo caso, al no poseer el agente toda la información

[3] *The Theory and Practice of Autonomy*, Cambridge University Press, 1988; p. 18

relevante (debido a que alguien que la poseía y tenía la obligación de dársela omitió hacerlo, o incluso activamente se la escondió), es posible que el agente tome decisiones que no habría tomado a no ser por esta interferencia, y actúe con base en ellas. En ambos casos, aunque el agente se identifique con sus motivos para actuar, tampoco está presente aquí la idea de autogobierno que es necesaria para la autonomía.

Aparte de la manipulación, otro caso que es importante mencionar es el del fracaso por parte del agente a responder a razones. Para afirmar que un agente es autónomo con respecto de sus acciones, no basta con que el agente se identifique con sus motivos. Parte de lo que se dice al afirmar que un agente es autónomo, es que él tiene la capacidad para elegir los motivos por los cuales actuará, evaluando sus respectivas ventajas y desventajas de manera razonable, y escogiendo entre ellos de acuerdo con sus propios valores superiores; por tanto, además de identificarse con sus motivos, es necesario que él haya sido capaz de examinar, evaluar y sopesar sus motivos y las razones con las que contaba para respaldarlos. En ciertos casos de adicción, por ejemplo, la persona puede seguirse identificando con su deseo de consumir droga aun cuando haya perdido la capacidad para evaluar sus motivos. Si no existe la posibilidad de que el agente inspeccione críticamente sus motivos y decisiones de acuerdo a las reglas de deliberación práctica que él mismo reconoce, el agente no posee la capacidad de autogobierno que la autonomía requiere.

Entenderé aquí el concepto de autonomía, entonces, como el ejercicio de la capacidad de un agente para evaluar y sopesar sus motivos, para decidir entre ellos,

modificarlos en razón de otros deseos y valores, libre de interferencias, y actuar en consecuencia.[4]

Volviendo al tema del autoengaño, según mi explicación de este fenómeno, el resultado del autoengaño es una creencia (de segundo orden). De acuerdo con Williams, a quien sigo, una creencia no puede ser adquirida a voluntad.[5] Esta imposibilidad no es empírica, sino conceptual: una creencia apunta siempre a ser verdadera, y si alguien pudiera creer alguna proposición a voluntad, tendría que aceptar que pudo haber creído alternativamente cualquier otra proposición inconsistente con la que cree, y de esa manera su creencia no podría ser considerada por él como verdadera. De esta manera, si existe alguna responsabilidad derivada del autoengaño, esa responsabilidad no se deriva del hecho de que el agente crea voluntariamente.

Sin embargo, si la postura que sostengo es correcta, el autoengaño (y esto es aplicable al fenómeno de las creencias ilusorias en general) no es algo que le sucede al agente; el agente comete el autoengaño. Las actividades de enfocar o desviar de manera selectiva la atención, recopilar selectivamente la evidencia e interpretarla deficientemente son actividades intencionales, y mientras las hace, el agente no está bajo ningún tipo de coerción externa —al menos comúnmente—, así que podemos asumir que también son voluntarias. Acerca de la actividad de enfocar o desviar la atención, puede decirse correctamente que en algunos casos esta actividad no es intencional ni voluntaria, puesto que hay cosas que atraen o desvían nuestra atención

[4] Sigo aquí, aproximadamente, la definición de autonomía de Gerald Dworkin: "la autonomía es una capacidad de segundo-orden para reflexionar críticamente sobre las preferencias y deseos de primer orden de uno mismo, y la habilidad para ya sea identificarse con éstos, o bien cambiarlos, a la luz de preferencias y valores de orden superior [autonomy is a second-order capacity to reflect critically upon one's first order preferentes and desires, and the ability either to identify with these or to change them in light of higher-order preferences and values]", *op. cit.*, p. 108).

