EL AUTOENGAÑO.
ANATOMÍA DE UNA PASIÓN HUMANA

Gustavo Fernández Acevedo

> Si fuésemos a dibujar un retrato del hombre en el cual se resaltara aquello que es más humano, ya sea noble o innoble, seguramente deberíamos emplazar en primer plano la enorme capacidad del hombre para el autoengaño (Herbert Fingarette, *Self-deception*, 1969).

A mi hijo Javier

Índice

Prefacio **4**
Capítulo I: ¿qué es el autoengaño? **6**
1. Algunos ejemplos para comenzar 6
2. Una caracterización inicial de trabajo 7
3. El autoengaño y otras formas de irracionalidad motivada 10
4. El problema en distintas disciplinas 18
5. Un esquema general de los principales problemas acerca del autoengaño 21
Capítulo II. Intenciones no conscientes, creencias contradictorias, mentes divididas: el autoengaño en la Filosofía contemporánea **28**
1. ¿Es posible el autoengaño? Las respuestas escépticas 32
2. El modelo filosófico "clásico" 40
3. El deflacionismo respecto del autoengaño 53
4. ¿Cuál explicación filosófica del autoengaño debe preferirse? 58
Capítulo III. Psicología, psicopatología, neurociencias. Perspectivas científicas sobre el autoengaño **62**
1. El autoengaño en la Psicología actual 62
2. Estudios empíricos tendientes a probar la existencia del autoengaño 63
3. La función defensiva del autoengaño 70
4. Autoengaño y psicopatología 80
4.1. Delirio, anosognosia, confabulación 80
4.2. Autoengaño y adicción 91
Capítulo IV. ¿Hemos sido diseñados por la evolución para autoengañarnos? **104**
1. El autoengaño como adaptación 107
2. El autoengaño como subproducto 111
3. La autopercepción y el "efecto ganador" 115
4. ¿Qué es el autoengaño, entonces, en términos evolutivos? 119
Capítulo V. Mentiras vitales. Las consecuencias prácticas del autoengaño **125**
1. Las implicaciones morales del autoengaño 125
2. Autoengaño, salud mental y felicidad 137
3. Autoengaño y creencias religiosas 154
4. ¿Autoengaño colectivo? 165
El porvenir de un problema **184**
Referencias bibliográficas **188**

Prefacio

Somos seres capaces de pensar y comportarnos de modo racional, pero nuestra racionalidad dista de ser perfecta. Está sujeta a múltiples sesgos, a la presencia de atajos cognitivos, al empleo de procedimientos veloces y poco seguros. Nuestro sistema cognitivo no funciona de manera aislada; nuestras emociones suelen interactuar fuertemente con los procesos por los cuales llegamos a formarnos creencias sobre algún tema particular. Sin embargo, el autoengaño es más que el mero sesgo, el atajo cognitivo o el procedimiento rápido e inseguro y, también, más que la mera influencia de las emociones sobre la racionalidad. El autoengaño se encuentra en el límite entre las formas "normales" de irracionalidad y las formas patológicas. Nos mentimos a nosotros mismos con convicción, con entusiasmo, incluso con pasión. Somos capaces de defender frente a otros, con una obcecación sorprendente, creencias que no resisten la más mínima confrontación con pruebas claras e indisputables. Esto no es casual; el autoengaño tiene lugar en áreas de nuestro mayor interés: las relaciones con otros significativos, nuestro estado de salud, nuestra autovaloración; difícilmente existirá autoengaño en áreas que no representen intereses vitales para el agente.

El origen de este libro también obedece a intereses vitales, así como a interrogantes teóricos y experiencias prácticas. La vertiente teórica de estos intereses y experiencias se relaciona con mi trabajo docente y de investigación en Filosofía y Psicología, y en las interacciones entre ambas. La vertiente práctica se relaciona con mi práctica profesional en la asistencia de personas que consultan por trastornos debidos al uso de sustancias psicoactivas. En este contexto me he encontrado con casos que, no por comunes en la experiencia clínica, resultan menos refractarios a la intuición: el padre que, luego de varios problemas de drogas con su hijo a causa del consumo de drogas de éste, acepta sus inverosímiles explicaciones cuando lo sorprende con grandes cantidades de esas sustancias; la alcohólica que luego de reconocer aparentemente múltiples problemas físicos, laborales y sociales generados por el uso de alcohol, expresa profundas dudas respecto de la posibilidad de padecer ese trastorno. El origen del libro también reconoce un antecedente en el interés de muchas personas a las que les comenté de mi trabajo sobre el tema. En muchos casos esta curiosidad parecía exceder claramente la inquietud intelectual; sospecho que muchos nos sentimos identificados cuando se habla de autoengaño, de un modo que no siempre nos resulta claro.

Mi propósito al escribir este libro fue el de presentar una visión abarcativa, que no se limitara a los aportes de una única disciplina. Hasta donde sé, no existen exposiciones sistemáticas en castellano acerca del conocimiento que tenemos sobre este fenómeno. Requiere del lector la apertura suficiente para aceptar no sólo que se está ante un problema cuya pertenencia parece oscilar entre distintas disciplinas, sino también, y más importante, que muchas preguntas permanecen aún sin respuesta. Con el fin de llegar a un público más amplio, he tratado de omitir el tratamiento de los aspectos más técnicos y complejos de los problemas examinados y las respuestas a estos problemas, aunque no espero haberlo logrado en todos los casos; la complejidad de algunos debates torna imposible un grado de simplificación tal que no redunde en una tergiversación de las posiciones y argumentos.

Varias personas han colaborado de una u otra manera en la preparación de este libro, y a ellas mi agradecimiento. A Manuel Comesaña, lector invariablemente agudo. A mis amigos y colegas Carlos Díaz-Lázaro, Mariana Cremonte, Fernando Poó, Roberto Sánchez, Silvana Montes y Teresa Bunge, quienes leyeron partes del texto y formularon muchas sugerencias útiles. A Alfredo Cosimi, quien me proporcionó una útil orientación respecto de diversos conceptos psicoanalíticos vinculados con el autoengaño. A mis colaboradores Valeria Nasarov, Boris Kogan, Mauro McIntyre y Agostina Vorano, con quienes compartí muchas jornadas de trabajo estimulantes a lo largo de los diversos proyectos de investigación que dieron origen a este libro. Por último, pero no menos importante, a mi familia y a mis amigos, quienes me acompañaron en los buenos momentos y en los que no lo fueron tanto. Sin pretensión alguna de originalidad, no puedo sino reiterar aquí que cualquier error contenido en el libro es de mi exclusiva responsabilidad.

Por último, al lector, una amable advertencia. Nadie esté exento del riesgo de autoengañarse. No sólo me apoyo al decir esto en mis intentos de lograr una perspectiva sistemática del fenómeno, o en mi experiencia como terapeuta; me baso también en mi propia trayectoria vital. Como veremos a lo largo del libro, no existe una forma de lograr inmunidad contra el autoengaño. Ni la inteligencia, ni el conocimiento y la cultura, ni la comprensión de los procesos que le dan origen constituyen recursos infalibles para evitar caer en él; por el contrario, a veces hasta pueden facilitarlo. Quizás, como señala Fingarette en la cita que da inicio al libro, debamos aceptar que la capacidad para el autoengaño es el rasgo más característicamente humano. Incluso, como veremos en algunos pasajes del libro, no deberíamos desestimar la posibilidad de que, en ciertos casos, constituya una contribución a nuestro bienestar. Sin embargo, sería imprudente subestimar los riesgos que implica. Una vida en la cual el autoengaño ocupe un lugar central sólo puede ser una existencia no auténtica, basada en la falsedad y en la distorsión de nuestra propia naturaleza.

Capítulo I: ¿qué es el autoengaño?

1. Algunos ejemplos para comenzar

1. Suponga que el hijo de un amigo suyo tiene un mal rendimiento escolar, escasa capacidad de lectura y no capta bromas que la mayoría de los chicos de su edad sí comprenden. Suponga también que su amigo insiste en que su hijo es muy inteligente, y de hecho es más inteligente que la mayoría de los chicos de su edad. Usted sabe que su amigo nunca miente y ciertamente nunca le mentiría, y también sabe que es por lo general sensato, por lo cual concluye que se está engañando a sí mismo (Van Leeuwen, 2007, p. 419).
2. Mitchell es un hombre vano, susceptible respecto de su raleado cabello. Ha adoptado el hábito de combar su cabello hacia un lado de un modo bastante exagerado y desagradable. Si bien adopta pasos tan obvios para ocultar su calvicie, no admite que es calvo. Sus amigos encuentran embarazoso el hecho de que a menudo formula observaciones burlonas en referencia a la calvicie de otros hombres, mientras no reconoce que él también lo es. Las instrucciones de Mitchell a su peluquero son muy precisas, e invariablemente posa en cierto ángulo cuando es fotografiado o cuando se observa en el espejo. Ni siquiera permite que su esposa despeine su cabello (Funkhouser, 2005, p. 296).
3. Joey es un hombre celoso. Observadores objetivos están convencidos de que es sumamente improbable que su esposa Marcia esté teniendo un *affaire*. Sin embargo, Joey piensa algo diferente, y se refiere a su esposa con expresiones que resultan chocantes a sus amigos. Le señalan que no tiene razones para pensar que ella se escabulle para encontrarse con otro hombre cuando visita a sus amigas. Comentan que el deseo sexual recientemente incrementado de ella probablemente no sea un acto para encubrir su culpa, pero Joey protesta de modo vehemente. De modo no sorprendente, a Joey no le agrada que le muestren que está equivocado y rechaza las afirmaciones de las amigas de Marcia que confirman la versión de los hechos que ella ofrece (Funkhouser, 2005, p. 297).
4. Ike es un olvidadizo bromista, hábil en la imitación de la escritura de otros, que ha engañado intencionalmente a sus amigos escribiendo entradas falsas en sus diarios. En una oportunidad decide engañarse a sí mismo escribiendo una falsa entrada en su propio diario. Consciente de su mala memoria, escribe bajo la fecha del día "estuve

particularmente brillante en clase hoy", contando con el hecho de que eventualmente olvidará que lo que escribió es falso. Semanas más tarde, cuando revisa su diario, lee esta afirmación y adquiere la creencia de que estuvo brillante en clase en esa fecha (Mele, 2001, p. 16).

Las descripciones precedentes ilustran, de acuerdo con los autores que las presentan, casos típicos de autoengaño. En todos ellos sus protagonistas parecen mantener creencias falsas que además, al menos en varios, están reñidas por la evidencia. Ahora bien, sin pretender ingresar de forma inmediata en los debates relativos a las características y variantes que este fenómeno puede adoptar, es posible señalar también algunas diferencias importantes entre ellos. El caso 1. y el caso 2. parecen ser claros ejemplos de estados en los cuales el pensamiento está afectado por el deseo: sus protagonistas parecen creer en que los hechos son diferentes a como en realidad son: sus creencias se ajustan más a lo que desean que lo que la evidencia a su disposición permite aceptar. El caso 3., a diferencia de los anteriores, resulta más difícil de interpretar. Si bien en este caso el pensamiento parece igualmente contrario a las pruebas de que la persona dispone, no parece, *prima facie*, que se encuentre influenciado directamente por el deseo. Por el contrario, sin duda muchos observadores externos estarían de acuerdo en que el sujeto autoengañado *no desea* que su esposa le sea infiel; no obstante, sostiene esa creencia contra las pruebas disponibles. El caso 4. difiere de los tres anteriores en su carácter directa y llanamente intencional. Mientras que en los tres primeros la intención de inducir en sí mismos la creencia falsa, si es que la hay (especialmente en el tercero), parece inconsciente o implícita, en el caso en cuestión tal intención aparece como algo claramente consciente. Podría pensarse, sobre la base de esas diferencias, que existen diferentes clases de autoengaño, suposición que parece plausible; sin embargo, también podría afirmarse que no todos esos casos ejemplifican adecuadamente el fenómeno del autoengaño. La existencia de estas interpretaciones alternativas conduce inevitablemente a examinar conceptualmente el autoengaño.

2. *Una caracterización inicial de trabajo*

Los ejemplos precedentes pueden resultar útiles para un primer acercamiento a lo que, hemos adelantado, es un fenómeno en alguna medida familiar para todos nosotros. Ahora bien, pese a la utilidad inicial de los ejemplos, resulta esperable que un libro que trata sobre el autoengaño comience con una caracterización al menos preliminar del fenómeno que se va a examinar, y no nos apartaremos de esta expectativa. Tal caracterización puede

hacerse de distintos modos. Es posible comenzar con una o más definiciones teóricas y examinar luego sus puntos en común y discrepancias, en un intento de encontrar un patrón descriptivo; también se puede partir de una caracterización preteórica del concepto, para luego pasar a un examen más detallado de sus notas definitorias; puede también iniciarse con el análisis de un conjunto de casos que presuntamente ejemplifiquen los rasgos fundamentales del fenómeno a explicar. Dado que, como veremos reiteradamente a lo largo del libro, la proliferación de estudios sobre el autoengaño no ha conducido a un acuerdo respecto de sus características fundamentales, y el recurrir a los ejemplos es una opción admisible, pero, como vimos, resbalosa, optaré por la segunda alternativa y comenzaré con algunas intuiciones sobre el fenómeno que parecen *prima facie* aceptables.

¿Cómo definir de la forma más preteórica posible al autoengaño? Podríamos decir lo siguiente. En ocasiones observamos ciertas conductas (verbales y no verbales) que nos hacen concluir que las personas que las ejecutan poseen una determinada representación falsa acerca de algún aspecto de sí mismas o del mundo externo; esto es, cuando afirman creer algo no están simulando o mintiendo para engañar a otros, sino que parecen actuar de buena fe al manifestar que sostienen una determinada afirmación. En segundo lugar, la posesión de la representación que es objeto de creencia no está apoyada por los elementos de juicio disponibles para la persona, sino que, por el contrario, parece haber pruebas que sustentan la representación opuesta de la que la persona dice poseer; además, parecería que tales elementos de juicio contrarios o bien son poseídos por la persona o bien están fácilmente disponibles para ella. Por último, tenemos razones para pensar que la adopción de tal representación falsa es *motivada*; nos parece que la adopción de la creencia por parte de la persona no es causada por un mero error intelectual generado por factores como la ignorancia, el cansancio o el error de cálculo, sino que se debe a la presencia de estados o procesos no cognitivos que poseen una valencia afectiva (como deseos y temores) que la conducen a aceptar la creencia en cuestión y no su opuesta.

Notemos, por otra parte, que existe un conjunto de expresiones en los lenguajes naturales que transmiten la idea de que, en ocasiones, distorsionamos o manipulamos nuestras representaciones de la realidad, tanto interna como externa: en nuestro idioma, afirmaciones como "se miente a sí mismo", "vio sólo lo que quería ver", "está negando la realidad" atestiguan esta idea. El empleo de algunas de estas expresiones no implica meramente que la persona posee una representación distorsionada de la realidad, sino que implica claramente que tal distorsión es *motivada*, esto es, la persona posee un motivo para adoptar esa representación. Ejemplos típicos de tales procesos son el de la esposa convencida de la fidelidad de su esposo, pese a las visibles señales de que éste tiene un

amorío, o el del padre que niega contra todas las pruebas disponibles que su hijo esté sufriendo de un trastorno debido al uso de drogas.

Sobre la base de las intuiciones anteriores propondremos algunos rasgos que parecen caracterizar al autoengaño; tal caracterización puede ser presentada como la conjunción de tres condiciones, a saber:

a. adquisición y/o mantenimiento de una creencia falsa,
b. frente a elementos de juicio contrarios a tal creencia,
c. motivados por procesos mentales no cognitivos (por ejemplo, deseos o emociones) que favorecen la adquisición y/o retención de esa creencia.

Ahora bien, todas y cada una de las notas de esta caracterización intuitiva pueden ser cuestionadas. Este cuestionamiento puede darse por varias razones. Por una parte, cada condición requiere de especificaciones importantes que son fuente de controversias posteriores. En segundo lugar, podría señalarse que las condiciones no son suficientes para una caracterización apropiada del autoengaño; se ha alegado, por ejemplo, que la *intención* de poseer una creencia es un componente fundamental del autoengaño (Pears, 1984; Davidson, 1986). Por último, también podría sostenerse que algunas de las condiciones o partes de ellas no son necesarias para atribuir tal estado a un agente; según algunos autores, por ejemplo, podría existir autoengaño sin que la creencia en cuestión fuese falsa.[1] Con la única finalidad de ilustrar las dificultades de esta caracterización intuitiva, presentaré algunos problemas y alternativas teóricas.

El primer punto de la caracterización da por sentado que el estado final resultante del autoengaño es una creencia. Esto es aceptado por una gran mayoría de autores, pero tal aceptación no es unánime. Para algunos (Audi, 1982; Rey, 1988) el producto del autoengaño no es una creencia, sino una manifestación [*avowal*]. Para estos autores una manifestación es, en términos llanos, una disposición a afirmar una proposición con "sinceridad", pero que carece de conexiones profundas con la acción, rasgo que sí caracteriza a las creencias. Asimismo, se ha sugerido que en algunos subtipos de autoengaño el estado resultante es un estado intermedio entre la creencia y un estado no cognitivo, como el deseo (Egan, 2008), y no una creencia *simpliciter*. Las motivaciones para proponer alternativas a la creencia son variadas, pero en general se proponen con el fin de evitar algunos de los enigmas conceptuales que el autoengaño genera. Si bien hay consenso en que el producto del autoengaño es una creencia, las posiciones minoritarias elevan objeciones y proponen alternativas de no menor importancia.

La segunda condición, relativa al requisito de elementos de juicio, también da lugar a controversias. Una de ellas es la referente a si, para que exista autoengaño, el agente debe

[1] Véase al respecto el análisis de Lynch (2010) en el §3 del capítulo V.

poseer *efectivamente* los elementos de juicio contrarios a la creencia a la que favorece, o bien es suficiente con que tales elementos sean fácilmente asequibles para él. Quienes sostienen la primera alternativa consideran que, para que exista autoengaño, el agente debe llevar a cabo un procesamiento sesgado de los elementos de juicio que posee, de un modo inconsistente con sus estándares epistémicos usuales (Davidson, 1986; McLaughlin, 1988). Por el contrario, otros han sostenido que, para atribuir autoengaño, basta con que el agente proceda a seleccionar de modo sesgado los elementos de juicio a su disposición, con lo que no llegará a poseer pruebas contrarias a la creencia a la que favorece (Mele, 2012).[2] Por supuesto, el grado de fortaleza que deben poseer las pruebas a disposición del agente o poseídas efectivamente por él es también un tema susceptible de debate.

Por último, existe una controversia importante acerca de los estados no cognitivos que motivan el autoengaño. Si bien parece claro que el autoengaño no es meramente un error cognitivo o sesgo intelectual, no hay acuerdo acerca de cuáles son los estados no cognitivos que lo generan. Entre los principales candidatos se encuentran el deseo (Pears, 1984; Davidson, 1986), la ansiedad (Barnes, 1999), y las emociones (Lazar, 1999). También hay modelos teóricos que no recurren a un solo factor motivacional, sino a varios de ellos (Mele, 2001). Este debate no sólo es importante por sí mismo, sino también porque la respuesta elegida tiene consecuencias respecto de la posibilidad de lograr una teoría unificada sobre el fenómeno (Mele, 2001).

Como se adelantó, estas y otras objeciones pueden ser elevadas contra la caracterización de trabajo que hemos propuesto. No obstante, ninguna especificación, agregado o supresión de las condiciones propuestas está libre a su vez de críticas de diversa importancia, por lo que será conveniente mantener la caracterización intuitiva al menos hasta que hayamos avanzado un poco más en la comprensión del fenómeno.

3. El autoengaño y otras formas de irracionalidad motivada

Suele señalarse que el autoengaño no es un fenómeno *sui generis*, sino que pertenece a una familia de fenómenos a la que suele hacerse referencia en el ámbito filosófico con la expresión "irracionalidad motivada". Esta familia de fenómenos incluye miembros bien conocidos, como la debilidad de la voluntad y el pensamiento desiderativo, y otros menos examinados y quizás aceptados, como la ceguera intelectual y el pensamiento sesgado (Bach, 1981), el engaño autoinducido (Audi, 1997) y las creencias obstinadas (Lynch,

[2] Examino esta cuestión en Fernández Acevedo (2011).

2012).[3, 4] El término "motivada" para caracterizar a esta familia es clave: el autoengaño y sus parientes cercanos no son meros sesgos implícitos y/o automáticos en el procesamiento de la información, ni errores causados por factores como el cansancio, la ignorancia o los estados alterados de conciencia (por ejemplo, estados debidos a una intoxicación). Debido al carácter motivado de la distorsión de la creencia, la categoría excluye algunos fenómenos en los que está presente la irracionalidad en el proceso que conduce a la adopción de la creencia y/o en el modo en que es mantenida, pero no los factores no cognitivos que conducen a ella. No incluye, por ejemplo, los conocidos sesgos cognitivos (sesgo de confirmación, falacia de la regresión, falacia de la conjunción, etcétera).[5] Tampoco incluye a los fenómenos patológicos cuyos estados resultantes parecen tener relación con el autoengaño, como las creencias delirantes y la confabulación, pero en los cuales el componente motivacional está ausente o, al menos, su presencia resulta dudosa. En este apartado nos limitaremos a examinar brevemente la relación del autoengaño con sus dos "parientes mayores", esto es, la debilidad de la voluntad y el pensamiento desiderativo; esto resultará útil para complementar, esperamos, la caracterización de trabajo que hemos hecho del fenómeno que nos ocupa en el primer apartado de este capítulo. Comenzaremos con la primera, que, como veremos, resulta comparativamente más fácil diferenciar del autoengaño.

El análisis de ejemplos básicos permite entender a qué fenómeno se hace referencia con la expresión "debilidad de la voluntad" (también llamada en ocasiones "*akrasia*" o

[3] El engaño autoinducido y las creencias obstinadas, en particular, se han propuesto no sólo como formas de irracionalidad motivada, sino como contraejemplos a ciertos modelos de autoengaño (Mele, 1997, 2001).

[4] Si bien las "creencias obstinadas" parecen tener un parentesco directo con el autoengaño, no todo fenómeno de formación y/o mantenimiento irracional de una creencia constituye automáticamente un miembro de la familia de la irracionalidad motivada. Un posible caso de esto es el constituido por las "creencias traviesas" [*naugthy beliefs*] (Huddleston, 2011). Huddleston sostiene lo siguiente: a veces parecería que "creemos" cosas que, a la vez consideramos falsas; esto es, en algunas circunstancias somos capaces de ser reflexivos y conscientes de la falsedad de nuestras creencias de primer orden, y aun así tales creencias pueden persistir. Esas creencias serán las denominadas "creencias traviesas"; ejemplo de ellas son fenómenos tales como las supersticiones, fobias y paranoias, que son autoconscientes (aunque no los casos típicos de paranoia, aclara Huddleston). Su existencia desafía un principio plausible que establece la siguiente imposibilidad psicológica: *p*, me percato conscientemente de que *p*, y sin embargo creo que no *p*. Si bien el caso del autoengaño parecería constituir un contraejemplo saliente para tal principio, Huddleston considera que esto no es el caso; el rasgo clave de los casos de autoengaño es que hay un *engaño* involucrado: el agente es capaz de engañarse a sí mismo acerca de que *p*, o empujar a *p* temporalmente fuera de la conciencia de modo de lograr la comodidad de creer en no *p*. Huddleston intenta mostrar que, aun cuando creer conscientemente algo contrario a lo que uno cree que es el caso es imposible en la mayoría de las circunstancias, no es imposible en *todas* las circunstancias. Ahora bien, aunque tales creencias son un notorio caso de irracionalidad, Huddleston no dice nada acerca de su origen; no es posible descartar que se trate de un caso de irracionalidad debido a pobres estándares epistémicos; en tal caso, el carácter "motivado" no estaría presente.

[5] Véase Thagard (2011) para un listado de los sesgos que generan creencias erróneas. Cabe mencionar, como veremos en el capítulo II, que estos sesgos tienen un rol importante en la producción del autoengaño en algunos modelos explicativos.

"incontinencia").[6] Muchas acciones corrientes no generan en nosotros ninguna clase de perplejidad; las consideramos el resultado de elecciones y preferencias quizás cuestionables, pero en modo alguno radicalmente irracionales. Este es el caso, por ejemplo, de alguien que elige adquirir un bien *a* en vez de un bien *b* con plena conciencia de que *a* es más costoso que *b*, simplemente porque según su evaluación la calidad de *a* sobrepasa a la de *b* y esta diferencia compensa el mayor costo. Similar es el caso de un agente que elige alimentarse con el producto *x* en vez de escoger el producto *z*, debido a que *x* es mucho más apetitoso que *z*, pese a que este último resulta ser más saludable.

Ahora bien, los casos descriptos resultan cualitativamente diferentes a una situación en la cual el agente hace *f* en vez de *e*, aun cuando está convencido de que *e* es lo mejor que puede hacer, considerando todas las circunstancias; esto es, *e* resulta ser superior a *f* no sólo en un aspecto e inferior en otro, como en los ejemplos precedentes, sino que resulta ser una mejor elección *en todo respecto*: no hay ninguna dimensión evaluativa en la que *f* no sea inferior a *e*, y, pese a eso, el agente escoge *f*. Este caso sí resulta genuinamente intrigante; no es posible asignarle un sentido de modo análogo a como puede hacerse en los casos previos. Tal intriga es generada por nuestras presunciones acerca de los juicios que fundamentan las acciones: si tales juicios involucran una evaluación superior en todo respecto acerca de un curso de acción determinado por sobre otros, esperamos que las acciones libres de un agente reflejen tal evaluación. Este juicio respecto de la superioridad absoluta de un curso de acción por sobre otros parece tener un carácter especial; es denominado en ocasiones el *mejor juicio*. En consecuencia, la perplejidad que casos como el descripto generan puede conducir a pensar que, en realidad, el agente no ha evaluado realmente que *e* es superior a *f* en todo respecto, esto es, que la descripción que se hace de ellos no es precisa y que, por el contrario, está eligiendo el curso de acción que le parece mejor. Pero si, en cambio, el agente evalúa efectivamente que un curso de acción es superior a sus alternativas en todo respecto, y pese a eso no elige tal curso, podemos decir que el agente ha actuado en contra de su mejor juicio. A este fenómeno es al que se hace referencia habitualmente con la expresión *debilidad de la voluntad*.

La descripción anterior del fenómeno no deja muchas dudas acerca de las diferencias que tiene con el autoengaño. Es una observación común en el ámbito filosófico que este fenómeno se distingue de la debilidad de la voluntad en que el resultado de ésta consiste en una intención o una acción intencional, mientras que el resultado del autoengaño consiste en una creencia. Si bien ambas constituyen formas de irracionalidad, el

[6] Cfr. Davidson (1970) para un detallado análisis de esta noción.

autoengaño estaría incluido en la categoría de irracionalidad teórica, mientras que la debilidad de la voluntad constituiría una forma de irracionalidad práctica.[7]

Si bien la debilidad de la voluntad es un miembro destacado de la familia de la irracionalidad motivada, su importancia para la comprensión del autoengaño es notoriamente inferior a la del otro de los miembros mayores de esta clase, esto es, el pensamiento desiderativo. Suele describirse a quien piensa desiderativamente del siguiente modo: S adquiere la creencia de que *p* porque quiere creer que es el caso que *p*.[8] Se han hecho varios intentos de caracterizar el autoengaño en términos comparativos con este fenómeno, como veremos a continuación.

Un influyente intento en esa línea de pensamiento es debido a B. Szabados (1973). Szabados considera que la diferencia entre uno y otro radica en que la persona a la que se atribuye un pensamiento desiderativo no pervierte los procedimientos por medio de los cuales se establece la verdad o falsedad de una afirmación. En caso de que esta persona sea confrontada con evidencia que entre en conflicto con su creencia será capaz de reconocerlo, aunque quizás con reticencia. En consecuencia, un aspecto crucial en el que se diferencian pensamiento desiderativo y autoengaño es que en el segundo la evidencia es contraria a la creencia sostenida. En caso de que una persona que sostiene un pensamiento desiderativo se enfrente a tal evidencia y proceda a resistir, por medio de tácticas ingeniosas, sus implicaciones naturales, sostiene Szabados, se podrá afirmar que está autoengañada. Su posición, en síntesis, sostiene una suerte de continuidad entre pensamiento desiderativo y autoengaño, pero mantiene la posibilidad de distinguirlos sobre la base del requisito de existencia de evidencia en contra de la creencia desiderativa.

Otros autores han sostenido posiciones muy similares respecto de la relación entre los dos fenómenos, como es el caso de K. Bach (1981). Bach observa que el autoengaño no es pensamiento desiderativo y, si lo es, se trata de un caso muy especial. La diferencia radica

[7] No carece de interés observar que la irracionalidad vinculada con una conducta no necesita ser atribuida a la conducta en sí misma (como es el caso de la debilidad de la voluntad) sino que puede ser *posterior* a una conducta que no sea en sí misma irracional. Este sería el caso de la racionalización, fenómeno que tiene lugar cuando una acción es explicada por motivos o razones distintas y más "aceptables" para el agente que aquellas por las cuales realmente es ejecutada. Elster (2007) señala que si bien el pensamiento desiderativo, el autoengaño y la racionalización se originan en la motivación del agente para sostener creencias específicas, la diferencia entre los dos primeros y la última radica en que, mientras que en la segunda la creencia ocurre con posterioridad a la conducta, en los primeros sucede lo opuesto.

[8] Se ha señalado correctamente, a nuestro entender, que el pensamiento desiderativo no opera de un modo tan simple como su descripción podría hacer pensar. Correia (s/f) ha observado que si bien en ocasiones este fenómeno parece reducirse a la inferencia "Deseo *p*, por lo tanto *p*", parece dudoso que el pensador desiderativo cometa una falacia en estos términos. En la mayoría de los casos de pensamiento desiderativo, los agentes no son conscientes de que llegan a la conclusión de que *p* es verdadero meramente porque desean que *p*. En vez de ello, la inferencia guiada por el deseo actúa por medio de un tipo más complejo e indirecto de falacia, el argumento por las consecuencias: "si *p*, entonces q. Deseo que suceda q. Por lo tanto, *p*". Pero aun en este caso, señala, la premisa que expresa el deseo tiende a permanecer implícita.

en que en el pensamiento desiderativo no hay involucrado ningún razonamiento o esbozo de razonamiento. La persona que piensa desiderativamente imagina algún estado de cosas, le agrada aquello que imagina, y supone que eso es el caso o lo será. No trata de justificar esta suposición, quizás por contentarse con la ausencia de evidencia en uno u otro sentido. Ahora bien, en caso que la persona sea consciente de la evidencia contraria a su creencia y la necesidad de lidiar con ella, estaríamos en presencia de un caso especial de autoengaño.[9] La posición de Bach acerca del pensamiento desiderativo, en consecuencia, se asemeja en dos aspectos a la de Szabados: la ausencia de evidencia pertinente respecto de la creencia, en el caso del pensamiento desiderativo, y la existencia de evidencia en contra, en el caso del autoengaño; en segundo lugar, en la posibilidad de que ocurra una "transición" del primer estado al segundo, en caso de que el agente enfrente evidencia contraria a su creencia.

En Davidson (1985) tenemos una descripción algo más compleja de las relaciones entre pensamiento desiderativo y autoengaño. En primer lugar, conviene introducir un concepto al que Davidson hace referencia, y que aparece en la caracterización tanto del pensamiento desiderativo como del autoengaño. Este es el error cognitivo al que denomina *debilidad de la justificación*. Este error sólo puede darse cuando una persona que posee evidencia tanto a favor como en contra de una hipótesis juzga que, en relación con toda la evidencia disponible, la hipótesis es más probable que su negación y, no obstante, no la acepta. En este caso, la persona no acata el principio normativo denominado por Hempel y Carnap *requisito de evidencia global en el razonamiento inductivo*. Esto es, cuando debemos decidir entre una serie de hipótesis mutuamente excluyentes, tal exigencia nos demanda adoptar la hipótesis que se encuentre mejor apoyada por toda la evidencia pertinente disponible.

Davidson señala también que el autoengaño *incluye* la debilidad de la justificación, ya que el sujeto no aceptaría la proposición con respecto a la cual se autoengaña si fuera liberado de su error; tiene mejores razones para aceptar la negación de la proposición. Además, agrega, como sucede en la debilidad de la justificación, la víctima del autoengaño sabe que tiene mejores razones para aceptar la negación de la proposición que de hecho acepta. Es en el punto anterior, observa Davidson, en el cual el autoengaño llega más lejos que la debilidad de la justificación, ya que la persona que se autoengaña tiene que tener una *razón* para su debilidad de justificación y, además, tiene que haber participado en la generación de esta razón. La debilidad de la justificación tiene siempre una causa, pero en el

[9] Bach agrega en este análisis otra distinción, que es la distinción entre autoengaño y ceguera intelectual. La ceguera intelectual consiste en el fracaso en percibir hacia dónde apuntan la evidencia o las razones de que dispone el agente; por el contrario, quien se autoengaña percibe esto de modo correcto, al menos inicialmente.

caso del autoengaño la debilidad de la justificación resulta ser autoinducida, esto es, uno mismo la produjo.

Con respecto a la distinción entre pensamiento desiderativo y autoengaño, Davidson mantiene de forma matizada la continuidad entre ambos, si bien advierte acerca de la posibilidad de casos en los que el primero no es parte del segundo. En una elucidación inicial, observa, el pensamiento desiderativo consiste en creer algo debido al deseo de que sea verdad. No considera que esta posibilidad sea irracional en sí misma, ya que en general no somos responsables de las causas de nuestros pensamientos. Sin embargo, advierte que el pensamiento desiderativo es a menudo irracional, por ejemplo cuando sabemos por qué tenemos la creencia y sabemos también que, si no fuera por el deseo, no la poseeríamos. Davidson agrega que no todo pensamiento desiderativo es autoengaño, ya que éste, a diferencia del primero, requiere de la intervención del agente; esto es, el autoengaño surge de la *intención* del agente de formar una creencia contraria a la evidencia de que dispone, intención que lleva a actuar de modo tal que conduzca a la formación de la creencia preferida.[10] No obstante, ambos se parecen en que en los dos tiene que actuar un elemento motivacional o evaluativo. A su vez, en este aspecto ambos difieren de la debilidad de la justificación, ya que en ésta el defecto determinante es cognitivo con independencia de su causa. Esto sugiere, agrega Davidson, que aun cuando el pensamiento desiderativo pueda ser más simple que el autoengaño, es siempre un ingrediente de éste. Sin embargo, aunque sin duda lo es con frecuencia, parece haber excepciones: en el pensamiento desiderativo la creencia toma la dirección del afecto positivo, nunca del negativo, mientras que en el autoengaño no sucede lo mismo. El pensamiento alimentado por el autoengaño puede ser doloroso; este es el caso, por ejemplo, del hombre erróneamente convencido de la infidelidad de su esposa, pese a la completa ausencia de pruebas a favor de su creencia.[11]

Por último, no todos los autores consideran que el pensamiento desiderativo pueda ser claramente distinguido del autoengaño. Mele (1987) discute brevemente el intento de Szabados de diferenciar ambos fenómenos. Advierte que, así como la ignorancia de la evidencia no es excluyente del autoengaño (esto es, que el agente ignore la evidencia contraria a su creencia no implica que no pueda estar autoengañado), tampoco es posible lograr una diferenciación nítida entre autoengaño y pensamiento desiderativo sobre la base de este requisito. Señala que, si hay alguna diferencia entre pensamiento desiderativo y autoengaño, ésta puede radicar simplemente en que el primero constituye un género

[10] Este criterio de demarcación entre pensamiento desiderativo y autoengaño es compartida por otros autores de la perspectiva intencionalista sobre el segundo (p. ej. Bermúdez, 2000).
[11] Véase el capítulo II, § 2 para un análisis de los "casos negativos" de autoengaño.

denotado por el término "autoengaño". Si Szabados está en lo correcto en lo referente a la ausencia de evidencia en el pensamiento desiderativo, esto puede deberse a que este fenómeno consiste en una clase de autoengaño en la cual, a causa de una conducta influenciada por el deseo, quien se autoengaña carece de buenas bases para rechazar la proposición que sostiene autoengañosamente. Si bien, agrega Mele, la expresión "pensamiento desiderativo" tiene un aire inofensivo del que carece el término "autoengaño", se trata simplemente de una cuestión terminológica, y se pregunta retóricamente que ocurriría si, en vez de "pensamiento desiderativo", habláramos de "falsa creencia desiderativa", expresión que le parece correcta si nos basamos en el análisis que hace Szabados.[12]

Parecería que las distinciones entre fenómenos cercanamente relacionados como el pensamiento desiderativo y el autoengaño son inevitablemente borrosas, y que los intentos de diferenciarlos nítidamente corren riesgos ciertos de enfrentar contraejemplos de manera inmediata.[13] Algunos de estos riesgos pueden disminuirse o eliminarse si se rechazan determinados supuestos teóricos respecto del autoengaño. Por ejemplo, no es necesario adherir a las concepciones que sostienen que el autoengaño requiere de una intención para producirse; como veremos en el capítulo II, existen teorías sobre este fenómeno que prescinden de las intenciones del agente como premisa explicativa. Ahora bien, otros problemas no pueden ser resueltos de la misma forma. Si bien la eliminación del componente intencional podría hacer más plausible el supuesto de continuidad entre pensamiento desiderativo y autoengaño, esta posibilidad enfrenta la objeción basada en la existencia de los casos negativos o retorcidos de autoengaño, que no parecen tener, *prima facie*, continuidad alguna con el pensamiento desiderativo. Una posible solución reside en plantear la existencia de casos de pensamiento "aprensivo"; esto es, así como existe el pensamiento desiderativo, habría casos de posesión de creencias sobre sucesos negativos no fundadas en ninguna clase de evidencia. Tales casos podrían tener, en consecuencia, una continuidad con el autoengaño negativo o retorcido. Sin embargo, la introducción de nuevos fenómenos genera a su vez problemas conceptuales adicionales no menores que los que presuntamente resuelve.

---

[12] Mele (1997) también critica los intentos de distinguir el pensamiento desiderativo del autoengaño sobre la base de que el segundo constituye un fenómeno intencional, rasgo que no caracterizaría al primero.

[13] Algunos autores han propuesto otros rasgos que diferenciarían al pensamiento desiderativo del autoengaño, aunque es difícil saber si tales intentos son más promisorios que los ya mencionados. Así, Elster (2007), sugiere que el autoengaño, a diferencia del pensamiento desiderativo, presupone la existencia de creencias inconscientes. Van Leeuwen (2013) señala que si bien tanto el pensamiento desiderativo como el autoengaño comparten la característica de estar motivados por deseos o actitudes conativas específicas, en el caso del pensamiento desiderativo de que *p* no se produce típicamente la situación de que el agente posea pruebas convincentes de que en realidad no acontece *p*; debido a esto, el pensamiento desiderativo está también libre (o libre en mayor medida) de la tensión epistémica que caracteriza al autoengaño.

Por último, no podemos dejar de mencionar aquí la relación del autoengaño con otro clásico problema filosófico relativo a la irracionalidad de las creencias: la posibilidad de creer a voluntad. La doctrina comúnmente conocida como "voluntarismo doxástico" sostiene que podemos ejercer un control voluntario sobre aquello que creemos, esto es, que podemos decidir adoptar libremente nuestras creencias; así como algunas de nuestras acciones (por ejemplo, levantar un brazo) son acciones básicas, que no requieren para su ejecución de la realización de ninguna acción adicional, algunos casos de adquisición de creencias constituirían acciones básicas: podríamos adoptar ciertas creencias de modo voluntario y directo, sin vernos forzados a ello por la particular relación entre las creencias y la realidad. El defensor del voluntarismo doxástico no necesita sostener que *cualquier* proposición sea candidata a la adquisición directa; sólo necesita sostener que *algunas* creencias pueden ser adquiridas de ese modo.

En cualquier caso, ha existido un amplio consenso en que esta doctrina es falsa; muchos autores han sostenido que es completamente imposible elegir a voluntad y de modo directo nuestras creencias (Adler, 2002; Elster, 1979; Williams, 1973). Para los críticos, un rasgo esencial de las creencias es que tienden o apuntan hacia la verdad; se imponen por el peso de la evidencia en su favor, y no por nuestro deseo de que el estado al que hacen referencia acontezca; esta tesis se ha denominado, consecuentemente, *involuntarismo doxástico*. Si fuera posible para nosotros creer a voluntad en algo a sabiendas, sería posible creer que *p* es verdadero, sabiendo todo el tiempo que la creencia no está justificada por la evidencia, que fue adquirida a voluntad y que las creencias pueden ser adoptadas sin consideración a su verdad o falsedad, todo lo cual es considerado imposible. Esta imposibilidad puede ser concebida como una imposibilidad conceptual/lógica/metafísica o como una imposibilidad psicológica y, aun cuando fuesen superables las objeciones relativas al primer tipo de imposibilidad, esto no implicaría que también lo serían las críticas referentes al segundo tipo. Se han presentado, no obstante, diversas objeciones a los argumentos contra el voluntarismo doxástico (Bennet, 1990; Funkhouser, 2003) e intentos de mostrar como las dificultades de esta posición pueden ser superadas (Ryan, 2003; Steup, 2008).

Ahora bien, la relación del autoengaño con el voluntarismo doxástico proviene de la distinción entre dos tipos de voluntarismo: el voluntarismo doxástico *directo* (que coincide con la noción arriba descripta) y el voluntarismo doxástico *indirecto*: sería posible, en principio, inducir en nosotros mismos determinadas creencias de manera indirecta, por medio de la modificación planificada de nuestras conductas y situaciones; esta alternativa suele ser ejemplificada por el consejo de Pascal de actuar como si se fuese creyente, de modo de inculcar en uno mismo la creencia en Dios. La posibilidad del voluntarismo

doxástico indirecto es admitida por los críticos más firmes del voluntarismo doxástico directo. Williams (1973), por ejemplo, concede que hay lugar para aplicar la decisión de creer por medio de rodeos. No obstante, considera que esta alternativa es profundamente irracional y que puede tener éxito sólo por medio del autoengaño.[14]

Como los casos anteriores atestiguan, el problema de las relaciones del autoengaño con otros tipos de irracionalidad motivada, como muchos otros problemas relativos al fenómeno que nos ocupa, no está resuelto ni mucho menos. No obstante, no creemos que la ausencia de una respuesta a este problema sea un grave déficit desde el punto de vista teórico; si bien su examen resulta útil para facilitar la comprensión de la naturaleza del autoengaño, esta cuestión de límites tiene una menor importancia comparativa en relación con los otros problemas relativos al fenómeno, como veremos en el último apartado de este capítulo.[15]

4. *El problema en distintas disciplinas*

Como ha ocurrido con un gran número de los problemas que hoy podemos encontrar en el campo de la ciencia, el estudio del autoengaño fue planteado originalmente en el ámbito de la Filosofía, y de él se han ocupado distintas ramas de esta disciplina: la gnoseología, la filosofía moral, y, posteriormente, la Psicología filosófica y la filosofía de la mente. Posiblemente la referencia más antigua a este fenómeno es debida a Platón, quien en el *Crátilo* afirma que "el autoengaño es la peor de todas las cosas", y agrega que cómo podría no ser terrible cuando quien engaña no se aparta de uno ni por un instante, sino que está siempre presente. Con similar preocupación se expresó posteriormente Kant, quien

[14] Pese al carácter claramente indirecto que parece distinguir al proceso que conduce a un estado de autoengaño, no siempre se ha dado por sentado a cuál variante del voluntarismo doxástico puede supuestamente prestar apoyo. Booth (2007) examina la posibilidad de que el autoengaño constituya una vía de justificación para el defensor del voluntarismo doxástico directo, que permitiría eludir los argumentos a favor de la imposibilidad psicológica (suponiendo, claro está, que la imposibilidad conceptual/lógica/metafísica haya sido superada). Esta posibilidad plantea el problema del modo en que el autoengaño debe ser concebido para lograr ese objetivo. Booth considera que, con ese fin, el fenómeno debe ser interpretado de modo tal que a) muestre cómo al menos algunos casos de autoengaño son logrados sin la intervención de ninguna mediación, y b) que tales actos sean conducidos deliberadamente y en plena conciencia. A partir del análisis de uno de los clásicos ejemplos de mala fe presentado por Sartre (1943), el de la joven que se niega a reconocer ante sí misma los intentos de seducción de su acompañante, Booth considera que no es posible satisfacer ambas condiciones simultáneamente: no es posible cumplir con la condición a) sin violar la condición b), y a la inversa; consecuentemente, dado que ambas deben ser cumplidas para que los casos de autoengaño constituyan casos de voluntarismo doxástico directo, parecería que ningún acto de autoengaño podría establecer la posibilidad psicológica de control doxástico directo. El autoengaño puede, concluye Booth, involucrar un control doxástico indirecto, pero no un control directo. Por otra parte, Cook (1987), considera que es posible decidir creer de modo indirecto sin que el autoengaño esté involucrado.

[15] Los lectores que deseen profundizar en la naturaleza del pensamiento desiderativo y en sus diferencias con el autoengaño podrán encontrar interesantes análisis en Elster (1983, 2007, 2010) y Scott-Kakures (1996, 2000, 2002).

reflexionó críticamente acerca de los efectos corrosivos del engaño, ya sea que alguien lo practique en su propia persona o en la de sus semejantes.

El problema del autoengaño también estuvo presente en el pensamiento cristiano. Así, encontramos el libro *The Mistery of Self Deceuing: or A Discourse and Discovery of the Deceitfulnesse of Mans Heart*, escrito por el ministro bautista Daniel Dyke y aparentemente publicado en 1614. Más recientes e influyentes que el libro de Dyke son las reflexiones del obispo Joseph Butler, expuestas en su "Sermon X. Upon Self-Deceit" de 1729. Butler, al igual que Kant, señaló las destructivas consecuencias del autoengaño sobre la condición moral y las acciones de quien se autoengaña.