[5] "Deciding to Believe", en *Problems of the Self*, Cambridge University Press, Cambridge, 1973.

sin que nos lo propongamos y sin que podamos evitarlo. Concediendo este punto, me parece difícil que un caso de autoengaño relativamente estable sea provocado por este enfoque o desviación de la atención involuntario, pues regularmente este fenómeno requiere no un desvío momentáneo y repentino de la atención, como lo es cuando sucede involuntariamente, sino un enfoque o desviación de la atención que sea continuo, sistemático, e inteligente, lo cual parece requerir que sea voluntario e intencional.

Por tanto, en los casos de autoengaño hay algún tipo de responsabilidad por parte del agente. Y, puesto que esta responsabilidad no puede consistir en el hecho de que el agente haya creído voluntariamente una proposición (pues eso es, según los argumentos presentados, imposible), creo que es correcto decir que el agente es responsable por haberse colocado en esa situación epistémica. Justo al momento del autoengaño, debido a la posición que el agente tiene con respecto a la evidencia (seleccionada, atendida e interpretada por él de una manera que está influenciada por sus deseos), el agente no puede evitar formarse las creencias que son resultado del autoengaño. Sin embargo, existen muchas cosas que el agente pudo haber hecho para evitarlo. En general, el agente pudo haber cultivado con mayor cuidado sus procesos formativos de creencias. Pudo haber fomentado un hábito de poner más atención a la evidencia a la hora de formarse una opinión tratándose de asuntos que involucran sus emociones y deseos, y de haber sido más diligente en la recolección de evidencia; pudo haber desarrollado mejor su habilidad, y alentado su valentía, para seguir el camino de la evidencia y los argumentos cuando cosas queridas para él estuviesen en juego, y haber reflexionado más acerca de sus patrones de razonamiento e inferencia en tales

casos, tomando en cuenta su propia actuación en casos anteriores similares y su conocimiento de lo que sucede a otras personas en situaciones de ese tipo; en otras palabras, el agente pudo haber examinado la influencia que sus deseos y miedos u otras emociones podían tener sobre su razonamiento, y las posibles consecuencias de esta influencia.

La responsabilidad del agente por haber caído en el autoengaño, entonces, consiste en un tipo de negligencia. El agente no es responsable directamente por poseer las creencias resultantes del autoengaño, pero sí es indirectamente responsable de ello, pues pudo haber modificado su propia relación con respecto a la evidencia que le era disponible. Podemos decir, por lo anterior, que el agente es epistémicamente responsable de las creencias que se forma. Según los diferentes usos de responsabilidad que mencioné, tal responsabilidad se traduce en que entendemos al agente como habiendo intervenido causalmente, y a través de acciones intencionales, en sus procesos de formación de creencias; en algún sentido importante, el agente es causa de sus creencias resultantes, que en el caso del autoengaño, son falsas. También es aplicable otro sentido de responsabilidad. La falsedad de una creencia es un defecto de la misma, pues como ya se dijo, las creencias siempre apuntan a ser verdaderas. Sin hablar todavía acerca de implicaciones morales, lo que el agente realizó es epistémicamente reprobable. Por tanto, el agente, en este sentido epistémico, es digno de recriminación por no haber intervenido de una mejor manera en su proceso de formación de creencias.

Neil Levy, en su artículo "Self-Deception and Moral Responsibility", argumenta que debemos abandonar el presupuesto de que los agentes autoengañados son

responsables de su autoengaño.[6] Siguiendo la postura de Mele para explicar el autoengaño, dice que este fenómeno debe asimilarse dentro de la categoría de los errores, y aunque la gente algunas veces es responsable de sus errores, no siempre es así. Una persona no es responsable de un error que comete, dice, si la persona en ningún momento tiene oportunidad para darse cuenta de que está cometiendo un error, o señal alguna de que lo está cometiendo. En contra de lo que han argumentado algunos teóricos del autoengaño, Levy afirma que tener sospechas, dudas o ansiedad con respecto de la creencia inducida por el autoengaño no es un requisito para estar autoengañado. Según Levy, entonces, el agente autoengañado, aunque hace cosas intencionalmente que sesgan su formación de creencias (como en la explicación de Mele), no tiene manera de saber que en ese momento sus acciones están sesgando su formación de creencias y por lo tanto no es responsable del autoengaño.

Según argumenté en el segundo capítulo, la explicación de Mele del autoengaño es equivocada, pues no toma en cuenta aquellos casos en los que existe comportamiento evasivo sistemático que sólo puede ser explicado mediante la atribución de creencias (verdaderas) del agente respecto a la proposición acerca de la cual trata el autoengaño. La explicación de Mele, dije, se aplica sólo a aquellos casos donde no existe la tensión típica del autoengaño, señalada por Audi, donde el sentido de la evidencia jala constantemente en contra de la verdad de *p*, mientras que los deseos del agente jalan hacia la verdad de *p*. Esta tensión se ve reflejada usualmente no sólo en el comportamiento evasivo sistemático del agente, sino también en sospechas, dudas y ansiedad por parte del agente. La explicación del autoengaño aceptada por Levy (la explicación de Mele), según argumenté, no da cuenta del fenómeno del autoengaño

[6] Neil Levy: "Self-Deception and Moral Responsibility", *Ratio*, 17 (3), 2004.

propiamente, y, siguiendo a Audi, dije que los casos que trata esa explicación podrían adecuadamente llamarse de otra manera —como se mencionó, Audi propone a tal efecto el término "auto-ilusión". Si todo lo anterior es correcto, el agente, en un caso de autoengaño genuino, típicamente tendrá algunas "señales de alarma" de su autoengaño: el sentido de la evidencia que lo jala constantemente hacia la verdad, su deseo de tener cierta creencia específica que lo jala en la dirección contraria, y las dudas y la ansiedad producidas por esta tensión constante. Estas señales pueden indicarle al agente, de manera prospectiva, que su proceso de formación de creencias puede estar sufriendo un sesgo o influencia tendenciosa y por tanto, epistémicamente, no es confiable. Cabe recordar aquí la primera de las máximas para la dirección de la mente propuestas por René Descartes[7], según la cual debemos apuntar a formar solamente juicios verdaderos y sólidos acerca de aquello que se presenta ante la mente. Al hacer caso omiso de estas señales de alarma el agente incurre en una violación de dicha regla.

Por otra parte, de manera retrospectiva, un agente puede darse cuenta, a partir de sus experiencias, de su propia propensión a caer en el autoengaño y contribuir a contrarrestar esa propensión. Si la tensión del autoengaño desaparece en algunos casos porque el agente termina por darse cuenta de la verdad (acerca de sus propias creencias), el agente se habrá dado cuenta de que tenía una creencia sesgada y probablemente, después de varios episodios de este tipo, entenderá que sus creencias estarán sesgadas motivacionalmente en casos importantes a menos de que él implemente algún tipo de cuidado especial al respecto.

Parece natural afirmar que un agente es responsable de cierta acción u omisión sólo si en algún sentido realista pudo haber hecho otra cosa. ¿Podría decirse entonces

[7] *Reglas para la dirección del espíritu*, ed. Alianza, trad. de Juan Manuel Navarro, Madrid, 1989.

que un agente que realiza cierta acción movido por ignorancia (como las acciones que resultan de la creencia inducida por el autoengaño), al no haber podido en ese momento hacer otra cosa, no es responsable de esa acción? Esta consideración parece amenazar la justificación de atribuir responsabilidad a un agente autoengañado por sus acciones. No es así, pues cabe distinguir entre distintos tipos de ignorancia. Carl Ginet[8], desarrollando una idea de Holly Smith, afirma que para que la ignorancia o el error de un agente en cierto momento sea culpable, tiene que haber en el pasado una omisión (por ejemplo, no haber investigado lo suficiente) o acción (por ejemplo, haberse drogado) acerca de la cual el agente sabía, o debía haber sabido, que previsiblemente lo llevaría o podría llevarlo a ese error culpable. Aun cuando, justo al momento de la acción, el agente no tiene otra opción más que actuar de acuerdo con sus deseos y creencias presentes, el agente autoengañado, desde esta concepción del autoengaño, debido a una influencia de sus deseos ha realizado acciones u omisiones que lo han llevado a colocarse en cierta posición con respecto a la evidencia que le es disponible, y eso lo ha llevado a adquirir creencias falsas que, a su vez, pueden haber determinado sus cursos de acción.[9] Dadas las señales de alarma que, según dije, el autoengaño presenta, no existe una justificación para decir que el agente nunca tuvo oportunidad para saber que sus creencias estaban influenciadas motivacionalmente y por lo tanto no es responsable de las acciones resultantes.