No sería justo, en esta laxa progresión histórica, soslayar la influencia de tres autores que reflexionaron, desde perspectivas muy diferentes, sobre fenómenos más o menos cercanos a lo que contemporáneamente se entiende por autoengaño. Estos autores son, por orden cronológico, Karl Marx, Sigmund Freud y Jean-Paul Sartre. Ninguno de ellos se ocupó del concepto de autoengaño tal como se lo comprende en la actualidad; no obstante, su influencia se ha hecho sentir en diversas dimensiones de la teorización sobre este fenómeno. Si bien no nos ocuparemos de analizar las contribuciones de estos autores para la comprensión del autoengaño, en capítulos posteriores examinaremos la influencia de su pensamiento sobre las distorsiones de las creencias en relación con distintas dimensiones del autoengaño. En particular, haremos referencia al concepto de *negación* en Freud al ocuparnos de los estudios contemporáneos sobre el autoengaño en Psicología, y a la noción de *ideología* de Marx al examinar los procesos de distorsión colectiva de creencias.

En las últimas décadas, además de los muchos estudios filosóficos sobre el tema, las investigaciones empíricas sobre el autoengaño se han multiplicado en distintas ramas de la Psicología (psicología cognitiva, social, de la personalidad y psicopatología), neurología y neuropsicología, teoría de la evolución y ciencias sociales (Sociología y Antropología). Ahora bien, esta multiplicación de estudios no ha conducido hasta el momento a una perspectiva unificada que integre los análisis conceptuales y lógicos más característicamente filosóficos con la teorización y la información empírica provenientes de las disciplinas científicas mencionadas. Si bien hay trabajos que ignoran los límites de la disciplina de origen del autor y avanzan en intentos explicativos más sistemáticos (p. ej., Mele, 2001), muchos análisis del autoengaño se han mantenido centrados dentro de las fronteras de una disciplina específica. Ahora bien, aunque no sea en absoluto objetable el examen de un problema desde la perspectiva de una disciplina específica, sí pueden serlo algunas de sus consecuencias; en particular, al examinar la bibliografía sobre el autoengaño es posible observar la existencia de enfoques que ignoran contribuciones pertinentes provenientes de otros campos, con la pérdida consecuente de perspectiva sobre el problema.

Una clara ilustración de esta proliferación no integrada de enfoques está dada por dos textos introductorios al tema: el de Paulhus (2007), quien examina el autoengaño desde la Psicología social, y el de Deweese-Boyd (2012), que lo hace desde la Filosofía.

Paulhus define al autoengaño simplemente como el "acto de mentirse a uno mismo", y pasa revista a distintas cuestiones teóricas relevantes sobre el fenómeno desde la perspectiva psicológica. Estas incluyen las condiciones de posibilidad del autoengaño (la aceptación inicial de que explicar casos de autoengaño requiere el reconocimiento de la existencia de partes inconscientes de la mente); los avances teóricos que facilitan el comprender cómo se produce; el posible fundamento evolucionista para el autoengaño; y las pruebas clínicas y experimentales que avalan la existencia de este fenómeno. En esta revisión, Paulhus toma posición sobre al menos dos cuestiones que son fundadamente objeto de debate desde hace tiempo y que merecerían mucho mayor análisis. En primer lugar, que la explicación del autoengaño debe incluir una división de la mente; en segundo lugar, al hacer referencia a los experimentos tendientes a probar la existencia del autoengaño, observa que "un experimento convincente tiene que mostrar que una persona cree algo y descree en el mismo momento". Ambas tesis son cuestionadas por algunos autores como más problemáticas que la propia existencia del fenómeno que pretende explicar.

La presentación de Deweese-Boyd, por el contrario, adopta un enfoque en el que enfatiza el carácter problemático del fenómeno, y admite desde el inicio que virtualmente todos los aspectos filosóficos del autoengaño son materia de controversia, incluyendo su propia caracterización y la identificación de casos paradigmáticos.[16] De este modo, pasa revista a controversias y problemas filosóficos "clásicos", tales como los debates que oponen a quienes defienden que el autoengaño es un fenómeno intencional y a los que consideran que el autoengaño no puede ser modelado a partir de la intencionalidad característica del engaño interpersonal; el problema del autoengaño "retorcido"; y las implicaciones morales del autoengaño, incluyendo el debate acerca de la responsabilidad moral de quien se autoengaña.

La dispersión conceptual representada por los dos ejemplos considerados no constituye una excepción, sino una regla. Como veremos, no se trata meramente de la coexistencia de teorías explicativas sobre la base de un acuerdo acerca del fenómeno a explicar, sino de perspectivas radicalmente diferentes sobre el problema, a menudo sin

[16] Las propias presentaciones filosóficas sobre el problema pueden presentar importantes variaciones entre sí. Guttenplan (1994), por ejemplo, se centra más en la discusión de las aparentes paradojas relativas al autoengaño y a sus relaciones con la agencia, la racionalidad y la acción, sin pasar revista por las distintas posiciones como lo hace Deweese-Boyd.

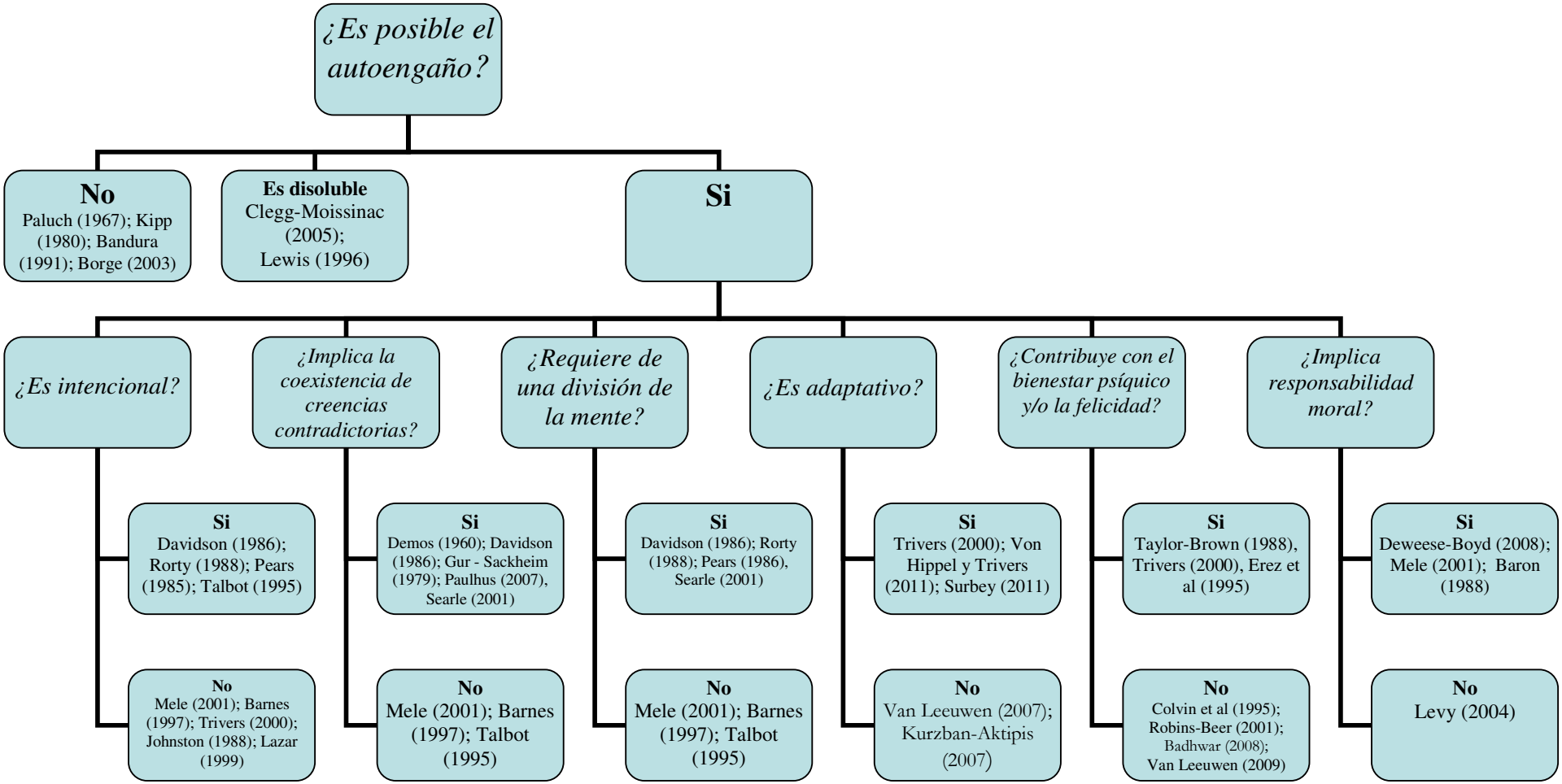
contacto entre sí. Esta situación constituirá el estándar al que deberemos habituarnos al examinar el problema del autoengaño.

5. *Un esquema general de los principales problemas acerca del autoengaño*

Hemos iniciado este capítulo señalando que son muy escasas las tesis sobre el autoengaño que gocen de acuerdo generalizado entre los estudiosos del tema. Es posible, incluso, que estas tesis se reduzcan a dos: primero, que el autoengaño es irracional, segundo, que esta irracionalidad es motivada, esto es, no se trata de un mero error debido a factores exclusivamente cognitivos. Sin embargo, existen mayores acuerdos acerca de cuáles son las preguntas pertinentes con respecto al autoengaño.

El primer y obvio interrogante que puede plantearse (dando por sentado que el autoengaño existe, cosa que algunos autores han cuestionado), es ¿qué es el autoengaño? Esta pregunta está inevitablemente ligada a otras dos: ¿cómo es posible el autoengaño? y ¿por qué existe? Estas dos preguntas son básicamente preguntas explicativas, referentes a la naturaleza del autoengaño y su posible papel en nuestro funcionamiento mental. Pero además de ellas nos vamos a plantear otras cuestiones referentes a este fenómeno, de igual interés que las anteriores. Entre ellas se cuentan las siguientes: ¿existe responsabilidad moral por el autoengaño?; ¿cumple el autoengaño alguna función de protección o defensa?; ¿es posible el autoengaño colectivo?; ¿existe relación entre el autoengaño y fenómenos patológicos, como el delirio y la confabulación?

Una revisión de la bibliografía sobre el problema posibilita arribar a una sistematización provisional de las preguntas que parecen haber concitado el mayor interés por parte de los especialistas. Presentaremos las preguntas anteriores en un esquema que reúna tanto esas preguntas como las posiciones principales en respuesta a ellas y algunos de los autores que las han defendido. El esquema no incluye obviamente *todas* las preguntas que pueden formularse sobre el autoengaño; incluye las que, a nuestro modo de ver, son aquellas a las que se les ha asignado más importancia en las últimas cinco décadas. Todas ellas, con la posible excepción de la pregunta por la responsabilidad moral por el autoengaño, que ha sido de manera más exclusiva de interés filosófico, han sido objeto en mayor o menor medida tanto de estudios filosóficos como empíricos.


¿Es posible el autoengaño?
No
Paluch (1967); Kipp (1980); Bandura (1991); Borge (2003)
Es disoluble
Clegg-Moissinac (2005);
Lewis (1996)
Si
¿Es intencional?
¿Implica la coexistencia de creencias contradictorias?
¿Requiere de una división de la mente?
¿Es adaptativo?
¿Contribuye con el bienestar psíquico y/o la felicidad?
¿Implica responsabilidad moral?
Si
Davidson (1986); Rorty (1988); Pears (1985); Talbot (1995)
No
Mele (2001); Barnes (1997); Trivers (2000); Johnston (1988); Lazar (1999)
Si
Demos (1960); Davidson (1986); Gur - Sackheim (1979); Paulhus (2007), Searle (2001)
No
Mele (2001); Barnes (1997); Talbot (1995)
Si
Davidson (1986); Rorty (1988); Pears (1986), Searle (2001)
No
Mele (2001); Barnes (1997); Talbot (1995)
Si
Trivers (2000); Von Hippel y Trivers (2011); Surbey (2011)
No
Van Leeuwen (2007); Kurzban-Aktipis (2007)
Si
Taylor-Brown (1988), Trivers (2000), Erez et al (1995)
No
Colvin et al (1995); Robins-Beer (2001); Badhwar (2008); Van Leeuwen (2009)
Si
Deweese-Boyd (2008); Mele (2001); Baron (1988)
No
Levy (2004)

i. ¿Es posible el autoengaño?

Como adelantamos, el debate acerca del autoengaño no se ha limitado a las preguntas sobre sus características fundamentales, sino que su misma existencia ha sido objeto de controversia. En consecuencia, tenemos aquí una primera división de perspectivas y autores según sea su posición respecto de la existencia del fenómeno.

Es posible distinguir tres posiciones. En primer lugar, tenemos las perspectivas que podemos llamar "escépticas" respecto del autoengaño. Diversos estudiosos han negado su existencia: Paluch (1967), Champlin (1977), Bok (1980), Kipp (1980), Bandura (1991), Borge (2003). Las razones para el escepticismo son diversas, pero un argumento recurrente se basa en la aparente existencia de paradojas presuntamente insolubles de distintas clases. Una de las más estudiadas es la llamada en ocasiones "paradoja doxástica" del autoengaño. Sintéticamente, esta paradoja puede expresarse de la siguiente forma: ¿cómo es posible que alguien crea simultáneamente *p* y no *p*? Como veremos, hay formas de enfrentar esta y otras paradojas pero, en opinión de los escépticos, tales intentos no tienen éxito, por lo que sólo cabe concluir que el autoengaño no es posible.

Una segunda posición es la que podría denominarse, "disolucionista" respecto del autoengaño (Harré, 1988; Lewis, 1996; Clegg y Moissinac, 2005). No se trata de una posición análoga a la anterior, que sostiene que el autoengaño es imposible; lo que se afirma es que el autoengaño es simplemente una cuestión de perspectivas diferentes, algo cuya existencia se debe más a la mente del observador que a supuestos procesos intrapsíquicos en el presunto autoengañado, un fenómeno cultural más que un proceso mental individual. Esta posición está muy estrechamente ligada con perspectivas provenientes del construccionismo social y del posmodernismo.

Por último, tenemos una abrumadora mayoría de autores que aceptan la existencia del autoengaño como un fenómeno genuino a explicar: Demos (1960); Fingarette (1969); Pears (1984); Davidson (1986); Mele (1987, 1997, 2001); Johnston (1988); McLaughlin (1988); Oksenberg-Rorty (1988); Barnes (1997); Talbot (1995); Lazar (1999); Trivers (2000, 2011); Funkhouser (2005); Van Leeuwen (2007a, 2007b, 2008, 2009); Scott (2000, 2001, 2009) y muchos otros. Para estos autores el problema del autoengaño no consiste en un seudoproblema originado o bien en un uso descuidado del lenguaje, que nos induce al error de suponer la existencia de un fenómeno donde sólo hay un término, o bien en una concepción tradicional y errónea acerca de la mente.

Sólo una vez que se admite la existencia del autoengaño es posible plantear los interrogantes que presentaremos a continuación. Estos seis interrogantes no son, como dijimos, las únicas preguntas pertinentes que pueden plantearse una vez que se ha aceptado la existencia del autoengaño, pero su importancia relativa dentro de la Filosofía y las ciencias que se ocupan del fenómeno justifica su inclusión.

ii. ¿Es intencional el autoengaño?

En primer lugar tenemos quienes consideran que el autoengaño es intencional (Davidson, 1986; Oksenberg-Rorty, 1988; Pears, 1991). Según los modelos intencionalistas, lo que ocurre cuando alguien se engaña a sí mismo es estructuralmente similar a lo que ocurre cuando una persona engaña intencionalmente a otra, con la obvia diferencia de que en el primer caso la persona se engaña *intencionalmente* a sí misma. En segundo lugar están aquellos que sostienen que es una clase de fenómeno no intencional (Mele, 1987, 2001; Barnes, 1997; Lazar, 1999). De acuerdo con las concepciones no intencionalistas, la formación de creencias autoengañosas puede ser explicada sin recurrir a la acción intencional. El autoengaño se explicaría por una combinación de sesgos cognitivos y factores motivacionales y afectivos; no hace falta postular una intención de autoengañarse por parte del agente. Cabe aclarar que los modelos intencionalistas no suponen una intención clara y consciente por parte del agente de cambiar una creencia verdadera por una falsa; esto último supondría que podemos decidir cambiar nuestras creencias a voluntad de modo directo, lo cual, como hemos visto en el apartado 2, es algo habitualmente reconocido como imposible. El componente intencional del autoengaño, en caso de existir, debe ser indirecto u oblicuo. El enfoque intencionalista, en consecuencia, concibe al sujeto autoengañado como alguien que actúa con la intención de que esa acción cause la formación de la creencia deseada; tal acción puede consistir en un direccionamiento intencional de la atención lejos de las pruebas que apoyan a *p* (la creencia que rechaza aceptar), o puede consistir en una búsqueda activa de elementos de juicio contra *p*.

iii. ¿Implica la coexistencia de creencias contradictorias?

Este problema está íntimamente relacionado con las concepciones intencionalistas, que han constituido durante muchos años la posición *standard* respecto del autoengaño. Si

el autoengaño puede comprenderse a partir del modelo proporcionado por el engaño interpersonal, entonces parecería que debemos admitir que, como ocurre en aquel, el agente que se autoengaña posee tanto la creencia verdadera, que rechaza, como la creencia falsa, que es la que adquiere y/o mantiene (Demos, 1960; Davidson, 1986; Gur y Sackheim, 1979). Esto, como es dable imaginar, conduciría a una situación cognitiva imposible: la coexistencia en la conciencia de una contradicción visible llevaría inevitablemente a la eliminación de una de las creencias. Una de las alternativas de solución es proponer, como se verá en el capítulo 3, una división de la mente. La otra es negar que las creencias contradictorias deban coexistir (Mele, 2001; Barnes, 1997; Talbot, 1995). Esto es, para que exista el autoengaño el agente sólo debe poseer la creencia falsa, en presencia de pruebas favorables a la creencia verdadera. Se han considerado al menos dos alternativas para defender esta perspectiva: o bien el agente sostuvo inicialmente y luego desechó la creencia verdadera en favor de la creencia falsa (con lo cual las creencias contradictorias no llegarían a coexistir), o bien no hace falta que el agente haya desestimado, y ni siquiera poseído inicialmente, la creencia verdadera, sino que adopta inicialmente la creencia falsa. Sean o no admisibles estas alternativas, la posible existencia de creencias contradictorias ha constituido un tema de debate recurrente en torno del autoengaño.

iv. ¿Requiere de una división de la mente?

Este debate tiene lugar entre quienes afirman que la existencia del autoengaño requiere de una división del psiquismo (Demos, 1960; Davidson, 1986; Rorty; 1988; Pears, 1991), y aquellos que sostienen que tal división no es necesaria (Mele, 2001; Barnes, 1997; Talbot, 1995). Como mencionamos, las posiciones divisionistas suelen recurrir a la tesis de una partición o división de la mente como forma de solucionar el problema generado por la presunta coexistencia de creencias contradictorias. Esta división de la mente puede coincidir con alguna teoría empírica acerca del funcionamiento mental (por ejemplo, alguna teoría cognitiva o neuropsicológica acerca de procesos conscientes e inconscientes), o bien constituir una teoría *ad hoc*. Un punto importante a destacar aquí es que muchas teorías divisionistas, como es dable imaginar, son cuestionadas por introducir distinciones que, en el mejor de los casos, solucionan un problema al costo de generar otro de igual o mayor dificultad. Por el contrario, las teorías no divisionistas parecen tener a su favor el ser más parsimoniosas, al no proponer ninguna división adicional de la mente a aquellas requeridas para explicar otros fenómenos mentales distintos del autoengaño.

v. ¿Es el autoengaño adaptativo?

Lo primero que es necesario hacer respecto de esta pregunta es aclarar en qué sentido se entiende el término "adaptativo". A veces, dentro del campo de la Psicología, se emplea el término "adaptación" en un sentido relacionado con el ajuste del individuo a su medio social.[17] También, como han hecho algunos psicólogos, puede emplearse para hacer referencia a alguna clase de proceso útil para la vida y productor de salud mental. Muchos autores, antes de que se comenzara a estudiar de manera científica el autoengaño, señalaron su carácter protector frente a las verdades dolorosas y a los sufrimientos vitales. Más allá de este uso posible del término "adaptativo", el significado del término al que queremos hacer referencia ahora es el proporcionado por la teoría de la evolución. "Adaptativo" hace referencia a un rasgo de un organismo que contribuye con su aptitud, esto es, con su capacidad para producir descendencia, perpetuando su patrimonio genético. La pregunta, entonces, es si la capacidad para el autoengaño constituye un rasgo seleccionado evolutivamente por su contribución con nuestra aptitud (Trivers, 2000; Von Hippel y Trivers, 2011; Surbey, 2011), o si se trata de alguna clase de subproducto estructural (Van Leeuwen, 2007; Kurzban-Aktipis, 2007), esto es, de un rasgo no seleccionado pero asociado con otros rasgos que sí han sido seleccionados por tal contribución.

vi. ¿Contribuye con el bienestar psíquico y/o felicidad?

Esta es quizás la pregunta más moderna y en apariencia paradójica de las siete que hemos seleccionado. Si bien, como se mencionó en el apartado precedente, el autoengaño parece tener una función de protección contra verdades dolorosas, tradicionalmente se ha considerado que es una clase de error censurable, tanto desde el punto de vista de la racionalidad como desde la perspectiva ética. Claramente, atenta contra la máxima tradicional del Oráculo de Delfos: "conócete a ti mismo". Sin embargo, hace ya más de tres décadas comenzaron a aparecer una serie de trabajos que cuestionaron la visión tradicionalmente negativa sobre el autoengaño. En particular, los trabajos de S. Taylor y sus

[17] Véase, por ejemplo, el empleo que se hace de "adaptativo" en el Manual Diagnóstico y Estadístico de los Trastornos Mentales, Cuarta Edición, en referencia al denominado "trastorno adaptativo": "La característica esencial del trastorno adaptativo es el desarrollo de síntomas emocionales o comportamentales en respuesta a un estresante psicosocial identificable. Los síntomas deben presentarse durante los 3 meses siguientes al inicio del estresante (Criterio A). La expresión clínica de la reacción consiste en un acusado malestar, superior al esperable dada la naturaleza del estresante, o en un deterioro significativo de *la actividad social o profesional (o académica)* (Criterio B)" (DSM IV, p. 639, cursivas nuestras).

colaboradores enfatizaron la importancia de lo que denominaron "ilusiones positivas" para hacer referencia a una serie de creencias que no son sostenidas por los elementos de juicio (autoevaluaciones positivas no realistas, percepciones exageradas de control o habilidad y optimismo no realista) y que pueden servir a una amplia variedad de funciones cognitivas, afectivas y sociales. Plantean también el intento de resolver una paradoja: ¿cómo pueden las percepciones erróneas de uno mismo y del entorno ser adaptativas cuando procesar la información con precisión parece ser esencial para el aprendizaje y el funcionamiento exitoso en el mundo? Este punto de vista (Taylor y Brown, 1988; Trivers, 2000; Erez et al, 1995) ha sido tan influyente como polémico; otros autores (Colvin et al, 1995; Badhwar, 2008; Van Leeuwen, 2009) han objetado que los presuntos bienestar y felicidad obtenidos por medio del autoengaño, en el dudoso caso de que fuesen posibles, serían ilusorios o efímeros.

vii. ¿Implica responsabilidad moral?

Como es fácil de imaginar, esta pregunta ha sido una de las de mayor interés dentro del ámbito filosófico, y sin duda es aquella a la que se le ha dado la respuesta que goza del mayor consenso entre los estudiosos del tema. Este consenso, prácticamente unanimidad, ha consistido en la afirmación de que quien se autoengaña es responsable y censurable por su estado; es digno de reproche desde el punto de vista moral (Baron, 1988; Mele, 2001; Deweese-Boyd, 2008). Es de imaginar, también, que existe una conexión muy directa entre las concepciones intencionalistas del autoengaño y la posibilidad de atribuir responsabilidad moral: si alguien se autoengaña intencionalmente (aun cuando sea de manera oblicua e indirecta), entonces tiene control sobre sus acciones y, en consecuencia, es responsable por ellas. Sin embargo, esta situación ha comenzado a no ser tan clara a partir de los enfoques "deflacionistas" del autoengaño. Si el autoengaño no es producto de alguna clase de intención del agente, sino de una compleja combinación de sesgos cognitivos no conscientes y estados emocionales y motivacionales, parece mucho más difícil, a excepción de ciertas circunstancias muy específicas, adjudicarle responsabilidad moral (Levy, 2004).

En los capítulos que siguen iremos presentando las respuestas a estos y otros problemas relativos al autoengaño. En vez de proceder a examinarlos siguiendo estrictamente el orden presentado, los estudiaremos en relación con las principales disciplinas que se han ocupado de ellos.

Capítulo II. Intenciones no conscientes, creencias contradictorias, mentes divididas: el autoengaño en la Filosofía contemporánea

Como hemos mencionado en la Introducción, la tarea de dilucidar la naturaleza del autoengaño ha sido tradicionalmente una empresa filosófica. El hecho de que, en las últimas décadas, cada vez más disciplinas científicas se hayan interesado en este fenómeno no ha eliminado las dimensiones del problema específicamente filosóficas,[18] y de ellas nos ocuparemos en el presente capítulo.[19]

Suele hacerse referencia al artículo de R. Demos "Lying to Oneself" (1960) como la piedra basal para el tratamiento filosófico contemporáneo sobre el tema. En este artículo podemos encontrar, analizados o sugeridos, muchos de los problemas relativos al autoengaño que ocuparán gran parte de la investigación sobre el tema en las siguientes décadas: el requisito de la coexistencia de creencias contradictorias, la adquisición autoengañosa de creencias indeseadas, la responsabilidad moral por el autoengaño, la división de la mente como forma de solución a las paradojas originadas por el requisito de creencias contradictorias, la posible distinción entre el autoengaño y el pensamiento desiderativo y la existencia de un conflicto interno en quien se autoengaña.

Demos comienza su análisis observando que en el lenguaje natural los términos mentir y engañar no son estrictamente equivalentes. Mientras que en "engañar" es el efecto lo que cuenta (puede inducirse a alguien a error sin tener la intención de hacerlo) en "mentir" la intención es parte del significado. Más aun, es posible mentir a otro sin haber tenido éxito en ese intento. Demos señala que una consecuencia extraña de este uso reside en que uno puede mentir a otro y de este modo inducir una creencia verdadera en el segundo. Se propone emplear el término "mentir" de un modo que evite esta rareza. De este modo, define "B miente (engaña) a C" del siguiente modo: B intenta inducir una creencia errónea en C, B tiene éxito en llevar adelante esta intención, y B sabe (y cree) que lo que ha dicho a C es falso. Los tres requisitos (intención, resultado y conocimiento) están incluidos. Sobre la base de lo anterior, Demos considera que existe autoengaño cuando

[18] Soy consciente de que la expresión "específicamente filosóficas" es problemática, entre otras razones porque las fronteras de la filosofía con otras disciplinas son notoriamente difusas; no obstante, creo que nadie negaría que el fenómeno del autoengaño implica problemas conceptuales y éticos característicos de aquella disciplina.

[19] El problema del autoengaño ha sido especialmente estudiado en la filosofía de orientación analítica, y es a esta corriente a la dedicaremos nuestra atención. Esto no implica, obviamente, desmerecer los importantes análisis del problema en la filosofía continental (el examen de Sartre de la "mala fe" es un ejemplo especialmente destacado).

> [U]na persona se miente a sí misma, esto es, se persuade a sí misma para creer que *sabe* lo que de hecho no ocurre. En pocas palabras, el autoengaño implica que B cree tanto *p* como no *p* en el mismo momento. Entonces, el autoengaño involucra un conflicto interno, quizás la existencia de una contradicción. Pero esto parece una imposibilidad (p. 588).

Creer y descreer, señala Demos, son actitudes en pro y en contra; al ser contrarias, es lógicamente imposible que coexistan en una persona en el mismo momento y en el mismo respecto.[20] Cuando B se miente a sí mismo termina por creer lo que sabe que es falso. Aceptar esto como una descripción de un hecho es admitir una violación de la ley de contradicción, con lo cual parecería que el autoengaño –mentirse a uno mismo- es lógicamente imposible. Podría ser el caso, observa Demos, que la descripción dada sea errónea. Una primera descripción alternativa, que evite la violación de la ley de contradicción podría adquirir la siguiente forma: en el autoengaño, creer que *p* y descreer que *p* ocurren en momentos diferentes y sucesivos. Una segunda posibilidad sería que, en el autoengaño, la creencia agradable ocupa la mente consciente mientras que la desagradable es reprimida en el inconsciente. Demos observa que ninguna de estas hipótesis concuerda con la evidencia. Concluye que ambas hipótesis son falsas; el creer y el descreer son *simultáneos y ambos existen en la conciencia de la persona.*

Para presentar su propia solución al problema del autoengaño, Demos propone emplear el análisis familiar de "percatarse", que ilustra por medio de un ejemplo. Tengo un dolor de cabeza y tomo una aspirina; como resultado de esto el dolor desaparece. Pero supóngase que la aspirina no está disponible, y acepto la invitación de un amigo para ir al cine, en donde mi atención es absorbida por una película excitante. Mientras estoy viendo la película no "siento" mi dolor de cabeza, pero tan pronto como finaliza, experimento nuevamente el dolor. Demos sugiere que las dos situaciones difieren en que, en el primer caso, el dolor simplemente desaparece, mientras que en el segundo continúa pero, absorto en la película, no me percato de él. Demos considera que lo mismo ocurre con el autoengaño, por ejemplo, cuando B cree que *p* mientras sabe que no *p*. Hay dos niveles posibles de percepción. Uno es la percepción simple, el otro la percepción acompañada de atención, o percatación. Se sigue que puedo estar consciente de algo sin que, al mismo

[20] Debe hacerse notar aquí una diferencia importante entre dos tipos de rechazo de la creencia de que *p*. Por una parte es posible negar la creencia de que *p*; en tal caso, el agente afirmará "no creo que *p*", pero no afirmará la falsedad de *p*. Pero además es posible negar que *p* sea verdadera; en tal caso, el agente afirmará "creo que no *p*". Esta diferencia puede ser ilustrada mediante la distinción entre el agnosticismo y el ateísmo; mientras que el agnóstico no posee la creencia en Dios, pero no se pronuncia respecto de su existencia, el ateo niega la existencia de Dios y, por lo tanto, considera falsa la creencia que afirma su existencia. Para un análisis de las diferencias entre estas dos clases de escepticismo respecto de una creencia *cfr.* Canfield y McNally (1961). Demos emplea el término "descreer" en el segundo sentido descripto.

tiempo, lo advierta o concentre mi atención sobre ello. Esto es posible porque puedo estar distraído por algo más, o porque puedo deliberadamente ignorarlo, o porque puedo no desear pensar acerca de él; en consecuencia, esta no percatación no necesita ser algo que simplemente me ocurre.

Demos observa que su propio examen del autoengaño sigue una línea similar al del análisis aristotélico de la *akrasia*. Como en el caso de la *akrasia*, hay un impulso que favorece una creencia a expensas de su contradictoria, y la persona que se miente a sí misma, porque cede ante el impulso, fracasa en advertir o bien ignora lo que sabe que es el caso. Tal análisis, en su opinión, "salva" el fenómeno mientras al mismo tiempo se ajusta a los requisitos de la ley de contradicción. Ciertamente, decimos que la persona que se miente a sí misma cree tanto *p* como no *p*, y es capaz de hacerlo porque está distraída respecto de la primera.

A la publicación del artículo pionero de Demos le sucedieron un enorme número de estudios filosóficos sobre distintos aspectos del fenómeno, tanto artículos,[21] como libros.[22] Estos estudios ha abarcado múltiples temas, que incluyen el escepticismo respecto del autoengaño, la diferencia de este fenómeno con otras formas de irracionalidad motivada, el producto del autoengaño, su presunto carácter intencional, las paradojas que genera, los distintos subtipos de este fenómeno y las implicaciones morales del autoengaño, entre otros. Si bien los debates filosóficos sobre el autoengaño han sido y siguen siendo en la actualidad una fuente fundamental para su comprensión, el propósito de este libro no es exclusivamente filosófico, por lo que nos limitaremos a presentar en este capítulo el debate fundamental, esto es, el que opone dos modelos radicalmente opuestos para la comprensión del fenómeno; asimismo, revisaremos uno de los principales desafíos filosóficos a los intentos de solución del problema, esto es, la posición que denominamos en el capítulo 1 *escepticismo respecto del autoengaño*. No obstante, en capítulos posteriores volveremos a plantear problemas específicamente filosóficos sobre el autoengaño, en particular, los relativos a sus implicaciones morales.

[21] Canfield y McNally, 1961; Siegler, 1963; Paluch, 1967; Gardiner, 1970; Hamlyn y Mounce, 1971; Szabados, 1974; Saunders, 1975; Champlin, 1977; Martin, 1979; Palmer, 1979; Bok, 1980; Kipp, 1980; Wilson, 1980; Bach, 1981; Audi, 1982; Davidson, 1982, 1986, 1998; Mele, 1983, 1997, 2003, 2004, 2006, 2007; Oksenberg-Rorty, 1985, 1988, 1994; Sorensen, 1985; Cook, 1987; Knight, 1988; Catalano, 1990; Pears, 1991; Hales, 1994; Dupuy, 1995; Scott-Kakures, 1996, 2000, 2001; 2002, 2009; Pataki, 1997; Fairbanks, 1999; Lazar, 1999; Bermúdez, 2000; Holton, 2000; Forrester, 2002; Nelkin, 2002; Borge, 2003; Levy, 2003, 2004; Noordhof, 2003, 2009; Funkhouser, 2005, 2009; Pedrini, 2005; Booth, 2007; Gendler, 2007; Martínez Manrique, 2007; Pihlström, 2007; Van Leeuwen, 2007a, 2007b, 2008, 2009, 2013; Lynch, 2009.

[22] Fingarette, 1969; Mele, 1987; 2001; McLaughlin y Oksenberg-Rorty (eds.), 1988; Barnes, 1997.

La postura filosófica canónica respecto del autoengaño ha combinado, hasta una fecha relativamente reciente, un conjunto de rasgos definitorios. Tal postura concibe al autoengaño como un fenómeno análogo al engaño interpersonal, lo que supone la existencia de intencionalidad en el proceso, la postulación de una coexistencia de creencias contradictorias y una división de la mente en alguna clase de sistemas más o menos claramente delimitados.[23] La asociación entre estos rasgos ha sido muy regular: quienes han defendido el carácter intencional del autoengaño han tendido a sostener también el requisito de creencias contradictorias y alguna clase de división de la mente (Pears, 1984; Davidson, 1986; Oksenberg-Rorty, 1988). Esta posición es conocida habitualmente como *intencionalismo*. Los autores que han negado la analogía entre el engaño interpersonal y el autoengaño, por otro lado, han tendido a rechazar también los otros rasgos mencionados, esto es, la condición de creencias contradictorias y la necesidad de postular una división de la mente (Mele, 1997, 2001; Barnes, 1999; Lazar, 1999). Estas últimas posiciones suelen ser rotuladas como *deflacionistas*, *no intencionalistas* o *motivacionistas*.[24]

No obstante, y como es frecuente en los debates filosóficos, existen posiciones intermedias, esto es, enfoques que aceptan algunas de las tesis del enfoque clásico del autoengaño, pero rechazan otras. De este modo, Talbott (1995) defiende el intencionalismo, pero no el requisito de coexistencia de creencias contradictorias ni la tesis de la división de la mente. Bermúdez (2000), a la vez que defiende el intencionalismo, rechaza el requisito de la coexistencia de creencias contradictorias, aunque admite alguna clase de partición de la mente para responder a algunas objeciones de los enfoques deflacionistas. Pese a estos casos intermedios, la oposición fundamental en los debates filosóficos sobre el autoengaño es la que enfrenta al modelo canónico, con sus tres rasgos distintivos, con los enfoques deflacionistas, caracterizados por el rechazo de todos y cada uno de tales rasgos, y de ella nos ocuparemos en lo que sigue. Sin embargo, antes de ocuparnos de los modelos antagónicos sobre el autoengaño, tendremos que examinar el primer desafío a la posibilidad de encontrar una elucidación filosófica adecuada del

[23] La tesis de la división ha calado profundamente no sólo entre filósofos que han dedicado especiales esfuerzos al problema, sino también en aquellos cuyo interés por éste ha sido sólo tangencial. Searle (2001), ilustra adecuadamente esta segunda posibilidad. Considera que el autoengaño típicamente posee la siguiente forma: "El agente posee el estado consciente: creo que no p. Posee el estado inconsciente: tengo evidencia abrumadora de que p y deseo fuertemente creer que no p" (p. 236). A su modo de ver, el autoengaño involucra irracionalidad y en algunos casos incluso inconsistencia lógica, y sólo puede existir si uno de los elementos es suprimido de la conciencia. Dos de los requisitos "clásicos" del autoengaño, en consecuencia, están presentes en la caracterización que Searle hace del fenómeno: el requisito de creencias contradictorias y el de la división de la mente.

[24] Por supuesto, también hay alternativas a las posiciones mayoritarias. Cfr. Fernández (2013) para una propuesta de esta naturaleza.

fenómeno; estas son las posiciones escépticas frente al problema, a cuyo análisis nos dedicaremos a continuación.

1. *¿Es posible el autoengaño? Las respuestas escépticas*

La mayoría de los autores que examinaron el problema con posterioridad a la publicación del artículo de Demos, pese a las discrepancias planteadas con el enfoque de éste, aceptaron la existencia del autoengaño. Sin embargo, para un grupo minoritario, el autoengaño no existe; simplemente, se trata de un seudofenómeno. Llamaremos a esta posición "escepticismo respecto del autoengaño". El atractivo de esta postura parece haber decaído en los últimos años, pese a lo cual hay trabajos recientes que siguen sosteniéndola. La lista de quienes adoptaron una u otra variante de esta posición incluye a Paluch (1967), Champlin (1977), Bok (1980), Kipp (1980) y, más recientemente, Borge (2003). La mayoría de los autores que niegan la existencia del autoengaño provienen del campo de la Filosofía; sin embargo, también hay psicólogos que sostienen argumentos similares, como es el caso de Bandura (1991).

Buena parte del escepticismo respecto del autoengaño se ha debido a la existencia de paradojas aparentemente insolubles. Estas paradojas revisten una importancia conceptual fundamental ya que, como veremos en el apartado 2. de este capítulo, algunas de las posiciones que pueden encontrarse entre los estudiosos del problema se originan justamente como intentos de evitar alguna de ellas. Este es el caso, por ejemplo, de las estrategias divisionistas ante el problema, cuya motivación reside en la necesidad de evitar la paradoja originada por la supuesta coexistencia de creencias contradictorias.

El carácter presuntamente paradójico del autoengaño aparece ya en una temprana réplica al artículo de Demos, debida a Canfield y McNally (1961), quienes objetan la falta de una distinción entre dos tipos de autoengaño, y las paradojas que se derivan de cada tipo. Diversos escritos posteriores al artículo de estos autores desarrollaron y profundizaron las objeciones relacionadas con estas y otras paradojas. Se ha llegado, incluso, a postular la existencia de siete paradojas del autoengaño (Correia, 2007), derivadas específicamente de los enfoques intencionalistas. Sin embargo, las que han recibido más atención son dos: la primera referida al producto del autoengaño, y la segunda relativa al proceso que conduce al autoengaño. De las distintas formulaciones que pueden encontrarse en la bibliografía voy a emplear la presentación de Mele (2001). Mele comienza por enunciar dos supuestos léxicos comunes:

a. Por definición, la persona A engaña a la persona B (B puede o no ser la misma persona que A) para que crea que *p* sólo si A sabe, o al menos cree verdaderamente que no *p*, y causa que B crea que *p*.
b. Por definición, engañar es una actividad intencional. El autoengaño no intencional es conceptualmente imposible.[25]

Cada uno de estos supuestos, observa Mele, está asociado con un rompecabezas particular acerca del autoengaño. Si el supuesto a. es verdadero, entonces engañarse a uno mismo para creer que *p* requiere que uno sepa, o al menos crea verdaderamente, que no *p*, y que cause en uno mismo la creencia de que *p*. Como mínimo, el agente cree inicialmente que no *p* y luego genera en sí mismo la creencia de que *p*. Algunos teóricos, señala, consideran que esto implica que quien se autoengaña cree tanto *p* como no *p*, y afirman que eso implica un estado mental imposible: la naturaleza misma de la creencia excluye la creencia simultánea en *p* y no *p*. En consecuencia, tenemos aquí un rompecabezas *estático* del autoengaño; el autoengaño, de acuerdo con la concepción en cuestión, implica un estado mental imposible. Este rompecabezas, como lo llama Mele, es conocido usualmente como la *paradoja doxástica* del autoengaño.

El supuesto b., prosigue Mele, genera un *rompecabezas dinámico* del autoengaño. En primer lugar, es difícil de imaginar cómo sería posible que una persona engañe a otra para que crea que *p* si la segunda persona sabe exactamente lo que la primera está intentado hacer, y también es difícil imaginar cómo tal objetivo puede ser logrado cuando el que intenta engañar y la presunta víctima son la misma persona. En segundo lugar, el engaño es facilitado normalmente por la posesión y ejecución exitosa, por parte de quien engaña, de una estrategia de engaño. Si, con el objetivo de evitar el fracaso de los propios esfuerzos para autoengañarse, el agente debe ejecutar de manera no intencional cualquier estrategia para lograr autoengañarse, ¿cómo puede esto tener éxito?, pregunta Mele. El desafío es, en consecuencia, explicar cómo el autoengaño en general es un proceso psicológicamente posible. Si quienes se autoengañan se engañan intencionalmente a sí mismos, cabe preguntar qué es lo que evita que la intención que guía el intento socave su propio funcionamiento efectivo. Y si el autoengaño no es intencional, ¿qué motiva y dirige el proceso del autoengaño? Esta segunda paradoja es conocida en la bibliografía como la *paradoja estratégica* del autoengaño.

[25] Tanto la condición de falsedad de la creencia como el carácter necesario de la intención para que sea posible hablar de engaño son objeto de controversia. Cfr. Barnes (1997) y Bermúdez (2000) respectivamente para una discusión acerca de estas cuestiones.

¿Tienen solución estos problemas? Como hemos adelantado, la respuesta ha sido claramente negativa para los escépticos: para ellos, no hay manera de dotar de significado al término "autoengaño" sin caer irremediablemente en paradojas o proponer estados lógica o fácticamente imposibles. Revisaremos brevemente algunos de sus argumentos en lo que sigue.

Paluch (1967), en una de las tempranas réplicas escépticas a Demos, pregunta si puede haber un ejemplo de sustitución para "sé que *p* pero creo que no *p*" (rasgo característico, se supone, del fenómeno en cuestión) que sea a) lógicamente coherente y b) capaz de ser consistente con una acusación de autoengaño. No hay ninguna manera, sostiene, que los modelos de autoengaño puedan, plausiblemente, tratar "saber" y "creer" del modo directo tal que, sea lo que fuere aquello que se diga que un agente sabe o cree, puede ser afirmado como conocimiento o creencia del agente. Para poder lograr esto es necesario introducir usos no estándar de "saber", como los que se encuentran en los modelos de autoengaño de Freud y Demos, que emplean expresiones como "conocimiento inconsciente" o "latente", pero respecto de los cuales no es en absoluto obvio que pueda decirse que reflejan casos reales de autoengaño.

Kipp (1980) examina lo que denomina concepción "literalista" del autoengaño, según la cual el acusado de encontrarse en tal estado sostiene y no sostiene su aparente creencia debido a que sus motivos lo convierten en alguna medida en su propio engañador; los literalistas tienden a describir el proceso de autoengañarse como un "persuadirse a sí mismo" o "engañarse a sí mismo para creer". Esta concepción, sostiene, parece requerir que el autoengañador sea simultáneamente engañado y engañador. Esto, a su vez, parece exigir que existan dos conciencias mutuamente opacas y autónomamente pensantes y volitivas dentro de la mente del sujeto. Sin embargo esas conciencias deberían existir también dentro de una conciencia unificada que base la identidad del autoengañador como un sí mismo. Sin la primera de esas dos condiciones, el ocultamiento propiamente engañoso de la creencia y la intención parece impensable y, sin la segunda, la reflexividad propiamente autoengañosa, dentro de la relación entre engañador y engañado parece igualmente inconcebible. Lo que todo esto requiere, en última instancia es que la conciencia no sea lo que, a su modo de ver, muestra ser más inexorablemente, esto es, algo cuyas "partes" están por naturaleza consustancialmente unificadas en un estado de auto transparencia perpetuamente "holístico".[26]

[26] Al igual que Paluch, Kipp sugiere y desecha la posibilidad de que recurrir a presuntos procesos inconscientes, como los propuestos por Freud, puedan constituir una solución al problema.

Ya en el campo de la Psicología, Bandura (1991), también se inclina decididamente por el escepticismo respecto del autoengaño. Dado que no es posible que alguien sea simultáneamente quien engaña y quien es engañado, observa, el autoengaño literal no puede existir; no es "lógicamente posible" actuar para inducir en uno mismo una creencia que se sabe que es falsa.[27] Los intentos de resolver la paradoja de la simultaneidad engañador-engañado que implican la postulación de varios yoes o sí mismos, uno de los cuales es inconsciente, sólo logran aniquilar al fenómeno en vez de explicarlo; no son capaces de explicar cómo es posible que un yo consciente pueda mentir a un yo inconsciente sin alguna conciencia o conocimiento de lo que el otro cree, requisito este último para que la parte del yo que engaña sea capaz de elaborar el engaño. La alternativa de proponer diferentes niveles de conciencia tampoco consigue superar estas dificultades: sólo vuelve a poner de manifiesto la paradoja relativa al hecho de ser quien engaña y quien es engañado al mismo tiempo. Es un hecho, señala, que las creencias y valores de las personas influyen sobre la información que se busca y cómo se la interpreta. La mayoría de las personas intenta mantener o incrementar su autoestima; en consecuencia, no buscan evidencia de sus fracasos o culpabilidad por los efectos negativos de sus acciones. Ahora bien, la autoexposición selectiva y las interpretaciones sesgadas de los sucesos, tendientes a mantener o fortalecer las creencias preexistentes, no constituyen autoengaño.[28]

Pese a que, como hemos adelantado, existen otros intentos de negar la existencia del autoengaño, lo expuesto debería bastar para ilustrar el núcleo de los argumentos escépticos ante el fenómeno, incluyendo los más recientes.[29] Ahora bien, lo anterior no es todo lo que puede decirse respecto de esta posición. Un intento bastante diferente de negar la existencia del autoengaño (al menos del modo en que habitualmente se lo concibe) es aquel en el cual la respuesta a la pregunta relativa a la existencia del autoengaño no es claramente afirmativa ni negativa. Este tipo de respuesta es característica de aquellas posiciones a las que en la Introducción hemos denominado "disolucionistas". Estas posiciones suponen que el autoengaño no existe tal como se lo ha concebido clásicamente,

[27] Tal imposibilidad, cabe aclarar, puede ser psicológica, pero nada hace pensar que es lógicamente imposible que alguien pueda intentar inducir en sí mismo una creencia falsa.