[8] En "The Epistemic Requirements for Moral Responsibility", en *Philosophical Perspectives*, 14, 2000.

[9] Puede trazarse una analogía interesante entre el famoso ejemplo de Ulises y las sirenas y el caso en discusión. El hecho de que Ulises no haya seguido el canto de las sirenas por no haber tenido libertad de movimiento no hace que lo despojemos de la responsabilidad (en este caso, el mérito) de no haber sucumbido a esa tentación, pues él mismo había pedido a sus marineros que lo ataran al mástil del barco con tal fin en mente.

Queda por examinar si puede hablarse de una obligación de los agentes por intervenir en sus procesos formativos de creencias de manera que las creencias resultantes sean más probablemente verdaderas. Según Davidson, el agente autoengañado, a causa de sus deseos, va en contra del principio de racionalidad conocido como exigencia de evidencia total para el razonamiento inductivo. Existe una larga tradición filosófica, desde autores clásicos como Aristóteles, Locke y Kant, hasta ciertos autores más contemporáneos como W. K. Clifford, que mantiene que existe una obligación (epistémica y moral) de creer sólo si el balance de la evidencia justifica la posesión de esa creencia. Clifford pone como ejemplo a un dueño de un barco que, sospechando que el barco en cuestión no está en condiciones para navegar, se autoengaña con base en un interés económico al efecto de creer que el barco está listo para la expedición y la aprueba sin tomar las medidas de seguridad necesarias, con el resultado de que el barco naufraga y toda la tripulación (el dueño, es de suponerse, no formaba parte de la tripulación) perece. En tal caso, dice Clifford, el dueño del barco "no tenía derecho" a creer que el barco estaba en condiciones de navegar, y por ello es responsable de la muerte de la tripulación y moralmente reprensible por ello. Para él, sin embargo, lo que es reprensible de la acción del dueño del barco, no depende de las consecuencias que de hecho tenga su acción (en este caso, el hundimiento del barco), pues pudo haber sucedido que, por suerte, el barco no zozobrara. Según Clifford, creer mal es tan reprochable como actuar mal, pues "una vez que uno cree, aunque uno no haya actuado con base en esa acción, uno ha cometido el mal en su propio corazón".[10] Este tipo de consideraciones lo llevan a enunciar su principio de que es malo siempre, en cualquier lugar y para cualquier persona, creer cualquier cosa con base en evidencia

[10] "The Ethics of Belief", p. 21.

insuficiente. Aunque tal vez sus expresiones suenen exageradas, en términos generales, estoy de acuerdo con Clifford en creer que es correcto atribuir responsabilidad a un agente por las posibles consecuencias dañinas producto de sus acciones (aun cuando estas consecuencias, por suerte, no se verifiquen) que hayan tenido origen en sus propios estados mentales defectuosos, cuando el agente haya sido responsable de tales defectos.

Es bien sabido que muchos casos de autoengaño pueden tener consecuencias positivas. En ciertos casos, estar autoengañado (positivamente) acerca de las cualidades de uno mismo hace que uno esté mejor preparado para enfrentarse a ciertas situaciones, promoviendo así que efectivamente adquiramos o desarrollemos esas cualidades. Hay quienes han argumentado que el autoengaño incluso, cuando es positivo y moderado, promueve la salud mental.[11] En efecto, se ha dicho, tener una percepción de la realidad clara y, a grandes rasgos, correspondiente de los hechos constituye sólo uno de los criterios para identificar a una persona mentalmente sana. Algunas ilusiones positivas acerca de las cualidades y el futuro de uno mismo y de los demás, aunque en principio vayan en contra de tal percepción de la realidad, probablemente contribuyan en mayor grado a la satisfacción de ciertos otros criterios propuestos para la identificación de la salud mental, como una visión positiva de uno mismo, la habilidad para estar tranquilo o satisfecho, la capacidad para el trabajo productivo y creativo y la capacidad para interesarse y preocuparse por los demás.

Lo anterior parece ser cierto, y bien puede ser que el mundo sea, en general, un mejor lugar gracias al autoengaño (con tales características) que lo que sería si el autoengaño no fuera posible. Sin embargo, desde el punto de vista de la racionalidad, el

[11] Shelley Taylor: *Positive Illusions: Creative Self-Deception and the Healthy Mind* (1989), p. 4-7.

autoengaño es un tipo de equivocación en la formación de creencias del agente. Si la explicación que ofrecí de este fenómeno es correcta, el autoengaño no es intencional ni se encuentra bajo el control del agente. Si tiene consecuencias positivas o negativas en cada caso específico dependerá de las circunstancias y no del agente. Lo que está bajo el control del agente es promover o no un hábito de revisar y monitorear frecuentemente sus creencias y las razones por las cuales las posee, especialmente si tiene sospechas acerca de su verdad. Al no promover este hábito, es cierto, el agente en algunos casos podría estar fomentando ciertas cosas positivas para él y para los demás. Sin embargo, él no controla este proceso, y podría a la vez estar también fomentando cosas terriblemente negativas. En cualquier caso, si el agente se propusiera promover y desarrollar su capacidad para el autoengaño (o simplemente no interferir con ella), podría estar indirectamente ayudando a mejorar algunos aspectos de su vida y de las de los demás, pero al mismo tiempo cedería necesariamente una parte considerable de su control sobre el mundo y, por tanto, perdería una parte de su autonomía.

De acuerdo con el concepto de autonomía que expuse al principio de este capítulo, una persona actúa de manera autónoma sólo si, además de identificarse con los motivos por los cuales actúa, tiene la capacidad para evaluar y sopesar sus motivos, para decidir entre ellos siendo sensible a razones, libre de interferencias de otras personas, y actuar en consecuencia. Me parece que en este aspecto es conveniente conservar, en un sentido débil, la dualidad de personajes o la partición del agente que tradicionalmente se atribuye al autoengaño: el agente por una parte es perpetrador del autoengaño, por otra parte es víctima de éste. La persona que cae en el autoengaño, en cuanto víctima, no está libre de interferencias "externas" en este sentido; ella misma, en

cuanto perpetradora del autoengaño, constantemente está interfiriendo con su capacidad para evaluar y sopesar sus motivos y creencias, y no se permite ser sensible a las razones (por ejemplo, la exigencia de evidencia total). Las acciones intencionales que conducen al autoengaño, aun cuando no sean realizadas con la intención de engañar, menoscaban la autonomía del agente al reducir sus opciones viables en la deliberación práctica (por no poseer él en tal caso las creencias verdaderas pertinentes) o presentar otras que son inviables (por basar sus razonamientos en creencias falsas), afectando de esta manera la forma de ejercer su capacidad para deliberar.

En el aspecto ético, el autoengaño se traduce en un impedimento para el auto-conocimiento y el entendimiento moral; si una persona no percibe claramente su propio carácter y las manifestaciones de éste en sus acciones, la persona pierde parte de su capacidad para actuar moralmente y para reparar actos inmorales previos. Un valor moral con larga tradición, como afirma Mark Platts, es el valor de la integridad, que él define "vaga y apresuradamente" como "el valor que le ordena a uno actuar en concordancia con uno mismo". [12] Pero, como él mismo dice, uno sólo puede actuar en concordancia con la *imagen* que uno tiene de uno mismo. El autoengaño, al presentarle al agente creencias falsas influenciadas motivacionalmente respecto de sus propias creencias, le da a éste una imagen errónea acerca de él mismo, y por tanto el agente se ve impedido en este aspecto a tener en mente una imagen correcta de sí mismo para actuar conforme a ella.