[28] Bandura ha mantenido esta posición escéptica respecto del autoengaño en escritos posteriores (cfr. Bandura, 2011).

[29] Baste al respecto esta cita de Borge (2003): "No existe algo así como el autoengaño. El autoengaño *qua* autoengaño es un mito. Cuando la noción de "autoengaño" es tratada seriamente resulta intrínsecamente incoherente. Los intentos filosóficos de evitar esto terminan renunciando o bien al elemento de engaño o bien al elemento del sí mismo [*self-element*] de aquello que se supone que constituye el autoengaño. Las explicaciones que *prima facie* parecen tener seriamente en cuenta ambos elementos del autoengaño hacen reaparecer la así llamada paradoja del autoengaño en algún otro nivel de análisis" (p. 1).

pero lo hacen a partir de una perspectiva teórica muy diferente a la de los escépticos hasta aquí considerados. Las examinaremos a continuación.

Para comprender cabalmente las características del tipo de respuesta que estamos considerando conviene señalar, en primer lugar, que el concepto mismo de autoengaño parece suponer como sustrato filosófico (al menos para la mayoría de los autores) la posibilidad de cierto ajuste entre las creencias y el mundo, esto es, parecería implicar cierta clase de realismo acerca del mundo externo y de teoría correspondentista de la verdad. En principio, sólo podemos engañar a otros (y eventualmente autoengañarnos) en caso de que exista un mundo independiente de nuestras representaciones de él y, además, en caso de que nuestras representaciones puedan ajustarse o corresponder en mayor o menor medida a ese mundo. Ahora bien, existen también perspectivas sobre el autoengaño que, enroladas en la corriente del construccionismo social, adoptan explícitamente supuestos filosóficos no realistas, subjetivistas y relativistas. Existen diversos trabajos en esta línea, entre los que se encuentran los de Gergen (1985), a menudo citados en los análisis posteriores, Harré (1988), Lewis (1996) y Clegg y Moissinac (2005). Estos dos últimos trabajos se caracterizan, en particular, por adoptar la perspectiva que hemos calificado de "disolucionista", que podría caracterizarse de la siguiente forma: el autoengaño no existe como fenómeno en la mente individual del autoengañado, sino que constituye una suerte de artefacto o construcción cultural, más ilustrativo de procesos sociales culturalmente determinados que de procesos psíquicos individuales. Debido a este apartamiento radical de los supuestos que fundamentan tanto las posiciones escépticas como las realistas acerca del autoengaño, la he considerado como una línea relativamente independiente ante el problema.

Dentro de los varios intentos en esta línea de pensamiento, describiremos brevemente aquí a los de Lewis (1996) y Clegg y Moissinac (2005). Vale aclarar que la denominación de "disolucionistas" no es empleada por los propios autores; en ningún momento consideran que lo que hacen sea una disolución del problema.

Un modelo relacional de la conciencia humana, sostienen Clegg y Moissinac, posibilita explicar el autoengaño, pero no lo hace por medio de la introyección de una experiencia fragmentada dentro del psiquismo (como los modelos que proponen una división de la mente), sino restaurando esa fragmentación hasta sus orígenes fenomenológicos en el así llamado mundo "objetivo". Lo que denominan "teoría relacional de la conciencia" afirma que los objetos de la experiencia no son reducibles más allá de sus relaciones primarias, relaciones que tienen lugar tanto dentro como entre interlocutores. El mundo objetivo (o quizás intersubjetivo) es a la vez único y común; único porque cada

"objeto" está irreductiblemente constituido en actos de experiencia particular y común porque otros interlocutores también constituyen cada "objeto". El mundo de los objetos no es independiente del discurso o la conciencia.

Este modelo de la conciencia, entienden, tiene consecuencias importantes para la teoría del autoengaño. En primer lugar, implica que todos los "objetos" son idiosincrásicos, esto es, no hay una realidad externa objetiva unificada que confiera autoridad epistémica a un observador externo. Más aun, implica que las discrepancias evidentes en los casos paradigmáticos de autoengaño no son el resultado de una conciencia fragmentada cegada a sí misma a los hechos inevitables de la realidad "objetiva", sino, más bien, el resultado de diferentes observadores que captan diferentes objetos idiosincrásicos, diferentes mundos fenomenológicos y hacen en consecuencia evaluaciones radicalmente diferentes. Tal supuesta naturaleza dual de los objetos de la experiencia, observan Clegg y Moissinac, es lo que hace posible al autoengaño.[30] En la misma línea, Lewis observa que si se considera seriamente la perspectiva según la cual es a través de nuestras historias que construimos nuestras experiencias del mundo, y que la experiencia de nosotros mismos en cada momento es una función de nuestras auto-historias, resulta entonces inevitable cuestionar la significatividad de la propia noción de autoengaño. El autoengaño existiría sólo como idea (historia) en la mente de un observador externo. Si bien la mente humana posee una notable capacidad para subsistir en un contexto de "infinita fragmentación" y crear un orden posteriormente percibido como unitario, uno de sus límites parece estar dado por la incapacidad de albergar más de una historia a la vez. Tal como cuando se lee una novela, prosigue Lewis, las historias requieren una inmersión del yo que necesita, al menos transitoriamente, una suspensión de las alternativas involucradas. La habilidad de cambiar las propias implicaciones, de sumergirnos a nosotros mismos en diferentes narrativas, requiere a la vez una suspensión de las alternativas.

[30] Clegg y Moissinac presentan un ejemplo típico de autoengaño y ofrecen una peculiar interpretación de éste: "Un padre amante advierte que su hija, normalmente animada, está crecientemente retraída, apática y gruñona. Ha perdido el apetito. Recibe llamadas a horas extrañas y a continuación abandona la casa abruptamente; pese a que no visita más a sus antiguos amigos. Comienza a vestir blusas con mangas largas incluso en verano y rechaza ir a la playa, la que ha sido su lugar favorito. Comienza a encerrarse en su habitación, algo que rara vez solía hacer. Él le pregunta ocasionalmente si se está sintiendo bien, pero ella lo despacha con un lacónico "sí". Un día la descubren muerta con una aguja en su brazo. Cuando la policía le informa lo ocurrido, dice que no puede creer que su hija fuera una drogadicta, que está atónito, que eso es totalmente imposible" (p. 103). Mientras que un enfoque de la conciencia fragmentada requiere al menos de dos "yoes", constructos hipotéticos extraexperienciales, el enfoque relacional puede explicar las mismas discrepancias sin postular tales constructos. Si se toma como dado que los objetos de la experiencia son idiosincrásicos, entonces se esperaría que el padre del ejemplo constituya objetos muy diferentes que el observador externo. Las discrepancias en la interpretación del padre y de un observador externo no requieren un homúnculo clandestino que ciegue al padre ante la verdad (esto es, la evaluación del observador externo); son inherentes a un mundo de objetos relacionalmente constituidos, un mundo en el cual lo único y lo común están en constante conversión.

Tanto para Clegg y Moissinac como para Lewis el supuesto fenómeno del autoengaño informa más acerca de procesos sociales y culturales que acerca del funcionamiento de nuestra mente. Los primeros, siguiendo a Gergen (1985), señalan que el autoengaño es un componente de la etnopsicología de la cultura, esto es, el sistema de creencias acerca de la naturaleza del funcionamiento humano en el nivel psíquico; a su vez, estas creencias de sentido común no son derivadas de la experiencia directa, sino de convenciones discursivas. En consecuencia, el término autoengaño posee poder ilocucionario como una herramienta retórica, más que como algo que hace referencia a un fenómeno mental específico. Por su parte, Lewis observa que, dentro de la perspectiva posmoderna "el autoengaño sigue siendo un fenómeno intrigante, no por lo que nos dice acerca del misterioso funcionamiento de la mente, sino por lo que dice acerca de la cultura en la que vivimos" (pp. 63-64).

Hasta aquí, algunas de las principales objeciones escépticas a la existencia del autoengaño. Una evaluación completa de tales posiciones requeriría una descripción mucho más exhaustiva que la que hemos realizado; en consecuencia, tal evaluación queda fuera de nuestras posibilidades. No obstante, sí es posible señalar algunos aspectos de interés de las distintas posiciones escépticas que hemos desarrollado, y, también, extraer algunas conclusiones tentativas respecto de las posiciones escépticas en general. Dadas sus diferencias, examinaremos por separado los argumentos escépticos propiamente dichos y los argumentos construccionistas.

Paluch, Kipp y Bandura coinciden en que el autoengaño es imposible, pero se basan en fundamentos distintos. Paluch se basa en la imposibilidad de dar un sentido aceptable a la afirmación de que alguien posee actitudes coexistentes incompatibles hacia una misma proposición; esto es, que sepa que *p* pero crea que no *p*. Kipp y Bandura, por su parte, argumentan en favor de la imposibilidad de que alguien sea simultáneamente el engañador y el engañado; el proyecto de inducir en uno mismo la creencia falsa está destinado al fracaso, y la estrategia de proponer partes de la mente que funcionen de manera aislada unas de otras están lejos de solucionar el problema. En Paluch, entonces, la tesis de la imposibilidad del autoengaño se basa en el supuesto de que este fenómeno debe satisfacer el requisito de creencias contradictorias; como vimos, la suposición de que el agente debe creer simultáneamente en *p* y en no *p*. En Kipp y Bandura por otro lado, la imposibilidad parece basarse en una consideración estratégica: no hay forma de que un agente pueda llevar adelante consciente y deliberadamente el proyecto de mentirse a sí mismo. Los argumentos

escépticos considerados, en síntesis, mantienen la línea expresada al inicio de este apartado: las paradojas estática y dinámica del autoengaño no pueden ser razonablemente resueltas de un modo tal que torne aceptable la existencia de este fenómeno.

Más en general, podemos observar que el procedimiento general que subyace a las posiciones escépticas consideradas parece ser el siguiente. Se construye (o se adopta uno preexistente) un modelo del autoengaño que tiene ciertas propiedades; en particular, un modelo que presenta analogías esenciales con los procesos de engaño interpersonal. Luego, se muestra como tal modelo, aun en sus interpretaciones más caritativas, presenta defectos o inconsistencias que lo hacen inviable. En consecuencia, se concluye que el autoengaño es imposible. La efectividad de esta estrategia depende, obviamente, de que hayamos elegido o construido un modelo plausible de autoengaño y que no haya otros modelos igualmente buenos a nuestra disposición. Si los hay, como ocurre si se renuncia a modelar el autoengaño sobre la base del engaño interpersonal, entonces los argumentos escépticos no logran su objetivo.[31]

Por su parte, y como hemos visto, la perspectiva construccionista se aparta marcadamente no sólo de las explicaciones usuales (ya sean divisionistas, intencionalistas, deflacionistas, etc.) del autoengaño, sino que cuestiona de forma radical los supuestos filosóficos mismos que usualmente subyacen al planteo del problema. Dado que el problema tradicional simplemente desaparece (recordemos la afirmación "Lo que los teóricos deben explicar no es el "autoengaño", sino las evaluaciones idiosincrásicas") y es sustituido por otro, parece preferible el rótulo de "disolucionistas", y no simplemente escépticos ante el autoengaño. El apartamiento de los planteos mayoritarios ante el problema es tan radical, que parece dudoso que sea posible hacer una crítica puntual ante la propuesta de solución construccionista. A nuestro modo de ver, la crítica tendría que sobrepasar el marco del problema del autoengaño y apuntar a los supuestos mismos del

[31] No carece de interés observar aquí que es posible hacer una observación similar respecto de aquellos que creen haber encontrado comprobaciones empíricas de la existencia del autoengaño. Esta posición puede encontrarse en Gomila (2007), quien considera que los estudios experimentales de Michael Anderson y sus colaboradores sobre el olvido voluntario muestran empíricamente la existencia de la represión en su sentido freudiano y, una vez aceptado eso, quienes niegan la existencia del autoengaño se encuentran en una posición indefendible: "Respecto a la primera cuestión, debe notarse de entrada que la demostración empírica de un fenómeno pone en una situación insostenible a quien afirme que ese fenómeno no es posible. En la medida en que esta afirmación se deriva del supuesto carácter paradójico de un determinado modelo de comprensión de ese fenómeno, lo que se requiere en todo caso es rechazar ese modelo, y proponer otro que evite tales paradojas" (p. 105). Más allá del dudoso carácter demostrativo de una prueba empírica, lo que sostiene este autor parece, en el mejor de los casos, discutible. Podría objetarse, en primer término, que no hay razones para sostener que el mecanismo de la represión deba ser el mecanismo explicativo básico del autoengaño. Sin embargo, la razón principal para dudar reside en que ninguna investigación empírica provee datos que tengan interpretaciones evidentes; como es sabido, un mismo conjunto de datos puede ser susceptible de muchas interpretaciones diferentes. Volveremos sobre los intentos de probar empíricamente la existencia del autoengaño en el apartado 2 del capítulo III.

construccionismo, tarea que excedería completamente los objetivos de este apartado y también de este libro. Sin embargo, no está demás señalar que el ejemplo que proporcionan Clegg y Moissinac parece singularmente inadecuado para mostrar las ventajas explicativas de la perspectiva construccionista, y tal vez, por la contraria, sirve para mostrar que carece de sentido hablar de autoengaño si no se adopta una perspectiva mínimamente realista. Parece plausible suponer que cualquiera admitiría que, aunque se acepte que puedan existir interpretaciones discrepantes de la situación ("construcciones", si se prefiere), es innegable que una de las dos construcciones tiene una mayor probabilidad que la otra de impedir un resultado indeseable (la muerte de la hija, en el ejemplo sugerido). Sólo podemos considerar en pie de igualdad a ambas construcciones si prescindimos de este resultado, cosa a todas luces absurda. Lo que pretendemos afirmar con esto es que parece inevitable pensar que algunas construcciones nos conducen al éxito en nuestro trato con el mundo, mientras que otras no, y parece forzoso concluir que alguna clase de realismo es la mejor explicación de este éxito. Esta observación no es más que la aplicación, en pequeña escala, de los argumentos que los defensores del realismo científico han esgrimido en contra de sus adversarios antirrealistas o relativistas: el realismo es la única filosofía que no hace del éxito de la ciencia un milagro. Podría afirmarse, en síntesis, que sólo cuando los construccionistas radicales sean capaces de explicar por qué algunas de las "narraciones", "relatos", "construcciones", o como se los desee denominar, tienen más éxito en nuestro trato con el mundo que otros, su perspectiva será digna de ser tomada seriamente en consideración.

2. *El modelo filosófico "clásico"*

Hemos señalado ya en más de una oportunidad que el modelo "clásico" del autoengaño, para muchos filósofos, concibe este fenómeno sobre la base proporcionada por los procesos de engaño interpersonal. El engaño interpersonal típico, en el que A engaña intencionalmente a B, parece requerir lo siguiente: que A crea que no es el caso que *p*, desee que B crea que *p*, crea que comportarse de una determinada manera probablemente causará que B crea que *p*, actúe de esa manera determinada y, en consecuencia, cause que B crea que *p*. El autoengaño, por consiguiente, debería caracterizarse de la siguiente manera: un agente A se engaña a sí mismo siempre y cuando crea que no es el caso que *p*, desee causar en sí mismo la creencia de que *p*, crea que comportarse de determinada manera producirá en él la creencia de que *p*, se comporte de esa determinada manera y, en consecuencia, cause en sí mismo la creencia de que *p*. En

vista de lo anterior, cualquier caracterización aceptable del autoengaño conservaría dos de los rasgos típicos del engaño interpersonal, esto es, la intencionalidad del proceso y el requisito de la coexistencia de creencias contradictorias; así como en el engaño interpersonal el engañador A posee la creencia de que no *p*, y el sujeto engañado B posee la creencia de que *p*, en el autoengaño el agente S creerá simultáneamente que *p* y que no *p*. La necesidad de alguna clase de partición o división, ya sea temporal o de la mente, tercer rasgo característico del modelo clásico, surgirá como una forma de evitar las paradojas creadas por las dos primeras condiciones.

Pasemos entonces a considerar el primer rasgo fundamental del modelo clásico: el carácter intencional del autoengaño. Como suele ocurrir, existen diversas formas de describir el intencionalismo. Mele (2001) presenta esta posición de la siguiente manera:

> La perspectiva de la agencia: todas las creencias motivacionalmente sesgadas son producidas o protegidas intencionalmente. En cada caso de creencia motivacionalmente sesgada de que p, tratamos de causar la adquisición o retención de p, o al menos tratamos de hacer más fácil para nosotros mismos la adquisición o retención de la creencia (p. 13).

Podemos también recurrir a la definición de autoengaño que proporciona Davidson (1986):

> Un agente A se autoengaña con respecto a una proposición P bajo las siguientes condiciones: A posee evidencia sobre la base de la cual cree que P es más verosímil que su negación; el pensamiento de que P, o de que debería creer racionalmente que P, ofrece a A motivos para actuar con vistas a causar en sí mismo la creencia en la negación de P. (…) Todo lo que el autoengaño exige de la acción es que el motivo tenga su origen en una creencia en la verdad de P (o en el reconocimiento de que la evidencia hace más probable la verdad de P que su falsedad) y *que se lleve a cabo con la intención* de producir una creencia en la negación de P (pp. 111-112. Cursivas nuestras).

Ahora bien, el carácter intencional del autoengaño parece *prima facie* paradójico. Si somos conscientes de que poseemos una creencia verdadera (o, como mínimo, una creencia que es la que recibe el mayor grado de apoyo por las pruebas disponibles), ¿cómo es posible que formemos de modo consciente y deliberado la intención de rechazar esa creencia, por los medios que fuese? El intento parece derrotarse a sí mismo, lo que convertiría al autoengaño en una ficción. Parecería entonces que el carácter intencional del autoengaño, como veremos, debe ser oblicuo o indirecto.

La estrategia más empleada para lidiar con paradojas como las descriptas ha consistido en proponer alguna clase de partición o división que permita eludirlas. Ahora bien, existen dos clases de particiones o divisiones: aquellas que llamaremos particiones

temporales y aquellas que denominaremos particiones mentales. Las examinaremos en ese orden y dedicaremos especial atención a las segundas.

La estrategia de la partición temporal descansa en el supuesto de que el autoengaño es un proceso complejo que con frecuencia se extiende en el tiempo, supuesto que coincide con las intuiciones mayoritarias sobre el problema.[32] Esta extensión temporal haría supuestamente posible que quien pretende engañarse pueda disponer las cosas de modo tal de inducir en sí mismo la creencia de que *p*, sabiendo que no *p* y, a lo largo del proceso, perder su creencia de que no *p*. Esto último puede ocurrir o bien por medio del completo olvido de la intención de engaño original, o bien tomándola como la causa, aunque accidental, de la creencia verdadera, a la cual se hubiera llegado de cualquier manera. Deweese-Boyd (2012) sugiere el siguiente ejemplo: un oficial de policía involucrado en alguna conducta ilegal podría destruir cualquier rastro de esta conducta y crear pruebas que lo encubrirían (entradas en un diario, correos electrónicos y otras), sabiendo que probablemente olvidará el haber ejecutado esas conductas en unos pocos meses. Cuando, un año más tarde, sus actividades son investigadas, ha olvidado sus esfuerzos manipuladores, y, basado en tales pruebas fraguadas, cree falsamente que no ha estado involucrado en las actividades ilegales de las que es acusado. En este caso, observa Deweese-Boyd, el agente en cuestión no necesita sostener nunca de modo simultáneo creencias contradictorias, aun cuando ha intentado generar en sí mismo la creencia de que *p*, la que considera falsa al inicio del proceso que conduce a su adopción, y verdadera al finalizar éste. Más aún, destaca, el agente ni siquiera necesita olvidar su intención original de generar el engaño; así, un no creyente que desea inducir en sí mismo la creencia en Dios, bien podría recordar tal intención al finalizar el proceso y considerar que, por la gracia de Dios, incluso ese erróneo camino lo condujo a la verdad. En tales casos resulta crucial tener en cuenta que el éxito de la intención depende de la actuación de lo que se ha denominado "medios autónomos" (por ejemplo, la degradación normal de la memoria o la tendencia a creer lo que uno practica), no la continua conciencia de la intención.[33]

¿Constituye la estrategia de la partición temporal una solución a las paradojas generadas por el enfoque intencionalista? La respuesta parece ser afirmativa: el agente es consciente de su proyecto de inducir en sí mismo la creencia falsa (con lo que se evitan en principio las dificultades supuestas por la existencia de intenciones inconscientes en el agente), pero no sostiene simultáneamente las creencias contradictorias (con lo que se evita

[32] Cfr. para una defensa de esta estrategia Sorensen (1985).

[33] Esta posibilidad se relaciona directamente con los debates relativos a la doctrina conocida como "voluntarismo doxástico indirecto", y a sus relaciones con el autoengaño (cfr. al respecto Williams, 1973).

la necesidad de proponer alguna estrategia *ad hoc* para lidiar con la imposibilidad psicológica supuesta por la coexistencia consciente de dos creencias incompatibles). Sin embargo, no parece satisfacer requerimientos básicos del enfoque clásico. Levy (2004) señala que no es éste el tipo de engaño del sí mismo que los defensores de la concepción tradicional tienen en mente. Si bien sostienen que el autoengaño es intencional, esta intencionalidad no es claramente perceptible para el agente; este no reconoce, incluso ante sí mismo, que está intentando engañarse. Para la concepción tradicional, quien se autoengaña se involucra en una conducta autoengañosa debido a que no desea enfrentar una verdad incómoda; pero admitir ante sí mismo que está tratando de engañarse implicaría el reconocimiento de la misma creencia que desearía evadir. En consecuencia, observa, "si el autoengaño es intencional, es una actividad intencional confusa y oscura" (p. 298). Pero el autoengaño que emplea medios indirectos es claramente percibido; involucra la decisión deliberada de cambiar creencias que son conscientemente reconocidas. No es este el tipo de engaño hacia sí mismo, concluye Levy, aquel que los partidarios de la concepción tradicional considerarían autoengaño.

La estrategia de la partición temporal, entonces, no parece ajustarse a algunas intuiciones básicas respecto del autoengaño. Resta evaluar, entonces, si el objetivo de evitar las paradojas puede ser satisfecho por alguna clase de partición mental, esto es, la suposición de una determinada estructura psíquica compleja que haga posible superar las paradojas estática y dinámica. Ahora bien, hablar de una división o partición de la mente suscita varios interrogantes. En primer lugar, el relativo a qué tipo de divisiones está proponiendo; nadie sostendría, es plausible suponer, que un objeto de la notable sofisticación de la mente carece de múltiples estructuras complejas relacionadas entre sí de maneras también complejas. La pregunta es, entonces, a qué tipo de división se está haciendo referencia cuando se propone una partición de la mente como manera de resolver problemas conceptuales relativos al autoengaño. Una segunda pregunta está relacionada con la necesidad conceptual de tales divisiones. La propuesta de división de la mente ha sido usualmente considerada como una consecuencia inevitable, no del requisito de intencionalidad, sino del de la coexistencia de creencias contradictorias. Esto es, suponer que coexisten sin conflicto en el agente dos creencias mutuamente excluyentes parece conducir de manera obligatoria a proponer una partición de la mente que mantenga tales creencias en subestructuras que carecen de contacto entre sí. Tal consecuencia, no obstante, no ha sido vista como inevitable por todos los autores (aunque sí por la gran mayoría). Un buen ejemplo de esta posición minoritaria es el análisis de Talbott (1995).

Este autor señala que, si bien los modelos *divisionistas* deben ser descriptos como intentos de evitar las paradojas que el fenómeno genera, es necesario distinguir entre diferentes clases de divisionismo. De este modo, distingue entre el divisionismo robusto de Rorty, el divisionismo moderado de Pears y el divisionismo débil de Davidson; sin embargo, reconoce que no es capaz de formular una caracterización positiva del divisionismo que supere la observación de que todos los divisionistas explican el autoengaño en términos de alguna clase de partición del yo que está más allá y por encima de los tipos de divisiones invocadas en la explicación de fenómenos no autoengañosos. En este sentido, Talbott emplea la expresión "divisiones inocentes" para hacer referencia a aquellas divisiones del yo que son necesarias para explicar conductas ordinarias, no engañosas. Un enfoque del autoengaño que sólo emplee divisiones inocentes no es divisionista en el sentido que el autor le da al término. Su propia perspectiva del autoengaño es anti-divisionista porque sólo emplea divisiones inocentes en la vida mental para explicar el autoengaño. Talbott considera que no hay esperanzas de evitar el divisionismo si el autoengaño involucra el poseer creencias contradictorias.[34] Entonces, si se pretende explicar el fenómeno sin postular una división, debe negarse, como él lo hace, que el autoengaño involucre creencias contradictorias. Considera que la única razón para suponer que el autoengaño involucra

[34] Otra manera de evitar caer en la paradoja derivada de la coexistencia de creencias contradictorias consiste en negar que el producto del autoengaño sea una creencia; en vez de ella, el producto de este fenómeno sería una manifestación [*avowal*] (Audi, 1982; Rey, 1988). De este modo, no habría creencias contradictorias que entrarían en conflicto. Para ambos autores una manifestación es básicamente una disposición a afirmar una proposición con "sinceridad"; sin embargo, tal proposición carece de conexiones profundas con la acción. Entonces, si quien se autoengaña meramente manifiesta que *p*, no sería necesario atribuir al agente tanto la creencia de que *p* como la creencia de que no *p*, con lo cual se evitaría el problema de explicar la coexistencia de creencias contradictorias. Cuando un agente está autoengañado respecto de que *p*, tanto Audi como Rey sostienen que el agente posee una creencia (que Rey llama "creencia central") de que no *p*. Como ha observado Van Leeuwen (2007) la disociación entre creencias y acciones es un componente esencial del enfoque de la manifestación; de otro modo resultaría completamente oscuro en qué respecto se supone que una manifestación es distinta de una creencia genuina. Para Rey, podría decirse que, *ceteris paribus*, una persona cree manifiestamente que *p* si afirmara sincera y decididamente que *p* si se le preguntara al respecto. Creencias manifestadas, para este autor, son aquellas que se atribuyen sobre la base de la conducta verbal, mientras que creencias centrales son aquellas que se atribuyen sobre la base de las acciones. Audi, a su vez, advierte que el hecho que S se encuentre autoengañado con respecto a *p* no implica su creencia –incluso consciente- de que *p*; lo que su posición requiere respecto de la actitud positiva de S hacia *p* es que S esté dispuesto sinceramente a manifestarlo. Otros autores mantienen similares actitudes críticas respecto de la creencia como candidata a producto del autoengaño. Dupuy (1995) observa que lo que denomina "realismo intencional" respecto de las creencias presente en la filosofía de la mente de orientación analítica (esto es, la doctrina según la cual estados mentales como las creencias poseen una realidad ontológica) no es compartido por otras perspectivas sobre el autoengaño. Lowe (2000), observa a su vez: "[N]uestro modo cotidiano de hablar acerca de "creencias", "conocimiento" y "deseos" posee una complejidad que hasta el momento ha escapado a los intentos de los filósofos de proporcionarle un fundamento claro. No necesariamente hemos de concluir, con los materialistas eliminativos, que los estados de actitud proposicional son una ficción nacida de una teoría de la mente "de sentido común" que es científicamente inadecuada (…). Pero tal vez deberíamos contemplar con mirada más escéptica el cómodo supuesto de que las creencias y los deseos claramente son "estados" de personas, o de sus mentes, en analogía con estados físicos de sus cuerpos, como la forma, la masa y la velocidad" (p. 255). Como otros debates filosóficos acerca del autoengaño, la viabilidad del enfoque de la manifestación permanece como una cuestión abierta. Para un examen acerca de si el resultado del autoengaño puede ser una manifestación, y no una creencia, cfr. Van Leeuwen (2007).

creencias contradictoras es la presunta analogía entre este fenómeno y el autoengaño interpersonal. La posición de Talbott, no obstante, es minoritaria entre los intencionalistas.

Como adelantamos, la estrategia de la partición mental ha sido considerada por los intencionalistas una alternativa más atractiva que la partición temporal, y también preferible a la renuncia al requisito de creencias contradictorias, como sugiere Talbott. En lo que sigue presentaremos brevemente la defensa de la estrategia de la partición mental que desarrolla un divisionista notorio: Donald Davidson, quien puede ser considerado el intencionalista paradigmático respecto del autoengaño. En "Engaño y división" (1986), al discutir el requisito de creencias contradictorias en el autoengaño, Davidson se pregunta cómo podría alguien abstenerse de reunir tales creencias; su respuesta es la siguiente:

> Lo importante es que las personas pueden, y a veces consiguen, mantener separadas creencias estrechamente relacionadas pero opuestas. En esa medida, hemos de aceptar la idea de que puede haber límites entre partes de la mente; allí donde hay creencias (obviamente) antagónicas, postulo la existencia de un límite entre ellas. Tales límites no son descubiertos por la introspección, sino que constituyen apoyos conceptuales para la descripción coherente de irracionalidades genuinas (…) No debemos concebir estos límites como barreras permanentes que demarcan territorios separados (…) Aunque han de pertenecer a territorios fuertemente imbricados, dos creencias contradictorias no pertenecen al mismo territorio; borrar la línea existente entre ellas conllevaría la destrucción de una de las dos. No veo ninguna razón obvia para suponer que uno de los territorios haya de estar cerrado a la conciencia, sea cual fuere el significado de esto, pero es claro, en todo caso, que el agente no puede inspeccionar el todo sin borrar los límites (p. 116).

Aunque el fragmento precedente ilustra de manera clara la justificación de la necesidad de una partición mental, es en un trabajo previo, "Paradojas de la irracionalidad" (1982), en el cual Davidson expone con mayor precisión cuál debe ser la naturaleza de la división.

Las paradojas de la irracionalidad, señala, surgen de aquello que está involucrado en nuestros modos más básicos de describir, comprender y explicar estados y sucesos psicológicos. Así, por ejemplo, explicamos la conducta de un hombre que intenta aprobar un examen memorizando el Corán por medio de su creencia de que memorizarlo incrementará sus posibilidades de aprobar el examen. La existencia de explicaciones racionalizadoras [*reason explanations*] de esta clase es un aspecto inherente a las intenciones, acciones intencionales y muchas otras actitudes y emociones. Tales explicaciones, prosigue, explican por medio de una razón: nos habilitan para percibir los sucesos o actitudes como razonables desde el punto de vista del agente. Un aura de racionalidad, o de ajuste a un patrón racional, es inseparable de estos fenómenos, al menos en la medida en que son descriptos en términos psicológicos. La explicación racionalizadora apela como mínimo a

dos factores: un valor, meta, deseo o actitud del agente, y una creencia de que por medio de la acción que es el objeto de la explicación es posible promover esa meta o valor. La acción, por un lado, y el par deseo-creencia que proporcionan la razón, por el otro, deben estar relacionados de dos modos muy distintos para conducir a una explicación. Primero, debe haber una relación lógica: las creencias y deseos tienen un contenido, y ese contenido debe ser tal que implique que hay algo valioso o deseable en relación con la acción. Así, un hombre que piensa que la salud es algo deseable, y cree que el ejercicio físico lo mantendrá saludable, puede concluir que hay algo deseable en el ejercicio, lo cual puede explicar por qué lo practica. Segundo, las razones de un agente para actuar deben jugar un rol causal en la ocurrencia de la acción.[35] Este análisis de la acción, considera Davidson, arroja claridad acerca de por qué todas las acciones intencionales, sean o no en algún sentido irracionales, poseen un elemento racional en su núcleo; esto es lo que conduce a una de las paradojas de la racionalidad. Puede concluirse, señala, que el mero hecho de rotular a un estado o suceso psicológico como algo que involucra lo que laxamente podría llamarse una actitud proposicional es una garantía de la pertinencia de una explicación racionalizadora y, en consecuencia, de un elemento de racionalidad; no obstante, tales estados y sucesos pueden ser irracionales, y el elemento de racionalidad no puede evitar que sean al mismo tiempo menos que racionales. Estos estados y sucesos incluyen variantes como el pensamiento desiderativo, el actuar en contra del mejor juicio propio y el autoengaño.

Para explicar cómo son posibles las creencias y las acciones irracionales Davidson adopta ciertas ideas de Freud que considera defendibles pese a las críticas filosóficas en su contra, y sobre la base de las cuales es posible brindar una explicación de ciertos fenómenos irracionales. Estas ideas son: la mente contiene un número de estructuras semi-independientes, que se caracterizan porque es posible atribuirles propiedades mentales como pensamientos, deseos y memorias; segundo, tales partes de la mente son, en importantes aspectos, como las personas, no sólo en que poseen creencias, deseos y otros rasgos psicológicos, sino en que esos factores pueden combinarse, como en la acción intencional, para causar sucesos en la mente o fuera de ella; por último, cuando algunas de las disposiciones, actitudes y sucesos que caracterizan las diversas subestructuras de la mente, afectan, o son afectadas por, otras subestructuras de la mente, deben ser concebidas sobre la base del modelo de fuerzas y disposiciones físicas.

[35] El carácter causal de las razones constituye una de las tesis centrales de la influyente posición de Davidson respecto de la explicación de la acción, expuesta en su célebre artículo de 1963 "Actions, Reasons and Causes".

En las explicaciones racionalizadoras estándar, señala Davidson, los contenidos proposicionales de varias creencias y deseos mantienen relaciones lógicas entre ellos y con los contenidos de las creencias, actitudes o intenciones que ayudan a explicar; los estados de creencia y deseo *causan* el estado o suceso explicado. Pero en el caso de la irracionalidad, a diferencia del anterior, la relación causal se mantiene, mientras que la relación lógica desaparece o es distorsionada. Esto es, hay una causa mental que no es una razón para aquello que causa. Su tesis es, entonces, que muchos ejemplos comunes de irracionalidad pueden ser caracterizados por el hecho de que existe una causa mental que no es, a la vez, una razón. Así, por ejemplo, en el pensamiento desiderativo un deseo *causa* una creencia, pero el juicio de que un determinado estado de cosas es deseable *no es una razón* para creer que exista. Esta solución, sin embargo, genera una segunda paradoja, que requerirá, para su solución, admitir la existencia de partes semiautónomas en la mente.

Si los sucesos están relacionados como causa y efecto, observa Davidson, mantienen esa relación independientemente del vocabulario que elijamos para describirlos; los sucesos mentales o psicológicos son tales sólo bajo una cierta descripción, ya que esos mismos sucesos son al mismo tiempo neurofisiológicos (físicos), si bien son identificables dentro de esos dominios sólo cuando se proporcionan descripciones neurofisiológicas o físicas. No hay dificultad en general en explicar sucesos mentales por medio de causas físicas o neurofisiológicas (por ejemplo, en los estudios sobre percepción o memoria). No obstante, cuando la causa es descripta en términos no mentales, necesariamente se pierde contacto con aquello que se necesita para explicar el elemento de irracionalidad. Esto se debe a que la irracionalidad sólo surge cuando la racionalidad es evidentemente apropiada, esto es, cuando causa y efecto tienen contenidos cuyas relaciones lógicas dan lugar a una razón o a su falta. Los sucesos, si son concebidos sólo en términos de sus propiedades físicas o fisiológicas no pueden ser juzgados como razones, ni como estando en conflicto, ni como relacionados con un tema.

Enfrentamos así, dice Davidson, el siguiente dilema: si pensamos en la causa de un modo que no tenga en cuenta su estatus mental (como creencia u otra actitud), esto es, si pensamos en ella meramente como una fuerza que actúa sobre la mente, sin que sea identificada como parte de ella, entonces fracasamos en explicar la irracionalidad. Las fuerzas ciegas se hallan en el dominio de lo no racional, no de lo irracional. De este modo, introducimos una descripción mental para la causa, la cual la convierte en candidata a ser una razón. Sin embargo, permanecemos aún fuera del único patrón claro de explicación que se aplica a lo mental, ya que ese patrón demanda que la causa sea más que una

candidata a ser una razón; debe *ser* (enfatiza Davidson) una razón, lo que no puede ser en el presente caso. Para una explicación de un efecto mental necesitamos una causa mental que sea también una razón para este efecto pero, si la tenemos, el efecto no puede ser un caso de irracionalidad.

Sin embargo, Davidson considera que existiría un modo en el cual un suceso mental podría causar otro suceso mental sin ser una razón para él, y en el cual no habría un rompecabezas ni, necesariamente, irracionalidad alguna. Esto podría ocurrir, sugiere, cuando la causa y el efecto ocurren en distintas mentes. Por ejemplo, deseando que otra persona entre en mi jardín, cultivo una hermosa flor en él. La persona anhela echar una mirada y entra en mi jardín. Mi deseo causa su anhelo y su acción, pero no es una razón para su anhelo, ni una razón por la cual actuó. Entonces, los fenómenos mentales pueden causar otros fenómenos mentales sin ser razones para ellos, y aun así conservar su carácter de mentales, siempre y cuando causa y efecto estén adecuadamente separados. Si bien la interacción social proporciona casos claros y obvios de esto, Davidson considera que la idea puede ser aplicada a una sola mente; de hecho, si se pretende explicar la irracionalidad, se debe suponer, afirma, que la mente puede ser particionada en estructuras cuasi-independientes que se encuentran en interacción.

Lo que es esencial para la explicación de la irracionalidad, entonces, es que ciertos pensamientos y sentimientos de la persona deben ser concebidos como algo cuya interacción produce consecuencias sobre los principios de las acciones intencionales, y esas consecuencias actúan como causas, pero no como razones, para posteriores estados mentales; en particular, cuando los sucesos relacionados como causas (pero no como razones) para una acción pertenecen a distintas mentes, y también a la mente de una sola persona.[36] Sólo mediante una partición de la mente de esta naturaleza, señala Davidson, parece posible explicar cómo un pensamiento o un impulso pueden causar otro con el cual no mantienen ninguna relación racional.

Como hemos adelantado, el divisionismo de Davidson es sólo una variante de esta estrategia intencionalista. Si bien debería resultar claro que no existe ninguna necesidad conceptual de adoptar una perspectiva específica para defender el intencionalismo, como lo hace Davidson al adherir a ciertos principios psicoanalíticos, [37] debería resultar igualmente

[36] No es esencial para la explicación de la irracionalidad, para Davidson, la idea de que partes de la mente deben ser no sólo no conscientes, sino tampoco fácilmente accesibles a la conciencia; la teoría es aceptable, observa, sin sucesos y estados mentales inconscientes.

[37] Oksenberg-Rorty (1988) presenta una posición divisionista alternativa y más fuerte que la sugerida por Davidson.

claro que cualquier elección teórica que introduzca divisiones en la mente, *ad hoc* o no, correrá el riesgo de generar tantos o más problemas que los que pretende resolver.

Hasta aquí, una caracterización del modelo intencionalista ("clásico") acerca del autoengaño. Variantes de de esta posición han sido defendidas (entre otros) por Bermúdez (2000), Davidson (1986, 1998), Demos (1960), Lynch (2009), Oksenberg-Rorty (1988), Pears (1984) y Talbott (1995). Como hemos señalado en más de una oportunidad, la atribución de intencionalidad al autoengaño ha constituido el estándar filosófico sobre el problema; sin embargo, nunca ha estado exento de objeciones, como veremos a continuación. Si bien existen muchos estudios que intentan demostrar las falencias y limitaciones del intencionalismo, seguiremos aquí el análisis de Lazar (1999) quien presenta una serie de objeciones sistemáticas a esta posición, que a su modo de ver lo tornan inviable como explicación del autoengaño.

Lazar señala que el intencionalismo tiene la ventaja de corresponder con un modo en el cual el autoengaño es discutido en los contextos no teóricos: a menudo se dice de alguien de quien sospechamos que está autoengañado, que se engaña a sí mismo. El acto de engañar involucra una intención de causar en otro la formación de una creencia considerada falsa por el engañador. En concordancia, esta perspectiva filosófica retrata al autoengaño como un fenómeno que involucra la intención de formar una creencia que es juzgada falsa por el sujeto. Pero esta no es la única razón de la popularidad de este enfoque. La razón principal es la siguiente: quien se autoengaña es altamente irracional y, a la vez, la presencia de la creencia irracional a menudo se corresponde con una meta del sujeto. Tal creencia falsa puede aliviar la ansiedad o impulsar la autoimagen, mientras que la creencia racional amenaza la satisfacción de esa meta. Esto sugiere que la creencia irracional es adquirida para lograr la meta que es frustrada por la creencia racional. Dado que se parte del supuesto de que el sujeto es competente para detectar la irracionalidad de su creencia, parecería que hay pocas explicaciones disponibles para su formación. La concepción según la cual la creencia irracional se ha formado intencionalmente presenta a la formación de tal creencia como la consecuencia de un razonamiento práctico: es el resultado de un proyecto emprendido por el agente para satisfacer su deseo.

Al insistir en que el autoengañado debe hacer algo para dar lugar a la formación de la creencia deseada, prosigue Lazar, Davidson y otros proponentes del enfoque intencionalista reconocen correctamente dos puntos. Primero, que es imposible formar una creencia directamente al mismo tiempo que se advierte que, considerando todas las

circunstancias, uno estaría mejor si la sostuviera. Segundo, se reconoce que la creencia producto del autoengaño corresponde a alguna de las razones prácticas del sujeto. Por esta razón, el enfoque intencionalista describe al sujeto autoengañado como a alguien que actúa con la intención de que esta acción cause la formación de la creencia deseada.

Lazar concede[38] que el enfoque intencionalista se aplica en algunos casos de autoengaño; sin embargo, sólo un pequeño conjunto de casos especiales deberían ser entendidos de este modo. Para apoyar su afirmación, Lazar presenta tres objeciones al enfoque intencionalista: el problema de las *elecciones locas*, el problema de los *casos negativos* y el problema de *cómo es hecho*. Haremos referencia aquí a los dos últimos problemas, que constituyen dos objeciones de peso al intencionalismo.

Algunos casos de autoengaño, observa Lazar, no garantizan la atribución de un deseo de creer.[39] Propone considerar el caso en el cual el autoengaño resulta en la formación de una creencia más negativa que lo que está justificado por la evidencia; este es, por ejemplo, el caso del marido celoso convencido de la infidelidad de su pareja sin pruebas adecuadas que fundamenten su creencia. Casos negativos como este, observa Lazar, presentan un problema para el intencionalista: parecen socavar la afirmación de que la formación de la creencia irracional en el autoengaño es siempre debida a un deseo de adoptarla. El intencionalista puede tomar una de dos rutas para defender su posición con respecto a tales casos: o bien puede insistir que la creencia irracional es debida a un deseo de creer, o bien denegar que tales casos ejemplifiquen autoengaño.

Con respecto a la primera opción, observa Lazar, no es necesario negar la posibilidad de que quien se autoengaña desee adoptar la creencia; por ejemplo, el marido celoso puede desear terminar el matrimonio y sabe que es más probable que reúna el coraje para hacerlo si cree que su esposa está teniendo un *affaire*. No obstante, agrega, esta estrategia no puede ser universalmente aplicada a los casos negativos de una manera convincente. Muchos casos de celos irracionales, observa, excluyen ésa y similares explicaciones que atribuyen un deseo de formación de la creencia.

---

[38] Al igual que lo hacen otros deflacionistas (Mele, 2001).

[39] El fenómeno de la adopción motivada de creencias "negativas" (creencias que no se desea, en principio, poseer) no justificada por la evidencia disponible para el agente ha recibido diversos nombres en la literatura: *twisted self-deception* (Mele, 1997, 2001); *unwelcome beliefs* (Barnes, 1997); *dreadful self-deception* (Van Leeuwen 2007). La presencia de esta variante del autoengaño en la literatura especializada no es nueva. Ya Demos señaló que una persona puede persuadirse a sí misma para creer en algo desagradable, y menciona el caso de una persona que camina sola por un bosque en tinieblas y se imagina bestias salvajes que acechan a su alrededor. No obstante, el estudio de la adopción motivada de creencias negativas ha sido comparativamente mucho menor que el dedicado a la modalidad más usual del fenómeno, esto es, aquellos casos en los cuales la creencia adoptada de manera motivada e irracional es consistente con los deseos del agente.

La segunda alternativa para responder a este desafío consiste en negar que tales casos ejemplifiquen autoengaño. Esta estrategia puede estar basada en dos premisas distintas. Primero, el intencionalista podría afirmar que, en contraste con casos genuinos de autoengaño, la irracionalidad de los casos negativos nunca es debida a actitudes evaluativas del sujeto autoengañado. Más bien, se diría, la irracionalidad es debida a la incompetencia intelectual del sujeto para detectar la el carácter no racional de su propia creencia, o es el resultado de un tipo diferente de error que no es atribuible a cualquiera de sus actitudes evaluativas. Esta alternativa, considera Lazar, tampoco conduce a una defensa efectiva de la posición intencionalista, ya que no se ajusta al fenómeno. Típicamente, observa, el autoengaño negativo es exhibido en áreas de importancia particular en la vida de la persona, áreas que están asociadas con fuertes deseos y emociones. Asimismo, es posible apelar a patrones de formación de creencias irracionales bien documentados que enfatizan la conexión causal entre deseo y creencias irracionales negativas. El segundo modo de basar la afirmación de que los casos negativos caen fuera del alcance del autoengaño puede consistir en la simple afirmación de que, en el discurso cotidiano, el término "autoengaño" se aplica sólo a casos que resultan en una creencia más positiva que lo que está justificado por la evidencia. En respuesta, observa Lazar, debería decirse que, incluso si los casos negativos no ejemplifican autoengaño, son efectivos para socavar la posición del intencionalista. La similitud entre los casos positivos y negativos del autoengaño es impactante: ambos tipos muestran la formación de una creencia que está a la vez subdeterminada por la evidencia y temáticamente conectada con las actitudes evaluativas del sujeto. A menos que se presente un argumento sólido a su favor, concluye Lazar, existirá una desventaja considerable en una teoría que considere que los casos positivos y negativos ejemplifican fenómenos psicológicos completamente diferentes.

La restante objeción al enfoque intencionalista (el problema de "cómo es hecho") se vincula con la naturaleza problemática del proyecto basado en la intención de formar una creencia. Lazar observa que, más que explicar la formación de la creencia irracional, la presencia de una intención o plan para formar la creencia *constituye un impedimento hacia su formación*, dado que es probable que interfiera con el efecto de los mecanismos de sesgo. Por esta razón, las explicaciones de la formación de la creencia irracional que no apelen a una intención de este tipo tienen *prima facie* una ventaja sobre la explicación intencionalista. El intencionalista puede no conceder este punto; podría afirmar que la intención de formar la creencia irracional no se interpone en el camino de la formación de la creencia en la medida en que el sujeto no sea consciente de ella. Lazar responde que esta réplica no puede

salvar el enfoque intencionalista, dado que el intencionalista no puede responder la pregunta de por qué el agente no es consciente de su intención de formar la creencia irracional. Hay dos opciones más o menos conocidas para explicar la carencia de conciencia del agente. O bien se le da a la correspondencia entre la falta de conciencia, por un lado, y la satisfacción del deseo, por el otro, una explicación *no temática*, o también puede explicársela por la intención del agente de "mantenerla fuera de la conciencia" como parte de un plan que está orientado hacia el logro de esa meta.

En el tipo de explicación no temático, el contenido de la intención de formar la creencia y, en especial, su relación con el logro de la meta no juegan ningún rol en la explicación de la falta de conciencia del agente de la intención (por ejemplo, mi carencia de intención de tomar una manzana mientras estoy involucrado en una conversación fascinante). El intencionalista enfatizará que la falta de conciencia de los estados mentales es un fenómeno bastante general. Sin embargo, esta opción determina que la carencia de conciencia por parte del agente fue debida no a su valor instrumental; es simplemente un "giro afortunado". En consecuencia, el autoengaño sería temático, mientras que la carencia de conciencia de la intención de formar la creencia sería no temática. Esto, observa Lazar, presenta algunos problemas. Primero, si la carencia de conciencia del agente autoengañado está a la par con la carencia de conciencia de tomar la manzana, deberíamos esperar que las personas serían regularmente conscientes de sus intenciones de formar creencias no justificadas. Sin embargo, uno no es nunca consciente de una intención de formar una creencia desacreditada, aun cuando el autoengaño es un fenómeno bastante común. Segundo, la intención de formar una creencia corresponde a una acción muy compleja, a diferencia de la acción de tomar una manzana. Es menos probable que intenciones de esta clase, en contraste con la otra, sean inconscientes por razones no temáticas. Tercero, si el intencionalista está en lo correcto, las personas que adviertan que están intentando formar una creencia en el autoengaño verían sus planes deshechos de inmediato.

El segundo candidato para explicar la falta de conciencia es mejor que el anterior en un respecto, observa Lazar: reconoce que la falta de conciencia de la intención de formar la creencia está relacionada con su contenido. De acuerdo con esta propuesta, el agente intencionalmente mantiene la intención fuera de la conciencia, solución que reitera el enfoque intencionalista del autoengaño. Sin embargo, sólo logra desplazar el problema del autoengaño un paso más allá: o esta intención de segundo orden es en sí misma inconsciente, en cuyo caso querremos una explicación de este hecho, o es consciente, en cuyo caso, una vez más, el proyecto parece autoderrotarse.

Huelga decir que los intencionalistas han respondido las críticas de los deflacionistas, así como han propuesto estrategias más efectivas en defensa del carácter intencional del autoengaño (Bermúdez, 2000; Nicholson, 2007). Como ocurre típicamente con los debates filosóficos, las preguntas relativas a la viabilidad del intencionalismo como explicación del autoengaño continúan abiertas. Volveremos a ellas en el apartado 4. de este capítulo.

3. El deflacionismo respecto del autoengaño

Como hemos adelantado, las posiciones deflacionistas con respecto al autoengaño, si bien más recientes que los modelos intencionalistas, han ido ganando terreno como explicación satisfactoria para el fenómeno que nos ocupa. Ejemplos de perspectivas deflacionistas pueden encontrarse en Johnston (1988), Mele (1997, 2001), Barnes (1999), Lazar (1999) y Nelkin (2002), entre otros. Si bien son variados los aportes de estos autores a la construcción de una perspectiva no intencionalista del autoengaño, me limitaré aquí, por razones de espacio, a la propuesta de Alfred Mele, quien es con seguridad el filósofo que más ha escrito sobre este fenómeno en las últimas dos o tres décadas (y, muy posiblemente, es el que más se ocupado del problema en sentido absoluto). En al menos dos libros[40] y en una serie de artículos que van por lo menos desde 1982 hasta la actualidad, Mele ha desarrollado una teoría compleja y sistemática del autoengaño, así como de otras formas de irracionalidad motivada. La teoría de Mele combina rasgos que la hacen una referencia obligada dentro de los estudios sobre el problema y, en particular, dentro del enfoque deflacionista; así como Davidson merece ser considerado el más destacado representante del intencionalismo, Mele podría ser reconocido con justicia como el más cabal defensor del deflacionismo. De la extensa bibliografía mencionada, tomaré como referencia el libro *Self-deception Unmasked*, que sintetiza lo esencial de su perspectiva sobre el problema.

Mele caracteriza la posición no intencionalista del siguiente modo:

> La perspectiva anti-agencial: ninguna creencia motivacionalmente sesgada es producida o protegida intencionalmente. En ningún caso de creencia motivacionalmente sesgada de que p tratamos de causar la adquisición o retención de p, o tratamos de hacer más fácil para nosotros mismos adquirir o retener la creencia (2001, p. 13).

[40] *Irrationality. An Essay on Akrasia, Self-Deception, and Self-Control* (1987) y *Self-deception Unmasked* (2001).

El enfoque adoptado por Mele, entonces, parte del rechazo (aunque matizado, como veremos) del supuesto de que el autoengaño puede ser adecuadamente comprendido sobre la base del engaño interpersonal; además, según su posición deflacionista, el autoengaño no es ni irresolublemente paradójico ni misterioso, y es explicable sin la asistencia de entidades mentales exóticas [*mental exotica*]. Asimismo, su teoría se fundamenta en una amplia base empírica, esto es, los hallazgos fácticos relativos a los procesos cognitivos y, en especial, sobre los factores que sesgan las inferencias. Si bien alguien cuyo interés en el autoengaño se limite a la posibilidad lógica o conceptual podría ver su posición, observa Mele, como un vaciamiento de la intriga conceptual del problema, considera que la fuente principal del interés relativo al autoengaño es la búsqueda de una comprensión de la conducta de seres humanos reales.

Pese a ser un deflacionista cabal, Mele no niega que pueda existir el autoengaño intencional consciente.[41] Considera, no obstante, que tal posibilidad está muy distante de los casos comunes de autoengaño, a los que denomina "variedad de jardín" (por ejemplo, el de la madre que cree falsamente que su hijo no consume drogas). La diferencia más visible es la naturaleza expresamente intencional del proyecto. Esto sugiere que en el intento de construir ejemplos hipotéticos que sean a la vez casos paradigmáticos de autoengaño y en los cuales los agentes se engañan intencionalmente a sí mismos, deberíamos imaginarnos, prosigue, que las intenciones de los agentes de engañarse a sí mismos deben estar de alguna forma ocultas para ellos. Mele no niega que las "intenciones ocultas" que estén actuando en casos de autoengaño puedan ser posibles. Esta posibilidad no le parece incoherente, aunque sí injustificada. Sin negar que las "intenciones ocultas" o "intentos ocultos" de autoengaño sean posibles, lo que debería preguntarse es, en su opinión, qué evidencia puede haber de que tales intenciones o intentos estén actuando en los casos de autoengaño de la variedad de jardín. ¿Pueden ser los datos explicables únicamente bajo la hipótesis de que tales intenciones o intentos están actuando en tales casos de autoengaño? La respuesta de Mele, que defenderá a lo largo del libro, es negativa. No niega, como vimos, que el autoengaño intencional, ya sea con intenciones abiertas y conscientes, ya sea con intenciones ocultas, sea posible; sin embargo, opta por intentar mostrar que los casos típicos de autoengaño pueden ser explicados sin recurrir a esos constructos problemáticos.

Mele, entonces, considera que los ejemplos estándar de autoengaño describen personas que creen falsamente en cosas que les gustaría que fueran verdaderas. Los

[41] Véase el ejemplo de Ike en el primer apartado del Capítulo 1. Cabe señalar, sin embargo, que no todos los autores están de acuerdo en que tal ejemplo ilustre un caso genuino de autoengaño intencional. Cfr. al respecto Audi (1997).

ejemplos de esta clase, pertenecientes a la "variedad de jardín" del autoengaño son reconocidos comúnmente como un fenómeno *motivado*. A este fenómeno puede dársele una interpretación consistente con el punto de vista antiagencial, que es más sutil y menos problemática que aquella basada en el modelo proporcionado por el engaño interpersonal. Para mostrar esto, Mele presenta una cantidad de procesos que pueden contribuir en una variedad de modos con el sesgo motivado de las creencias:

- Interpretación errónea negativa. Nuestro deseo de que *p* puede conducirnos a interpretar erróneamente datos que, en ausencia de tal deseo, consideraríamos adversos a *p*; por el contrario, el deseo de que *p* nos lleva a considerar esos datos como favorables a *p*.
- Interpretación errónea positiva. Nuestro deseo de que *p* puede conducirnos a interpretar erróneamente datos que, en ausencia de tal deseo, consideraríamos favorables a *p*; por el contrario, el deseo de que *p* nos lleva a considerar esos datos como desfavorables a *p*.
- Concentración/atención selectiva. Nuestro deseo de que *p* puede conducirnos a no prestar atención a evidencia que sería contraria a *p* y a enfocarnos en evidencia que sugiere que *p*.
- Recolección selectiva de la evidencia. Nuestro deseo de que *p* puede conducirnos tanto a soslayar evidencia fácilmente obtenible de que no *p* como a encontrar evidencia mucho menos accesible de que *p*.

En ninguno de estos casos, observa Mele, la persona sostiene la creencia verdadera de que no *p* y luego intencionalmente causa en sí misma la creencia de que *p*. Sin embargo, si damos por sentado que estos agentes hipotéticos adquieren creencias falsas y no justificadas en los modos descriptos, entonces tales ejemplos constituyen casos de autoengaño de la variedad de jardín. Sobre la base de lo anterior, considera, hay al menos una captación intuitiva de cómo un deseo de que *p* puede activar y sostener cada uno de los cuatro procesos anteriores y conducir a una creencia sesgada de que *p*. Sin embargo, y a excepción del proceso de concentración/atención selectiva (que puede ser intrínsecamente placentero), ¿cómo los deseos de que *p* provocan y sostienen los otros tres tipos de procesos, conduciendo de esa forma a la creencia sesgada de que *p*?

Se ha observado que la creencia, en los casos de autoengaño, es una creencia sesgada y motivada. Ahora bien, pueden existir fuentes de sesgo para las creencias que sean

no motivadas o "frías", y que se combinen con los procesos mencionados más arriba. Mele da tres ejemplos de tales fuentes:

- El carácter vívido [*vividness*] de la información. La vivacidad de un dato, para un individuo particular, es a menudo una función del interés individual, el carácter concreto del dato, su poder para "provocar imágenes" o su proximidad sensorial, temporal o espacial. Es más probable, entonces, que se preste más atención a los datos vívidos, y que sean más reconocidos y recordados que aquellos que no lo son.
- La heurística de la disponibilidad. Cuando formamos creencias acerca de la frecuencia, probabilidad o causas de un suceso, a menudo somos influenciados por la disponibilidad relativa de los objetos o sucesos, esto es, por su accesibilidad en los procesos de percepción, memoria o construcción basada en la imaginación.
- El sesgo de confirmación. Las personas que testean una hipótesis tienden más a menudo a buscar (en la memoria y en el mundo) casos que la confirmen que casos que la disconfirmen, y a reconocer más fácilmente los primeros, aun cuando la hipótesis sea sólo tentativa.

El aspecto que Mele destaca respecto de estos sesgos "fríos" o no motivados es que si bien pueden funcionar independientemente de la motivación, pueden también ser provocados y sostenidos por ésta en la producción de creencias particulares motivacionalmente sesgadas. Por ejemplo, la motivación puede incrementar la vivacidad o saliencia de ciertos datos; los datos que avalan la verdad de una hipótesis que uno quisiera que fuera cierta pueden resultar más vívidos o salientes dado el reconocimiento de la dirección en la que apuntan.

Mele considera que no hay ningún misterio en la manera en que el sesgo de confirmación, la heurística de la disponibilidad y la vivacidad de la información pueden contribuir con los dos tipos de interpretación errónea y las dos clases de selectividad que ha identificado. Los efectos predecibles de la motivación sobre el sesgo de confirmación incrementan la probabilidad de una interpretación sesgada y de una atención y recolección selectivas de la evidencia. Lo mismo es verdadero acerca de los efectos de la motivación sobre la disponibilidad y vivacidad de los datos.

Sin embargo, Mele advierte que hay mucho más para decir respecto de los procesos motivacionales que sesgan las creencias, y presenta con ese fin dos modelos de testeo de hipótesis cotidianas que han sido diseñados para acomodar la evidencia relativa a tales procesos. Uno de ellos es el presentado por James Friedrich, denominado "Primary Error

Detection and Minimization" (PEDMIN). Este autor argumenta que la detección y minimización de errores cruciales es el principio organizador central en el testeo de hipótesis comunes; las personas, según este modelo, son razonadores pragmáticos más interesados en minimizar errores cruciales que en llevar a cabo procesos de testeo tendientes a conocer la verdad. El segundo modelo es el presentado por Yaacov Trope y Akiva Liberman, modelo que es combinado con el de Friedrich sobre la base de sus semejanzas con éste; denomina a esta fusión "modelo FTL".

En el modelo de Friedrich, quienes testean hipótesis comunes proceden como lo hacen en parte a causa de su deseo de evitar o minimizar "errores costosos". Esto no implica que, en el proceso de testeo, quienes ponen a prueba tales hipótesis normalmente tengan el propósito consciente de minimizar ciertos errores, o que incluso inconscientemente traten de hacerlo. Friedrich conjetura que ciertas estrategias de testeo automático, derivadas de presiones evolutivas, pueden activarse toda vez que se detecta la posibilidad de un error significativamente perjudicial. Un examen adicional del sesgo de confirmación, observa Mele, resulta útil para esclarecer la conexión entre el modelo de testeo de hipótesis cotidianas y las discusión previa acerca de los mecanismos de sesgo "fríos" motivacionalmente desencadenados. Dada la tendencia implicada en el sesgo de confirmación, un deseo de que *p* –por ejemplo, de que nuestro hijo no está consumiendo drogas- puede, dependiendo de otros deseos y de la calidad de la evidencia que se posee, promover la adquisición o retención de una creencia sesgada de que *p* conduciéndonos a testear la hipótesis de que *p* (en tanto que opuesta a la hipótesis de que no *p*). El rol del deseo en la producción de la creencia sesgada será el de impulsar y ejecutar tal test.

Este sesgo tiene una aplicación bastante directa al fenómeno del autoengaño, como Mele muestra basándose en sugerencias de Friedrich. Este autor considera que, cuando se testean hipótesis que tienen un peso en la propia autoestima o autoimagen, un candidato principal a error primario será el considerar como verdadero algo que reduce nuestra autoestima o nos conduce a autocriticarnos de un modo erróneo. Estos costos son, por lo general, muy importantes, y conllevan inmediatamente un malestar psicológico. Cuando los costos asociados con un posible autoengaño son escasos (una preservación o incremento erróneos de la autoimagen), una revisión errónea a la baja de la autoimagen o la omisión en incrementarla de modo apropiado debería ser, consecuentemente, el error focal.

Sobre la base de lo expuesto, es posible comprender la caracterización de la adquisición autoengañosa de una creencia que proporciona Mele, en la cual se pueden apreciar las diferencias con los enfoques que asimilan el autoengaño al engaño

interpersonal. Mele sugiere que las condiciones que se exponen a continuación son conjuntamente suficientes para que alguien adquiera de manera autoengañosa una creencia de que *p*:

1. La creencia de que *p* que *S* adquiere es falsa
2. *S* trata datos relevantes, o al menos aparentemente relevantes, respecto del valor de verdad de *p*, de un modo motivacionalmente sesgado.
3. El tratamiento sesgado es una causa no desviada de que *S* adquiera la falsa creencia de que *p*.[42]
4. El cuerpo de datos poseídos por *S* en ese momento provee mayor garantía para ¬ *p* que para *p* (Mele, 2001, p. 50-51).

La caracterización anterior, reiteramos, no requiere de la intencionalidad, ni consciente ni inconsciente. Como dijimos, no la rechaza por completo, pero constituye un constructo del cual es posible prescindir sin pérdida explicativa. Notemos también que no requiere de ninguna división especial de la mente, más que lo que sea requerido para explicar procesos de sesgo cognitivo. Al no proponer intenciones ni divisiones *ad hoc* de la mente, no hay duda de que el modelo de Mele es más parsimonioso que sus rivales intencionalistas.

4. ¿Cuál explicación filosófica del autoengaño debe preferirse?

Como es fácil imaginar, los enfoques deflacionistas, al igual que los intencionalistas, no están exentos de objeciones. Deweese-Boyd (2012) señala tres cuestionamientos fundamentales al deflacionismo.[43] En primer lugar, la dificultad de esta posición para distinguir adecuadamente el autoengaño del pensamiento desiderativo; en segundo lugar el llamado "problema de la selectividad"; por último, la carencia, por parte de las perspectivas deflacionistas, de una explicación convincente de por qué típicamente quienes se

[42] El concepto de cadena causal desviada suele ser presentado mediante ejemplos, uno de los cuales es el siguiente: supóngase que, en un concierto, un espectador pretende alterar al director de la orquesta, su enemigo personal. Para lograr esto tiene la intención de toser ruidosamente, pero esta misma intención causa en él un estado de nerviosismo que lo hace toser. El espectador tenía la intención de toser, y esa intención causó que tosiera, pero esto último no ocurrió de modo intencional. Suele señalarse que este tipo de ejemplos supone una dificultad para las teorías causales de la acción intencional, ya que tales teorías consideran que, para que alguien actúe intencionalmente, es tanto necesario como suficiente que la persona en cuestión se encuentre en algún estado mental que represente el objetivo de su acción, y que el estar en tal estado cause que logre su objetivo.

[43] Se ha objetado también al modelo deflacionista de Mele el hecho de que no incluya como condición la tensión típica que el autoengaño suele traer aparejado, reflejada en fenómenos tales como sospechas o dudas recurrentes acerca de la creencia favorecida, o la evitación de ciertos pensamientos. Cfr. al respecto Audi (1997), y la respuesta de Mele (2001).

autoengañan son considerados responsables por su estado y sujetos a la crítica en muchos casos. En lo que sigue me concentraré en la primera y segunda objeción y pospondré el análisis de la tercera hasta el capítulo V.

La primera objeción no constituye, a mi modo de ver, una crítica de peso para el deflacionismo. Como hemos visto con cierto detalle en el capítulo I, la presencia de intencionalidad en el autoengaño, y no en el pensamiento desiderativo, ha sido considerada un criterio de demarcación entre ambos (Bermúdez, 2000). Sin embargo, como algunos de los mismos intencionalistas han admitido (Davidson, 1986), puede existir una continuidad entre ambos, e incluso que el segundo sea un ingrediente frecuente del primero. En cualquier caso, se trata de un problema meramente clasificatorio, esto es, si el autoengaño y el pensamiento desiderativo deberían ser claramente incluidos o no en categorías diferentes; el logro de una respuesta satisfactoria a esta pregunta no ha formado parte de las principales demandas filosóficas impuestas a las teorías sobre el autoengaño. La supuesta importancia que este problema pudiera tener palidece en comparación con las posibles dificultades explicativas del deflacionismo, que examinaremos a continuación.

El denominado "problema de la selectividad" es posiblemente la crítica más difícil de responder por los deflacionistas. Esta objeción es presentada por Bermúdez (2000) en los siguientes términos. El autoengaño es paradigmáticamente selectivo; en consecuencia, cualquier explicación de un caso determinado de autoengaño necesita explicar por qué los sesgos motivacionales actúan en *esa* situación particular. Sin embargo, el deseo de que *p* sea el caso es insuficiente para motivar los sesgos cognitivos en favor de la creencia de que *p*; existen toda clase de situaciones en las cuales, independientemente de cuán fuerte sea nuestro deseo de que *p* sea el caso, no experimentamos de ningún modo un sesgo en favor de la creencia de que *p*. La pregunta es entonces, concluye Bermúdez, cómo distinguir esas situaciones de otras en las cuales nuestros deseos resultan en un sesgo motivacional.

Mele (2001), intenta dar una respuesta al problema mediante el siguiente ejemplo. Un agente de la CIA ha sido acusado de traición. Si bien sus padres y sus colegas de su equipo de inteligencia tienen acceso a básicamente a la misma información acerca de su supuesto crimen, y todos desean que sea inocente, arriban a diferentes conclusiones. Mientras que los padres del agente concluyen que es inocente, sus colegas concluyen que es culpable. Mele pregunta cómo podría ser esto posible, dado que tanto unos como otros tienen acceso a la misma información, y, además, tienen deseos coincidentes respecto de su inocencia. Esto se debe, sugiere, a que el costo de creer falsamente que *p* (la inocencia del agente) es mucho más elevado para sus colegas que para sus padres; el deseo de sus colegas

de que el agente sea inocente es sobrepasado por su deseo de no ser traicionados, y sus umbrales de aceptación y rechazo difieren de manera acorde con los umbrales de aceptación y rechazo de los padres, quienes se sentirían reconfortados al creer falsamente que su hijo es inocente.

El análisis de costos y beneficios que lleva a cabo Mele, responde Bermúdez, provee de una explicación plausible respecto de lo que podría ocurrir en el caso del agente de inteligencia, pero no resuelve el problema de la selectividad. Este problema no consiste en responder por qué dos personas en situaciones similares adquieren diferentes creencias, sino que surge del hecho de que poseer el deseo de que *p* sea verdadero no es suficiente para generar sesgos cognitivos, incluso si todos los otros factores se mantienen igual. Simplemente no es el caso que, prosigue Bermúdez, toda vez que mi conjunto de motivaciones haga descender el umbral de aceptación de una hipótesis particular, terminaré aceptando esa hipótesis de modo autoengañoso; en consecuencia, el problema de la selectividad reaparece.

Por el contrario, los intencionalistas, señala Bermúdez, tienen una respuesta clara y directa al problema de la selectividad. La adquisición autoengañosa de que *p* requiere más que el simple deseo de que *p* sea el caso y de un umbral bajo para la aceptación y alto para el rechazo de la hipótesis de que *p*. Requiere por parte de quien se autoengaña la intención de producir la adquisición de la creencia de que *p*. El hecho de que el intencionalista pueda resolver de este modo el problema de la selectividad parece, concluye Bermúdez, al menos una razón para pensar que no es posible abandonar enteramente una aproximación intencionalista acerca del autoengaño.

Como suele ocurrir con los debates filosóficos, la polémica entre intencionalistas y deflacionistas no parece arrojar un ganador claro. Los intencionalistas (al menos la mayoría de los defensores del modelo "clásico") parecen no poder librarse por completo de las paradojas del autoengaño y de las complicaciones generadas por la necesidad de proponer divisiones *ad hoc* en la mente; los deflacionistas, por su parte, parecen no tener una respuesta satisfactoria al problema de la selectividad y, quizás, al planteado por la arraigada presunción relativa a la responsabilidad moral por el autoengaño. La sugerencia avanzada por Bermúdez quizás podría constituir una conclusión plausible para este debate: aunque los enfoque deflacionistas consigan erigirse finalmente en la mejor opción explicativa para el problema del autoengaño, es posible que no pueda prescindirse por completo de algún componente intencionalista.

Cabe agregar, aunque éste sea un aspecto menos visible del debate descripto, que las diferencias entre intencionalistas y deflacionistas parecen no limitarse a los rasgos que atribuyen al autoengaño; tienen también ciertas diferencias metodológicas. Los deflacionistas tienden en mayor medida a basarse en conocimientos proporcionados por diversas ciencias fácticas (lo que se pone claramente de manifiesto en el enfoque de Mele), mientras que los intencionalistas se han mantenido, por lo general, dentro de los límites del análisis conceptual más característicamente filosófico (Davidson sería, una vez más, el ejemplo paradigmático). Es plausible suponer que tales diferencias metodológicas (o, quizás mejor, pluralidad de enfoques), redundarán en una mayor complejidad en las teorías sobre el problema y, de algún modo, hacen más difícil prever los derroteros que adoptará el debate filosófico.

Capítulo III. Psicología, psicopatología, neurociencias. Perspectivas científicas sobre el autoengaño

Hemos anticipado en capítulos anteriores que, en las últimas décadas, los estudios filosóficos sobre el autoengaño comenzaron a coexistir con estudios empíricos provenientes de distintas ciencias. No nos consideramos en condiciones de realizar una enumeración exhaustiva de las disciplinas científicas que se han ocupado de examinar este problema; no obstante, sin duda las siguientes se destacarían dentro de ese listado: Psicología, biología evolucionista, neurología y neuropsicología y ciencias sociales (Sociología y Antropología). Ahora bien, como ya hemos comentado, la proliferación de estudios empíricos sobre el fenómeno no ha conducido a dos consecuencias que, *prima facie*, podrían resultar esperables. En primer lugar, la existencia de tales estudios no ha implicado la desaparición, ni mucho menos, de las reflexiones filosóficas sobre el problema; en segundo lugar, no se ha trazado ninguna línea nítida entre los análisis conceptuales característicos de la reflexión filosófica y el empleo de métodos empíricos típicos de la investigación científica. De hecho, y como hemos visto en el capítulo anterior, existen intentos explicativos del autoengaño que sin abandonar el campo de la Filosofía aspiran a integrar conocimiento empírico relativo a los procesos psíquicos que contribuyen a su generación. En este capítulo y en el siguiente pasaré revista a lo que entiendo son los principales avances en estos campos, con especial énfasis en los tres primeros. Dado que las perspectivas evolucionistas del autoengaño se han enfocado en cuestiones explicativas bastante diferentes (en particular, al problema relativo a las causas de su surgimiento en el proceso evolutivo) les dedicaremos un capítulo aparte.

1. *El autoengaño en la Psicología actual*

Si bien el interés por el fenómeno del autoengaño dentro de la Psicología ha tomado un impulso importante en las últimas décadas, existen antecedentes bastante lejanos en este respecto. En un artículo publicado en el año 1939 titulado "Mechanisms of Self-Deception", Else Frenkel-Brunswik describe una investigación tendiente a profundizar en el conocimiento de lo que denomina "ilusiones acerca del sí mismo"; en particular, al

descubrimiento de criterios formales que pudieran ser empleados en el diagnóstico de esas ilusiones.

Pese a estos antecedentes relativamente remotos, no ha sido sino aproximadamente hasta las últimas cuatro décadas que el problema del autoengaño pasó a constituir un terreno explorado de manera sistemática por los psicólogos. Como veremos, las investigaciones de Sackheim y Gur (1979) constituyeron un importante precedente para la proliferación posterior de estudios empíricos, no sólo tendientes a elucidar la naturaleza del fenómeno, sino también a examinar sus relaciones con diversos constructos psicológicos, entre los que se cuentan la responsividad al dolor (Jamner & Schwartz, 1986), el bienestar subjetivo (Erez, Johnson & Judge, 1995)[44], el autoconocimiento (Greenwald, 1996), la mentira (Ekman, 1997), la represión (Boag, 2007; Garssen, 2007), la sugestionabilidad y conformidad (Gudjonsson & Sigurdsson, 2004), el rendimiento (Johnson, Vincent & Ross, 1997), el aprendizaje, la meticulosidad y la autoeficacia (Lee & Klein, 2002), el optimismo cognitivo (Metcalfe, 1998), la defensividad y el optimismo (Norem, 2002), la religiosidad (Norris, Powell & Hickson 2007), loa delirios y la intencionalidad (Shean, 1993), la cooperación y depresión (Surbey, 2011), la psicoterapia (Westland & Shinebourne, 2009), e incluso la competición deportiva (Starek & Keating, 1991).

La revisión de estos estudios requeriría, como es fácil de observar, un volumen entero dedicado al tema. Me limitaré, en lo que sigue, a examinar algunos de ellos, ya sea por haber sido especialmente influyentes (como es el caso de los estudios tendientes a probar empíricamente la existencia del autoengaño) como por su interés teórico o práctico.

2. *Estudios empíricos tendientes a probar la existencia del autoengaño*

Habida cuenta de las múltiples discusiones no sólo acerca de la naturaleza del autoengaño sino también de las dudas sobre su misma existencia, no debería resultar extraño que algunos destacados estudios en Psicología estuvieran destinados a mostrar la realidad del fenómeno. Así como el artículo de Demos (1961) constituyó un texto pionero en los estudios filosóficos contemporáneos sobre el autoengaño, el artículo publicado por Ruben Gur y Harold Sackheim en 1979 y sugestivamente titulado "Self-Deception: A Concept in Search of a Phenomenon" es innegablemente un punto de referencia para la investigación empírica en Psicología sobre el tema en las últimas décadas.

[44] Este punto será objeto de un análisis detallado en el § 2 del capítulo V.

Gur y Sackheim comienzan por señalar que el concepto de autoengaño ya había recibido, al momento de la realización de sus estudios, mucha atención en textos literarios y filosóficos, y también en el ámbito de la Psicología. Respecto de este último ámbito mencionan a Meehl y Hathaway, por una parte y a Anastasi, por la otra, quienes argumentaron que el autoengaño contribuye más a la carencia de validez de los inventarios de personalidad por autoinforme que el engaño a otros; a Hilgard, quien afirmó que el autoengaño constituye una característica definitoria de todos los mecanismos de defensa; a Wason y Johnson-Laird, quienes invocaron el concepto de autoengaño como una posible explicación de la persistencia en mantener hipótesis frente a la disconfirmación; y a Murphy, quien relacionó el concepto de autoengaño con posibles interpretaciones de hallazgos experimentales sobre defensas perceptuales.

Pese a este considerable rol asignado al autoengaño en varias áreas de estudio, señalan Gur y Sackheim, hasta ese momento no habían existido intentos de demostrar que algún determinado conjunto de conductas se ajustara a lo que se entiende por autoengaño. Los filósofos, observan, han destacado frecuentemente que determinar lo que se entiende por autoengaño y mostrar que las personas se engañan a sí mismas tiene importantes implicaciones para la estructura de la conciencia; si bien dentro de la Psicología ha habido una larga tradición de suponer que las personas son necesariamente conscientes de sus cogniciones, al momento de la realización de la investigación ya habían aparecido estudios que arrojaban fuertes dudas acerca de la validez de tal concepción. En este sentido, la evidencia de que la falta de conciencia selectiva de la cognición puede ser motivada –esto es, la esencia de lo que implica el concepto de autoengaño- es particularmente controvertida.

Gur y Sackheim señalan que cuando se ha supuesto que las personas son necesariamente conscientes de sus cogniciones, el concepto de autoengaño resulta paradójico. Por otro lado, el rechazo del supuesto de que la cognición es necesariamente objeto de conciencia está implícito en el uso común del término autoengaño; también está implicado en este uso común que el individuo autoengañado, que sostiene creencias contradictorias, no hace esto último gratuitamente o por ignorancia: más bien, una creencia no accede a la conciencia con el objetivo de proveer alguna ganancia psicológica. De acuerdo con lo anterior, Gur y Sackheim proponen los siguientes criterios como condiciones necesarias y suficientes para el autoengaño: "1. el individuo sostiene dos creencias contradictorias (*p* y no *p*). 2. Esas dos creencias contradictorias son sostenidas simultáneamente. 3. El individuo no es consciente de que sostiene una de esas creencias. 4.

El acto que determina cual creencia es objeto de la conciencia y cual no es un acto motivado" (p. 149). Sobre la base de esta caracterización, Gur y Sackheim diseñaron dos estudios con el objetivo de testear si un fenómeno particular, la identificación errónea de las voces de otros y de uno mismo, se ajusta a estos criterios.

La experiencia de autoconfrontación proveyó a los autores de un probable candidato que se ajuste al concepto de autoengaño. Existe evidencia de que cuando los sujetos son confrontados con grabaciones de audio o video de sí mismos y de otros, el *feedback* proveniente del sí mismo estaba asociado con una considerable reactividad psicofisiológica y con cambios en el afecto y el autoconcepto. Las diferencias en la dirección y la intensidad de la autoconfrontación habían sido asociadas con manipulaciones experimentales de la autoestima y con diferencias individuales en la personalidad. Estos hallazgos condujeron a la conclusión de que los individuos que sostienen actitudes negativas hacia el sí mismo y puntúan alto en las mediciones de discrepancia cognitiva (definida por Gur y Sackheim como la medida en la cual los individuos sostienen actitudes y creencias discrepantes hacia sí mismos) encontrarán la autoconfrontación aversiva y tenderán a evitarla. Por otro lado, las personas con baja discrepancia cognitiva no encontrarán la autoconfrontación aversiva, y de hecho la buscarán.

La cuestión planteada para la investigación, señalan Gur y Sackheim, es si los individuos se encuentran en un estado de autoengaño cuando evitan la autoconfrontación por medio de la identificación errónea del sí mismo y cuando buscan la autoconfrontación identificando erróneamente a otros como sí mismos. Ambos errores involucran distorsiones de la realidad. Su hipótesis, en consecuencia, es que al menos algunas identificaciones erróneas del sí mismo y de otros son casos de autoengaño.

En un primer estudio se les administró a los sujetos una simple tarea de identificación en la cual se les pidió que indicaran si los estímulos de audio que escucharían eran grabaciones de sus propias voces o de voces de otras personas. Se registraron la respuesta galvánica de la piel (gsr) y el tiempo de reacción a cada voz, y los sujetos completaron una batería de inventarios de personalidad antes de la sesión experimental.

Los resultados del primer estudio, señalan Gur y Sackheim, apoyan la afirmación de que hay un fenómeno que se ajusta al concepto de autoengaño, dado que se encontró evidencia confirmatoria para los cuatro criterios para su adscripción. Cuando los sujetos identificaron erróneamente las voces propias y de otros mostraron que en algún nivel de procesamiento fue realizada una identificación correcta; sus niveles de gsr no difirieron de aquellos cuando identificaron correctamente las voces y, en consecuencia, sostuvieron

simultáneamente creencias contradictorias. Más aun, los sujetos no fueron conscientes del error en la identificación de la voz propia y en ocasiones tampoco fueron conscientes de la identificación incorrecta de las voces de otros. Finalmente, hubo evidencia inicial de que las identificaciones erróneas de las voces propias y de otros fueron motivadas.

Sin embargo, Gur y Sackheim señalan que los resultados del primer estudio son menos convincentes con respecto a la satisfacción del cuarto criterio, el referente a la motivación; la concepción de que las identificaciones erróneas de la propia voz y de voces de otros es un fenómeno motivado requiere justificación adicional. Además, la evidencia ofrecida a través del primer estudio es de naturaleza correlacional, y no responde completamente la pregunta de si las identificaciones erróneas tienen un propósito, un punto implícito en las atribuciones de motivación. En consecuencia, Gur y Sackheim diseñaron una segunda investigación destinada a responder a estos interrogantes. Observan que estudios previos mostraron que manipulaciones experimentales de la discrepancia cognitiva influencian la exposición selectiva al yo. Después de experiencias de retroalimentación negativa, la autoestima desciende y la confrontación con el yo se vuelve más aversiva; por otro lado, la retroalimentación positiva aumenta la autoestima y la autoconfrontación se vuelve menos aversiva. Si las identificaciones erróneas que hacen los sujetos son motivadas, podría esperarse que los sujetos que han experimentado un fracaso o que han visto disminuida su autoestima mostrarán mayores dificultades en realizar identificaciones de sí mismos; deberían ser más lentos en sus tiempos de reacción, mostrar menos certidumbre en realizar tales identificaciones y, en particular, cometer más errores por falsos negativos. Por otro lado, los sujetos que han experimentado éxitos deberían mostrar menos dificultad en identificar el yo y, más importante, deberían cometer un número mayor de errores por falsos positivos. En su análisis del segundo estudio, Gur y Sackheim sostienen que, cuando las personas identifican erróneamente las voces propias y de otros, están implicados en una conducta autoengañosa. En su opinión, parte de la atribución de autoengaño necesita demostrar que tales errores son motivados. También afirman que los individuos que están insatisfechos consigo mismos encuentran la confrontación con el sí mismo más aversiva. Por otro lado, los individuos que se tienen a sí mismos en alta estima no encuentran aversiva la autoconfrontación, y de hecho la buscan de manera narcisista. Agregan que si su afirmación relativa a que los errores en la identificación del sí mismo y de otros es motivada, entonces cabría esperar que las manipulaciones de la autoestima influencien la dificultad para hacer identificaciones del sí

mismo y de la tasa de los tipos de errores cometidos en la identificación. Los resultados del segundo estudio apoyaron esas predicciones.

Gur y Sackheim observan que su hallazgo de que algunas identificaciones erróneas del yo y de otros son casos de autoengaño demuestra colateralmente que debería atribuirse a la conciencia la propiedad de no transparencia selectiva y motivada. Afirman que, a veces, tal falta selectiva de conciencia puede estar determinada por demandas motivacionales. Dada esta evidencia inicial de que el autoengaño es un fenómeno experimentalmente real, surge un número adicional de preguntas, que conciernen a la naturaleza del autoengaño, a la posibilidad de diferencias individuales en la frecuencia de la conducta autoengañosa, y a los procesos y estructuras que subyacen al autoengaño. Una cuestión importante para la elucidación de la naturaleza del autoengaño es si los actos de autoengaño son respuestas típicas de individuos a, por ejemplo, estímulos amenazantes, o si son específicos y limitados a situaciones y estímulos.

Las tentativas de probar de modo empírico la existencia del autoengaño no se limitaron a los experimentos de Gur y Sackheim. Un segundo y destacado intento en esa dirección, algo posterior al de éstos, fue llevado a cabo por George Quattrone y Amos Tversky y descripto en su artículo "Causal versus Diagnostic Contingencies: On Self-Deception and on the Voter's Illusion" (1984). En este experimento los autores argumentan en favor de los cuatro criterios para el autoengaño postulados por Sackheim y Gur. Este estudio consistió en lo siguiente. Se pidió a treinta y ocho estudiantes que sumergieran un brazo en agua fría hasta un punto en el que ya no podían tolerarla. Después de períodos de 5 segundos los sujetos informaban su malestar a través de un número de 1 a 10; el número 10 expresaba el punto en el cual los sujetos no toleraban más el agua fría. Luego, los sujetos pedalearon en una bicicleta de ejercicio por un lapso de un minuto. A continuación, durante un breve período de descanso, se les dio a los sujetos una breve charla por la cual fueron inducidos a creer que las personas tienen uno entre dos posibles complejos cardiovasculares, designados como corazones de Tipo 1 y de Tipo 2. Los sujetos fueron asignados luego al azar a uno de dos grupos. A la mitad se les informó que un corazón Tipo 1 (no saludable) incrementaría la tolerancia al agua fría luego del ejercicio, mientras que en el caso del Tipo 2 (saludable) decrecería la tolerancia. A la mitad restante se les dijo lo opuesto, esto es, que la posesión de un corazón no saludable (Tipo I) disminuiría la tolerancia al agua fría, mientras que el poseer un corazón saludable (Tipo 2) conduciría a un incremento de la tolerancia. Por último, los participantes fueron sometidos nuevamente

a la prueba del frío, luego de lo cual se les preguntó, entre otras cosas, lo siguiente: ¿trató usted de alterar adrede el lapso en el que mantuvo el brazo en el agua luego del ejercicio? Como se predijo, los sujetos a los que se les había dicho que la disminución de la tolerancia era diagnóstica de un corazón sano mostraron una tolerancia significativamente menor en la segunda prueba, mientras que los sujetos a los que se les había dicho que el incremento en la tolerancia era indicador de salud mostraron una tolerancia significativamente mayor. Veintisiete de treinta y ocho sujetos mostraron el cambio predicho. Sólo nueve sujetos indicaron que trataron de modificar su tolerancia, y sólo dos de los 9 (22%) que admitieron esto infirieron que tenían un corazón de Tipo 2 (sano), mientras que veinte de los veintisiete "negadores" (69%) infirieron que tenían un corazón de Tipo 2.

Quattrone y Tversky consideran que la mayoría de los sujetos *trataron* de modificar su tolerancia en la segunda prueba. La mayoría de los sujetos negó esto, y los autores sostienen que muchos de los sujetos creyeron que no habían tratado de modificarla mientras simultáneamente creían que habían tratado de hacerlo. Argumentan también que esos sujetos no eran conscientes de que sostenían la última creencia, y que la "carencia de conciencia" es explicada por su deseo de aceptar el diagnóstico implicado por su conducta.

Las investigaciones empíricas descriptas presentan sin duda un innegable interés para el estudio del autoengaño, tanto por su diseño como por sus resultados, y las de Gur y Sackheim, además, han cosechado algunas defensas enfáticas.[45] Sin embargo, esas defensas no son compartidas por todos los autores interesados en el autoengaño; por el contrario, lejos han estado los experimentos de Gur y Sackheim y de Quattrone y Tversky de gozar de aceptación generalizada como pruebas empíricas de la existencia del fenómeno. La interpretación de los resultados experimentales, en particular, ha sido objeto de una serie de críticas que describiremos brevemente.

Douglas y Gibbins (1983), mediante el empleo de un procedimiento experimental muy similar al desarrollado por Gur y Sackheim, presentan y defienden una interpretación alternativa de los resultados de éstos. Si bien la respuesta galvánica de la piel es empleada para la detección de mentiras, señalan, esta reacción tiene lugar no sólo ante la presencia de mentiras, sino que es un indicador de cualquier tipo de excitación. En consecuencia, sugieren, la respuesta observada por Gur y Sackheim puede ocurrir toda vez que las personas se enfrenten a la tarea de identificar cualquier voz conocida. De esto se sigue que

[45] Trivers (2010) considera que los experimentos de Gur y Sackheim constituyen comprobaciones empíricas de la existencia del autoengaño; además, afirma que estos experimentos constituyeron un gran adelanto metodológico y que es de lamentar que no se haya desarrollado con posterioridad la línea de trabajo completamente nueva creada por estos autores.

si el problema que deben enfrentar los sujetos consiste en identificar alguna otra voz específica, se obtendrían resultados similares a los logrados por Gur y Sackheim. Con el objetivo de testear esta hipótesis modificaron el diseño original; adicionaron a la tarea de reconocer la propia voz una segunda tarea experimental, consistente en la identificación de voces de personas conocidas. Los resultados, afirman Douglas y Gibbins, sostienen su interpretación alternativa de los hallazgos de Gur y Sackheim: si bien tales hallazgos son replicables y confiables, el hecho de que el mismo patrón de respuestas haya sido logrado en los mismos sujetos en la condición de identificación de voces de otros muestra que la explicación sugerida por Gur y Sackheim, esto es, el autoengaño, no es la explicación correcta.[46]

Mele (2001), posteriormente, realiza una serie de críticas a la interpretación de los resultados tanto de los experimentos de Gur y Sackheim como de los realizados por Quattrone y Tversky. Respecto del estudio de los primeros, Mele señala que no es claro que las respuestas fisiológicas demuestren la existencia de una *creencia*. Plantea si, además de la creencia de que la voz escuchada no era suya, los sujetos también *creían* que era la suya, o meramente exhibían respuestas fisiológicas que a menudo acompañan la percepción de la propia voz; esto es, quizás se trataría meramente de una sensibilidad subdoxástica; a esta posibilidad se suman los resultados obtenidos por Douglas y Gibbins. Entonces, aun si las respuestas fisiológicas fueran indicativas de creencia, no establecerían que los sujetos sostienen creencias "contradictorias". Quizás, observa Mele, los sujetos creyeron que la voz no era la propia mientras también "creían" que era una voz familiar.

Respecto del estudio de Quattrone and Tversky, Mele señala que el estudio no ofrece ninguna evidencia directa de que los negadores sinceros estaban intentando cambiar su tolerancia; también señala que tampoco el supuesto de que creían esto es necesario para explicar su conducta, y presenta un caso hipotético en apoyo de esta tesis. Ana, una joven que desea conscientemente el amor de sus padres, cree que ellos la querrían si fuese una abogada exitosa. En consecuencia, se inscribe en la facultad de Derecho. Al hacerlo, está tratando inconscientemente de complacer a sus padres. Pero Ana no cree, a ningún nivel, que al inscribirse en la facultad de Derecho esté tratando de complacer a sus padres, ni cree que su deseo de obtener el amor de sus padres sea en modo alguno responsable de su decisión. Ana cree que su acción es debida únicamente a su deseo de ser abogada. Mele observa que en su descripción del caso simplemente ha *estipulado* que Ana carece de la

[46] Véase también la réplica de estos autores (Sackheim y Gur, 1985).

creencia en cuestión, pero lo que pretende mostrar con esto es que tal estipulación no transforma a la descripción en algo incoherente.

Si Mele está en lo correcto, ninguno de los experimentos descriptos logra su objetivo de probar empíricamente la existencia del autoengaño. El balance, sin embargo, no debería ser completamente negativo a la hora de albergar esperanzas de que intentos experimentales diferentes permitan sustentar empíricamente la realidad del fenómeno. A nuestro modo de ver, es posible aplicar algo similar a lo que afirmamos respecto de los intentos de mostrar que el autoengaño es imposible: no logran probar esto último sino que, en el mejor de los casos, consiguen mostrar que una determinada definición del autoengaño conduce a paradojas o problemas conceptuales aparentemente insolubles. Análogamente, tal vez podamos decir que los experimentos descriptos no logran mostrar la existencia del autoengaño si la caracterización de este fenómeno incluye las condiciones de coexistencia de creencias contradictorias y de intencionalidad. No obstante, como hemos visto en el capítulo II, no hay razones para pensar que tales requisitos son ineludibles para una caracterización rigurosa del autoengaño. Queda pendiente, hasta donde sabemos, el diseño de experimentos que intenten mostrar la existencia del fenómeno a partir de una definición que prescinda de tales condiciones.