Las relaciones expuestas a lo largo de este capítulo entre la explicación del fenómeno del autoengaño y los conceptos de responsabilidad y autonomía son

[12] Mark Platts: "Réplica a 'Engaño y división'", en *Quinto Simposio Internacional de Filosofía*, Enrique Villanueva (ed.), Universidad Nacional Autónoma de México-Instituto de Investigaciones Filosóficas, México; 1992; p. 122-123

consecuentes con la bien conocida postura de Davidson, que comparto, de que ser un agente implica aceptar y normalmente seguir ciertos principios básicos de racionalidad. Ir en contra de esos principios es perder, en algún grado, la calidad de agente. No creo que haya sido demostrada hasta ahora la existencia de una obligación (moral o racional), para cualquier agente racional, de continuar siendo un agente racional. Pero es ahora claro que existen exigencias de la racionalidad para cualquier agente que desee conservar el estatus de agente racional.

Conclusiones:

1. Las dos paradojas de la irracionalidad formuladas por Davidson se disuelven si desarrollamos correctamente la idea de que un agente puede de hecho actuar, creer o desear con base en razones sin que éstas sean necesariamente las mejores razones del agente.

2. La controversial estrategia de Davidson de explicar la irracionalidad en general recurriendo a particiones o barreras mentales se vuelve innecesaria una vez que se han disuelto las paradojas por él propuestas.

3. De manera alternativa, la irracionalidad puede ser explicada, en términos amplios, a partir de una distinción entre razones aléticas (razones basadas en un interés por la verdad) y razones oréticas (razones basadas en un interés práctico distinto). Una persona puede tener numerosas razones oréticas (es decir, motivos), para creer, desear o actuar de diversas maneras, pero sólo aquellas que están apoyadas por razones aléticas (es decir, por enunciados verdaderos que toman en cuenta el balance total de la evidencia del agente, en el caso de razones para creer, o de los deseos y las creencias del agente, en el caso de razones para actuar y desear) son autorizadas por los estándares de la racionalidad práctica.

4. La aproximación léxica del autoengaño conduce a una visión que presenta a este fenómeno como intencional y que requiere de creencias contradictorias en la mente del agente. Donald Davidson, adoptando esta aproximación pero evadiendo sus dificultades a través de la estrategia de postular divisiones mentales, logra presentar una propuesta

coherente para explicar el autoengaño. Sin embargo, existen buenas razones para dudar de la aproximación léxica, por lo cual queda abierto el camino para encontrar una explicación del autoengaño que no precise de las controversiales particiones mentales propuestas por Davidson.

5. Alfred Mele presenta una propuesta simple para explicar el autoengaño que no recurre a las divisiones mentales, pero al final su propuesta enfrenta serios problemas dado que, desde su explicación, no es posible distinguir entre el autoengaño y el fenómeno relacionado, pero distinto, de las creencias ilusorias, además de que los casos que él presenta como autoengaño carecen de la característica filosóficamente más problemática de requerir una explicación para las acciones sistemáticas e inteligentes del agente autoengañado con las que éste evade la evidencia que haría desaparecer al autoengaño.

6. La propuesta de Audi hace justicia a tales casos filosóficamente más problemáticos, y explica el autoengaño a partir de una jerarquización de las creencias que participan en éste; la creencia inicial resulta ser una creencia de primer orden y la creencia resultante del autoengaño es una creencia de segundo orden, por lo cual no existe una contradicción entre ambas.

7. El deseo que mueve el proceso del autoengaño es un deseo de creer cierta proposición con independencia de su verdad. El resultado de este deseo, cuando el autoengaño se consuma, es que el agente termina creyendo que cree dicha proposición. Esta creencia de segundo orden (falsa) se refleja principalmente en el comportamiento lingüístico del

autoengañado, quien puede sinceramente afirmar que cree en dicha proposición, mientras que la creencia de primer orden (verdadera), que el agente conserva, se ve manifestada principalmente en su comportamiento evasivo no-lingüístico.