3. *La función defensiva del autoengaño*

Existe una larga tradición en la Psicología, sobre la cual la influencia del psicoanálisis no ha sido menor, de atribuir a ciertas distorsiones motivadas en nuestra concepción de la realidad una función de defensa contra información externa o interna amenazante. La función de estas distorsiones sería, según esta tradición, la de proteger al individuo de padecimientos de los que, en principio, no tiene manera de librarse. Un destacado psicólogo social contemporáneo ha expresado esta idea en los siguientes términos:

> El concepto [de autoengaño] es central a la necesidad humana de un compromiso o, al menos, de un equilibrio entre dos motivaciones fundamentales. Las personas quieren información precisa acerca de su mundo y su complejidad; al mismo tiempo, *necesitan defenderse contra la información que destruiría las ideas sobre las cuales están construidas sus vidas* (Paulhus, 2007, p. 804. Cursivas nuestras).

Sin embargo, la idea de que el autoengaño implica alguna clase de beneficio para quien se encuentra en tal estado también ha sido defendida dentro del ámbito de la

Filosofía.[47] Davidson ha sido uno de los filósofos que han sugerido esta tesis de manera más clara:

> Normalmente, el autoengaño no representa un gran problema para el que lo practica, sino que, por el contrario, tiende a aligerarle, en parte, de la pesada carga de pensamientos dolorosos cuyas causas se hallan más allá de su control (Davidson, 1986, p. 99).

Estas optimistas afirmaciones respecto de la presunta función o las ventajas del autoengaño no parecen tener en cuenta el caso del autoengaño negativo. Davidson (1986), sin embargo, reconoce que la creencia generada por el autoengaño puede ser dolorosa; de este modo, una persona celosa puede encontrar en todas partes "pruebas" que confirman sus sospechas, así como el que anhela una vida privada puede llegar a creer que hay espías escondidos muy cerca de él. Este reconocimiento de que el pensamiento generado por el autoengaño puede producir malestar, no obstante, no conduce a Davidson a revisar la idea de que "normalmente" el autoengaño no representa un gran problema para quien lo practica.[48]

La posibilidad de atribuir una función defensiva al autoengaño requiere de varias aclaraciones. En primer lugar, no está demás observar que la atribución de una función no significa necesariamente que esta función sea la de *defender* al sujeto autoengañado de fuentes de malestar interno o externo. Como veremos en el capítulo V, el biólogo evolucionista Robert Trivers ha sostenido que la función principal del autoengaño es la de hacer más eficiente el engaño a otros, lo que implica indirectamente un incremento en la aptitud[49] de quien posee la capacidad para autoengañarse. En segundo lugar, el concepto mismo de *función* requiere de algunas consideraciones. Como se expondrá en el capítulo V en relación con los debates acerca de las explicaciones evolucionistas del autoengaño, un rasgo (sea orgánico o mental) no necesariamente debe su existencia a que haya cumplido, durante el proceso evolutivo, una función útil para la supervivencia y capacidad reproductiva de un organismo. Es posible que tal rasgo sea un subproducto estructural de otras estructuras o rasgos que sí han sido seleccionados por los beneficios adaptativos que confieren al organismo; en tal caso, no podría afirmarse que ese rasgo posee una *función*.

---

[47] "El autoengaño y la *akrasia* son generalizados debido a que están estrechamente conectados con procesos psicológicos sumamente útiles: los intentos por erradicar la *akrasia* y el autoengaño pondrían en riesgo tales procesos" (Oksenberg-Rorty, 1980, p. 919, citada en Bok, p. 932).

[48] Podría alegarse que el autoengaño negativo constituye una variante comparativamente mucho menos frecuente que la variante positiva; de este modo sería posible salvar la afirmación relativa a la reducción del malestar debida al autoengaño positivo.

[49] El término "aptitud", hace referencia, en el contexto de la biología evolucionista, a la capacidad reproductiva de un individuo.

En lo sucesivo nos mantendremos neutrales acerca de si el autoengaño existe debido a que posee alguna clase de función (sea cual fuere ésta); nos limitaremos a examinar si este proceso *puede* implicar alguna clase de beneficios para el funcionamiento psíquico de quien lo experimenta.

Hemos observado que la afirmación de que el autoengaño posee una función defensiva cuenta con importantes defensores tanto dentro del campo de la Filosofía como dentro del ámbito de la ciencia. Sin embargo, no han faltado detractores de la idea de que el autoengaño posee tal función, aun cuando no se pretenda explicar su existencia a partir de ella. Trivers (2010) señala que existe una analogía muy popular en Psicología (aunque cita una sola fuente en respaldo de la existencia de esta analogía "muy popular") consistente en observar que, así como nuestro cuerpo está bajo la constante amenaza de los parásitos, nuestro psiquismo está bajo la amenaza de factores que reducen la felicidad. En consecuencia, tenemos mecanismos de defensa psicológicos, así como tenemos mecanismos de defensa inmunológicos. Los primeros son para mantenernos saludables y libres de enfermedades, mientras que los segundos tienen la función de mantenernos felices. Se dice, observa Trivers, que esta analogía entre las respuestas inmunes y el grado de defensividad psicológica es "inusualmente apropiada" debido a que ambas comparten la característica de que tanto el exceso como el defecto son perjudiciales (por ejemplo, muy poco oxígeno es malo, y también lo es demasiado oxígeno). La *razonabilidad* (término muy elástico y de laxa definición) sería el factor clave para el funcionamiento del mecanismo psíquico: tenderíamos a buscar la perspectiva más favorable de la realidad, mientras que simultáneamente nos impulsaría a mantenernos razonablemente cercanos a los hechos. De este modo, observa Trivers, nos mantenemos a nosotros mismos felices en buena parte vía autoengaño: negación, proyección, disociación, entre otros mecanismos. Cocinamos los hechos, sesgamos la lógica, soslayamos las alternativas –en pocas palabras, nos mentimos a nosotros mismos-. Al mismo tiempo, un "centro de razonabilidad" determinaría, por medio de criterios desconocidos, la medida en que nos está permitido mantener nuestra felicidad por medio del autoengaño.

Trivers se pregunta por qué la evolución fue incapaz de producir un modo más sensato de regular una emoción importante como la felicidad. Luego de describir el antiquísimo y complejo funcionamiento "real" del sistema inmune, plantea la posible relación entre ese funcionamiento y la necesidad del yo de defenderse a sí mismo contra diversas "amenazas", y cuáles serían exactamente esas "amenazas". Si la amenaza reside en

las opiniones acerca de uno mismo, observa, la analogía falla: a diferencia de las amenazas a la vida constituidas por los parásitos, no parece que esta cuestión constituya algo de un valor apremiante para la supervivencia. Por otra parte, habría que explicar por qué adoptar algo tan dudoso como el autoengaño para resolver este problema: en el sistema inmune no existe ninguna mejora en su funcionamiento por medio de la mentira hacia sí mismo.

Por último, Trivers se plantea si es posible que la Psicología haya cometido durante tanto tiempo este "terrible error" simplemente debido a haber tomado en serio una aproximación a la vida que va desde lo interno hacia lo externo, una aproximación en la cual la introspección constituye la guía, y en la que se ha elegido a nuestros procesos de autoengaño como las unidades básicas para la construcción de nuestra teoría. La Psicología social, finaliza, se ha comprometido con una concepción defensiva del autoengaño, concepción que congenia con una autopercepción inflada: no nos mentimos a nosotros mismos para engañar mejor a otros (la que es, según él, la función principal del autoengaño), sino que nos mentimos a nosotros mismos para defendernos de ataques contra nuestra integridad personal y nuestra propia felicidad.

Es posible que, como sostiene Trivers, la analogía sea incorrecta y las similitudes entre el sistema inmunológico y el autoengaño no sean las supuestas por el autor al que cita. No obstante, parece plausible suponer que podrían existir elementos de juicio empíricos que indiquen la existencia de una dimensión defensiva para el autoengaño, más allá de que no sea ésta la función para la cual fue seleccionado por el proceso evolutivo. No hay, hasta donde sabemos, estudios sistemáticos tendientes a explorar empíricamente el rol defensivo del autoengaño. Sin embargo, quizás pueda encontrarse una indicación de la plausibilidad de este rol en áreas cercanamente relacionadas; en particular, quizás sea posible extraer alguna conclusión a partir de la exploración de un fenómeno en principio estrechamente emparentado con el autoengaño: la negación. Este concepto, como se verá enseguida, se encuentra envuelto en debates tan amplios como los que rodean el estudio del propio autoengaño.

El concepto de negación tiene una larga historia dentro del campo de la Psicología, y se han propuesto distintos enfoques teóricos y metodológicos para su estudio. Sin embargo, una parte no menor de los estudios sobre este fenómeno han estado influidos por las corrientes de inspiración psicodinámica y en particular por los trabajos de Freud. Dentro de la obra de este autor (y también en la de algunos de sus continuadores) no sólo es posible encontrar una de las primeras caracterizaciones del concepto de negación, sino

también una red de conceptos más o menos cercanos al autoengaño.[50] Sin pretensiones de exhaustividad, estos son los conceptos de negación, renegación, racionalización, escisión del yo e intelectualización.[51] Asimismo, el concepto de "ilusión", que Freud emplea en su escrito "El porvenir de una ilusión" (1927) es asimilable al concepto contemporáneo de pensamiento desiderativo.[52]

En su breve trabajo "La negación" (1925) Freud expone una explicación la negación como mecanismo de defensa y su importancia como origen de la función intelectual homónima. Señala que un contenido de representación o pensamiento reprimido pueden hacerse presentes en la conciencia sólo en el caso de que puedan ser negados. Tal posibilidad ocurre, por ejemplo, cuando en el curso del proceso psicoanalítico un paciente niega ante el analista que una persona que aparece en su sueño sea su madre; en tal caso, observa Freud, corresponde pensar que efectivamente lo es. En consecuencia, la negación posibilita adquirir conocimiento acerca de lo reprimido, constituye en sí misma una cancelación de la represión; no obstante, no es equivalente a una aceptación de aquello que se encuentra reprimido. La negación implica una separación entre la función intelectual y el proceso afectivo; este mecanismo sólo implica la eliminación de una de las consecuencias del proceso represivo, esto es, que el contenido de la representación no pueda arribar a la conciencia. Este proceso redunda en que se produzca una aceptación intelectual de lo reprimido, pero que persista lo esencial de la represión. Más aun, Freud señala que, en el curso del análisis, puede producirse una variante importante del proceso descripto. Es posible triunfar sobre la negación y lograr en el paciente una aceptación intelectual completa de lo reprimido, pese a lo cual el proceso represivo mismo no queda, de todos modos, cancelado. De este modo, la negación complementa el trabajo de la represión: si un contenido reprimido logra pasar a la conciencia, la negación (al igual que otros mecanismos defensivos) impide la plena aceptación consciente de ese contenido.

En la misma línea de pensamiento psicoanalítico, el trabajo de Freud fue continuado por su hija Anna (1936), quien concibió a la negación como un mecanismo de

[50] De hecho, un sistema teórico que parte del principio según el cual una parte sustancial de la vida mental está esencialmente oculta para el individuo, que esta parte oculta es totalmente determinante para sus pensamientos y conductas conscientes, y que el individuo yerra de manera sistemáticamente sesgada al intentar comprender los orígenes de éstos, podría ser considerado un sistema que hace del concepto de autoengaño su piedra basal; en otros términos, como ha afirmado Heinz Hartmann, "…una gran parte del psicoanálisis puede ser descripto como una teoría del autoengaño" (Citado en Nesse & Lloyd, 1992, p. 256).

[51] Para la relación entre diversos conceptos psicoanalíticos y la noción de autoengaño, véase Pataki (1997).

[52] En esta obra Freud postula que las creencias religiosas no son decantaciones de la experiencia ni resultados de una actividad intelectual, sino *ilusiones*, cumplimientos de los más antiguos, intensos y urgentes deseos de la humanidad. La categoría de "ilusión" freudiana puede asimilarse a la de pensamiento desiderativo, como señalamos, dado que Freud considera que la ilusión no constituye un mero error, y no es necesario que sea falsa, irrealizable o contradictoria con la realidad, sino que se trata de una creencia en cuya génesis cumple un rol esencial el *deseo*.

defensa primitivo, previo a la actuación de mecanismos más maduros. A estas concepciones iniciales de la negación como mecanismo de defensa se fueron adicionando con posterioridad no sólo otras perspectivas psicodinámicas, debidas a autores como Fenichel, Klein y Kernberg, sino también cognitivas. Puede resultar útil, dada esta proliferación teórica, examinar un análisis abarcativo de los distintos aspectos teóricos del concepto de negación.

En un artículo de revisión teórica e histórica sobre la negación, Livneh (2009a, 2009b), sintetiza un importante número de estudios empíricos y teóricos sobre este fenómeno, síntesis que puede servir de orientación para el objetivo de este apartado. Si bien el autor está especialmente interesado en la relación entre la negación, por una parte, y las enfermedades crónicas y la discapacidad, por la otra, algunas de sus conclusiones pueden ser válidas para una perspectiva más general del primer fenómeno. Por otro lado, hemos afirmado, en consonancia con algunos autores (Lazar, 1999) que el autoengaño surge en áreas de importancia vital para la persona. El inicio de una enfermedad discapacitante o crónica se ajusta perfectamente al tipo de episodios vitales que ponen en juego la estructura vital misma de la persona y que podrían generar distorsiones motivadas de la realidad.

Livneh destaca, desde el inicio, la naturaleza controvertida y compleja del fenómeno, que se refleja en las múltiples definiciones que se han dado de él. Tales definiciones incluyen, como hemos adelantado, perspectivas psicodinámicas (la negación como un mecanismo de defensa o una estrategia de afrontamiento), perceptuales (la negación entendida como un déficit neurológico/anosognosia) y cognitivas (la negación como un conjunto de procesos para manejar el estrés interno y externo). Livneh señala que, pese a esta multiplicidad de perspectivas teóricas, casi todas las definiciones enfatizan los siguientes elementos de la negación: a) se trata de una estrategia protectora que defiende al individuo de la ansiedad y la amenaza percibida; b) constituye un esfuerzo por evitar una realidad estresante; c) es un proceso inherentemente inconsciente, con limitada penetración consciente; y d) en el contexto de la enfermedad crónica y discapacidad, una tendencia a negar o repudiar la modificación de la realidad y el afecto doloroso asociados con el inicio y el subsecuente reconocimiento de las implicaciones a largo plazo del trastorno.

Respecto del punto que nos interesa en este apartado, las supuestas funciones de la negación, no están rodeadas de menos controversia que su caracterización. Sin embargo, Livneh, a partir de las concepciones de Lazarus y Vandereycken, observa que, en relación con la aparición de una enfermedad crónica o discapacidad, la negación sirve a diversas

funciones, como aliviar, minimizar o evitar el reconocimiento completo de la realidad dolorosa desencadenada por el inicio del suceso traumático, reducir el efecto o extensión de la ansiedad, combatir los sentimientos de impotencia e indefensión, lo que ayuda a mantener la esperanza y la motivación, y preservar la autoestima y una autoimagen positiva.

La idea de que la negación sirve a alguna función debe ser contrastada, a nuestro modo de ver, con otras consecuencias de su acción. En este sentido, puede resultar útil examinar los costos y beneficios de este proceso. Livneh (2009) observa que la negación está asociada con resultados tanto positivos (beneficios) como negativos (costos), y es el contexto, duración y su uso particular lo que determina su valor. Siguiendo una vez más a Lazarus, enumera las ventajas y desventajas de este proceso.

La negación tiene consecuencias potencialmente nocivas si la acción directa es necesaria para modificar interacciones dañinas entre la persona y el entorno, si su empleo impide el dominio de ciertas situaciones estresantes (especialmente aquellas situaciones que se encuentran repetidamente en la vida), si impide la asimilación de importantes consecuencias aprendidas que están asociadas con situaciones y sucesos estresantes y si afecta de manera adversa a las relaciones interpersonales mediante un incremento de la distancia entre el individuo y su entorno social.

Livneh señala además diversas situaciones positivas que pueden estar también asociadas con la negación. Estas incluyen: a) la negación que sigue de modo inmediato al inicio de una enfermedad crónica o discapacidad (CID, en lo sucesivo) o su diagnóstico puede hacer posible una asimilación gradual de la información dolorosa y la aceptación de la situación estresante, evitando la desintegración psicológica y permitiendo un tiempo adicional para movilizar recursos para la resolución de problemas más adaptativos; b) en situaciones en las que la acción directa para confrontar la crisis (el inicio de la CID) es inconducente, la negación no tiene efectos negativos y puede ser de valor como un mecanismo más confiable de reducción del estrés; c) permite al individuo mantener su autoestima y esperanza; y d) tiene la ventaja de no requerir que el individuo manipule directamente su entorno y los estímulos que producen estrés; la energía, entonces, está mayormente dirigida hacia el control de las percepciones personales de la CID y sus consecuencias potenciales.

En suma, considera Livneh, puede concluirse que la negación es perjudicial cuando es empleada como una estrategia extrema, generalizada y continua para defenderse de situaciones estresantes generadas por el inicio de la enfermedad crónica y la discapacidad;

sin embargo, cuando la negación es parcial, de duración limitada y mínima en alcance, carece de sus elementos perniciosos y puede incluso servir a un rol beneficioso en la reducción del estrés disparado por tal tipo de enfermedad.

La negación, en síntesis, parece estar asociada tanto a beneficios como a costos.

Evidencia adicional de las posibles consecuencias defensivas de la negación es aportada por Baumeister, Dale y Sommer (1998), en una revisión del concepto psicoanalítico de mecanismos de defensa a la luz de los conocimientos provistos por la Psicología social y de la personalidad contemporáneas. Tal revisión parte de un examen del presunto objeto de la defensa: a diferencia de la formulación freudiana según la cual la función de los mecanismos defensivos es defender al yo de impulsos internos inaceptables, ya sea sexuales o agresivos, los autores observan que las modernas Psicología social y de la personalidad no aceptan en general la concepción según la cual la personalidad está fuertemente basada en esfuerzos para enmascarar los propios impulsos sexuales y agresivos. Sin embargo, admiten que subsiste fuertemente la necesidad de mecanismos de defensa; resulta mucho más consistente con las modernas formulaciones de aquellas disciplinas la afirmación de que la función de los mecanismos de defensa es en realidad la protección de la autoestima. En este sentido, muchos investigadores en las áreas de la Psicología social y de la personalidad reconocen que las personas defienden su autoconcepto contra las amenazas. Incluso observan que el objetivo de protección de la autoestima puede no contradecir las concepciones de Freud, sino que puede concebirse como un cambio de énfasis.

El método empleado por los autores es el de revisar una lista de siete mecanismos de defensa (formación reactiva, proyección, desplazamiento, anulación, aislamiento, sublimación y negación) que Freud describió y que han sido influyentes en el trabajo subsecuente. Nos extenderemos aquí únicamente sobre la negación, nuestro interés en este apartado.

La negación presenta, en opinión de estos autores, varias dificultades teóricas. En primer lugar, la distinción entre negación y represión es difusa y difícil de articular de un modo significativo. Para sus fines, consideran que la negación puede ser suficientemente caracterizada como *el rechazo a enfrentar ciertos hechos*. Por otra parte, observan que la negación puede ser comprendida de una manera amplia o de una manera estrecha. La manera amplia, sostenida por algunos teóricos, concibe a la negación como una categoría de mecanismos de defensa más que como una defensa singular; en este caso podría incluir, entre otros, a

fenómenos tales como defensas perceptivas, construcción de fantasías personales, negación, minimización, maximización y el ridículo.

Una primera conclusión es la ausencia de evidencia de que las personas rechacen sistemáticamente aceptar la realidad física de sucesos reales, especialmente cuando son confrontadas con pruebas palpables. Por otro lado, hay evidencia abundante de que las personas rechazan implicaciones e interpretaciones que encuentran amenazantes.

La más común de las formas de negación es, probablemente, el desechar las respuestas a fracasos u otras formas de retroalimentación negativa. Cuando las personas reciben evaluaciones negativas, a menudo rechazan sus implicaciones, en lugar de incorporarlas a sus autoconceptos. Realizar atribuciones externas del fracaso, como apuntar a la mala suerte o a la dificultad de la tarea, es uno de los patrones comunes y bien documentados para denegar las implicaciones de aquél. Una variante en la respuesta de atribución externa es la búsqueda de defectos o deficiencias en el método de evaluación que conduce a la retroalimentación negativa; otra variante consiste en desechar tal retroalimentación como algo motivado por el prejuicio. Todas estas variantes son apoyadas por la investigación empírica.

Un segundo ámbito en el que se ha encontrado alguna evidencia comparable con la hallada respecto de las amenazas a la autoestima es el de la Psicología de la salud. La noción de que las personas emplean la negación en respuesta a las amenazas relacionadas con la salud se remonta a los trabajos de Elizabeth Kübler-Ross de 1969, que conciben a la negación como una de las "etapas" o tipos de respuesta ante el conocimiento de que una enfermedad será fatal. Trabajos posteriores han demostrado la existencia de mecanismos de negación ante amenazas menos extremas. Por ejemplo, se ha mostrado que los usuarios de cafeína tienden a criticar (selectivamente) y desestimar evidencia de un nexo entre el consumo de esa sustancia y la enfermedad fibroquística, mientras que los no usuarios no muestran tal sesgo.

Una esfera diferente en la cual se ha encontrado evidencia de negación, señalan Baumeister et al, es en la relativa a las proyecciones de las personas acerca de su futuro. Existe evidencia de que las personas tienden a ser optimistas de una manera no realista, patrón que la investigación posterior ha confirmado repetidamente. La persona promedio cree que es menos probable para ella que para la persona promedio llegar a sufrir diversas desgracias, tales como fracasos en su carrera, enfermedades debilitantes, o accidentes discapacitantes. Asimismo, se ha acuñado el término "ilusión de invulnerabilidad única" para hacer referencia a la creencia de la persona promedio de que no le ocurrirán cosas

malas. Esta ilusión no permanece como una conjetura abstracta o vaga, sino que se manifiesta en comportamientos tales como conductas sexuales promiscuas sin protección.

Si bien observan que la negación puede ser en ocasiones desadaptativa (por ejemplo, se has observado respuestas desfavorables relacionadas con el empleo de la negación en estudios sobre respuestas ante exámenes y en pacientes oncológicos), en otras circunstancias puede ser bastante adaptativa. Mencionan aquí el trabajo de Taylor y Brown (1988) relativo a la forma en que la salud mental y la elevada autoestima fueron asociados con patrones de procesamiento sesgados que negaban la responsabilidad personal por malos resultados, mientras reclamaban el crédito por los buenos. Hay también sugerencias de que la negación podría ser especialmente adaptativa a continuación de un trauma, porque permitiría que los procesos de reinterpretación procedan gradualmente.

Por último, existe alguna evidencia para formas más elementales de la negación, como la percepción (se ha sugerido que algunas personas tienen defensas que trabajan minimizando el reconocimiento de estímulos amenazantes) y la memoria (las personas no depresivas recuerdan más los adjetivos favorables al yo que los desfavorables y, en general, las personas rememoran la retroalimentación acerca de sus rasgos positivos mejor que al relativa a sus fracasos y defectos).

En vista de la heterogeneidad de los hallazgos descriptos, concluyen los autores, un marco conceptual más diferenciado podría ser útil, lo que podría incluir que el concepto de negación fuera reemplazado por varios mecanismos más específicos y particulares.

En ausencia de una caracterización precisa del autoengaño, y también de la negación, no resulta posible una simple extrapolación al primero de los hallazgos relativos a la función defensiva de la segunda. Sin embargo, y sin perder de vista las diferencias entre autoengaño y negación, parece indudable la idea de que ambas pertenecen a la categoría de distorsiones motivadas en nuestra percepción de la realidad. Habida cuenta de los elementos de juicio favorables a la función defensiva de la negación, parece razonable suponer que el autoengaño posee, al menos en ocasiones, una función similar. Hablar de las "ocasiones" en las que el autoengaño desempeña un rol defensivo se debe no sólo, como se examina en el apartado que sigue, a que en ciertas condiciones puede ser disfuncional, sino también a tener en cuenta los casos de autoengaño negativo, contraparte que parece estar ausente en la negación. Parece ser claro que, tanto en el caso del autoengaño como en el de la negación, ciertas variantes de estos fenómenos resultan disfuncionales, o al menos no desempeñan la función defensiva que parecen tener en algunas oportunidades. No podría

decirse que los casos disfuncionales *prueben* que no se trata de una función. Pese a las objeciones de Trivers (cuyas tesis respecto del autoengaño son sumamente discutibles, como veremos en el capítulo IV), la analogía con algunos sistemas biológicos *puede* ser apropiada. Al igual que lo que ocurre con el funcionamiento de ciertos sistemas orgánicos, por ejemplo, el sistema endócrino, cierto grado de negación o autoengaño pueden contribuir con el bienestar; por el contrario, un grado demasiado elevado, como ocurre, por ejemplo, con la secreción en exceso de una hormona, se torna disfuncional. Cabe hacer dos consideraciones respecto de esta posible analogía. En primer lugar, su postulación no equivale a una prueba del carácter contingentemente benéfico del autoengaño y la negación, y tampoco prueba de la existencia de la analogía; se deben proveer elementos de juicio fácticos que sustenten ambas tesis. En el apartado 2 del capítulo V nos ocuparemos de analizar algunas posibles pruebas en favor de la primera tesis. En segundo lugar, la afirmación de que tanto el autoengaño como la negación poseen funciones defensivas no significa que puedan ser explicados por la existencia de esta función; esto es, no implica que se trate de rasgos mentales seleccionados evolutivamente por tal contribución. Nos ocuparemos de este problema en el capítulo IV.

4. *Autoengaño y psicopatología*

4.1. Delirio, anosognosia, confabulación

> - Eres muy cruel. Pero tienes la sartén por el mango (…) Me quieres. Estás enamorado de mí, y yo no tengo más remedio que corresponderte. (…) No comprendo por qué me has elegido a mí. Lo único que sé es que ahora yo también te quiero, y que hay una razón, una finalidad (…)
> - No le conozco, no sé dónde vive, ni a qué se dedica, ni quién es usted. Y no tengo especiales deseos de saberlo. Sólo le he visto una vez y puedo asegurarle que no albergo sentimientos de ninguna clase hacia usted (…)
> - No hagas esto, por favor… No tiene por qué ser así, de verdad. No debes hacerme esto a mí (…) Yo no puedo controlar mis sentimientos tan bien como tú. Sé que eso te da poder sobre mí, pero no puedo hacer nada para evitarlo
> - No tengo ningún sentimiento que controlar, créame (…)
> - Si es una broma, ya está bien. Nos está haciendo daño a los dos.

El fragmento que encabeza este apartado pertenece a la novela *Amor perdurable*, de Ian McEwan, en la cual uno de sus personajes centrales sufre lo que comúnmente se denomina delirio erotomaníaco (o Síndrome de De Clerambault): quien lo padece está convencido contra toda evidencia de que alguien (a menudo una persona de un estatus social superior) está enamorado de él. Aunque se ha sugerido que se trata de un ejemplo de autoengaño,[53] es más común considerar que éste y otros casos similares no constituyen ejemplos de este fenómeno; por el contrario, se los suele incluir dentro de la categoría de los trastornos delirantes. Sin embargo, y dado que ambos conjuntos de fenómenos parecen compartir una característica fundamental, esto es, la existencia de una o más creencias firmemente sostenidas pese a la evidencia en contrario a disposición del agente, surge naturalmente la pregunta relativa a las relaciones entre ellos, cuestión de la que nos ocuparemos en lo que sigue.

La idea de que existen diferencias entre los fenómenos "normales" de sesgo de las creencias y fenómenos extremos o patológicos de distorsión tiene antecedentes, como en virtualmente todas las cuestiones relativas al autoengaño, en los escritos filosóficos. Demos (1960) señala la existencia de lo que normalmente se denomina "delirios" [*delusions*], casos en los cuales la persona que los experimenta no es encontrada responsable. El supuesto es que la causa es externa a la propia persona; este sería el caso, por ejemplo, de una persona alcoholizada que manifiesta delirios de grandeza. En tales casos, observa, nos inclinamos a decir que la operación de la razón ha sido obstruida, no que es intrínsecamente deficiente. Una importante diferencia que Demos encuentra entre delirio y autoengaño es que la persona que padece del primero no experimenta conflicto, como sí ocurre en el segundo caso.[54]

Ahora bien, el delirio *stricto sensu* no es el único fenómeno patológico que, a primera vista, podría tener alguna cercanía con el autoengaño; por el contrario, el espectro de fenómenos patológicos caracterizados fundamentalmente por la presencia de creencias ya sea falsas (implausibles o no), ya sea radicalmente incompatibles con la evidencia a disposición del agente incluye trastornos de clases muy diferentes. Estos trastornos han sido ordenados según diversos criterios, entre ellos su contenido (o algún aspecto o dimensión determinada de éste), la firmeza o tenacidad con la que son defendidas, y su etiología (la presencia o no de alguna clase de déficit, el rol de la motivación en su producción, y otras posibles causas). Radden (2011) señala que ya sea que los trastornos

[53] Cfr. Trivers, 2000.

[54] La idea del carácter egosintónico de las creencias falsas como rasgo típico de las creencias patológicas será, como veremos, retomada posteriormente.

delirantes compartan o no algún origen en una disfunción cerebral, como fenómenos clínicos parecen constituir una colección heterogénea, unida no tanto por la presencia de rasgos esenciales comunes a todos como por compartir normas sociales relativas al alivio del sufrimiento y la disfunción, a la evitación del dolor, y a la racionalidad.[55] Bortolotti y Cox (2009) han observado algo similar acerca de las dificultades en lograr una definición abarcativa de la confabulación, fenómeno caracterizado igualmente por la presencia de creencias que no se ajustan a los hechos o están pobremente apoyadas por la evidencia disponible. Dadas estas manifiestas dificultades en encontrar un sistema clasificatorio que goce de algo parecido al consenso respecto de los trastornos que nos interesan, adoptaremos la postura de considerar conjuntamente diversos fenómenos que suelen diferenciarse, aunque no nítidamente, tanto dentro del ámbito clínico como del psicopatológico.

Comencemos por el concepto de *delirio*. Dado que la naturaleza de los delirios, pese a que han sido reconocidos como entidades clínicas hace mucho tiempo, sigue siendo objeto de controversias, convendrá comenzar con una definición estándar como punto de partida. El Manual Diagnóstico y Estadístico de los Trastornos Mentales, Quinta Edición (DSM 5, en lo sucesivo), propone la siguiente definición de delirio [*delusion*]:

> Los delirios son creencias inamovibles no susceptibles de modificación a la luz de evidencia que se encuentre en conflicto con ellas. Sus contenidos pueden incluir una variedad de temas (por ejemplo, persecutorias, referenciales, somáticas, religiosas, grandiosas). (…) Los delirios son considerados *extraños* [*bizarre*] si son claramente implausibles y no comprensibles para personas de la misma cultura y no derivan de experiencias comunes de la vida. Un ejemplo de delirio extraño es la creencia de que una fuerza externa ha removido los órganos de la persona y los ha reemplazado por los de alguien más, sin dejar heridas o cicatrices. Un ejemplo de delirio no extraño es la creencia de que uno se encuentra bajo la vigilancia de la policía, pese a la ausencia de evidencia convincente. (…) En ocasiones es difícil trazar una distinción entre un delirio y una idea fuertemente sostenida, y depende en parte del grado de convicción con el que se sostiene la creencia, pese a evidencia clara o razonablemente contradictoria respecto de su veracidad (p. 87).

La categoría anterior incluye distintos subtipos, caracterizados por el contenido de la ideación delirante: extraña (idea que implica un fenómeno que la cultura a la que pertenece la persona consideraría físicamente imposible); celos delirantes (la idea de que el compañero sexual es infiel, también denominado "síndrome de Otelo"); erotomaníaca (idea de que otra persona, usualmente de estatus superior, está enamorada del sujeto); de

[55] Radden sugiere también otros fenómenos que considera dentro del espectro de los trastornos delirantes, como las percepciones delirantes y los casos de *folies à deux* y trastornos delirantes compartidos. Los fenómenos examinados en el presente apartado, empero, nos parecen suficientes para explorar las relaciones entre el autoengaño y esta clase de fenómenos patológicos.

grandeza (idea de valor, poder, conocimientos o identidad exagerados, o de una relación especial con una deidad o una persona célebre); de ser controlado (idea de que ciertos sentimientos, impulsos o actos son experimentados como si estuvieran bajo el control de alguna fuerza externa más que bajo el propio control); de referencia (idea cuya temática consiste en que ciertos sucesos, objetos o personas del ambiente inmediato del sujeto adoptan una significación particular e inusual, que suele ser de naturaleza negativa o peyorativa, pero también puede ser de grandiosidad); persecutoria (idea cuyo tema central consiste en que el sujeto, o alguien cercano a él, está siendo atacado, acosado, engañado, perseguido o se conspira contra él); somática (idea cuyo principal contenido pertenece a la apariencia o al funcionamiento del propio cuerpo); difusión del pensamiento (idea de que los propios pensamientos están siendo difundidos en alta voz de modo que pueden ser percibidos por otros); inserción del pensamiento (idea de que ciertos pensamientos de la persona no son propios, sino que más bien son insertados en su mente) (pp. 819-820).

Ahora bien, los anteriores no son los únicos tipos de delirio que aparecen regularmente en la literatura. Cabe mencionar, entre otros, el delirio de Cotard (en el cual la persona cree que está muerta), el de Otelo revertido (la creencia delirante en la fidelidad de la pareja pese a la evidencia en contrario), y los llamados "síndromes de identificación errónea delirante". Estos son el síndrome de Capgras, el síndrome de Frégoli, el síndrome de intermetamorfosis y el síndrome de los dobles subjetivos. El Síndrome de Capgras se caracteriza por la creencia inamovible del paciente que las personas cercanas a él (como su pareja) han sido sustituidas por dobles; el Síndrome de Frégoli se caracteriza por una identificación delirante de familiares en diversos extraños; el síndrome de intermetamorfosis, que consiste en la convicción delirante de que personas cercanas a él modifican su aspecto a voluntad intercambiándose por otros; el síndrome de dobles subjetivos tiene lugar cuando se cree que un extraño se ha transformado física pero no psicológicamente en el propio paciente, es decir, cuando el paciente cree que hay un doble de sí mismo que actúa independientemente de él.

Sin embargo, los delirios no son los únicos fenómenos patológicos que *prima facie* presentan similitudes con el autoengaño: la anosognosia y la confabulación son, sin duda, muy buenos candidatos para el intento de establecer vínculos entre fenómenos normales y patológicos de distorsión de las creencias.

En algunas caracterizaciones la confabulación aparece asociada a la presencia de algún tipo de deficiencia o patología. Hirstein (2000) señala que cuando se formula a una persona que confabula una pregunta que guarda alguna relación con el déficit que

experimenta, en vez de reconocerlo proporciona una respuesta falsa o irrelevante, como si pretendiera encubrirlo.[56] Otras definiciones parecen tener un alcance más amplio. Bortolotti & Cox (2009), consideran que, de modo típico, las personas confabulan cuando hacen afirmaciones o narran historias que o bien son erróneas o bien son mal sustentadas por la evidencia disponible. Quienes confabulan no mienten; por el contrario, creen sinceramente las historias que cuentan, que pueden ser sostenidas con convicción y en presencia de contraargumentos. Esta caracterización de la confabulación ha generado controversias debido a su excesiva vaguedad, a menos que pueda ser cualificada de acuerdo con las preferencias que se tengan con respecto al modo en que surgen las confabulaciones. Agregan que comúnmente se piensa que la confabulación ocurre en conjunción con la amnesia, y que hay cuatro ideas o explicaciones de su ocurrencia: 1) como un fenómeno que sucede a una falla en la recuperación de información almacenada en la memoria; de este modo, la confabulación constituiría un intento de completar los vacíos en aquella; 2) la confabulación ocurriría cuando sucesos previamente experimentados son mal ubicados en tiempo o espacio; 3) la confabulación involucra una conjunción inapropiada de sucesos independientes; 4) la confabulación es el resultado de sucesos que nunca han sido experimentados, pero que son erróneamente considerados reales.

Ahora bien, la ausencia de definiciones estrictas de los fenómenos mencionados no ha impedido el intento de establecer relaciones con el autoengaño. Reseñaremos brevemente dos perspectivas relativas a los puntos de contacto entre ellos.[57]

Bortolotti (2013) destaca la ausencia de un consenso respecto del posible solapamiento entre el delirio y el autoengaño. Mientras que el autoengaño ha sido caracterizado tradicionalmente como un fenómeno producido por factores motivacionales, el delirio ha sido explicado primariamente en términos neurobiológicos, y las teorías sobre este fenómeno implican una referencia a deterioros perceptivos y cognitivos. Sin embargo, agrega, los factores motivacionales pueden jugar un rol importante en la explicación de

[56] Hay distintas patologías cerebrales que pueden causar la confabulación, como tumores o traumatismos en el cerebro, aneurismas y demencia. Según Hirstein (2000), los tres síndromes que más a menudo producen confabulación son el del cerebro dividido, la anosognosia y el síndrome de Korsakoff. El síndrome del cerebro dividido es debido a la comisurotomía, una lesión inducida quirúrgicamente por la cual el cuerpo calloso es seccionado. "Anosognosia" significa falta de conciencia de la enfermedad, y es exhibida por muchos tipos de pacientes neurológicos, pero ocurre con mayor frecuencia a continuación de un daño producido por un derrame en la corteza parietal del hemisferio derecho. El síndrome de Korsakoff es una forma de amnesia mayoritariamente causada por una vida caracterizada por un consumo importante de alcohol. El déficit de la memoria afecta a la memoria episódica, un sistema que almacena la información acerca de episodios autobiográficos, no la memoria semántica, nuestro conocimiento de conceptos, incluyendo el significado de las palabras.

[57] El lector interesado en ampliar este punto puede encontrar provechosa la lectura de Mele (2007), artículo en el cual este autor examina posibles conexiones entre el autoengaño y tres tipos de delirio.

algunos delirios, por ejemplo, mediante la determinación parcial del contenido específico del estado delirante. De este modo, una visión plausible es que el autoengaño y el delirio son fenómenos distintos que pueden solaparse en algunas circunstancias, y existen tres posibles perspectivas acerca de este solapamiento.

La primera perspectiva sostiene que el solapamiento entre delirio y autoengaño ocurre debido a que ambos involucran un tratamiento motivacionalmente sesgado de la evidencia. Si se acuerda con los deflacionistas en que tal tratamiento es un rasgo clave del autoengaño, entonces puede decirse que las personas que deliran se encuentran autoengañadas en tanto buscan evidencia, o tratan la que tienen a su disposición, de un modo motivacionalmente sesgado. No obstante, esto no parece ser lo que generalmente ocurre, por lo que resulta útil distinguir entre diferentes tipos de delirio. Algunos delirios de identificación errónea, de acuerdo con enfoques neuropsicológicos, no parecen ser similares al autoengaño, dado que no hay ningún rol fundamental para los sesgos motivacionales en la explicación de cómo la persona adquirió o retuvo la creencia delirante. Un análisis diferente podría ser adecuado para otros delirios, como los delirios celotípicos o de persecución.

La segunda perspectiva es que algunos delirios son casos extremos de autoengaño y poseen una función protectora y adaptativa. Bortolotti sugiere como ejemplo de esto las hipótesis propuestas por Ramachandran (1996) acerca de la posible función defensiva de ciertas ideas delirantes en un caso de anosognosia y somatoparafrenia. Nos detendremos algo más en esta concepción.[58]

Ramachandran observa que la anosognosia es usualmente observada en pacientes que han tenido un ataque en el hemisferio derecho del cerebro, el que da como resultado una parálisis en el lado izquierdo del cuerpo (aclara en nota al pie que restringe el uso del término "anosognosia" a la negación de la hemiplejia, no al uso genérico que indica negación de otros tipos de déficits). Algunos de los pacientes con este trastorno niegan vehementemente la parálisis y, en casos extremos, pueden atribuir el brazo a otra persona (somatoparafrenia). Luego de describir dos casos clínicos que ilustran las respuestas de pacientes anosognósicos ante la presencia del déficit (ignorar el hecho de que el brazo no se mueve y afirmar que sí lo hace, o afirmar que no se lo mueve voluntariamente debido al dolor producido por la artritis), observa que hay una impactante similitud entre las estrategias empleadas por los pacientes y lo que el psicoanálisis denomina "mecanismos psicológicos de defensa". Estos mecanismos son empleados por la gente normal cuando

[58] En *Phantoms in the Brain* (1998) Ramachandran vuelve a explorar y profundizar en el estudio de estos problemas.

son confrontados con hechos perturbadores acerca de sí mismos; ahora bien, señala, las personas que sufren de anosognosia hacen las mismas cosas, pero de una manera *groseramente exagerada*. El modo correcto de formular el problema implica considerar las dos cuestiones siguientes: a) ¿por qué los individuos normales tienen mecanismos psicológicos de defensa? Esto es, ¿por qué deberían sostener falsas creencias acerca de ellos mismos?; b) ¿por qué esos mecanismos están groseramente exagerados en la anosognosia?

Ramachandran señala que los mecanismos de defensa freudianos son esencialmente *falsas creencias* acerca de uno mismo, pero se pregunta cuál podría ser el posible beneficio que eso conferiría al organismo, ya que parecerían realmente desadaptativas.[59] La razón real para la evolución de los mecanismos de defensa como la confabulación y la racionalización, conjetura, es crear un sistema de creencias coherente con el fin de imponer una estabilidad a la propia conducta.

Con el fin de comprender esta hipótesis se debe recurrir al concepto de especialización hemisférica. Ramachandran sugiere que cada uno de nosotros tiene una fuerte necesidad de imponer consistencia, coherencia y continuidad a nuestra conducta. El hemisferio izquierdo es primariamente responsable por imponer consistencia a la historia argumental, lo que correspondería rústicamente a lo que Freud llama el Yo. Con el fin de actuar, el cerebro debe tener algún modo de seleccionar la superabundancia de información que llega al organismo y organizarla en un "sistema de creencias" consistente, una historia que dé sentido a la evidencia disponible. Sin embargo, cuando algo no encaja en el guión, raramente se abandona la historia entera y se comienza de cero. Lo que se hace, de hecho, es negar o confabular con el fin de hacer que la información se ajuste al cuadro general. Lejos de ser desadaptativos, tales mecanismos de defensa cotidianos mantienen el cerebro libre de una indecisión no direccionada causada por la "explosión combinatoria" de posibles historias basadas en el material disponible a los sentidos.

Lo que resta por explicar, dice Ramachandran, es por qué esos mecanismos de defensa están *groseramente* exagerados en los pacientes. Aquí es, señala, cuando el hemisferio derecho entra en escena. La idea básica consiste en que las *estrategias de afrontamiento* de los dos hemisferios son fundamentalmente diferentes. El trabajo del hemisferio izquierdo es crear un modelo y mantenerlo a cualquier costo. Si es confrontado con alguna información nueva que no encaja en el modelo, descansa en los mecanismos de defensa freudianos para negar, reprimir o confabular; cualquier estrategia que proteja el *statu quo*. La estrategia del hemisferio derecho es actuar como "detector de anomalías". Cuando la información

[59] Ramachandran desestima la teoría propuesta por Trivers, según la cual engañarse a uno mismo es la manera más efectiva de engañar a otros. Cfr. el capítulo IV para una descripción de la concepción del segundo.

anómala alcanza cierto umbral, el hemisferio derecho decide que es tiempo de forzar al hemisferio izquierdo a revisar el modelo entero y comenzar de cero. El hemisferio derecho, entonces, "*fuerza un cambio kuhniano de paradigma en respuesta a las anomalías, mientras que el hemisferio izquierdo siempre trata de aferrarse al modelo original*" (p. 352, cursivas del autor). En los pacientes que presentan anosognosia, el hemisferio izquierdo está llevando a cabo toda la confabulación y la negación como si se tratase de una persona normal. La diferencia es que esos pacientes han perdido los mecanismos del hemisferio derecho que los fuerza a efectuar un cambio de paradigma en respuesta al conflicto informacional. Esto fuerza a los pacientes a una trampa delirante y continúan confabulando sin cambiar de paradigma.

Ramachandran agrega que es necesaria una nota de prudencia acerca de la especialización hemisférica; debemos tener en cuenta no sólo que la especialización es *relativa* más que absoluta, sino también que el cerebro humano tiene incontables subdivisiones.

La tercera y última perspectiva presentada por Bortolotti acerca del solapamiento entre delirio y autoengaño sugiere que la misma existencia de los delirios, que muestra que el conflicto doxástico es posible, puede ayudar a reivindicar los enfoques tradicionales acerca del autoengaño, de acuerdo con los cuales la persona posee dos creencias contradictorias, pero es consciente sólo de una de ellas, debido a que posee una motivación para no tomar conciencia de la restante. Levy (2008), si bien considera que las condiciones requeridas por los modelos tradicionales del autoengaño no son necesarias, afirma que casos como los descriptos por Ramachandran constituyen una prueba viviente de que una persona puede, simultáneamente, creer que su brazo está paralizado y creer que puede moverlo. Más aun, es la creencia de que su brazo está paralizado lo que causa que adquiera la creencia de que no lo está. Levy reconstruye el caso típico de una persona con anosognosia de la siguiente manera: 1) el sujeto cree que su brazo está sano; 2) No obstante, también posee simultáneamente la creencia (o fuerte sospecha) de que su brazo está significativamente dañado, creencia (o sospecha) que lo perturba fuertemente; 3) la condición 1) es satisfecha debido a que la condición 2) es satisfecha, esto es, el sujeto está motivado a adquirir la creencia de que su brazo está sano debido a que posee la creencia (sospecha) concurrente de que en realidad está dañado de modo significativo, creencia que lo perturba fuertemente. Si este análisis es correcto, al menos un caso de delirio (la anosognosia) involucra conflicto doxástico.

Todo lo anterior implica, observa Bortolotti, que puede ser muy difícil justificar una separación clara entre delirio y autoengaño. Resulta útil, tanto del punto de vista diagnóstico como del científico, mantener una distinción entre síntomas de condiciones tales como la amnesia, la demencia o la esquizofrenia, por una parte, y las creencias irracionales que caracterizan la cognición normal, pero no por esto debería dejar de reconocerse que también hay muchos elementos de genuino solapamiento.

Bayne y Fernández (2009), consideran que hay múltiples puntos de contacto entre autoengaño y delirio. Ambos aparecen fundamentalmente como ejemplos de "creencias patológicas", creencias que son erróneas en algún sentido. En el caso del autoengaño es bastante claro, al menos en términos generales, qué es lo que es errado: los estados motivacionales y afectivos de los sujetos los han llevado a descuidar ciertas normas de formación de creencias. En el caso del delirio, es menos claro por qué el sujeto ha adquirido creencias patológicas. Si bien muchos de los análisis clásicos del delirio en la bibliografía psicoanalítica eran de naturaleza fuertemente motivacional, el foco de la teorización mucho más reciente acerca del delirio ha estado en los factores "fríos" más que en los "calientes". No obstante, este foco ha llevado a perder de vista importantes *insights* acerca del delirio, y hay mucho que aprender acerca de las creencias delirantes por medio del examen del rol de los procesos motivacionales y afectivos en la formación de creencias.

El enfoque estándar acerca del delirio y el autoengaño como patologías de la formación de creencias apela a la noción de racionalidad epistémica. De acuerdo con este punto de vista, las creencias delirantes y engañosas son patológicas en el sentido que el sujeto desprecia la norma epistémica de creer sólo aquello que la evidencia autoriza, aproximación que forma parte de la caracterización del delirio del DSM. El enfoque epistémico apunta a una conexión profunda entre delirio y autoengaño, debido a que la creencia autoengañosa también involucra una falla en creer de acuerdo con la evidencia de que uno dispone. Podría querer reservarse el término "delirio" para las fallas groseras de la racionalidad epistémica, y sostener que el autoengaño resulta delirante sólo cuando la creencia en cuestión enfrenta lo que el DSM denomina pruebas incontrovertibles y obvias en contrario. Así, en esta concepción tendríamos un solapamiento parcial entre las categorías de delirio y autoengaño: ciertos casos de autoengaño calificarían como delirantes,

pero también habría casos de delirio que no serían casos de autoengaño y casos de autoengaño que no serían casos de delirio.[60]

Pese a que puede decirse mucho en favor y en contra de la concepción epistémica del delirio y el autoengaño, Bayne y Fernández prefieren explorar otra ruta a través de la cual conceptualizar las patologías de la formación de creencias. Más que pensar en las patologías doxásticas en términos de *apartamiento de las normas de racionalidad epistémica*, pueden pensarse en ellas en términos de *apartamiento de las normas operativas de formación de creencias*; estas últimas son normas que especifican cómo un sistema psicológico debería funcionar. Este enfoque, observan, es a menudo soslayado debido a la suposición implícita de que las normas operativas de formación de creencias y las normas de racionalidad epistémica deben converger. Según esta concepción, un ser humano que no crea sólo en lo que la evidencia permite manifestará también anormalidades en la formación de creencias. Esta imagen, enfatizan, debería ser cuestionada. Si se dejan a un lado los efectos de la motivación sobre la formación de creencias, hay amplia evidencia para la concepción según la cual las normas operativas de formación de creencias implican un significativo apartamiento de las de racionalidad epistémica; esto se pone de manifiesto en sesgos tales como la falacia de conjunción. La creencia en un mundo de hechos morales objetivos, entidades sobrenaturales e inmortalidad personal parecen ser rasgos casi universales de la condición doxástica humana, pero el estatus epistémico de esas creencias es en gran medida objeto de debate.

Cabe preguntar, entonces, si el delirio y el autoengaño involucran apartamientos de las normas operativas de la formación de creencias. El autoengaño –al menos en su forma cotidiana- no necesita implicar un apartamiento de esas normas. Hay evidencia abrumadora de que los seres humanos normales tienen un autoconcepto sistemáticamente distorsionado.[61] Poseer una autoimagen excesivamente positiva parece ser parte del perfil funcional del sistema doxástico humano. No hay razón para considerar la formación de creencias motivacionalmente orientada como patológica *en tanto que tal*. Respecto del delirio, parecería obvio que los delirios involucran apartamientos –típicamente muy radicales- de

[60] Bayne y Fernández observan que si bien hay mucho de recomendable en el análisis para el cual autoengaño y delirio son patologías de la creencia, hay ciertos problemas con él. Debe notarse, en primer lugar, que deberíamos aceptar que el delirante y el autoengañado pueden poseer *alguna* evidencia para su creencia, incluso para los delirios más extraños. Para ilustrar este punto, mencionan una perspectiva según la cual muchos delirios pueden estar fundados en experiencias inusuales. Siguiendo a otros autores, denominan a esto "enfoque empirista" del delirio. Incluso creencias delirantes, como la de que los propios movimientos están controlados por fuerzas extraterrestres, pueden estar basados en evidencia (evidencia experiencial) de cierta clase. Si esto es correcto, entonces no es ya obvio que esos delirios *son* sostenidos "pese a lo que constituye incontrovertibles y obvias pruebas o evidencia en contrario".

[61] Bayne menciona aquí a los estudios de S. Taylor relativos a los sesgos "positivos" de la autovaloración, de los que nos ocuparemos en el capítulo V.

las normas operativas de formación de creencias. Los delirios sobresalen como especímenes exóticos en el jardín de la creencia, como ejemplos de lo que ocurre cuando colapsan los mecanismos de formación de creencias.

En cualquier caso, las perspectivas para enfoques puramente empiristas del delirio parecen poco prometedoras. Por un lado, hay muchos delirios que serían difíciles de explicar para cualquier tipo de enfoque empirista (menciona aquí los delirios floridos y politemáticos de la esquizofrenia y también los delirios de persecución y celotípico). Más aun, la idea de la experiencia inusual no provee una explicación completa del autoengaño. A la luz de lo anterior, muchos teóricos han argumentado que es necesario un *factor no experiencial* – un denominado "segundo factor"- para explicar por qué es que la experiencia inusual impulsa al paciente a desarrollar (y retener) su creencia delirante. Los candidatos propuestos para este segundo factor son (pero no se limitan a): la tendencia a privilegiar datos observacionales sobre las creencias de trasfondo, la posición de un estilo atribucional particular, una disposición para saltar a las conclusiones y una preferencia por explicaciones personales más que subpersonales.

Habida cuenta de la variedad de fenómenos que muy razonablemente, merecen ser incluidos dentro de la categoría de delirio, por una parte, y la diversidad de teorías explicativas sobre ellos, por la otra, resulta especialmente difícil establecer conclusiones siquiera plausibles relativas a sus vínculos con el autoengaño. Como hemos visto, parecería que estamos bastante lejos de una teoría aceptable que goce de consenso respecto de los procesos más extremos de distorsión de las creencias, como es el caso de los delirios y la confabulación. Cabe sospechar, sin embargo, que es posible que ocurra con estos fenómenos un proceso análogo al que ha ocurrido con otros conceptos de la psicopatología, esto es, el reemplazo de un concepto muy general por un conjunto de categorías más específicas y de menor alcance, que requieren de teorías explicativas de un nivel de generalidad correlativamente inferior. Si esto es así, el intento de establecer un nexo conceptual general con el fenómeno del autoengaño tendrá una probabilidad de éxito considerablemente menor.

4.2. Autoengaño y adicción[62]

La relación entre las adicciones y el pensamiento sesgado o distorsionado es un tema profusamente investigado, si bien no es muy frecuente que las distorsiones características del pensamiento del adicto aparezcan agrupadas en la bibliografía especializada bajo el rótulo de "autoengaño" y sí, como veremos, bajo otros tales como "negación" o "creencias distorsionadas".

Una observación extremadamente usual referente a los adictos es la de afirmar que son mentirosos, manipuladores, explotadores y egoístas. Es verdad que muy frecuentemente lo son, pero muchas veces las mentiras y manipulaciones están íntimamente relacionadas con fenómenos psíquicos muy particulares, fenómenos que son típicos de lo que puede denominarse "pensamiento adictivo", y debido a los cuales el adicto no sólo engaña a los demás, sino que, suele decirse, se engaña a sí mismo. En el examen del pensamiento adictivo será importante tener en cuenta la siguiente observación: si bien el alcohólico o adicto puede (y frecuentemente lo hace) mentir al negar la existencia del problema, sucede muy a menudo que la negación explícita y externa del problema es sólo la manifestación de sus propias distorsiones de pensamiento: induce al error a otros pero sólo como parte del proceso de engañarse a sí mismo.

Sobre la base de lo anterior, si hubiera que elegir una característica definitoria del pensamiento del adicto, es altamente probable que el acuerdo mayoritario entre los especialistas sería sin duda que la negación es tal característica. No es en absoluto casual que una de las frases más conocidas del programa de Alcohólicos Anónimos es la que sostiene que "el alcoholismo es la enfermedad de la negación". De este modo, es extremadamente frecuente que los familiares, amigos o profesionales se encuentren, ante el intento de señalar al alcohólico o adicto la existencia de un problema, con respuestas como las siguientes: "no hay ningún problema con mi modo de beber", "puedo dejar esa sustancia (alcohol, cocaína, marihuana, etc.) cuando quiera", "controlo perfectamente lo que consumo", "no es para tanto, consumo sólo ocasionalmente", "sólo lo hago en situaciones sociales", "es verdad que he estado bebiendo demasiado, pero he estado bajo mucha presión en el trabajo", etc. Ahora bien, este uso del término "negación" involucra varios tipos distintos de estrategias tendientes a desestimar el problema: desde la negación

[62] Los términos "adicción" y derivados ("adicto", "adictivo" y otros) han sido desplazados gradualmente en el ámbito científico y sanitario para hacer referencia a los trastornos debidos o relacionados con el uso de sustancias psicoactivas. Dada su amplia difusión y el uso que de ellos siguen haciendo algunos autores analizados mantendremos su empleo, aun cuando no suscribamos algunos de sus significados y connotaciones.

abierta ("no hay ningún problema con mi modo de beber"), hasta la minimización ("no es para tanto, consumo sólo ocasionalmente"), pasando por la racionalización ("es verdad que he estado bebiendo demasiado, pero he estado bajo mucha presión en el trabajo") y por una ilusión de control ("puedo dejar esa sustancia cuando quiera"). En consecuencia, aunque genéricamente suele emplearse el término "negación" cuando se hace referencia a fenómenos como éstos, es conveniente tener presente que se está agrupando bajo un término de alcance impreciso a fenómenos psíquicos diferentes.[63] Conviene aclarar, por otro lado, que este estilo de pensamiento no sólo es extremadamente común entre los adictos a sustancias, sino también entre quienes sufren las denominadas "adicciones conductuales", como el juego patológico o la adicción al trabajo.

En el examen del pensamiento adictivo y, en particular, al analizar el fenómeno de la negación, convendrá no olvidar, una vez más, la útil observación de Mele: la ignorancia no excluye al autoengaño; por el contrario, hay casos en los que parece contribuir con éste. Debe tenerse en cuenta que, por muy difundidos socialmente que estén los términos "adicción" y "alcoholismo", con mucha frecuencia las personas no tienen conocimiento del significado técnico de estos términos (significado, por otra parte, en absoluto carente de controversias), y confunden o ignoran los rasgos definitorios que caracterizan a las entidades clínicas respectivas. Y este desconocimiento, sin duda, es a menudo extensivo para los mismos alcohólicos o adictos.[64]

Dado que, como dijimos, las distorsiones en el pensamiento del adicto suelen caer bajo el rótulo de "negación", convendrá comenzar con un análisis de la aplicación de este concepto a los trastornos adictivos. Tal como hemos adelantado en el apartado referente a la relación entre autoengaño y negación, el estudio de este último concepto ha estado durante mucho tiempo fuertemente influido por las teorías psicodinámicas. Esta influencia se extendió, previsiblemente, a los análisis acerca de las distorsiones del pensamiento de los

[63] Trivers (2011) observa que las "drogas recreativas" (como las denomina) y el autoengaño están estrechamente relacionados. Dado que el efecto de las sustancias, ya sean legales o ilegales, puede ser grave, considera que la persona tiene que autoengañarse para justificar ante sí misma el riesgo de consumirlas, y luego justificarlo ante los demás. De ahí que, dice, el autoengaño sea casi imprescindible cuando se consumen drogas. Tal como detallaremos en el capítulo IV, Trivers adopta el termino "autoengaño" como una denominación "paraguas" que abarca fenómenos tan distintos como la racionalización, la proyección, y otros. Desde esta perspectiva, el uso no puede ser cuestionado. Sin embargo, esto sí puede hacerse si se adopta una definición más acotada y estricta del fenómeno.

[64] He tenido oportunidad de observar, por ejemplo, que una paciente con indicadores inequívocos de alcoholismo afirmara que ella no se consideraba alcohólica porque no consumía alcohol todos los días. No obstante, esta paciente sí reconocía abiertamente que había perdido todo control sobre su consumo, que las cantidades que consumía habían ido incrementándose de manera marcada, y que el consumo le producía un malestar derivado de algunas de sus consecuencias negativas, como el deterioro de su estado y aspecto físico. Todo esto la llevaba a pensar que debía hacer algo al respecto; no negaba la existencia de un problema, simplemente se equivocaba al categorizarlo.

adictos, incluso en el caso de autores que no adhieren estrictamente a las concepciones psicoanalíticas acerca de la adicción. En Metzger (1988), por ejemplo, encontramos una descripción que podríamos llamar "clásica" acerca de la negación como mecanismo característico de funcionamiento de los alcohólicos. El autor comienza señalando que la interacción humana es la primera fuente de nuestra ansiedad; para protegernos a nosotros mismos, desarrollamos mecanismos de defensa, los cuales son una parte normal de nuestro proceso de desarrollo. Haciendo referencia a los trabajos de Sigmund y Anna Freud, señala que los mecanismos de defensa son empleados por todas las personas para reducir la ansiedad cuando la identidad es amenazada, empleo que es fundamentalmente inconsciente. La negación, que es la que interesa especialmente en el contexto del estudio del alcoholismo, está localizada en la base de una empinada jerarquía. Sobre la base de los desarrollos de G. Vaillant, expuestos en su libro *Adaption to Life* (1977), Metzger adopta una jerarquía de las defensas estructurada en cuatro grupos, a saber:

Nivel I: mecanismos psicóticos. Negación (de la realidad externa), distorsión, proyección delirante.

Nivel II: mecanismos inmaduros (comunes en las depresiones severas, trastornos de personalidad y en la adolescencia). Fantasía (supresión esquizoide, negación a través de la fantasía), proyección, hipocondría, conducta pasivo-agresiva, *acting out*.

Nivel III: mecanismos neuróticos (comunes en todas las personas). Intelectualización (aislamiento, conducta obsesiva, anulación, racionalización), represión, formación reactiva, desplazamiento, disociación.

Nivel IV: mecanismos maduros (comunes en adultos "saludables"). Sublimación, altruismo, supresión, anticipación, humor.

Metzger adhiere a la concepción según la cual el empleo de defensas maduras está asociado con una vida más saludable y exitosa y, conversamente, el empleo de defensas en los niveles inferiores de la jerarquía y con mayor rigidez está asociado a vidas más empobrecidas y problemáticas. A la vez, asocia el "camino descendente" de la adicción con un descenso en la jerarquía de las defensas. Señala que miembros de Alcohólicos Anónimos reconocen este fenómeno cuando afirman respecto de alguien que tiene una "borrachera seca", expresión que refiere a un uso continuo de razonamientos defectuosos que emplean mecanismos de negación, racionalización y proyección, aun cuando el uso del alcohol haya cesado. Con la ayuda de psicoterapia u otras estrategias similares, las defensas rígidas pueden ser abandonadas y reemplazadas por medios de afrontamiento más flexibles.

Esta concepción de la negación como fenómeno típico del pensamiento de los adictos no es compartida por otros autores, algunos de ellos muy influyentes. Miller y Rollnick (1991), en su conocido estudio acerca de los conflictos motivacionales característicos de la adicción, analizan brevemente el problema de la negación. El examen y cuestionamiento las estrategias agresivas de ciertos programas terapéuticos diseñados para el tratamiento de las adicciones los conduce a afirmar que el empleo de tales estrategias está basado en el presupuesto clave de que los adictos "-*como grupo*, y de forma inherente a su condición- poseen niveles extraordinariamente elevados de ciertos mecanismos defensivos, que los convierten en personas inaccesibles si se utilizan los métodos habituales de terapia y persuasión. Se ha creído que estos mecanismos están profundamente enraizados en la personalidad y el carácter de estas personas" (p. 29) Encontramos en la descripción de estos autores, en consecuencia, dos ideas muy difundidas y polémicas acerca de la naturaleza y orígenes de la adicción: la negación y la personalidad adictiva. Miller y Rollnick consideran que tales creencias parecen haber surgido del pensamiento psicodinámico, el cual considera a las adicciones como síntomas de un trastorno de personalidad. Una característica de este trastorno sería el empleo excesivo de algunos de los mecanismos de defensa más primitivos descriptos por Anna Freud, punto de vista que fue adoptado por profesionales de peso en el campo del alcoholismo. Miller y Rollnick citan a una influyente psiquiatra, Ruth Fox, según la cual el alcohólico "levanta un elaborado sistema defensivo en el que niega que sea alcohólico o esté enfermo, racionaliza que necesita beber por razones laborales, de salud o sociales, y proyecta la culpa de los problemas por los que está atravesando" (p. 29). La difusión de tales ideas constituyó el consenso acerca de estas características como universales e inherentes a la estructura del carácter de alcohólicos y adictos.

Estas ideas, observan Miller y Rollnick, no son sostenibles; la idea de una personalidad adictiva o alcohólica no está corroborada por los escritos de Alcohólicos Anónimos ni tampoco por cinco décadas de investigación psicológica. Cuando los mecanismos de defensa han sido definidos de forma operacional y estudiados en detalle, se ha encontrado que la negación no es una característica más distintiva de los adictos que del resto de las personas; incluso las medidas de las diferencias individuales de la negación arrojaron resultados curiosos (por ejemplo, se han relacionado resultados terapéuticos positivos con niveles *más altos* de negación al inicio del tratamiento). No existe ni ha existido, enfatizan, evidencia en favor de la afirmación de que los alcohólicos o adictos manifiestan un patrón de personalidad común y coherente caracterizado por un excesivo

uso de mecanismos de defensa concretos. La negación no debe ser definida, sostienen, como un rasgo de personalidad, sino que se la debe concebir como un rechazo a admitir los problemas, un engaño consciente y una actitud mendaz. En resumen, la investigación no sostiene la creencia de que existe una característica central de personalidad o un conjunto de defensas intensas, y tampoco que ésta sea una característica de las personas que sufren de alcoholismo o un problema con otras drogas. En palabras de los autores, "A partir de la investigación de que disponemos, la hipótesis de la negación –de que los alcohólicos o las personas con "dependencias químicas", *como prototipo*, presentan alteraciones de la personalidad concretas o altos niveles de ciertas defensas- no es más que un mito" (p. 34).

El rechazo de Miller y Rollnick a las teorías acerca de la existencia de una personalidad adictiva y a la negación como mecanismo de defensa distintivo de los adictos no elimina, sin embargo, la presunción de que el pensamiento de los adictos se caracteriza por un elevado grado de distorsión y por la manifestación de procesos cognitivo-afectivos característicamente sesgados. Abraham Twerski, en su libro *Addictive Thinking: Understanding Self-deception* (1997), narra una historia que ilustra hasta qué punto el pensamiento del adicto puede distorsionar los modos de razonamiento más elementales:

> Alan, un alcohólico en recuperación, no era consciente de los efectos de su modo de beber, pese a lo que otras personas le decían. Dado que sólo bebía cerveza, estaba seguro de no tener un problema con el alcohol. Con el paso del tiempo, Alan se enfermó físicamente y ya no pudo seguir negando que algo estaba mal. Concluyó que al beber medio barril de cerveza por día estaba consumiendo demasiado líquido. Cambió entonces por el escocés con soda. Cuando los síntomas físicos empeoraron, le echó la culpa a la soda y cambió por whisky con agua. Como sus síntomas empeoraron aún más, eliminó el agua (p. 13).

En la búsqueda de una identificación de los rasgos típicos del pensamiento adictivo, algunos autores, entre ellos el propio Twerski, remarcan las similitudes entre esa clase de pensamiento y ciertos tipos de trastornos mentales, como la esquizofrenia.[65] Twerski apunta a esta clase de similitud señalando que en ocasiones las personas que sufren trastornos adictivos son erróneamente diagnosticadas como esquizofrénicas. Este diagnóstico erróneo se basa en la presencia de los mismos síntomas, entre ellos delirios, alucinaciones y conducta marcadamente anormal. Ahora bien, Twerski agrega que una diferencia crucial entre el pensamiento esquizofrénico (abiertamente absurdo) y el pensamiento adictivo reside en que éste posee una lógica superficial que puede llegar a ser

[65] Este texto es abundante en ejemplos clínicos y en general carece de elaboración teórica o de intentos de elucidación de los términos empleados (entre ellos, el de autoengaño), pero no carece de interés para señalar algunos puntos y también limitaciones del enfoque y pensar algunas cuestiones en su relación con la práctica.

muy seductora y desorientadora. Especialmente en las primeras etapas de la adicción, señala Twerski, la perspectiva y explicación del adicto respecto de lo que está ocurriendo pueden parecer muy razonables, y muchas personas son persuadidas por su razonamiento.

Esta razonabilidad superficial del pensamiento del adicto tiene consecuencias de mayor alcance. Es posible, por ejemplo, que la familia del adicto pueda ver las cosas en "el modo de pensamiento adictivo" por un tiempo prolongado, esto es, el pensamiento autoengañoso puede "infectar" a los miembros codependientes de la familia tanto como a la persona químicamente dependiente. Twerski señala también que las similitudes entre la conducta de un adicto y la conducta de un codependiente son impactantes. Los adictos usualmente buscan nuevos modos de continuar su uso de químicos, mientras tratan de evitar sus consecuencias destructivas. Cuando los esfuerzos de control fallan, los adictos no concluyen "no puedo controlar mi uso". En vez de eso se dicen a sí mismos "Este método no funciona. Debo encontrar otro método que funcione". Del mismo modo, los codependientes no concluirán que, dado que los esfuerzos para detener al adicto han sido fútiles, no hay modo de controlarlo. Más bien buscan nuevas maneras que funcionen. Los rasgos autoengañosos del pensamiento adictivo y la codependencia, enfatiza Twerski, tienen mucho en común: en ambos hay a menudo negación, racionalización y proyección, pueden coexistir ideas contradictorias y hay una fuerte resistencia al cambio propio combinada con un deseo de cambiar a otros. También están presentes una ilusión de control y baja autoestima. Todos los rasgos del pensamiento adictivo están presentes en ambos tipos de dependencia, y el único rasgo que haría posible distinguirlos sería el uso de químicos.[66]

Un concepto que usa el autor, muy empleado en el tratamiento de las adicciones, es el de "pensamiento distorsionado". Si bien es obvio que las distorsiones del pensamiento no son exclusivas de los trastornos adictivos y no están necesariamente relacionadas con el uso de químicos, el autor considera que la intensidad y regularidad de este fenómeno son más comunes entre los adictos. Podemos agregar que la necesidad de lidiar con esta distorsión del pensamiento ocupa un lugar central en los enfoques cognitivistas de la adicción, por ejemplo, los trabajos de Beck et al (1993), quienes hablan de "creencias adictivas".

Más allá de la presunta similitud del pensamiento adictivo con la esquizofrenia, hay un aspecto que es importante resaltar y que puede pasar inadvertido en este análisis. Tal

[66] Esta observación es interesante, ya que concierne a las relaciones entre engaño y autoengaño, y confirma una vez más la dificultad de hablar de autoengaño como proceso único o fundamental cuando hay más de una persona involucrada en el proceso. Volveremos sobre esta clase de procesos en el capítulo V, al examinar el fenómeno del autoengaño colectivo.

como señalaron algunos autores (Oksenberg-Rorty, 1988), el autoengaño tiende a ramificarse y desarrollarse. Si bien es perfectamente posible hablar de autoengaño con respecto a una creencia, no hay duda de que, cuando se trata de ciertos procesos, es muy probable que las creencias autoengañosas tiendan a constituir una estructura o entramado complejo y autosustentante.

Otra observación perspicaz hecha por Twerski es que el pensamiento adictivo no es afectado por la inteligencia. Las personas que funcionan en los niveles intelectuales más altos, señala, son tan vulnerables a esas distorsiones del pensamiento como cualquier otra. Más aun, las personas con intelectos inusualmente altos a menudo tienen grados más intensos de pensamiento adictivo. Entonces, las personas altamente intelectuales pueden ser los pacientes más difíciles de tratar: a mayor brillantez de la persona, más ingeniosas son sus razones para beber, para no abstenerse y para considerar organizaciones como AA como carentes de valor. Las personas más inteligentes y educadas son capaces de elucubrar teorías justificatorias más complejas y difíciles de desarmar, tanto para negar su adicción como para negar la necesidad de ciertos tipos de ayuda, y en muchos otros respectos.

El análisis de Twerski, más allá de sus sin duda agudas observaciones sobre la naturaleza del pensamiento adictivo, adolece de la misma imprecisión respecto del término "autoengaño" a la que hemos hecho referencia al principio de este apartado. En efecto, en este autor el término abarca un conjunto de operaciones mentales de muy distinta naturaleza, que van desde la negación abierta del trastorno hasta la racionalización. Por otro lado, algunas de sus observaciones se basan en consideraciones sobre el funcionamiento cognitivo que no se sustentan en lo que sabemos acerca de estos procesos. Twerski sostiene, por ejemplo, que el pensamiento adictivo es diferente del pensamiento lógico en que no llega a una conclusión basado en la evidencia de los hechos de la situación, sino justo a la inversa; de este modo, el adicto llega a la conclusión "necesito un trago" (o una droga), y a partir de allí construye una justificación para esa conclusión, sea o no lógica y esté o no apoyada por los hechos. Parece razonable sostener que esta descripción no se ajusta a una concepción apegada a lo que sabemos actualmente del funcionamiento cognitivo. Como se ha observado en el capítulo precedente, las normas operativas a través de las cuales realizamos inferencias y nos formamos creencias acerca del mundo y de nosotros mismos muy a menudo se apartan de lo que la racionalidad epistémica recomienda. Si bien es posible, como señalamos anteriormente, que el adicto haga uso de procedimientos tales como adoptar la conclusión preferida y luego ver de qué manera puede ser sostenida mucho más que lo que lo hace el razonador "normal", la diferencia con

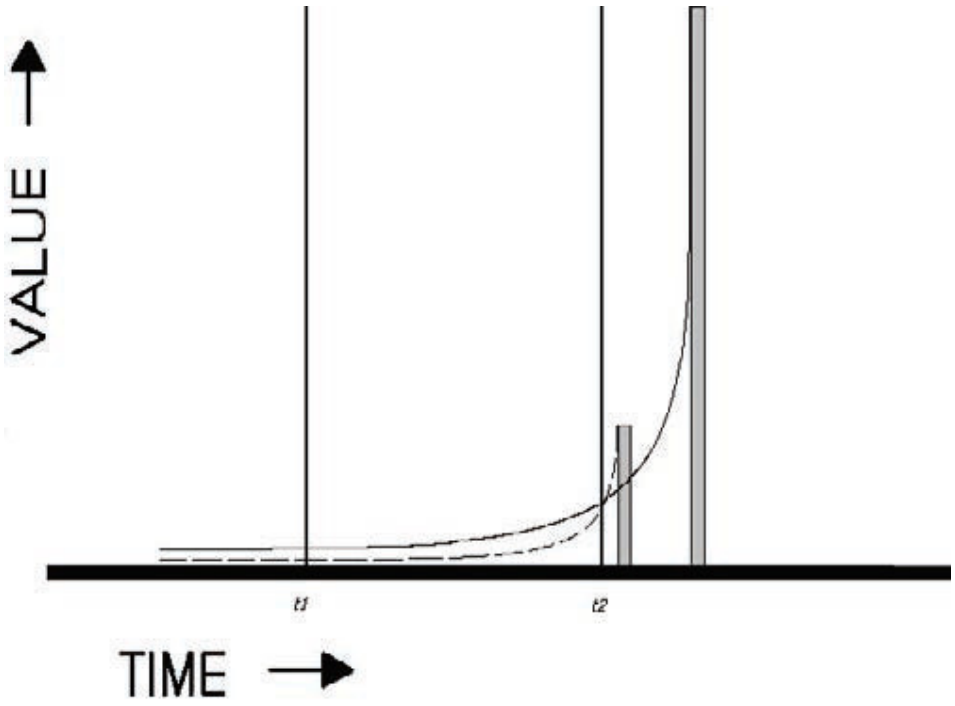
éste será de grado, y no de clase. La relación entre autoengaño y adicción, en síntesis, sigue careciendo de una elucidación conceptualmente precisa.

Un interesante intento de tal elucidación puede encontrarse en Hanson (2009), quien ofrece una perspectiva novedosa acerca de la relación entre el autoengaño y la adicción. Esta perspectiva está basada en la adopción del denominado "Modelo del Descuento Hiperbólico", desarrollado por G. Ainslie. Según la descripción del modelo de Ainslie que proporciona Hanson, los seres humanos sufrimos crónicamente de una "inversión de las preferencias", esto es, un modo de actuar irracional en el que se privilegian recompensas menores a corto plazo a costa de la pérdida de recompensas mayores en un plazo más largo. Tal inversión de las preferencias no está reservada exclusivamente para unas pocas personas que presentan fallas morales o racionales, sino que es endémica a la condición humana. La conducta humana, bajo el modelo de Ainslie, es profundamente akrática; por el contrario, la conducta consistente es un logro notable. Los seres humanos actuamos de acuerdo con un esquema de valuación futura (el Modelo del Descuento Hiperbólico) que necesariamente conduce a la irracionalidad. Cuando es presentado gráficamente, el Modelo muestra que la valuación de bienes a futuro sigue una forma profundamente curvada, "hiperbólica", en vez de un crecimiento gradual consistente. El Modelo puede describirse esquemáticamente de la siguiente forma (véase la figura 1). El agente se encuentra ante dos bienes objeto de valuación entre los que debe elegir: x_1 (por ejemplo, el consumo de una sustancia adictiva)[67] y x_2 (por ejemplo, el mantenimiento de la abstinencia). El agente asigna un valor mayor a x_2 que a x_1; sin embargo, a medida que el momento de la elección se acerca, la "tentación" comienza a jugar un rol, lo que provoca que la valuación de x_1 alcance su máximo. Lo que ocurre, entonces, es que cuando el momento de la elección está relativamente lejano, el agente asignará un mayor valor a x_2 que a x_1, pero cuando se aproxima el momento de la elección las preferencias se invierten y tiene lugar la acción akrática: el agente elegirá la recompensa más cercana y menor antes que la recompensa más lejana y mayor. Supongamos, propone Hanson, que mientras estoy sentado tranquilamente en mi oficina asigno un mayor valor a abstenerme de fumar que a hacerlo. Sin embargo, al mismo tiempo, comienzo a sentir el *craving* de nicotina, y comienzo a valorar más el fumar que el abstenerme de hacerlo. El valor que asigno al fumar, a diferencia del valor que le he asignado en un momento pasado, supera la línea de valuación de la abstinencia; dado que generalmente hacemos aquello que

[67] Cabe aclarar que el Modelo puede aplicarse también a una serie mayor de fenómenos, incluyendo las llamadas "adicciones conductuales".

preferimos hacer, la conducta final será la de fumar. La racionalidad, como se ha señalado más arriba, es un logro; no somos en ningún sentido seres "naturalmente" racionales.

Fig. 1. La línea continua representa la conducta de abstenerse, mientras que la discontinua representa la de fumar. [68]



Hanson señala que una faceta del modelo de Ainslie digna de especial atención es que implica (si bien enfatiza que este autor generalmente rechaza el empleo de esta expresión) la existencia de un yo múltiple, en el sentido en que cada persona contiene un conjunto de yoes diacrónicamente sucesivos, cada uno de los cuales reemplaza al precedente. El modelo supone un escenario en el que cualquier yo con un conjunto de preferencias se encuentra "en guerra" con cualquier otro que no comparta esas preferencias. El yo futuro amenaza con hacer elecciones que no están en línea con las que el yo presente prefiere; lo que el yo presente debe hacer, en consecuencia, es emprender un curso de acción que frustre la capacidad del yo futuro para hacer una elección que no sea la que el yo presente prefiere. De este modo, por ejemplo, el yo presente de un alcohólico que ingiere Disulfiram (fármaco aversivo que produce una serie de síntomas desagradables

[68] Extraída de Hanson (2009), p. 23.

cuando se consume alcohol, como cefalea, náuseas y dolor en el pecho), privará al yo futuro del placer de tomar alcohol.[69]

Ahora bien, Hanson considera que una explicación completa de la adicción requiere ampliar el modelo de Ainslie, ya que por sí mismo no alcanza a explicar varios aspectos distintivos de este trastorno. Hanson argumenta que un problema central para los adictos, no contemplado en el modelo, es el autoengaño, y aplica a la explicación de este fenómeno la concepción del yo múltiple desarrollada previamente.

Con el fin de presentar su propia versión del fenómeno del autoengaño en el adicto, Hanson presenta en primer lugar las explicaciones de este fenómeno proporcionadas por Davidson y Mele.[70] Ambos modelos, pese a sus enormes diferencias, comparten un rasgo que Hanson va a modificar: los dos (aun en el caso de Davidson, que supone alguna división de la mente) sostienen una concepción de un yo unitario. Si, por el contrario, observa Hanson, una concepción del yo múltiple como la que ha supuesto es correcta, entonces la concepción del autoengaño que modela este fenómeno sobre la base del engaño interpersonal puede ser capaz de explicarlo. Esto es, siguiendo en este punto a Davidson, es posible decir que el autoengaño consiste en un yo que engaña a otro. Esta alternativa, no obstante, tiene vedada la posibilidad de sugerir que tal engaño ocurra de un modo directo. Esto se debe a que Hanson ha afirmado con anterioridad que la memoria es transitiva: incluso si mi yo presente cree falsamente algo que un yo anterior le ha transferido, mi yo actual tendría acceso al recuerdo de las intenciones del yo pasado. De este modo, no resulta posible una solución directa al problema del autoengaño a través de la concepción de múltiples yoes.

La estrategia seguida por Hanson para proponer su modelo alternativo del autoengaño se basa en la presentación de ejemplos de aquel fenómeno, ejemplos que pueden proveer el método para comprender la totalidad de las "variedades de jardín" del autoengaño. Estos ejemplos, que por razones de espacio no describiré aquí, describen tres situaciones distintas: un adicto a la heroína que tras un gran conflicto interno decide inyectarse la droga "por última vez", el mismo adicto luego de que los efectos han comenzado a disiparse, convenciéndose a sí mismo de que la adicción es su destino, y un padre que grita enfurecido a su hijo porque ha volcado un vaso de leche.

Al examinar el tercer caso de autoengaño (en el cual el padre inventa razones *ex post* para justificar su conducta, luego de haber actuado contra su mejor juicio), Hanson

[69] Hanson plantea para una serie de requisitos para su teoría de los yoes múltiples; no obstante, creo que no hace falta profundizar en esos requisitos para comprender el núcleo de su teoría.
[70] Cfr. Al respecto el capítulo II.

considera que tiene un parecido tanto con la racionalización como con el pensamiento desiderativo. Consiste en una forma de racionalización en cuanto involucra una búsqueda activa de explicaciones falsas de la conducta que, no obstante, son psicológicamente confortables. Es pensamiento desiderativo en el sentido en que esas explicaciones son precisas debido a que representan el mejor escenario normativo, dada la cercanía temporal de la conducta que, *prima facie*, lamentamos. La creación de justificaciones racionales y morales resulta natural para nosotros, señala Hanson, y afirma que este fenómeno es, cuando se aplica a un contexto intrapersonal, un ejemplo teóricamente crucial de autoengaño.

Luego de examinar el segundo caso, Hanson examina la concepción resultante del autoengaño. Considera que Ainslie apoya una concepción "desguionada" [*dehyphenated*] del autoengaño. El autoengaño no es tanto "auto-engaño" [*self-deception*] como "auto engaño" [*self deception*]: se trata del fenómeno de creer erróneamente que las acciones ejecutadas en el pasado fueron acciones realizadas bajo el mismo conjunto de preferencias que se posee actualmente; es esta última creencia la que es falsa. Las acciones que una persona ejecutó en el pasado, que resultan inconsistentes con lo que la persona valora hoy, fueron acciones realizadas bajo un conjunto distinto de preferencias. Pero decir que una acción fue ejecutada bajo un conjunto distinto de preferencias es afirmar, si se emplea el aparato conceptual del modelo de Ainslie, que fueron ejecutadas por un sí mismo diferente. Presumiblemente, observa Hanson, la razón gramatical por la cual empleamos un guión en el término "auto-engaño" [*self-deception*] es la de ligar reflexivamente el acto de engañar con el engañado. Al remover el guión, prosigue, no está intentando ejecutar algún truco lingüístico, sino que está intentando señalar que el mejor camino para comprender el autoengaño en un sistema de descuento hiperbólico es entenderlo como un sistema que divorcia el engañador del engañado. El agente que es engañado por el sí mismo es un descontador cuyo yo actual intenta dar sentido (o "acomodar") las acciones ejecutadas bajo un conjunto diferente de preferencias en términos que serían coherentes con el conjunto de preferencias que actualmente tiene. Esta concepción "desguionada", considera, es capaz de acomodar las variedades de jardín de autoengaño de un modo consistente con la perspectiva de los yoes diacrónicamente distintos.

Hanson considera que el modelo de "auto engaño" que ha desarrollado presenta una consecuencia indeseable: implica que el auto engaño es abundante. El auto engaño ocurrirá toda vez que un yo intente explicar una conducta que cree (erróneamente) que le pertenece, pero que en realidad es la conducta de otro. Deberíamos suponer

razonablemente, observa Hanson, que el auto engaño ocurre con la misma frecuencia con la que ocurren la inversión de las preferencias y del control. Bajo esta concepción, concluye, las personas son generalmente auto engañadores crónicos, dado que son generalmente descontadores crónicos.

Si bien no emprenderemos aquí una crítica sistemática de la propuesta de Hanson respecto de las relaciones entre autoengaño y adicción, a nuestro modo de ver, su concepción del autoengaño presenta varios puntos débiles. Tales debilidades no requieren cuestionar ni el Modelo del Descuento Hiperbólico ni (aunque quizás constituya el eslabón más débil de la cadena argumentativa) la concepción del yo múltiple. Estas falencias son, principalmente, la limitación del análisis a casos de racionalización de las conductas y afirmar, sin una justificación apropiada, que la creación de justificaciones racionales y morales constituye un ejemplo teóricamente crucial de autoengaño. Concedamos, en primer lugar, que los límites entre el autoengaño y otras distorsiones motivadas de las creencias no son nítidos: múltiples ejemplos atestiguan la existencia de casos dudosos, que parecen oscilar entre el autoengaño y fenómenos tales como el pensamiento desiderativo, la ceguera intelectual, la racionalización, la negación y, como hemos mencionado en capítulos previos, la mera ignorancia no motivada. Sin embargo, el carácter borroso de las fronteras entre tales fenómenos no las elimina por completo. Muchos casos de autoengaño no involucran la generación de explicaciones que justifiquen una conducta que resulta cuestionable para el propio agente; más aun, el propio autoengaño parece impedir o dificultad conductas que resultarían adecuadas a la vista de observadores externos. Este es el caso, por ejemplo, del hombre que niega la infidelidad de su pareja pese a las claras pruebas en contrario a su creencia, creencia que le impide realizar las conductas que resultarían adecuadas en tal situación. El modelo de Hanson, en suma, si bien sumamente provocativo y novedoso, adolece de algunas limitaciones conceptuales que limitan su interés como explicación de las relaciones entre el pensamiento y la conducta adictivas, por una parte, y el autoengaño, por la otra.

En este apartado hemos examinado algunas posiciones respecto de la relación entre autoengaño y adicción. Podemos, en consecuencia, sugerir algunas conclusiones, seguramente provisionales, habida cuenta tanto de los avances en el estudio de los procesos relativos a los trastornos por dependencia de sustancias como los relativos al autoengaño. En primer lugar, las distorsiones observables en el pensamiento del adicto exceden lo que técnicamente denominamos autoengaño, a menos que demos al término un alcance mucho

mayor para que incluya fenómenos como la racionalización o la minimización. Este es el caso de autores, como Trivers y Hanson (especialmente el primero), que le asignan un alcance mayor que el usual. Esta multiplicidad de variantes de pensamiento sesgado no constituiría un inconveniente para el estudio de las relaciones entre ambos fenómenos, en principio, si se hiciera el intento de distinguir el autoengaño de otros fenómenos con los que mantiene un "aire de familia". En ausencia de este intento, sin embargo, la relación permanece inevitablemente indeterminada. El vínculo entre las adicciones y el autoengaño, en síntesis, sigue pendiente de un esclarecimiento conceptual y empírico apropiado.

Capítulo IV. ¿Hemos sido diseñados por la evolución para autoengañarnos?

Uno de los aspectos más intrigantes del autoengaño, como hemos señalado en capítulos previos, es la propia existencia de una capacidad para producir en nosotros mismos una distorsión en las creencias acerca de algún aspecto de la realidad. Lo anterior puede traducirse en dos preguntas fundamentales: cuáles son los mecanismos mentales que hacen posible tal distorsión, y por qué se encuentra presente en nuestro sistema psíquico. Como hemos visto en el capítulo precedente, esta última pregunta ha sido respondida por algunos autores sobre la base de la supuesta función defensiva que el autoengaño tendría dentro de nuestro funcionamiento mental. Sin embargo, esta función del autoengaño, o su presunta contribución a nuestro bienestar, es una tesis que ha resultado a muchos autores como mínimo polémica.[71] No resulta extraño, en consecuencia, que varios teóricos hayan explorado otros caminos en la búsqueda de una explicación para la existencia de un fenómeno tan elusivo. Una línea especialmente destacada ha estado basada en la presunción de que la Teoría de la Evolución podría dar una respuesta satisfactoria a esos y otros interrogantes.

En las últimas décadas el pensamiento evolucionista ha ido ganando cada vez más terreno como fundamento para la explicación de fenómenos psicológicos y sociales, desde el primitivo y controvertido programa de la sociobiología en la década del '70 (Wilson, 1975), hasta los posteriores y más elaborados intentos de basar nuestra comprensión de lo mental en los principios de la evolución. El programa teórico conocido como "psicología evolucionista", [72] en particular, se ha caracterizado por la tentativa sistemática de mostrar cómo nuestras estructuras mentales son el resultado de un proceso evolutivo. Tales estructuras fueron seleccionadas por su contribución para la supervivencia y el éxito reproductivo de los organismos que las poseían; en otros términos, constituyen *adaptaciones*.[73] En los últimos veinte años se han publicado multitud de ensayos que

[71] Van Leeuwen (2007), de quien hablaremos en este capítulo y el siguiente, observa que la explicación del autoengaño sobre la base dela evitación del dolor no constituye una explicación apropiada para su existencia, ya que la mala información parecería inevitablemente disminuir la aptitud a largo plazo. El dolor físico y psicológico, señala, existe por razones biológicas.

[72] En otra parte (Fernández Acevedo, 2008) identificamos tres compromisos teóricos fundamentales de la psicología evolucionista: adaptacionismo, computacionalismo/modularidad e innatismo.

[73] No debe confundirse el sentido del término "adaptativo" para la Teoría de la Evolución del que frecuentemente se le asigna en la psicología contemporánea, que haría referencia a las ventajas que una determinada función o rasgo mental puede poseer para el bienestar o la salud mental. Véase al respecto el § 3 del capítulo III.

examinan los principios teóricos y hallazgos fácticos en psicología evolucionista, ya sean los producidos por los principales autores que promueven y desarrollan esta perspectiva (Tooby y Cosmides, 1992; Symons, 1992; Buss, 1995; Cosmides y Tooby, 1997; Pinker, 1997), como aquellos escritos por partidarios o simpatizantes críticos (Wright, 1994; Caporael, 2001; Kennair, 2002; Durrant y Ellis, 2003). La perspectiva evolucionista en Psicología ha sido aplicada a un muy amplio campo de investigación, que incluye entre muchos otros temas el lenguaje (Pinker y Bloom, 1992), el intercambio social (Cosmides y Tooby, 1992), la elección de parejas sexuales (Buss, 1992), el homicidio (Daly y Wilson, 1988), la psicopatía (Mealey, 1995), y las emociones (Cosmides y Tooby, 2000; Nesse, 1998). No resulta extraño, en consecuencia, que distintos autores interesados en el autoengaño hayan orientado sus intentos de explicación del fenómeno a la búsqueda de sus orígenes en los desafíos que nuestra especie debió enfrentar para sobrevivir; así, han visto la luz diversos intentos de lograr una comprensión evolucionista del autoengaño (Ramachandran, 1996; Moomal y Henzi, 2000; Surbey y McNally, 1997), incluso en territorios teóricos bastante más complejos que los comparativamente "simples" procesos de autoengaño individual (Wrangham, 1999; Johnson, Wrangham y Rosen, 2002).

Dados los avances que el pensamiento evolucionista posibilitó para las ciencias biológicas, *prima facie* parecería plausible pensar que este pensamiento constituiría un cimiento sólido para el logro de una comprensión del autoengaño. No obstante, tampoco encontraremos en este terreno una explicación unificada. Esto se debe fundamentalmente a que la aplicación de la perspectiva evolucionista a la explicación de los fenómenos humanos ha sido terreno de fuertes controversias.[74] Si bien no todas estas controversias internas se han trasladado a los intentos de explicar el autoengaño en términos evolucionistas, esto sí ha ocurrido con un debate fundamental, que es el que opone a quienes consideran que el principio adaptacionista es la clave para comprender tanto la mente como las conductas humanas con aquellos que rechazan este supuesto. Sin ánimo de aspirar a una descripción detallada de estas posiciones, resulta necesario describir brevemente este debate, que se reflejará en las explicaciones del autoengaño.