8. El resultado del deseo de creer inicial en el proceso del autoengaño es exactamente el mismo que podría esperarse en un caso análogo de creencias ilusorias donde el deseo inicial es un deseo de creer. Al poseer características idénticas, el autoengaño puede ser considerado entonces una forma especial del fenómeno de las creencias ilusorias, y sugiero que ciertos mecanismos psicológicos descritos por Mele para explicar el autoengaño funcionan más bien para explicar el fenómeno de creencias ilusorias en general.

9. Un agente es siempre responsable de caer en el autoengaño, a pesar de no ser este fenómeno intencional ni de estar bajo el control del agente al momento que sucede. El agente es responsable porque puede notar ciertas señales de alarma cuando el autoengaño está sucediendo, y tomar precauciones fomentándose a sí mismo hábitos de racionalidad epistémica que eviten el autoengaño. La responsabilidad derivada del autoengaño es entonces una forma de negligencia.

10. El agente pierde parte de su autonomía siempre que cae en el autoengaño, pues en tales casos el agente ve obstaculizado su ejercicio de la habilidad para evaluar y sopesar toda la información relevante que le es disponible con respecto al asunto del autoengaño, y sus decisiones y acciones subsecuentes relacionadas sufren una interferencia que es incompatible con la autonomía.

Bibliografía:

Audi, Robert: *Moral Knowledge and Ethical Character*, Oxford University Press, EUA, 2002.

Audi, Robert: "Self-Deception and Rationality", en Mike W. Martin (ed.): *Self-Deception and Self-Understanding*, 1985; pp. 169-194.

Audi, Robert: "Self-Deception, Rationalization and the Ethics of Belief", en su libro *Moral Knowledge and Ethical Character*, 2002.

Clifford, W. K.: "The Ethics of Belief", publicado originalmente en *Contemporary Review*, enero de 1877; reimpreso en Stephen, Leslie y Pollock, Fredrick (eds.): *W. K. Clifford: Lectures and Essays*, 1986.

Davidson, Donald: "Actions, Reasons, and Causes", en su libro *Essays on Actions and Events*, 1980. Traducido al español en Donald Davidson: *Ensayos sobre acciones y sucesos*, 1995.

Davidson, Donald: *Essays on Actions and Events*, Oxford, Clarendon Press, 1980. Traducido al español en *Ensayos sobre acciones y sucesos*, Crítica/Instituto de Investigaciones Filosóficas, Barcelona, 1995.

Davidson, Donald: "How is Weakness of the Will Possible?", en su libro *Essays on Actions and Events*, 1980. Traducido al español en Donald Davidson: *Ensayos sobre acciones y sucesos*, 1995.

Davidson, Donald: "Paradoxes of Irrationality", en Wollheim, R. y Hopkins, J. (eds.): *Philosophical Essays on Freud*, 1982; pp. 289-305.

Davidson, Donald. "Deception and Division", en Elster, Jon (ed.): *The Multiple Self*, Cambridge, 1986. Traducido al español en Villanueva, Enrique (ed.): *Quinto Simposio Internacional de Filosofía*, 1992.

Davidson, Donald: "Who is Fooled?", en Dupuy, J. P. (ed.): *Self-Deception and Paradoxes of Irrationality*, 1998.

Descartes, René: *Reglas para la dirección del espíritu*, ed. Alianza, trad. de Juan Manuel Navarro, Madrid, 1989.

Dupuy, J. P. (ed.): *Self-Deception and Paradoxes of Irrationality*, CSLI Publications, Stanford, 1998.

Dworkin, Gerald: *The Theory and Practice of Autonomy*, Cambridge University Press, Cambridge, 1988.

Elster, Jon (ed.): *The Multiple Self*, Cambridge, Cambridge University Press, 1986.

Frankfurt, Harry: "Freedom of the Will and the Concept of a Person", *The Journal of Philosophy*, vol. 68, no. 1, enero de 1971; pp. 5-20.

Frankfurt, Harry: *The Importance of What We Care About*, Cambridge University Press, Cambridge, 1988.