El supuesto adaptacionista, dominante en las perspectivas evolucionistas en Psicología, puede ser caracterizado como el principio regulativo según el cual los sistemas mentales han surgido como adaptaciones, y se debe buscar ante todo el beneficio para la aptitud por el cual fue seleccionado un determinado rasgo. Desde el campo de la biología evolucionista, el supuesto adaptacionista ha enfrentado críticas referentes a que presenta

[74] Cfr. Gould, 1997, Fodor, 1998; Panksepp, 2000; Tattersall, 2001, para una serie de críticas a la perspectiva evolucionista en psicología y ciencias sociales.

una imagen poco plausible de los mecanismos de la evolución, al centrar la atención exclusivamente en las adaptaciones debidas a la selección natural y desestimar o minimizar la importancia de otros tipos de fenómenos. Entre estos otros fenómenos se cuentan los *spandrels*[75] (véase fig. 2), esto es, subproductos no adaptativos surgidos como consecuencia de la adquisición de rasgos adaptativos, y las exaptaciones, estructuras que contribuyen a la aptitud del organismo pero que evolucionaron por otras causas y fueron luego cooptadas para ese fin.[76]

Sobre la base de lo expuesto, presentaré aquí tres perspectivas evolucionistas sobre el autoengaño. La primera de ellas afirma que este fenómeno es una adaptación, esto es, es un rasgo seleccionado por su contribución a la aptitud del organismo que posee esta capacidad. La segunda sostiene que el autoengaño es un *spandrel* o subproducto estructural de otros rasgos de la arquitectura mental; en consecuencia, el autoengaño no contribuye, ni lo ha hecho en el pasado, a nuestro éxito reproductivo. La tercera, diferente pero de algún modo parcialmente compatible con las posiciones anteriores, considera que el autoengaño provee de beneficios para la aptitud, pero no necesariamente ha evolucionado como una adaptación.

Fig. 2

Spandrel: espacio de forma aproximadamente triangular resultante de la superposición de un domo sobre una estructura de arcos. No cumple ninguna función dentro del conjunto, sino que es un subproducto de necesidades estructurales.

[75] El uso del término "spandrel" para hacer referencia a esta clase de fenómeno se debe a Gould & Lewontin (1979).

[76] Cfr. Gould (1997) y Gould & Lewontin (1979), para una crítica ya clásica al adaptacionismo. Véase también Buss y otros (1998), para una evaluación de estas críticas. Cabe aclarar, no obstante, que el supuesto adaptacionista de la psicología evolucionista no debe ser interpretado como la tesis extrema según la cual todos los rasgos fenotípicos de un organismo son adaptaciones; en diversos escritos los psicólogos evolucionistas aceptan la existencia de *spandrels* (cfr. Cosmides & Tooby, 1997; sobre distintos modos del concebir el adaptacionismo, cfr. Godfrey-Smith, 2001; Atran, 2005).

1. El autoengaño como adaptación

Si bien varios autores se han inclinado hacia una explicación adaptacionista para el autoengaño (Byrne y Kurland, 2001; Stevens et al, 2006; Surbey, 2011) sin duda el más influyente y consecuente defensor de la tesis de que el autoengaño constituye una adaptación ha sido el biólogo evolucionista Robert Trivers. Trivers, reconocido por sus aportes a la comprensión de la evolución social, la cooperación y el conflicto, ha defendido esta posición desde la década del '70, mucho antes del surgimiento de la psicología evolucionista. En su prólogo al célebre libro de Richard Dawkins *El gen egoísta* (1976), Trivers sostuvo que:

> [S]i (según argumenta Dawkins) el engaño es fundamental en la comunicación animal, entonces debe existir una rigurosa selección destinada a detectar el engaño, y ello implica, a su vez, una selección que favorezca el autoengaño permitiendo que algunos hechos y motivos permanezcan en la esfera de la inconsciencia para no traicionar, mediante las sutiles señales del conocimiento de sí mismos, el engaño que se esté cometiendo. Así, el punto de vista convencional, que afirma que la selección natural favorece los sistemas nerviosos que producen imágenes cada vez más precisas del mundo, es una concepción muy ingenua de la evolución mental (pp. viii-ix).

En diversos artículos posteriores (1982, 2000, 2002, 2010) y en un libro (2011), Trivers ha mantenido esta posición, si bien con algunos cambios que no la han alterado en lo sustancial. Ha defendido una caracterización amplia (quizás excesivamente amplia) del autoengaño; para él, el término abarca no sólo los fenómenos intrapsíquicos a los que suele hacer referencia comúnmente, sino también a fenómenos limítrofes entre lo biológico y lo psíquico, y a hechos en los cuales los componentes sociales y culturales juegan un rol fundamental (como los desastres aéreos y las guerras). Dado que nuestro propósito en este capítulo es examinar específicamente las explicaciones evolucionistas acerca de la existencia del autoengaño, me limitaré a este aspecto de la teoría de Trivers; con este fin, tomaré como referencia la versión más sistemática y "académica" de su teoría, que aparece expuesta en un artículo en colaboración con el psicólogo William von Hippel (2011).

von Hippel y Trivers comienzan planteando diversas preguntas referentes al autoengaño: ¿por qué las personas se engañan a sí mismas?; ¿cuál es la arquitectura mental que hace posible que la misma persona sea tanto quien engaña como quien es engañado?; ¿cómo se manifiesta psicológicamente el autoengaño en sí mismo? Estas tres preguntas son

abordadas desde una perspectiva que sostiene que el autoengaño evolucionó básicamente como un medio para facilitar el engaño a otros y para la obtención de beneficios en la interacción social.[77]

Los autores parten de una concepción del autoengaño como un concepto que abarca varios tipos de procesos distintos, que son directamente comparables a los procesos involucrados en el engaño interpersonal; incluyen estrategias sesgadas de búsqueda de información y procesos sesgados de interpretación y de memoria. Lo que caracteriza a estas variedades de autoengaño es que las personas favorecen la información bienvenida por sobre la información indeseada de un modo que refleja sus metas o motivaciones.[78] También examinan casos de autoengaño a los que caracterizan como "clásicos", tales como la racionalización y el convencerse a uno mismo de que una mentira es verdadera. Las "variedades de autoengaño" a las que hacen referencia incluyen:

a. Búsqueda de información sesgada, en tres variantes: el monto de la información recolectada, la búsqueda selectiva y la atención selectiva.
b. Interpretación sesgada.
c. Recuerdos erróneos.
d. Racionalización.
e. Convencimiento al sí mismo de que una mentira es verdadera, variante que admite dos versiones: autoengaño acompañado de daño neurológico y autoengaño no acompañado de daño neurológico.[79]

El enfoque consistente en tratar al autoengaño como sesgos en el procesamiento de la información que dan prioridad a la información bienvenida por sobre la indeseable difiere, observan, de los enfoques clásicos del autoengaño que sostienen que el individuo autoengañado debe poseer dos representaciones de la realidad: una verdadera,

[77] El interés de von Hippel y Trivers no se limita a la función del autoengaño, sino también a los mecanismos que lo hacen posible. Respecto de esta cuestión, los autores examinan tres divisiones de la mente: memoria explícita versus memoria implícita, actitudes explícitas versus actitudes implícitas y procesos automáticos versus procesos controlados. Esos "dualismos mentales", como los llaman, no involucran por sí mismos autoengaño, pero cada uno de ellos juega un rol importante en posibilitar el autoengaño.

[78] Notemos aquí, aunque volveremos sobre este punto, el hecho de que esta caracterización no abarca al autoengaño negativo o retorcido.

[79] Esta enumeración resulta útil para recordar una limitación de algunos estudios psicológicos sobre el autoengaño, limitación que se reproduce en este trabajo de un modo particular. Si bien el artículo incluye un gran número de referencias a trabajos empíricos que los autores emplean para apoyar sus tesis, la elaboración conceptual de la categoría de autoengaño es sumamente limitada. El listado anterior es un claro indicador de ello: incluye una serie de fenómenos (en particular, el autoengaño acompañado de daño neurológico) que generan una categoría de límites sumamente difusos. Si bien es obvio que no puede prohibírsele a nadie que emplee el término con estos alcances, no parece forzoso que debamos aceptar que se está explicando de esta manera algún aspecto relevante del autoengaño.

preferentemente almacenada en la mente inconsciente, y una falsa, almacenada en la mente consciente. Por el contrario, las personas pueden engañarse a sí mismas evitando que la información indeseada sea inicialmente almacenada, y es posible mostrar claramente que este acto puede ser motivado. von Hippel y Trivers consideran importante destacar que no todos los sesgos en el procesamiento de la información son autoengañosos. Desde su perspectiva, sólo pueden ser considerados autoengañosos cuando favorecen la información bienvenida sobre la información indeseada de un modo que refleje las metas de la persona.

Como hemos señalado, los autores encuentran una cercana conexión explicativa entre el autoengaño y los procesos de engaño interpersonal. En la lucha por la obtención de recursos, una estrategia que ha surgido a lo largo de la evolución es el engaño. Las prácticas engañosas instigaron una lucha coevolucionista, dado que la selección favoreció en el engañado el desarrollo de nuevos medios para detectar el engaño, y nuevos medios de engaño en el engañador. En el engaño entre seres humanos, señalan, hay al menos cuatro categorías generales de señales que las personas pueden emplear para detectar intentos engañosos: los signos de nerviosismo, la supresión, la carga cognitiva y las fuentes idiosincrásicas. Los signos de nerviosismo resultan típicamente de la consideración de los costos potenciales de la detección del engaño. La supresión refiere al esfuerzo por evitar los signos no verbales de nerviosismo que revelarían el engaño por medio del control del rostro, torso y miembros; esto, a su vez, da lugar a otros indicadores, como un incremento en el tono de voz. La carga cognitiva tiene lugar cuando las personas mantienen simultáneamente dos tipos de contenidos en la memoria de trabajo; en este caso el engaño puede ser detectado a través de pistas asociadas a la carga, como los intervalos entre oraciones y la simplificación de la estructura de éstas. Por último, respecto de las fuentes idiosincrásicas, cada persona revela sus estados mentales de diferente forma, de modo que una persona familiarizada con quien intenta engañar también puede detectar indicadores específicos de engaño.

von Hippel y Trivers consideran que los seres humanos somos detectores competentes del engaño cuando se cumplen ciertas condiciones (por ejemplo, que se pueda interrogar a quien intenta engañar, o que el engaño tenga un costo para quien lo intenta). En consecuencia, la tesis central de su trabajo es que, por medio de engañarse a sí mismas, las personas pueden engañar mejor a otros, ya que cesan de emitir las señales de engaño mediado conscientemente que podrían revelar su intento. A la vez, el autoengaño evita el costo de la carga cognitiva, por lo que un corolario de la tesis anterior es que engañándose a sí mismas, las personas son capaces de evitar los costos cognitivos del engaño

conscientemente mediado. Un segundo corolario deriva de la existencia del castigo merecido para quien engaña en caso de que su intento sea descubierto. Este castigo parece tener, señalan los autores, profundas raíces evolucionistas, dado que las personas reaccionan con fuertes sentimientos de enojo y otras emociones negativas cuando advierten que están siendo engañadas. Ahora bien, dado que hay muchas razones por las cuales una persona puede no comportarse de acuerdo con lo esperado o deseado, una solución para la amenaza de castigo cuando un engaño es descubierto es la de cooptar tales razones mediante una apelación a la ignorancia o a la ineptitud en vez de al engaño. La atribución de un intento de engaño es crítica a la hora de determinar si la víctima del engaño siente enojo y busca castigar o está dispuesto a perdonar. El segundo corolario de la tesis central es que, engañándose a sí mismas, las personas pueden reducir el castigo si su engaño a otros es descubierto.

Además de lo anterior, los autores enuncian una segunda forma más general en la cual el autoengaño puede facilitar el engaño a otros, esto es, puede ayudarnos a convencer a otros de que somos mejores de lo que realmente somos. Entonces, los beneficios del autoengaño van más allá de convencer a otros de mentiras específicas, en la medida en que puede ayudarnos a incrementar las ventajas sociales más generales del automejoramiento [*self-enhancement*]. Las personas son impresionadas por la confianza que observan en los demás, y tal confianza es también determinante de la influencia social; de este modo, las personas que mayor confianza muestran resultan más creíbles, y es más probable que su consejo sea seguido en comparación con el proporcionado por personas que carecen de tal confianza. Las pruebas muestran, según los autores, que los sesgos de automejoramiento son visibles en una amplia variedad de dominios y estrategias y dentro de una amplia variedad de poblaciones; asimismo, las personas tienden a estar convencidas de que son mejores que el promedio. Pero además del sesgo de automejoramiento, las personas también subestiman a otros. De hecho, el automejoramiento y la subestimación a otros son caras de una misma moneda, en la medida en que las personas construyen su autoimagen mediante la comparación social.

Como síntesis del punto anterior, los autores enuncian una segunda tesis, la que sostiene que, por medio del engaño a sí mismos acerca de sus propias cualidades positivas y las cualidades negativas de otros, las personas son capaces de desplegar mayor confianza de la que de otra manera sentirían, posibilitándoles en consecuencia avanzar social y materialmente.

La existencia del autoengaño resulta así, en el enfoque de Trivers, explicada por sus beneficios adaptativos; no sólo proporciona ventajas comparativas en términos de la aptitud, sino que, justamente, tal capacidad ha sido seleccionada por proveer esta clase de ventajas para su poseedor. Esta posición, que adjudica claros beneficios a la capacidad para el autoengaño, sin reparar en sus posibles costos, ha sido objeto de una serie de objeciones muy incisivas.[80] Por mencionar sólo una, la omisión de un examen siquiera superficial del autoengaño negativo (determinada por la restrictiva definición inicial del fenómeno) genera fundadas dudas respecto de la adecuación de la explicación propuesta. Pospondremos, no obstante, una evaluación crítica de la teoría de Trivers hasta el final de este capítulo, donde la compararemos con las explicaciones evolucionistas alternativas.

2. El autoengaño como subproducto

Los autores que abogan por una perspectiva no adaptacionista acerca del autoengaño, esto es, quienes sostienen no sólo que la capacidad para el autoengaño no fue seleccionada por su contribución para la aptitud, sino que tampoco provee un valor adicional para aquella, parecen constituir un grupo minoritario (Kurzban y Aktipis, 2007; Mercier, 2011). Sin embargo, puede encontrarse una sólida defensa de esta posición en dos artículos de N. Van Leeuwen (2007, 2008).

Van Leeuwen comienza planteando el problema de la existencia del autoengaño de la siguiente manera. El conocimiento, observa, es *prima facie* crítico para el éxito evolucionista que nuestra especie ha tenido. Ahora bien, la extendida capacidad humana para el autoengaño socava el conocimiento. La existencia de tal fenómeno, en consecuencia, crea un dilema relativo a la manera de comprender el valor que el conocimiento en general posee para la aptitud. Si el conocimiento mejora la aptitud, entonces la capacidad para el autoengaño no debería existir, dado que socava el conocimiento. Pero, de hecho, existe. Si el conocimiento no incrementara la aptitud, entonces la capacidad para el autoengaño no constituiría un problema para la teoría evolucionista, pero en tal caso nos privaríamos de la más obvia explicación de nuestros rasgos más complicados e interesantes.

La respuesta de Van Leeuwen a este dilema (como él lo denomina) es defender una posición respecto del autoengaño que lo concibe como el subproducto de un número de habilidades humanas críticas que nos hacen posible conocer y comportarnos racionalmente

[80] Véanse al respecto los comentarios críticos al artículo de Trivers y von Hippel (2011) contenidos en el mismo número de la revista *Behavioral and Brain Sciences*.

dadas nuestras capacidades mentales finitas. Asimismo, tal concepción considera que la capacidad para el autoengaño constituye lo que Gould y Lewontin denominan un *spandrel*: es un subproducto estructural de rasgos que no fueron seleccionados por su rol en la producción de tal capacidad. Van Leeuwen no afirma que los rasgos que hacen posible al autoengaño sean en sí mismos adaptaciones; le resulta suficiente afirmar que existe una justificación independiente para postular su existencia. Ambas tesis son idénticas, observa, en la afirmación relativa al carácter de subproducto estructural del autoengaño, pero difieren en lo que sostienen respecto de tales rasgos: la primera tesis afirma que son críticos para la conducta y cognición racional finitas; la segunda, que no han surgido en la selección natural para producir el autoengaño. Van Leeuwen considera que su enfoque tiene ventajas sobre el de Trivers. En primer lugar, en la medida en que la capacidad para el autoengaño surge a partir de rasgos de la mente cuya existencia es plausible de modo independiente, su solución es más parsimoniosa debido a que no postula funciones adicionales; el trabajo que debe hacerse se limita a explicar la manera en que tales rasgos dan origen al autoengaño. En segundo lugar, su enfoque no descansa en un postulado empírico dudoso, esto es, que el autoengaño incrementa la aptitud.

Conviene comenzar por la definición que Van Leeuwen propone del autoengaño. Un agente se encuentra en un estado de autoengaño si y sólo si: 1) posee una creencia; 2) esa creencia es contraria a aquello que sus normas epistémicas, en conjunción con la evidencia que posee, usualmente dictarían; y 3) un deseo, que posee un contenido adecuadamente relacionado con la creencia en cuestión, genera una diferencia causal tal que la creencia es sostenida de un modo epistémicamente ilegítimo.

Sobre la base de lo anterior, el argumento central de Van Leeuwen en favor de la tesis del autoengaño como *spandrel* consistirá en definir lo que concibe como el complejo esencial de rasgos de la mente a partir de los cuales el autoengaño (el desiderativo en particular) surge. Tal complejo incluye siete rasgos:

EC1. Los deseos tienen un "aguijón" [*sting*] característico que acompaña la anticipación o la evidencia de no satisfacción.

EC2. Los seres humanos tienen la habilidad de atender selectivamente a las entradas de evidencia.

EC3. Los seres humanos tienen una inclinación general a evitar el malestar.

Estos tres rasgos pueden conducir al autoengaño; no obstante, Van Leeuwen observa que es importante destacar que el funcionamiento normal de EC1-EC3 no tiende a producir autoengaño. EC1 es conducente al logro de metas. EC2 es un rasgo esencial de

cualquier sistema cognitivo finito; sin él, la mente estaría invadida por una miríada de estímulos, muchos de los cuales serían irrelevantes para las metas del organismo. EC3 típicamente tiene la función de mantener al organismo lejos de situaciones que puedan ser dañinas y es generalmente conducente al logro de metas.[81]

EC4. La evidencia está estructuralmente organizada en la mente humana. No seríamos capaces de prestar atención a la evidencia de un modo selectivo si no estuviera organizada de forma tal que nos permitiera buscar en ella.

EC5. Los humanos forman creencias sobre la base de la evidencia en conjunción con normas epistémicas. Existe una profunda ironía, observa Van Leeuwen, en el hecho de que este rasgo racional de la mente es tanto subvertido como implicado por el propio autoengaño: es subvertido con respecto a la evidencia total, pero implicado en la formación de la creencia autoengañosa por la evidencia a la que se atiende selectivamente.

EC6. Los humanos experimentan placer ante la evidencia de que sus deseos serán satisfechos.

EC7. Los seres humanos buscan el placer. Debería resultar claro, señala Van Leeuwen, que este rasgo y el anterior contribuyen al logro de las propias metas, pero también están implicados en la modulación de la atención guiada por la búsqueda de comodidad [*comfort*] que produce el autoengaño.

Además de los rasgos enumerados, en la base del autoengaño se encuentran también los facilitadores y las fuentes de deseo. Un facilitador es un rasgo mental que hace más fácil la falla de las normas epistémicas y la evidencia en el autoengaño; una fuente de deseo da origen a deseos apropiados para generar un estado de autoengaño. Este segundo grupo incluye:

F1. La red de creencias tiene inercia. Un sistema de creencias no cambia globalmente con facilidad debido a la existencia de hechos que son anómalos desde la perspectiva de creencias particulares. Este rasgo es ampliamente ventajoso para la coherencia de nuestro sistema de creencias, ya que sin él éste experimentaría una revolución con cada descubrimiento de hechos anómalos. No obstante, también puede hacer posible al autoengaño.

F2. Los deseos y otras emociones generan conjeturas y pensamientos. En la medida en que las emociones y los deseos sugieren pensamientos, conjeturas e hipótesis al agente, también influencian la formación de creencias (razonamiento teórico). Este modo de

[81] En este apartado señala Van Leeuwen que el autoengaño ocurre muy raramente respecto de las creencias perceptivas; ocurre mucho más a menudo en relación a lo que se podría llamar intangibles.

influencia de los deseos y emociones sobre la formación de creencias es beneficioso para el logro de metas tanto cognitivas como prácticas, ya que sin la guía de la emoción dirigida hacia aquello a lo que atendemos nuestras capacidades cognitivas serían arrastradas a la consideración de una infinidad de información inútil.

F3. Los humanos pueden aplicar diferentes grados de escepticismo a diferentes proposiciones. El escepticismo –apartamiento de la creencia en una proposición que no satisface cierto nivel de justificación- admite diferentes grados. Ser escéptico es una capacidad racional, que ayuda a descartar falsas creencias. Pero aplicar el escepticismo de manera diferencial a diferentes proposiciones puede conducirnos a buscar en menor medida la verdad, toda vez que la aplicación diferencial del escepticismo es causada por deseos distintos que el deseo por la verdad. El malestar con la evidencia en favor de la creencia de que no *p*, causada por un deseo de que *p*, puede llevar a una persona a desear reclasificar esa evidencia; la aplicación diferencial del escepticismo puede ayudar en la reclasificación de la evidencia y consecuentemente en el autoengaño de que *p*.

F4. Los humanos pueden suprimir memorias no deseadas. Van Leeuwen apela aquí a las investigaciones de Michael Anderson, que muestran que esto es posible, así como al conocimiento de los procesos neurobiológicos subyacentes. La supresión no es equivalente al autoengaño, dado que puede tener la función de minimizar la distracción cuando debemos manejar una situación a la cual no pertenece la memoria suprimida. No obstante, la supresión puede facilitar el autoengaño, y que puede ser usada deliberadamente en el autoengaño para debilitar el recuerdo de evidencia contraria a la creencia deseada.

Por último, tenemos las fuentes de deseo, tercer conjunto de rasgos que se encuentran en la base del autoengaño:

DS1. Los humanos forman intenciones complicadas. Una intención juega tres roles en la vida mental. Primero, causa que las personas se involucren en razonamientos medios-fines acerca de cómo lograr la meta pretendida. Segundo, que se desestimen opciones que socavarían tal meta. Tercero, que uno siga la pista de la meta pretendida. Van Leeuwen sostiene que las intenciones, en virtud de jugar esos roles, engendran lo que llama *deseos subsidiarios*, esto es, deseos de estados de cosas que serían conducentes a la realización de la meta buscada. Por lo general los deseos subsidiarios no conducen al autoengaño, pero en ocasiones, especialmente en los casos de intenciones a largo plazo, nos orientarán hacia objetos no inmediatamente tangibles, pero aún así percibidos como importantes.

DS2. Los humanos experimentan disonancia cognitiva cuando su conducta difiere de las normas que rigen su autoconcepto. La disonancia cognitiva es el malestar que las personas sienten cuando sus conductas no se ajustan a sus concepciones de sí mismos como seres morales, competentes y consistentes; como un estímulo aversivo, queremos librarnos de ella. Tal deseo puede ser beneficioso para el logro de proyectos prácticos, dado que puede motivar al cambio de conductas inmorales, incompetentes o inconsistentes. Sin embargo, tal deseo puede ser también conducente al autoengaño, pues involucra contenidos no cercanos a la periferia sensorial de la red, que son igualmente percibidos como importantes.

El sostener que el autoengaño es un *spandrel* derivado de los rasgos descriptos, señala Van Leeuwen, no sólo es una explicación más parsimoniosa que la que lo considera una adaptación, ya que no postula funciones adaptativas adicionales; es también perfectamente consistente con la intuición que da origen a uno de los dilemas relativos al autoengaño, esto es, que el conocimiento incrementa la aptitud.

La capacidad para el autoengaño, concluye Van Leeuwen, no es una adición incidental a nuestra estructura cognitiva; proviene de la racionalidad en un contexto de finitud. Tenemos deseos que tienen la capacidad de generar autoengaño; si a esto se suma una evidencia mixta, nos encontraremos en un contexto propicio para este estado. ¿Significa esto que el autoengaño es inevitable para los humanos? Van Leeuwen observa que la propensión al autoengaño no puede ser eliminada; sin embargo, puede ser evitado en alguna medida mediante el cultivo de hábitos cognitivos que neutralicen los aspectos de la mente que lo hacen posible. Podemos confrontar evidencia que produce malestar y aceptarla tal cual es; también podemos reconocer el potencial de generación de autoengaño de nuestros deseos. El autoengaño, en síntesis, no es un fenómeno que se encuentre por completo fuera de nuestras posibilidades de control.

3. La autopercepción y el "efecto ganador"

Kido Lopez y Fuxhager (2012) sugieren una tercera explicación evolucionista alternativa a las propuestas por Trivers y Van Leeuwen. Sobre la base de las concepciones del primero, centran su análisis en el rol que juega el autoengaño en la modulación de la autopercepción, y no en el argumento según el cual el autoengaño hace más eficiente el engaño a terceros. La razón que aducen para este cambio de énfasis es que consideran

haber encontrado una conexión importante entre un fenómeno conductual llamado *efecto ganador* [*winner effect*] y el autoengaño, y que esta conexión no sólo defiende la tesis de von Hippel y Trivers de ciertas críticas presentadas por Van Leeuwen, sino que extiende y enriquece en gran medida sus alcances y mecanismos evolucionistas. Su argumento es que la conexión del autoengaño con el efecto ganador muestra que el primero puede *incrementar* la aptitud, o poseer *valor* o *beneficios adaptativos*. Si bien el argumento presentado deja abierta la posibilidad de que el autoengaño sea una adaptación, no es ésta su conclusión; en vez de ello, su objetivo es argumentar que el autoengaño, independientemente de cómo haya llegado a existir, posee la *capacidad* de incrementar el éxito reproductivo individual.

Mediante la adopción de la definición de autoengaño de Van Leeuwen (a la que consideran preferible), señalan que el argumento de von Hippel y Trivers que desean defender de las críticas de ese autor tiene dos pasos: a) el autoengaño conduce a una autopercepción positiva y b) la autopercepción positiva puede incrementar la aptitud. Dado que no es posible afirmar que el autoengaño es *necesario* para una autopercepción positiva, el debate deviene acerca de *cuán a menudo* conduce a tal autopercepción. Si fuese posible mostrar que el autoengaño conduce más frecuentemente al automejoramiento que a su contrario, entonces a) debería ser aceptado. Existe, sin embargo, un escollo para este proyecto, que son los casos de *autoengaño retorcido*. Sustentar la afirmación de que el autoengaño resulta más a menudo en creencias positivas acerca de uno mismo requiere dar razones para creer que el autoengaño no es usualmente del tipo retorcido. Los autores proporcionan tres razones tendientes a sustentar esta conclusión. En primer lugar, los ejemplos que Van Leeuwen emplea para generar su caracterización del autoengaño son ambos de la variedad no retorcida; segundo, en una búsqueda en diez artículos y libros sobre el tema, sobre un total de cuarenta y cinco ejemplos de autoengaño identificados, treinta y ocho pertenecen a alguna variedad estándar, por lo que queda sustentada la tesis de la preeminencia de tal variedad de autoengaño; por último, el autoengaño manifestado en estudios empíricos (como los de Quattrone y Tversky) es usualmente de la variedad estándar, aun cuando los sujetos tuvieron la oportunidad para engañarse a sí mismos de un modo retorcido. Una vez que las consideraciones precedentes se anexan a los argumentos según los cuales el autoengaño necesariamente debe ser acerca de uno mismo (Holton, 2001), puede concluirse que el autoengaño *usualmente* conduce a creencias positivas acerca de uno mismo.[82]

[82] Los autores desean permanecer agnósticos acerca del valor adaptativo del autoengaño retorcido, si bien hay razones para pensar que versiones más modestas de este tipo de autoengaño incrementan la aptitud (Hartung,

Sobre la base de la defensa de la tesis a), los autores consideran que están en posición de defender la segunda tesis, esto es, que las creencias positivas acerca de uno mismo conducen a beneficios adaptativos. von Hippel y Trivers argumentan, basándose en algunos estudios empíricos, que el automejoramiento causado por el autoengaño conduce a numerosos beneficios sociales. Como se señaló, las personas resultan impresionadas por la confianza en sí mismos que otros demuestran, y la confianza a menudo determina qué personas son elegidas como líderes o parejas románticas. Esto constituye un buen punto de partida para mostrar por qué el autoengaño puede ser útil a las personas en la sociedad contemporánea; no obstante, persiste la necesidad de elaborar un argumento adicional para sostener que el autoengaño puede tener valor adaptativo.

El fundamento para un argumento como el buscado puede hallarse en un fenómeno conductual conocido como "efecto ganador". Tal efecto es definido como el incremento en la capacidad para vencer en conflictos sociales que sigue a la obtención de victorias previas. Si bien se encuentra presente en diversas especies de mamíferos, peces, pájaros e invertebrados, su presencia y magnitud varía sustancialmente entre especies, e incluso entre especies estrechamente relacionadas. Tales datos sugieren que el efecto ganador es un fenómeno con el cual la selección interactúa. De hecho, lo anterior es respaldado por investigaciones más amplias que muestran tanto una refinada variabilidad intraespecífica como una heredabilidad evidente en los fenotipos agresivos y sus correspondientes mecanismos neurales y fisiológicos. Dado lo anterior, los autores consideran que es posible afirmar que el efecto ganador es capaz de haber evolucionado en sí mismo como muchos otros rasgos conductuales; no es meramente el resultado de diferencias individuales en la capacidad intrínseca para el combate, sino que es debido exclusivamente a un cambio en la experiencia relativa a las habilidades para vencer.

Uno de los roles del efecto ganador es ayudar a la formación de jerarquías sociales. Este nexo es, para los autores, el aspecto más importante puesto de manifiesto por las investigaciones descriptas acerca del efecto. La razón para esto es que la formación de jerarquías es un mecanismo a través del cual el efecto ganador incrementa la aptitud individual. Por lo tanto, la selección probablemente actúa sobre los mecanismos que subyacen al efecto ganador.

Los autores consideran que es esencial para su argumento general comprender las bases próximas del efecto ganador; esto es, cómo se forma el efecto ganador dentro de un individuo dado. Casi toda la investigación apoya la idea de que el efecto ganador es un

1988). Ahora bien, observan que si autoengañarse acerca de creencias negativas sobre uno mismo es beneficioso, entonces esto apoyaría su tesis más amplia de que el autoengaño tiene valor adaptativo.

producto de cambios dependientes de la victoria para el estado intrínseco de un individuo, y esos cambios alteran subsecuentemente el modo en que un individuo estima sus propias habilidades para la lucha. En consecuencia, esos cambios están enraizados en la alteración concurrente de la habilidad real para el combate y la habilidad percibida. El segundo de esos factores es de particular interés para los autores; indica que hay un efecto ganador psicológico que causa que los individuos se perciban a sí mismos como contendientes superiores respecto de otros en la misma población, independientemente de su habilidad real.

Un estudio que sugiere que existe un efecto ganador en los seres humanos constituye una ayuda adicional de gran importancia para la idea de que el mantenimiento de esas creencias puede ser funcionalmente importante de un modo que probablemente incremente la aptitud. En conjunto, ese estudio y los trabajos relativos al efecto ganador en vertebrados proveen un fuerte marco conceptual en el cual el acto de adquirir creencias positivas acerca de uno mismo puede ser considerado como un fenotipo adaptativo que puede ser modificado a través del tiempo por diversas presiones de selección. Debería notarse, observan, que el nexo entre el efecto ganador, la autopercepción positiva, el autoengaño y la aptitud es consistente con muchos de los estudios de Psicología humana citados por von Hippel y Trivers. Sin embargo, consideran que muchos de los estudios de la conducta animal que ellos resaltan en su trabajo extiende el alcance evolucionista de esas conexiones de un modo en que von Hippel y Trivers no lo hacen; esto es, conectan el efecto ganador con índices reales de aptitud.

Una dimensión de importancia no menor del argumento de Kido Lopez y Fuxhager es el intento de mostrar su posible compatibilidad tanto con la perspectiva del origen adaptativo del autoengaño como con la perspectiva que lo concibe como un *spandrel*. Específicamente, si bien Van Leeuwen sostiene que el autoengaño es un subproducto estructural, no desestima la posibilidad que su valor positivo para la aptitud evite que sea deseleccionado (y por lo tanto desaparezca); no obstante, considera que esta posibilidad carece de sustento. Las investigaciones sobre el efecto ganador, afirman, proveen de apoyo para esta tesis. En teoría, el mecanismo del autoengaño que han esbozado es incluso consistente con la idea de que el autoengaño evolucionó originalmente como un *spandrel* o como un subproducto de otros fenotipos cognitivos. Existen fuertes antecedentes respecto de que un cierto rasgo que aparentemente carece de valor adaptativo pueda coevolucionar con otro rasgo que sí lo tiene. Sin embargo, tal rasgo coevolucionado puede resultar

subsecuentemente adaptativo y por lo tanto ser favorecido por la selección bajo condiciones ecológicas nuevas. De este modo, parecería que el autoengaño es similar a una *exaptación*: un rasgo que evolucionó por una función que es cooptado para una nueva. Los autores consideran la posibilidad de que el autoengaño fuese originalmente un *spandrel* que luego fue cooptado para ayudar a los individuos a lograr autopercepciones positivas adaptativamente beneficiosas. Si esto es correcto, observan, el autoengaño no es una exaptación *per se*, pero sí algo similar. Si bien Van Leeuwen puede estar en lo correcto respecto de que el autoengaño fue originalmente un subproducto de otros rasgos, no necesariamente se sigue que nunca haya tenido valor adaptativo; sería incorrecto concluir que su razonamiento constituye un apoyo a la afirmación de que el autoengaño es un *spandrel*. Lo que les interesa mostrar es que su perspectiva es consistente con esta última perspectiva; el autoengaño, independientemente de sus orígenes evolutivos, tiene un valor adaptativo a causa de su conexión con el efecto ganador.

4. ¿Qué es el autoengaño, entonces, en términos evolutivos?

Tenemos entonces tres explicaciones evolucionistas acerca de la capacidad humana para el autoengaño. La posición posiblemente mayoritaria, expresada por von Hippel y Trivers, considera que la capacidad para el autoengaño es una adaptación, esto es, un rasgo seleccionado por su contribución para la aptitud del organismo que posee tal capacidad. Tal contribución es doble: por medio de engañarse a sí mismas, las personas pueden engañar mejor a otros, ya que cesan de emitir las señales de engaño mediado conscientemente que podrían revelar su intento; en segundo lugar, por medio del engaño a sí mismos acerca de sus propias cualidades positivas y las cualidades negativas de otros, las personas son capaces de desplegar mayor confianza de la que de otra manera sentirían, posibilitándoles en consecuencia avanzar social y materialmente.

El enfoque de von Hippel y Trivers deja varias cuestiones sin resolver. En primer lugar, su caracterización del autoengaño es tan abarcativa que incluye procesos de muy distintas clases, tanto normales como patológicos. En segundo lugar, no responde satisfactoriamente a una serie de críticas, algunas de las cuales preexistían a su artículo de 2011.[83] Algunas de estas críticas fueron formuladas por el propio Van Leeuwen:

[83] McKay y Dennet (2009) evalúan la tesis según la cual, dada la presencia sistemática de creencias falsas en nuestro sistema cognitivo, tales creencias deben ser adaptativas en el sentido evolucionista del término. En este examen revisan brevemente la posibilidad de que el autoengaño, como sostienen von Hippel y Trivers,

a. La habilidad para mentir de manera convincente tiene tan grandes beneficios para la aptitud que la contribución del autoengaño a ella pesa más que la pérdida de la aptitud que esperaríamos que acompañara a algo que reduce el flujo de información confiable. El problema es que el autoengaño debería ser altamente especializado (por ejemplo, asociado casi únicamente con situaciones en las cuales mentir es beneficioso) de modo tal que posea suficientes beneficios para la aptitud que posibiliten que sea seleccionado. Incluso si tal rasgo cognitivo fuese posible parece muy disímil al autoengaño que encontramos en el mundo real.[84]
b. Una segunda crítica a la concepción del autoengaño de Trivers se basa en el denominado "*grain problem*", esto es, la dificultad para determinar con precisión el problema para resolver el cual surgieron las supuestas adaptaciones que se desea explicar.
c. Por último, el que Van Leeuwen considera quizás el mayor de los problemas: Trivers concibe al autoengaño como una suerte de módulo (o módulos) que surgen en la historia de la evolución cognitiva humana. Presumiblemente, entonces, este módulo se activaría toda vez que los beneficios adaptativos de engañarse a sí mismo fuesen convenientemente cercanos. Tal punto de vista es problemático debido a que no logra relacionar el autoengaño con otros aspectos de la cognición.

Además de los argumentos de Van Leeuwen hay que recordar la falta de análisis, por parte de Trivers, de un examen de los casos de autoengaño retorcido. Como hemos visto en relación a las objeciones a las perspectivas intencionalistas sobre el autoengaño, podría negarse que los casos negativos sean casos de autoengaño, pero las similitudes entre ambos fenómenos son tan grandes que tal negativa debería contar con argumentos o pruebas muy sólidos, argumentos o pruebas que están ausentes en los trabajos de Trivers.

La posición de Van Leeuwen, como hemos visto, concibe al autoengaño como un *spandrel*, un subproducto estructural de un número de habilidades críticas que nos han

---

constituya una adaptación. Además de observar que la tesis de estos autores no cuenta con evidencia empírica que la respalde de modo directo, existe evidencia preliminar que contradice la afirmación del carácter adaptativo del autoengaño.

[84] Entiendo que a esta clase de objeción apunta Ramachandran (1996) de un modo más colorido. Sugiere considerar la siguiente situación. Supóngase que un chimpancé (chimpancé A) observa en una ocasión el lugar en el que el cuidador del zoológico deposita un gran racimo de bananas. El chimpancé A indica entonces al chimpancé B una dirección errónea, de modo tal de que sólo él pueda comer las bananas. Sobre la base del argumento de Trivers, supóngase ahora que el chimpancé A, que desea asegurarse que el chimpancé B no detecte la mentira, se involucra en un proceso de autoengaño, esto es, realmente cree que las bananas están en la errónea ubicación que ha indicado al otro chimpancé. Pero si esto es así, concluye, entonces el chimpancé A también buscaría las bananas en la ubicación errónea, lo que destruiría el propósito del engaño y resultaría obviamente desadaptativo.

posibilitado conocer y comportarnos racionalmente dadas nuestras mentes finitas, y que no ha sido seleccionados por su rol en la producción de esa capacidad. No afirma que los rasgos de los que deriva el autoengaño sean adaptaciones en sí mismos; es suficiente para él mostrar que hay justificación independiente para su existencia. El autoengaño, de esta forma, carece en sí mismo de valor adaptativo. Pese a estas cruciales diferencias, el enfoque de Van Leeuwen puede ser objeto de una objeción similar a la que hemos formulado respecto de la tesis adaptacionista de Trivers: no proporciona una explicación evolucionista para la existencia del autoengaño retorcido. La definición de autoengaño de Van Leeuwen, a diferencia de la de von Hippel y Trivers, no excluye *a priori* la existencia de esta clase de autoengaño. Sin embargo (aunque no de un modo definitivo), Van Leeuwen desiste de explicar la forma en la que el modelo que propone podría extenderse para acomodar los casos de autoengaño retorcido y mostrar cuáles serían los tipos de deseo constitutivamente involucrados que pudieran desencadenar la atención selectiva y otras clases de proceso que ha identificado en la producción del autoengaño en general. Permanece como una cuestión abierta, entonces, la pregunta relativa a si tal empresa puede tener éxito. No obstante, *prima facie* tal intento parece tener una ventaja respecto del adaptacionismo: dado que el autoengaño no es una adaptación, parece plausible la posibilidad de que tanto su variante desiderativa como su variante retorcida sean meramente subproductos que, de modo contingente, pueden resultar útiles en algunas ocasiones y perjudiciales en otras.

La perspectiva de Kido Lopez y Fuxhager, por último, sostiene que el autoengaño posee un valor adaptativo ligado al efecto ganador, pero no se pronuncian acerca de si constituye un rasgo seleccionado por su contribución a la aptitud (esto es, una adaptación), o si se trata de un rasgo que originalmente constituyó un subproducto estructural de otros componentes fundamentales del funcionamiento mental, y que luego fue cooptado para cumplir una función adaptativa. La capacidad humana para el autoengaño podría ser, en consecuencia, tanto una adaptación como un fenómeno similar a las exaptaciones. El argumento de Kido Lopez y Fuxhager, como hemos señalado, descansa en una premisa que puede pasar inadvertida debido al escaso espacio que se le dedica en el texto, esto es, el supuesto según el cual el autoengaño *necesariamente* incluye creencias sobre el sí mismo. Existen buenas razones para pensar que tal premisa es como mínimo dudosa (Fernández Acevedo, 2015); es muy cuestionable que todo caso de autoengaño efectivamente se trate de un caso de engaño respecto del sí mismo. Esta objeción puede sintetizarse de la siguiente manera.

Habitualmente se dice de alguien que está autoengañado respecto de *una* creencia, pero esta manera de describir el fenómeno podría generar la impresión engañosa de que quien se encuentra en estado de autoengaño sólo posee *una* creencia falsa. Ahora bien, como ha observado Oksenberg-Rorty (1988), el autoengaño se multiplica. El autoengaño parece requerir de algunas actitudes de segundo orden, en sí mismas también autoengañosas; cierto reconocimiento de los conflictos entre las creencias y algunas estrategias *ad hoc* para reconciliar estos conflictos. En otras palabras, es posible distinguir entre diversos niveles de creencias, ordenados según se trate de creencias principales sobre el tema del autoengaño (por ejemplo, una creencia falsa acerca de nuestro buen estado de salud), o creencias derivadas de la necesidad de dar coherencia y sustentar las creencias principales sobre el tema del autoengaño (por ejemplo, la creencia en que los médicos que nos han diagnosticado una enfermedad grave no son profesionales competentes). El hecho de que el autoengaño involucre un conjunto de creencias, por consiguiente, no elimina la posibilidad de establecer una organización o jerarquía dentro de ese conjunto. Parece útil hacer una distinción entre creencias centrales y creencias periféricas en los casos de autoengaño. Sin negar que el autoengaño implique la existencia de un sistema complejo de creencias, parece razonable suponer que no todas las creencias tienen la misma jerarquía. La que llamaremos "creencia central" es la creencia falsa fundamental del autoengaño (por ejemplo, que nuestra pareja nos es infiel, cuando no lo es). Las creencias periféricas serán el conjunto de creencias desarrolladas alrededor de la creencia central, consistentes y mutuamente sustentantes (por ejemplo, que el hecho de que nuestra pareja alguna vez se haya demorado más de lo normal en volver del trabajo es prueba sólida de engaño). Si se admite la distinción anterior, es posible clasificar en términos generales a los casos de autoengaño en dos tipos: aquellos en los cuales el objeto de la creencia central es algún estado del mundo, y aquellos en los cuales el objeto de la creencia central es el sí mismo. Es obvio que en los casos de autoengaño en los que la creencia central versa sobre nosotros mismos podremos decir que el estado de autoengaño incluye creencias erróneas sobre nosotros mismos. Sin embargo, esto es mucho más dudoso en el caso de que la creencia central verse sobre algún objeto del mundo externo. La admisión de que el autoengaño implica la existencia de un sistema de creencias erróneas no significa que las actitudes de segundo orden, el reconocimiento de conflictos entre las creencias y las estrategias *ad hoc* para reconciliar estos últimos impliquen la existencia en acto de todas las creencias correspondientes derivadas de esos procesos. Muchas creencias erróneas surgirán como consecuencia de diversos procesos subsecuentes a la adquisición de la creencia central falsa:

entre otros factores posibles, la confrontación de tal creencia por parte de personas cercanas, el hallazgo de evidencia que la contradice, o una personalidad tendiente a la duda y al autocuestionamiento. Supongamos una vez más el caso del individuo celoso convencido falsamente de la infidelidad de su pareja. Más allá de la dificultad en determinar las motivaciones de semejante estado, parece aceptable la idea de que los mismos mecanismos cognitivos que hacen posibles los casos de autoengaño "positivos" conducen al sujeto a adoptar una creencia falsa contraria a la evidencia disponible. Ahora bien ¿posee el sujeto efectivamente creencias erróneas sobre sí mismo, o estas creencias son un producto ramificado y no necesario de procesos cognitivos posteriores? Imaginemos un proceso en el cual la creencia del celoso es puesta en tela de juicio por un interlocutor externo. En particular, el interlocutor cuestiona la manera en que el celoso ha evaluado la evidencia a su disposición. Ante el cuestionamiento, el celoso responde afirmando su ecuanimidad y equilibrio tanto en la búsqueda como en la interpretación de la evidencia. Cabe preguntarse ¿poseía el celoso creencias falsas relativas a sí mismo, o estas creencias surgen en el proceso, proceso por otra parte no siempre presente? Salvo que se recurra a la dudosa (y posiblemente *ad hoc*) noción de que las creencias preexistían en forma inconsciente en la mente del sujeto celoso, y son simplemente explicitadas ante el cuestionamiento, parece más plausible la respuesta que sugiere que tales creencias no son más que una ramificación doxástica del proceso de autoengaño. Se puede admitir, también, que es verdad que un proceso de autocuestionamiento puede conducir a la formación de creencias falsas sobre sí mismo, pero este proceso no es necesario; al menos entiendo que la carga de la prueba recae en quien diga que los procesos de autoengaño *siempre* implican el autocuestionamiento o alguna otra clase de proceso generador de creencias de segundo orden y la consecuente formación de creencias falsas acerca de sí mismo. Es posible afirmar, en síntesis, que es muy plausible la idea de que el autoengaño en muchos casos (quizás incluso la mayoría) involucra la existencia de creencias falsas acerca de uno mismo, de una falla en el autoconocimiento; no obstante, parece poco plausible la idea de que *siempre* deben estar presentes estas creencias para poder hacer una atribución de autoengaño. Si la objeción anterior es correcta, entonces perdería gran parte de su apoyo la tesis de Kido Lopez y Fuxhager según la cual el autoengaño estándar conduce mayormente a creencias positivas acerca de uno mismo y, consecuentemente, el nexo entre el autoengaño y el beneficio para la aptitud.