Eric Funkhouser: "Do the Self-Deceived Get What They Want?", *Pacific Philosophical Quarterly*, 86, no. 3, Sept. 2005; pp. 295-312.

Ginet, Carl: "The Epistemic Requirements for Moral Responsibility", en *Nous. Philosophical Perspectives*, vol. 34, supl. 14, octubre de 2000; pp. 267-277.

Hahn, Lewis Edwin (ed.): *The Philosophy of Donald Davidson*, The Library of Living Philosophers, Open Court Publishing, Chicago, 1999.

Hansberg, Olbeth y Platts, Mark (eds.): *Responsabilidad y libertad*, Instituto de Investigaciones Filosóficas/Fondo de Cultura Económica, México, 2002.

Lazar, Ariela: "Akrasia and the Principle of Continence or What the Tortoise Would Say to Achilles", publicado en Hahn, Lewis Edwin (ed.): *The Philosophy of Donald Davidson*, 1999; pp. 381-400.

Levy, Neil: "Self-Deception and Moral Responsibility", *Ratio*, vol. 17, no. 3, 2004; pp. 294-311.

Martin, Mike W. (ed.): *Self-Deception and Self-Understanding*, University Press of Kansas, EUA, 1985.

Mele, Alfred: *Self-Deception Unmasked*, Princeton University Press, Princeton, 2001.

Mellor, D. H.: "Conscious Belief", en *Proceedings of the Aristotelian Society*, vol. 77, 1977-78; pp. 87-101.

Platts, Mark: "Réplica a 'Engaño y división'", en *Quinto Simposio Internacional de Filosofía*, 1992.

Platts, Mark: "Introducción: Responsabilidades" en Hansberg, Olbeth y Platts, Mark (eds.): *Responsabilidad y libertad*, 2002.

Smith, Michael: *The Moral Problem*, Blackwell, Oxford, 1994.

Smith, Michael: "Is There a Nexus Between Reasons and Rationality", publicado en Tenembaum, Sergio (ed.): *New Trends in Moral Psychology*, 2006.

Stephen, Leslie y Pollock, Fredrick (eds.): *W. K. Clifford: Lectures and Essays*, Macmillan and Co., EUA, 1986.

Tenembaum, Sergio (ed.): *New Trends in Moral Psychology*, Rodolphi, Amsterdam, 2006.

Taylor, Shelley: *Positive Illusions: Creative Self-Deception and the Healthy Mind*, Basic Books, Nueva York, 1991.

Villanueva, Enrique (ed.): *Quinto Simposio Internacional de Filosofía*, Universidad Nacional Autónoma de México-Instituto de Investigaciones Filosóficas, México, 1992.

Williams, Bernard: "Deciding to Believe", en su libro *Problems of the Self*, 1973. Traducido al castellano en *Problemas del yo*, trad. José N. Holguera, UNAM-Instituto de Investigaciones Filosóficas, México, 1983.

Williams, Bernard: "Internal and External Reasons", en su libro *Moral Luck*, 1981. Traducido al castellano en *La fortuna moral*, trad. Susana Marín, UNAM-Instituto de Investigaciones Filosóficas, 1973.

Williams, Bernard: "Internal Reasons and the Obscurity of Blame", en su libro *Making Sense of Humanity*, 1995.

Williams, Bernard: *Making Sense of Humanity*, Cambridge University Press, Cambridge, 1995.

Williams, Bernard: *Moral Luck*, Cambridge University Press, Cambridge, 1981. Traducido al castellano en *La fortuna moral*, trad. Susana Marín, UNAM-Instituto de Investigaciones Filosóficas, México, 1993.

Williams, Bernard: *Problems of the Self*, Cambridge University Press, Cambridge, 1973. Traducido al castellano en *Problemas del yo*, trad. José N. Holguera, UNAM-Instituto de Investigaciones Filosóficas, México, 1983.

Wollheim, R. y Hopkins, J. (eds.): *Philosophical Essays on Freud*, Cambridge University Press, Cambridge, 1982.