¿Cuál es la evaluación que puede hacerse de estas posiciones? Creo que la posición según la cual el autoengaño es un subproducto es más sólida que las alternativas. Independientemente de las críticas que pueden formulársele a cada una de ellas, el enfoque propuesto por Van Leeuwen presenta dos claras ventajas respecto del adaptacionismo, y también respecto de la posición que postula beneficios adaptativos para el autoengaño.

En primer lugar, si bien la teoría no ha sido desarrollada aún con ese propósito, parece capaz, *prima facie*, de explicar al autoengaño negativo de una manera en la que las otras dos no parecerían poder hacerlo. Si, como hemos visto en el capítulo II, la adopción motivada de creencias negativas falsas es, indiscutiblemente, una variante de autoengaño, entonces parecería que toda teoría evolucionista comprehensiva sobre este fenómeno debería explicarlo, o al menos sentar las bases conceptuales para este logro. Sin embargo, dado el vínculo del autoengaño con el automejoramiento presente tanto en la teoría adaptacionista de Trivers como en la de Kido Lopez y Fuxhager, que reconoce beneficios adaptativos a aquel fenómeno, parece plausible la suposición de que una forma de autoengaño que en principio tiene consecuencias negativas para la imagen del sí mismo no puede ser explicado de modo satisfactorio.

En segundo lugar, la concepción de que el autoengaño es un subproducto parece ajustarse mucho mejor a los casos de autoengaño que es posible observar en el mundo real, independientemente de los casos hipotéticos (e incluso ellos mismos cuestionables, como vimos a través del ejemplo de Ramachandran) propuestos por los adaptacionistas. El autoengaño puede, de modo contingente, representar una ventaja a corto plazo para el organismo que es capaz de producirlo, pero también puede representar una fuente de costos mucho mayores a mediano y largo plazo.

Capítulo V. Mentiras vitales.[85] Las consecuencias prácticas del autoengaño

En los capítulos anteriores hemos examinado una serie de cuestiones teóricas relativas al autoengaño: la caracterización de este fenómeno, los mecanismos psíquicos que permitirían explicarlo y sus presuntos orígenes en la evolución de nuestra especie, entre otros. No obstante, al considerar su posible función defensiva, también hemos avanzado en la consideración de sus posibles consecuencias para nuestro bienestar. En este capítulo final examinaremos preguntas de una índole muy diferente a las primeras y en línea con las relativas a la posible función defensiva, esto es, nos ocuparemos aquí de las implicaciones y consecuencias que el autoengaño tiene para nuestra vida y para la de nuestros semejantes. De este modo, nos ocuparemos de las consecuencias del autoengaño para nuestra constitución moral y bienestar subjetivo, por una parte, y para nuestros sistemas de creencias, tanto individuales como colectivos. Tales implicaciones y consecuencias, sin embargo, dependerán en parte tanto de la caracterización del autoengaño como de las explicaciones que de él se consideren adecuadas. Esto es, las preguntas más puramente conceptuales y explicativas se encuentran íntimamente relacionadas con aquellas relativas a las consecuencias concretas del autoengaño.

1. Las implicaciones morales del autoengaño

En este apartado examinaremos el que, podría decirse, es el problema filosófico por excelencia sobre el autoengaño, y sobre el cual la filosofía mantiene su posición de privilegio.[86] Este es el problema relativo a las implicaciones morales del autoengaño, que ha interesado a pensadores de perspectivas muy disímiles, que van desde las reflexiones del Obispo Butler hasta el análisis de la "mala fe" de Sartre, pasando por Adam Smith y Kant. Las agudas observaciones volcadas por Butler (1726) en sus sermones sirven como útil punto de partida para el examen de las dimensiones morales del autoengaño.

[85] La expresión "mentira vital" fue acuñada por el dramaturgo noruego Henrik Ibsen en su obra "El pato salvaje". Con ella se hace referencia a una clase de mentiras que hacen posible la continuidad de nuestras vidas.

[86] Cfr. al respecto Fingarette (1969), Linehan (1982), Darwall (1988), White (1988), Tenbrunsel y Messick (2004), Kinghorn (2007) y Deweese-Boyd (2007, 2008), entre otros.

Butler, al igual que muchos autores que se han interesado por las implicaciones éticas del autoengaño, consideraba a este fenómeno como una severa amenaza para la moralidad. En la base del autoengaño se encuentra, en su opinión, una "auto-parcialidad" que explica comportamientos de otra manera incomprensibles: "no hay ninguna cosa relativa a los hombres y sus características más sorprendente e inexplicable que esta parcialidad hacia sí mismos, que es observable en muchos". Muchos hombres, observa, parecen perfectos extraños respecto de sus propias características; piensan, razonan y juzgan de manera muy diferente respecto de cuestiones relativas a sí mismos y a asuntos en los cuales no están interesados, ignorancia y parcialidad hacia sí mismos que puede darse en muy diferentes grados. Esta parcialidad, a su vez, conduce al autoengaño y a los males mayores que éste hace posibles. La interferencia del autoengaño sobre nuestra conciencia (que constituye una guía respecto de las deliberaciones y acciones morales) no sólo conlleva un perjuicio en sí mismo, sino que haría posible al individuo en estado de autoengaño actuar de maneras malvadas sin tener conciencia de sus defectos morales y la naturaleza de sus actos. Esto es, el autoengaño no sólo es negativo por el deterioro de la capacidad para conocer nuestra propia naturaleza sino, y en gran medida, por las graves consecuencias negativas que puede acarrear. La "ignorancia" que el autoengaño genera en el hombre hace posible que ejecute acciones que no elegiría en caso de ser consciente de sus verdaderos motivos. El autoengaño, en síntesis, destruye la moralidad y corrompe el carácter moral.

Contemporáneamente, el problema relativo a las implicaciones morales del autoengaño se ha escindido en una serie de preguntas diferentes, si bien, no cabe duda, la relativa a la responsabilidad moral por tal estado sigue siendo central. Van Leeuwen (2013) sugiere tres cuestiones centrales. Primero, cómo debería evaluarse la responsabilidad moral de un agente por las acciones ejecutadas sobre la base de sus creencias autoengañosas.[87] Segundo, si el autoengaño promueve la felicidad, cuestión de la que nos ocuparemos en el siguiente apartado. Tercero, si existe algo intrínsecamente incorrecto en el autoengaño. Supuesta la pertinencia de estas tres cuestiones, nos concentraremos principalmente en la primera, pero previamente haremos algunas consideraciones sobre la tercera.

Van Leeuwen (2013) introduce este problema de la siguiente manera. El agente virtuoso no tiene nada que ocultar ante otros o ante sí mismo, mientras que el malvado sí

[87] Esta formulación del problema, si bien legítima, no parece ser equivalente al problema de la responsabilidad moral *por* el autoengaño. Si bien es razonable suponer que el autoengaño estará, en la mayoría de los casos, acompañado por ciertas acciones destinadas a su mantenimiento, también es posible plantear qué responsabilidad tiene un agente por su estado de autoengaño, independientemente de las acciones (benévolas o malvadas) ejecutadas a partir de tal estado.

lo tiene, y el autoengaño puede ayudar a ocultar o incluso a sostener la maldad. Más aun, prosigue, existen estudios empíricos que apoyan la existencia de una conexión entre el autoengaño y la hipocresía moral o el juicio hacia las acciones de otros mediante el empleo de estándares que el agente no aplica para evaluar sus propias conductas. En consecuencia, la siguiente conclusión condicional parece razonable: si deseamos ser agentes morales, deberíamos evitar el autoengaño. Pero, prosigue, la conclusión sólo afirma que el autoengaño es a menudo instrumental para actos censurables; no responde a la pregunta relativa a si existe algo *intrínsecamente* malo o erróneo en el autoengaño.

Baron (1988) sugiere la siguiente respuesta al problema. Comienza por observar que muchas personas piensan que existe algo objetable en el autoengaño, pero no resulta claro qué es. El examen de algunos casos de autoengaño, incluso, torna menos clara la suposición de que haya algo *prima facie* malo en él. Por ejemplo, una persona enferma de HIV que se autoengaña al pensar que sus probabilidades de recuperación son del cincuenta por ciento puede estar manejando una situación difícil de un modo adecuado. Sin embargo, la existencia de estos casos no elimina la presunción de que a menudo el autoengaño es cuestionable, y la pregunta a responder es qué es lo que lo hace así. Podría pensarse que el autoengaño es incorrecto por las mismas razones que es malo el engaño interpersonal. Sin embargo, Baron considera que el autoengaño y el engaño interpersonal son incorrectos por diferentes razones, y que el estatus moral del autoengaño es algo superior al del engaño interpersonal.

El engaño a otros y el autoengaño parecen compartir algunas características. En primer lugar, ambos involucran una misma falta de consideración hacia la verdad, si bien en el segundo tal falta de consideración no es abiertamente adoptada. En segundo lugar, ambos se asemejan en que en ocasiones, aunque no siempre, causan un perjuicio a otras personas. En el caso del autoengaño quizás esto no parezca tan visible, pero resulta innegable que sí puede ocurrir en casos en los que el autoengaño impide que el agente ejecute acciones que podrían resolver o mejorar problemas dentro de su alcance; así, por ejemplo, si un hombre cree de modo autoengañoso que su hijo es un poco inmaduro para su edad, y no que padece un retraso, podría omitir las acciones correctivas pertinentes. Asimismo, tanto el autoengaño como el engaño a otros no siempre resultan perjudiciales. Dadas estas semejanzas, podría suponerse que el saber qué es lo que está mal en el autoengaño requiere la determinación de aquello que está mal en el engaño a otros, con las modificaciones menores que se requieran. Sin embargo, si se examinan con más detalle las semejanzas, emergen diferencias importantes entre ambos. La primera de estas diferencias

reside en la importancia de la indiferencia por la verdad: según Baron, esta indiferencia es más importante para la naturaleza incorrecta del autoengaño que en la del engaño, en parte porque hay mucho más de malo en el engaño que en el autoengaño. El engaño a otros es malo aproximadamente por las mismas razones por las que la manipulación a otros lo es; el hecho de que el engaño, a diferencia de la manipulación, involucre el conducir a alguien a que crea que algo es falso sólo añade algo más de incorrección. La segunda diferencia, estrechamente relacionada, se vincula con la naturaleza del daño causado a otros. En el caso del engaño, el otro es tratado no como un agente sino como un sujeto a ser conducido en la dirección que el engañador prefiere; éste se arroga el poder sobre otro sin revelar su propósito. Sin duda, hay mucho de cierto en la idea de que cuando el autoengaño daña a otros, el daño es menos directo que el daño causado a otros por medio del engaño.

Baron considera que lo que hay de malo en el autoengaño, a diferencia del engaño, es que el primero deteriora gradualmente la agencia.[88] Esto ocurre de dos modos. En primer lugar, el autoengaño puede convertirse en un hábito al cual el agente recurre demasiado a menudo; en segundo lugar, el autoengaño requiere, para ser eficaz, de más autoengaño. La necesidad de ver la realidad de cierto modo, pese a la evidencia en contrario, resulta crecientemente demandante, lo que conduce al agente a interpretar lo que percibe de un modo que apoye el engaño. Al hacer esto, corrompemos nuestros procesos de formación de creencias y gradualmente deterioramos nuestra naturaleza como agentes responsables. Si bien no siempre el autoengaño es tan fecundo, observa Baron, en la medida en que la motivación para éste no desaparezca mayor será la estructura de creencias y actitudes que resulta necesario sostener para evitar considerar seriamente la creencia verdadera, y menor disposición existirá para renunciar a la creencia falsa.

Baron agrega una segunda razón por la cual el autoengaño resulta malo. Existen casos en los que el autoengaño sirve de "escudo" al agente, impidiendo el reconocimiento de algo a lo que, moralmente, debería prestar atención. Aquello a lo que debería prestarse atención puede ser algo relativo a la propia conducta o naturaleza (por ejemplo, una adicción o la tendencia a implicarse en relaciones afectivas con personas dominantes y abusivas) o a otras personas (por ejemplo, el caso de una mujer que se niega a aceptar que su pareja ha abusado de su hija).

[88] Baron considera que el deterioro de la agencia, en el caso del engaño a otros, no ocurre de modo gradual, sino de modo episódico y, en general, no acumulativo. Esta afirmación parece como mínimo discutible, debido a que ciertos tipos de engaño a otros (por ejemplo, mentiras destinadas a encubrir acciones tales como una infidelidad prolongada en el tiempo) parecerían, *prima facie*, tener ese mismo efecto perjudicial de modo gradual. No obstante, el punto que nos interesa aquí es aquello que hace que el autoengaño sea erróneo, y no aquello que lo diferencia del engaño a terceros.

El autoengaño, en síntesis, corrompe los procesos de formación de creencias (y, con ello, la propia agencia) y socava el sentido de la responsabilidad hacia uno mismo y hacia los otros. No obstante, advierte Baron, esto no implica que todo de autoengaño sea objetable en sí mismo.

El análisis de Baron deja, a mi modo de ver, algunas preguntas abiertas respecto de la naturaleza moralmente cuestionable del autoengaño. Ella y otros filósofos admiten que, en ciertas oportunidades, el autoengaño puede servir para proteger a la persona de sufrimientos innecesarios. Ahora bien, Baron observa que la existencia de casos "aceptables" de autoengaño no elimina la presunción de que hay algo malo o moralmente cuestionable en este fenómeno. Cabe preguntar, dada la existencia de casos "aceptables", ¿deberíamos admitir que la naturaleza moralmente cuestionable del autoengaño es suficiente para condenarlo? ¿O deberíamos tener en cuenta, como cuestión fáctica, si los casos en los que el autoengaño es beneficioso son muchos menos que aquellos en los que, como sostiene Baron, deteriora la agencia y debilita nuestra responsabilidad hacia nosotros mismos y hacia otros (o, como diría Butler, corrompe el carácter moral)? La duda anterior podría fortalecerse si se tiene en cuenta la presunta contribución positiva que ciertos tipos de autoengaño podrían tener para nuestra salud mental y felicidad (Taylor y Brown, 1989). Entonces, si el autoengaño es negativo porque corrompe la agencia y nos ciega ante situaciones que deberíamos atender por razones morales, pero a la vez tiene consecuencias positivas, ya que contribuye con nuestra salud mental y felicidad, ¿podría decirse que hay algo malo en el autoengaño? Si el autoengaño puede constituir una ventaja en estos últimos sentidos, ¿no sería posible decir que simplemente es un fenómeno que implica ciertos riesgos morales?

Podemos conjeturar aquí que la respuesta de muchos autores (lo que incluye seguramente a la propia Baron), aun en caso de que reconocieran los presuntos beneficios que el autoengaño tiene en general para nuestra salud mental y felicidad,[89] sería igualmente negativa: el autoengaño es intrínsecamente cuestionable desde una perspectiva moral, más allá de cualquier contribución positiva que pueda acarrear. Lo que implica la presunción anterior, creo, es que el debate acerca de la naturaleza moral presuntamente cuestionable del autoengaño conduce a una controversia fundamental de la teoría ética, esto es, la que opone a las perspectivas deontológicas y consecuencialistas respecto de las acciones morales. Para las primeras, si el autoengaño es moralmente erróneo, lo es

[89] Contribuciones, como veremos en el próximo apartado, en absoluto exentas de controversias.

independientemente de cualquier consecuencia positiva que pudiese tener, ya que implica la violación de algún principio absoluto (por ejemplo, el *dictum* "conócete a ti mismo"). Para los consecuencialistas, por el contrario, la naturaleza moral del autoengaño debe ser evaluada sobre la base de sus consecuencias positivas; de este modo, los beneficios efectivos que el autoengaño conlleva para nuestra existencia podrían justificar su carácter moral. Como ocurre con otros debates éticos fundamentales, no existe una respuesta concluyente dentro de la teoría ética; la naturaleza moral del autoengaño, por ende, quedará abarcada por el mismo interrogante.

Pasemos ahora a la consideración del tercer problema moral relativo al autoengaño. Como se ha mencionado, la posición clásica respecto de este problema es que quien se autoengaña es moralmente responsable por su estado, y es fácil comprender por qué. Si el autoengaño es concebido bajo el modelo del engaño interpersonal, en el cual quien engaña lo hace intencionalmente, quien se autoengaña lo hace también intencionalmente, y es en consecuencia moralmente responsable por su estado. Demos (1961), por ejemplo, considera que un hombre que se miente a sí mismo es censurable porque actúa con conocimiento de los hechos y en consecuencia puede ser considerado responsable de su creencia errónea. Usualmente, una persona se engaña a sí misma dado que encuentra displacentero creer lo que de hecho ocurre; por ejemplo, una madre que se engaña a sí misma respecto de su hijo, creyendo que es un buen chico, cuando en realidad no lo es. Demos observa aquí "tal mentira a uno mismo es un ejemplo de lo que es llamado pensamiento desiderativo" (p. 589). Sin embargo, agrega, esto no es necesariamente así; una persona puede persuadirse a sí misma para creer en algo desagradable, y menciona el caso de una persona que camina sola en el bosque en tinieblas y se imagina bestias salvajes que lo acechan.[90] Sería más exacto, entonces, hablar de impulsos o pasiones como influencias sobre la creencia. Demos dice hablar de "influencia" *prudentemente*; sería erróneo decir de alguien que está "abrumado" por la pasión. Se cede a un impulso en el sentido de que uno no intentó resistir suficientemente a él cuando, de hecho, se podría haber intentado resistir y haberlo logrado. Entonces: a) se dice que una persona es responsable por una acción debido a que ha decidido o elegido hacerla; b) Se sostiene que una persona es responsable cuando simplemente consiente el acto; c) un tercer caso, más débil, ocurre cuando el

[90] Este podría ser un ejemplo de lo que se podría llamar "pensamiento aprensivo", como fenómeno opuesto al pensamiento desiderativo. El hombre del ejemplo no tiene (al menos bajo esa descripción) evidencia en favor o en contra para considerar que efectivamente animales salvajes lo están acechando; su pensamiento parece ser infundado. Cfr. Van Leuween (2008) para la distinción entre autoengaño retorcido y "pensamiento aprensivo". Esta distinción, no obstante, no quita pertinencia al análisis de Demos.

hombre "se deja hacer a sí mismo"; por ejemplo, cuando claudica ante un impulso momentáneo. Aun en este caso es responsable por su acto. Tanto intentar como no intentar resistir el impulso están dentro de los propios poderes de la persona. Lo mismo ocurre con el autoengaño: un hombre desea (o está inclinado) a creer que *p*, contrariamente a lo que sabe que ocurre. Eventualmente llega a creer que *p*, debido a que ha cedido al impulso de obtener una ventaja; podría haber resistido al impulso, pero no lo intentó con la suficiente firmeza.

La autocausación, continúa Demos, no es condición suficiente para la responsabilidad humana; debemos ser capaces de decir que la persona sabía lo que estaba haciendo, y podría haber obrado de otra manera. En consecuencia, a) no sostendríamos que un cleptómano es responsable por haber robado comida de un almacén, aunque podemos afirmar que su "manía" es debida a alguna afección en su cerebro; b) encontramos a una persona responsable por sus creencias erróneas cuando el error es debido a su negligencia en consultar a las autoridades apropiadas o a las fuentes pertinentes en una biblioteca. Este último caso debe ser distinguido de c), caso en el cual una persona practica el engaño sobre sí misma. Mientras que en el caso anterior la persona es directamente responsable por su negligencia, en el cual la creencia errónea es sólo un resultado accidental, en éste el nexo causal con la creencia es directo.

La perspectiva según la cual somos responsables por nuestro autoengaño parece seguir siendo dominante hoy en día, pese al surgimiento de enfoques radicalmente opuestos sobre el fenómeno, como los sugeridos por los deflacionistas. Sin embargo, hay autores que sostienen que, bajo estos nuevos enfoques, quien se autoengaña no debe ser considerado responsable por su estado. Neil Levy (2004) es quizás el más fuerte defensor de esta tesis, y a sus argumentos nos dedicaremos en lo que sigue.

Levy observa inicialmente que quien se autoengaña es ampliamente considerado culpable por su estado. Si bien todos cometemos errores, no se presupone que somos responsables cuando los cometemos. Pero, agrega, el autoengaño es un tipo particular de error, en el cual la víctima tiene mucho de responsabilidad. La perspectiva según la cual quienes se autoengañan son responsables por su estado tiene sus raíces en una concepción del autoengaño que puede rastrearse hasta el siglo XVII. Esta concepción fue desarrollada contra el trasfondo del yo entendido como transparente para sí mismo. Hoy, con el eclipse de tales certezas cartesianas, esta concepción del autoengaño está siendo gradualmente reemplazada por perspectivas que sostienen que el autoengaño es algo que nos acontece sin nuestra intención. Levy considera que, sobre la base de estas concepciones, se debe

abandonar el supuesto de la responsabilidad típica de quien se autoengaña. El autoengaño debe ser asimilado a la categoría de los errores, en los que no hay característicamente una presuposición de responsabilidad.

Levy señala que casi todos los autores, lo que incluye a los filósofos que sostienen concepciones deflacionistas del autoengaño, continúan creyendo que el autoengañado es característicamente culpable por su estado, que el autoengaño es típicamente malo, y que quienes se autoengañan son censurables por su engaño. Aclara que la responsabilidad que está en juego en este debate es la responsabilidad *moral* y no la responsabilidad meramente causal. Alguien es moralmente responsable por una acción (o una creencia) si es un objetivo apropiado para las actitudes reactivas con respecto a esa acción. El autoengaño, se supone usualmente, es moralmente erróneo; en consecuencia, quienes se autoengañan deben ser censurados. Sin embargo, objeta Levy, no es obvio que quienes se autoengañan son siempre, o incluso típicamente, responsables por su autoengaño. Si la concepción tradicional fuese correcta, y el autoengaño una actividad intencional, entonces la atribución de responsabilidad sería sencilla. No obstante, dado que el autoengaño no es intencional, requiere de cierto trabajo mostrar que quienes se autoengañan son, pese a eso, responsables por su estado.

Levy observa que las acciones que conducen o sostienen el autoengaño deben ser intencionales, al menos contrafácticamente, si hemos de ser responsables por nuestro autoengaño. La responsabilidad requiere *control*. Puede apelarse, señala, a una conocida concepción de la responsabilidad moral según la cual ejercemos "control guía" sobre nuestras acciones si reconocemos razones, incluyendo razones de índole moral, para actuar de otro modo, y lo haríamos (o intentaríamos hacerlo) en alguna otra situación. Entonces, tenemos control sólo cuando podemos intentarlo; no es una condición necesaria para tener control sobre una acción *a* que esa acción sea intencional, sólo que podríamos haber intentado *a*, o que podríamos haber omitido intencionalmente *a* (incluso nuestros reflejos podrían estar bajo nuestro control intencional en este sentido si, por ejemplo, está bajo nuestro control el ser parte de una situación en la cual nuestros reflejos pueden ser provocados). Para que un agente sea moralmente responsable por su autoengaño debe poseer un grado de control real o contrafáctico sobre él.

Sobre la base de lo anterior, Levy examina dos cuestiones. Primero, la responsabilidad por episodios individuales de autoengaño; segundo, la responsabilidad por ser un tipo de persona que tiene una disposición a autoengañarse. Nos ocuparemos exclusivamente de la primera.

Levy señala que, al igual que las creencias autoengañosas, las creencias en general están en cierta medida bajo un control indirecto. Es posible, por ejemplo, involucrarse en ciertas actividades que están diseñadas para producir creencias. Hay que advertir, no obstante, que no es posible en general intentar formarse creencias *falsas*, como no sea mediante el uso de medios autónomos.[91] La atribución de culpa por las falsas creencias es apropiada, observa, cuando puede ser rastreada hasta un acto de negligencia epistémica deliberado (por ejemplo, el agente consultó a sabiendas una fuente obsoleta). Por consiguiente, quien se autoengaña es responsable por su estado en la medida en que es el resultado de un acto u omisión conocidos por él. Sin duda esta condición es satisfecha *en ocasiones* por quienes se autoengañan.

Pese a lo anterior, observa Levy, muchos teóricos (incluyendo a Mele) continúan sosteniendo que quien se autoengaña es *típicamente* responsable por su estado. Mele, por ejemplo, señala que quienes se autoengañan son típicamente responsables debido a que el tipo de mecanismos de distorsión usualmente en funcionamiento están en alguna medida bajo nuestro control. Por ejemplo, si conocemos la existencia del sesgo de confirmación, podemos instruirnos a nosotros mismos de modo de minimizar sus efectos sobre nuestro pensamiento. Por supuesto, no podemos librarnos por completo de su actuación, pero cuando hay cuestiones importantes en juego, podemos recordarnos a nosotros mismos de su potencial de sesgo, y actuar en consecuencia.

Levy objeta que algunos de los mecanismos de distorsión descubiertos por los psicólogos sociales son considerablemente contraintuitivos, al menos en algunas de sus operaciones. En esa medida, parecería que no podemos ser culpables por no lograr enseñarnos a nosotros mismos el modo de evitarlos. Es verdad, reconoce Levy, que a veces nos recordamos a nosotros mismos el considerar ambos lados de una cuestión antes de formarnos una opinión. Sin embargo, lo hacemos sólo bajo ciertas condiciones. Esas condiciones son justamente las condiciones bajo las cuales alguien es culpable de engañarse a sí mismo. En su opinión, tales condiciones son: a) la materia relativa a la creencia es importante (por sus implicaciones morales, o de alguna otra clase), y b) tenemos dudas acerca de su verdad; las denomina condición de importancia y de duda. Levy afirma que ambas condiciones, y no sólo una, deben ser satisfechas antes de que podamos encontrar a alguien moralmente responsable por su creencia autoengañosa. Advierte, además, que los casos de autoengaño que satisfacen estas dos condiciones son muy inusuales. En

[91] Para el concepto de "medios autónomos" véase el capítulo II, § 2.

consecuencia, los casos en los cuales podemos hacer una atribución de responsabilidad moral por el autoengaño son, también, muy inusuales.

Esta posición, como el mismo Levy reconoce, es la posición minoritaria respecto de la responsabilidad moral de quien se autoengaña,[92] y por supuesto no ha escapado a las críticas. Van Leeuwen (2013) sugiere considerar los siguientes ejemplos para examinar esta cuestión. Jeff, un político que es candidato para un cargo electivo, ataca a un oponente por promover medidas que dañarían a las escuelas públicas. El ataque estaría justificado en caso de que fuese verdad que las medidas propuestas por su oponente fueran a dañar las escuelas públicas; sin embargo, esto no lo que ocurre: en realidad las medidas propuestas por el otro candidato ayudarían a las escuelas. Jeff, que desea tener algún elemento para atacar a su oponente, resulta autoengañado acerca de las medidas propuestas por éste. ¿Es moralmente censurable por su ataque? Considérese ahora, propone, dos casos relacionados. A) El jefe de campaña de Jeff, inmediatamente antes de un debate, hace llegar a éste un informe en apariencia creíble que describe el modo en que las políticas de su oponente resultarían perjudiciales. Si Jeff cree sinceramente en el informe y ataca a su oponente de manera acorde, el jefe de campaña, y no Jeff, sería el culpable. B) Jeff miente de manera voluntaria y consciente. Aunque sabe que las políticas de su adversario serían beneficiosas, miente afirmando lo contrario. En este caso, Jeff es claramente censurable. El problema en cuestión, afirma Van Leeuwen, puede ser planteado en estos términos: ¿es el caso de autoengaño de Jeff más parecido a A) o a B)? Claramente, señala, la respuesta a esta pregunta depende de aquello que sea el autoengaño. La línea de pensamiento que identifica con la propuesta de Levy y otros filósofos, sostendría que Jeff no es responsable por su autoengaño. Su caso sería más parecido a A), situación en la cual Jeff es engañado y su creencia falsa no ha sido el resultado de su intención. No obstante, la pregunta relativa a su falta de culpa no debería recibir aún una respuesta afirmativa. Levy observa que el pensamiento según el cual sólo somos culpables por aquellas cosas que yacen clara y completamente dentro de nuestros pensamientos conscientes es un dogma erróneo de la teoría moral. Contraejemplos familiares a esta teoría, observa, son la negligencia y la perversidad. En esta línea de pensamiento, Van Leeuwen señala que la mala postura

[92] En el capítulo II señalamos que una de las objeciones al deflacionismo es, justamente la carencia de una explicación convincente de por qué típicamente quienes se autoengañan son considerados responsables por su estado y sujetos a la crítica, al menos en muchos casos. Quizás podría aplicarse aquí la observación de Susan Haack relativa a que lo que es un *modus ponens* para un filósofo es un *modus tollens* para otro: si, para el intencionalista, la imposibilidad de atribuir responsabilidad moral debe conducir al rechazo del deflacionismo, para el deflacionista (al menos para Levy) debe concluirse que quien se autoengaña no es responsable por su estado.

constituye una útil analogía. La mala postura es un hábito con consecuencias negativas para la salud. ¿Puede ser alguien que posee tal hábito responsable por sus consecuencias negativas? Tal persona podría argumentar que no es su culpa, ya que nunca pretendió tenerlo. Sin embargo, aun sin pensar en su postura cada minuto del día, podría ser suficientemente consciente de ella para corregirla. Corregir la tendencia hacia el autoengaño, señala Van Leeuwen, consistiría en algo similar.

La pregunta por la moralidad del caso de Jeff, prosigue, sería entonces más parecido al caso siguiente: A') El jefe de campaña de Jeff tiene el hábito de hacerle llegar documentos que distorsionan las políticas de su oponente, y Jeff tiene es en cierta medida consciente de este hábito. Jeff no se molesta en evaluar tales documentos. ¿Es, entonces, censurable por su ataque? Parecería que la respuesta es afirmativa, dado que está a su alcance la posibilidad de evaluar los documentos. Permanece como una cuestión ética abierta, dice Van Leeuwen, si la moral de Jeff en el caso A') es tan mala como la de Jeff en el caso B).

El autoengaño, concluye, aun cuando no sea intencional, es el resultado de malos hábitos cognitivos, de los cuales es posible tomar conciencia. Podría ser evitado mediante la elección de no soslayar evidencia incómoda, la adhesión de modo activo a definiciones consistentes o el cultivo del hábito de considerar como patrones aquello que tomamos por excepciones. De modo similar, es posible comprometerse con estándares probatorios consistentes para creer en una proposición. En la medida en que exista alguna posibilidad de control sobre el autoengaño, habrá también alguna responsabilidad por sus consecuencias.

Las diferencias entre la posición minoritaria, expresada por Levy, y la mayoritaria, defendida por Van Leeuwen, parecen claramente una cuestión de grado y no de clase. Levy admite que existen casos en los cuales somos responsables por nuestro autoengaño (si bien son muy raros). Van Leeuwen considera que, en la medida en que tengamos cierto control sobre nuestros hábitos cognitivos, es posible una atribución de responsabilidad moral por el autoengaño; no sostiene, hasta donde podemos ver, que *siempre* es posible realizar esta atribución. Ahora bien, y de modo condicional, Levy podría tener razón en un punto: si es verdad que los mecanismos cognitivos de distorsión de la formación de creencias son (además de múltiples),[93] en muchos casos, notoriamente contraintuitivos, la posibilidad de que seamos capaces de controlar su actuación de manera irrestricta parece mucho más

[93] Cfr. Thagard (2011).

remota. En tal caso, y más allá de que tengan razón quienes sostienen que el deflacionismo no elimina por completo la responsabilidad de quien se autoengaña, podría decirse que sin duda la responsabilidad parece ser mucho más difusa que en el caso en el cual el autoengaño fuese intencional.

Sin embargo, es posible pensar una alternativa intermedia entre las dos posiciones, pero no en términos meramente cuantitativos. Podría ser el caso, como con frecuencia se observa en ciertos fenómenos cercanos al autoengaño en el campo de la psicopatología, que un agente tienda de modo sistemático a formar creencias falsas referidas a un área vital específica. En tal caso, la tendencia al autoengaño sería "idiosincrásica" y temática: el agente no se autoengañaría respecto de *cualquier* creencia, sino sobre ciertos tipos de creencias, restringidas según su contenido. Así, si un agente tiende de modo sistemático, por ejemplo, a formarse creencias negativas acerca de su estado de salud (que van desde simples pensamientos aprensivos hasta creencias autoengañosas firmemente arraigadas acerca de padecer una enfermedad grave) podría ser posible, en principio, que tome gradualmente conciencia acerca de las distorsiones que plagan sus procesos de formación de creencias y lo conducen a aceptar creencias mucho más negativas que lo que está justificado en creer en virtud de la evidencia que posee. Lo anterior podría no ser una posibilidad solamente teórica; es posible pensar que, si una persona tiende de modo sistemático a formar creencias falsas acerca de un campo determinado, existirán más posibilidades de que su autoengaño sea advertido y comunicado por quienes actuarían, en ese caso, como jueces externos de la verdad de sus creencias. Si el agente efectivamente tomara conciencia de las tendencias idiosincrásicas de su autoengaño, podría ser posible en principio considerarlo responsable por su estado.

La posibilidad anterior no excluye la existencia de autoengaño en áreas del pensamiento que el agente no ha identificado como afectadas por sesgos firmemente establecidos. Esto es, no hay nada que impida pensar que un agente que posee tendencias sistemáticas a cierto tipo de autoengaño se encuentre, también, autoengañado respecto de creencias ajenas a tales tendencias. En tal caso, resultaría mucho más difícil atribuirle una responsabilidad por la generación de creencias falsas en las que intervienen, como hemos señalado, sesgos que no sólo se encuentran regularmente fuera de nuestro control consciente sino que, además, resultan muy escasamente cercanos a la intuición. Podría concluirse, en consecuencia, que el deflacionismo parece minar la posibilidad de realizar de modo general una atribución de responsabilidad moral por el autoengaño, pero no permite eliminarla por completo.

2. *Autoengaño, salud mental y felicidad*

En el apartado precedente hemos examinado las implicaciones morales del autoengaño, y encontramos que las posiciones más "benevolentes" respecto de tales implicaciones se limitan a eximir de responsabilidad a quienes se encuentran en tal estado, pero no encuentran en él nada digno de encomio o eventualmente beneficioso. Puede resultar extraño, en consecuencia, que en los últimos años uno de los debates más destacados en torno de las consecuencias prácticas del autoengaño sea el relativo a su potencial contribución para nuestra salud mental y felicidad. Esta extrañeza, no obstante, no estaría del todo justificada, en vista de las consideraciones que diversos autores han hecho con respecto a ciertas consecuencias positivas que el autoengaño podría conllevar. Martin (2012) señala que mientras los filósofos clásicos tendieron a condenar todo autoengaño como deshonestidad, cobardía e hipocresía, los filósofos contemporáneos han destacado la existencia de casos permitidos e incluso deseables de autoengaño, como esperanzas autoengañosas que reafirman significados y habilidades para lidiar con enfermedades que amenazan la vida. Un contraste paralelo se ha presentado en las discusiones de los psicólogos respecto del autoengaño: mientras los psicoanalistas tradicionales enfatizaron cómo este fenómeno (entendido como mecanismo de defensa) resulta dañino para el yo, los actuales psicólogos cognitivos y evolucionistas destacan la manera en que promovería la felicidad y la adaptación.

Desde una perspectiva que podríamos denominar "clásica", entonces, el autoengaño no parece poseer aspecto positivo alguno: atenta tanto contra la sabiduría filosófica tradicional (expresada en el *dictum* "conócete a ti mismo" y en las exigencias relativas a nuestra constitución y acciones morales) como contra la comprensión científica de la Psicología contemporánea, para la cual una percepción ajustada de la realidad constituye un criterio básico para la adaptación de los organismos a su medio. Gordon Allport, uno de los defensores de este enfoque, señaló que una actitud objetiva e imparcial hacia uno mismo es una virtud primaria, básica para el desarrollo de todas las restantes. En su opinión, el autoengaño y las autojustificaciones y racionalizaciones que hace posibles evitan la adaptación y el desarrollo. Y, concluye, si existe un rasgo de personalidad intrínsecamente deseable, es la disposición para percibirse a uno mismo en perspectiva. Otros psicólogos destacados, como C. Rogers y A. Maslow, se expresaron en parecidos

términos respecto de las virtudes implicadas tanto en el autoconocimiento como en el conocimiento preciso de la realidad externa.

Sin embargo, y en línea con lo expresado por Martin, los mismos filósofos (p. ej., Davidson, 1986) han señalado también que el autoengaño positivo puede eventualmente liberar a quien lo practica de pensamientos o realidades dolorosas de distinta clase. De este modo, el hombre que se niega a aceptar, contra toda evidencia, que su hijo ha muerto en un accidente en el mar (Williams, 1973), se protege a sí mismo contra una realidad insoportablemente dolorosa. Esta observación de los filósofos relativa a que el autoengaño positivo podría redundar en ocasiones en una mejora en el bienestar individual ha sido sistemáticamente investigada por los psicólogos sociales, que han examinado la contribución que podría representar para la salud psíquica y la felicidad. El estudio de estas posibles consecuencias positivas del autoengaño se extendido hasta el examen de los vínculos entre el optimismo y el autoengaño (Norem, 2002) y también entre el optimismo cognitivo y el autoengaño (Metcalfe, 1998). Aunque con una atención menor, no han faltado los estudios relativos a los posibles beneficios de la adopción autoengañosa de creencias más negativas de lo que la evidencia permite concluir. [94, 95] Sin embargo, han sido

[94] En esta dirección se encuentra la hipótesis propuesta por Hartung (1988), quien sugiere que las personas emplean el autoengaño para disminuir su autoestima cuando esta estrategia es ventajosa para mantener cierta satisfacción con una posición que de otro modo sería percibida como injusta. Este sería el caso, por ejemplo, de un hombre cuyo trabajo tiene una jerarquía inferior a la que sabe que merece. Si no tiene esperanza de mejorar puede, a través del autoengaño, convencerse a sí mismo de que él está a la altura del estatus de su trabajo, en vez de percibirse como demasiado bueno para éste; esta forma de autoengaño lo habilita para reconciliarse con la disparidad entre su autoimagen y la realidad. Esto le posibilitará ver a sus jefes como realmente superiores y mejorar su habilidad para comportarse de modo subordinado. A su vez, todos se sentirán más cómodos con su presencia, y él incrementará su probabilidad de conservar el empleo. De acuerdo con lo anterior, entonces, el ajuste "hacia abajo" de la autoestima puede facilitar la seguridad psicológica, social y económica que de otro modo estaría en riesgo. El autoengañarse "hacia abajo" [*self-deceiving down*], sostiene Hartung, es la imagen en espejo del autoengañarse "hacia arriba" [*self-deceiving up*]; este último implicaría la elevación de la propia autoestima de modo de hacer posible el logro de una posición para la cual uno está inicialmente subcalificado. Ya sea que el autoengaño sea "hacia abajo" o "hacia arriba", la manipulación de la autoestima puede ser una profecía autosatisfactoria. Cabe observar aquí (cuestión sobre la que volveremos sobre el final de este apartado) que en realidad el caso descripto por Hartung podría no tratarse de un caso de autoengaño, sino de una "ilusión negativa" leve.

[95] Correia (s/f) examina la hipótesis según la cual las ilusiones negativas son realmente beneficiosas en el largo plazo debido a que funcionan como una suerte de "medicina amarga". Por un lado, son displacenteras y pueden impulsar a acciones irracionales, las cuales representen un costo considerable (de aquí su carácter "amargo"). Por otro lado, motivan una conducta que puede probarse como adaptativa con respecto al logro de metas primarias (de aquí su carácter de "medicina"). Posibles ejemplos de esto son el "síndrome de Otelo", que podría ser adaptativo en el sentido de incrementar la vigilancia contra potenciales rivales y desalentar la infidelidad de la pareja, y la explicación de Freud de la paranoia como un fenómeno motivado por el deseo inconsciente de ser el centro de atención. De modo similar, y con respecto a casos de pesimismo irrealista, algunos teóricos han sugerido que podría ser adaptativo tener una tendencia a suponer lo peor, en la medida en que la disposición mental pesimista parecería minimizar el impacto de futuras desilusiones y estimular la evitación de riesgos. De acuerdo con algunos autores, señala Correia, este aspecto es particularmente obvio en individuos depresivos, quienes discutiblemente "se ajustan mejor" en el trato con problemas complejos y son menos vulnerables a las alteraciones. El principal problema con la hipótesis de la "medicina amarga", observa Correia, es que está basada en una premisa que nunca ha sido probada, esto es, que los beneficios de sostener una ilusión negativa superan en última instancia a los costos (como es el caso

los estudios llevados a cabo por Shelley Taylor y sus colaboradores (Taylor y Brown, 1988, 1994; Taylor et al, 2000) los que han llevado más lejos la tesis de que ciertas formas "suaves" de autoengaño positivo harían posible, para quienes las experimentan, una mayor felicidad y salud mental en comparación con aquellas personas cuyas visiones de la realidad externa y de sí mismos son más precisas o ajustadas. Dada la influencia que han tenido los trabajos de Taylor y sus colaboradores, no sólo dentro del campo de la Psicología,[96] sino también en otros campos,[97] nos concentraremos en sus investigaciones y, en particular, en su artículo pionero, "Illusion and Well-Being: A Social Psychological Perspective on Mental Health" (1988).

Taylor y Brown comienzan por caracterizar lo que ha sido la perspectiva generalmente aceptada respecto de la percepción correcta de la realidad: ésta constituye un sello de la salud mental. Se ha sostenido habitualmente que la persona bien ajustada se involucra en testeos precisos de la realidad, mientras que el individuo cuya visión está nublada por la ilusión es percibido como vulnerable a (si no ya una víctima de) una enfermedad mental. Mediante la adopción de una perspectiva diametralmente opuesta, examinan evidencia que sugiere que ciertas ilusiones pueden ser adaptativas[98] para la salud mental y el bienestar; en particular, se examina evidencia de que un conjunto de ilusiones positivas interrelacionadas pueden servir a una amplia variedad de funciones cognitivas,

---

de los costos de los celos mórbidos). No hay evidencia de que los pacientes que sufren de trastornos delirantes sean más felices o más exitosos que otras personas, y una afirmación similar puede hacerse respecto del pesimismo que caracteriza a los individuos depresivos. Desde una perspectiva costo-beneficio, la idea de que las ilusiones negativas tienden a ser beneficiosas parece aún más cuestionable si tenemos en mente que no son los únicos ni los mejores medios para lograr las metas deseadas; por ejemplo, la creencia de que nuestra pareja está teniendo un *affaire* no es la única estrategia para incrementar la vigilancia y tomar medidas preventivas. En síntesis, la hipótesis de que las ilusiones negativas pueden ser beneficiosas en términos del bienestar parece incluso menos plausible que las hipótesis correlativas de que las ilusiones positivas son globalmente adaptativas. Correia concede que puede haber casos excepcionales en los que una ilusión negativa accidentalmente puede dar lugar a un resultado positivo; sin embargo, en referencia a la pregunta de si las ilusiones negativas son globalmente beneficiosas en las sociedades actuales, la respuesta parece ser que son desadaptativas tanto a corto plazo (por la presencia de pensamientos y emociones negativos) como a largo plazo (por las conductas irracionales a las que dan origen).

[96] Norem (2002) formula una observación que revela la medida en la que las tesis de Taylor y sus colaboradores parecen haber influido en la psicología actual: "En su influyente trabajo, Taylor y Brown (1988) argumentaron que el optimismo ilusorio y los procesos de automejoramiento ("ilusiones positivas") son fundamentales para la salud mental, *una posición que constituye uno de los pilares del floreciente movimiento de la psicología positiva*" (p. 549, énfasis nuestro).

[97] Algunos filósofos (cfr. Elga, 2003) parecen haber sido impresionados por los argumentos y evidencia expuestos por Taylor y sus colaboradores al punto de perder de vista sus limitaciones y deficiencias, como veremos más adelante. Pero no sólo algunos filósofos han adoptado las afirmaciones de Taylor y otros por su valor nominal: Trivers (2000) cita de modo aprobatorio las investigaciones de Taylor y Brown en apoyo de su posición adaptacionista respecto del autoengaño.

[98] Conviene recordar aquí que este uso de los términos "adaptativo" y "función" no tiene el mismo significado que posee para los psicólogos evolucionistas. Cfr. al respecto las observaciones relativas a esta distinción formuladas en el capítulo precedente.

afectivas y sociales. Plantean también el intento de resolver una paradoja: cómo pueden las percepciones erróneas de uno mismo y del entorno ser adaptativas cuando procesar la información con precisión parece ser esencial para el aprendizaje y el funcionamiento exitoso en el mundo.

Taylor y Brown consideran conveniente diferenciar entre *errores* y *sesgos*, por una parte, de *ilusiones*, por la otra, en la medida en que estas últimas implican patrones de error o sesgo (o ambos) más duraderos, que adquieren una forma o dirección particular; las ilusiones constituyen imágenes o concepciones mentales falsas que o bien pueden ser una interpretación errónea de algo real, o bien pueden ser producto de la imaginación. Pueden ser placenteras, inofensivas, o incluso útiles en algunos casos.[99] Pueden ser de distintas clases, y si bien una taxonomía de ellas es, en alguna medida, arbitraria, consideran que tres tipos de ilusiones que afloran de manera consistente dentro de los hallazgos empíricos son las siguientes: concepciones positivas no realistas del sí mismo, exagerada percepción de control personal y optimismo no realista.

La existencia de concepciones positivas no realistas acerca del sí mismo es sustentada por una serie de hallazgos empíricos, que incluyen (entre otros), un fuerte predominio de los adjetivos positivos sobre los negativos para describir la propia personalidad, un procesamiento y recuperación más eficientes de la información positiva sobre la negativa respecto de la propia personalidad, un recuerdo más pobre de los fracasos que de los éxitos, una mayor atribución al sí mismo de resultados positivos que de resultados negativos, una desestimación o minimización de los aspectos negativos del yo, y una creencia en la mejora en habilidades que son importantes para la persona, en presencia de un rendimiento sin cambios. Este desequilibrio entre percepciones positivas y negativas del sí mismo, señalan Taylor y Brown, no prueba que tales concepciones sean irrealistas o ilusorias. No obstante, hay evidencia de este carácter. En primer lugar, existe una tendencia dominante a percibir al yo como mejor que otros; los individuos juzgan a los atributos de personalidad positivos como más descriptivos de sí mismos que de la persona promedio, y a los atributos negativos como menos descriptivos de sí mismos que de la persona promedio. Sin embargo, dado que es lógicamente imposible que la mayoría de las personas sea mejor que el promedio, estas concepciones pueden ser vistas como evidencia de su naturaleza irrealista e ilusoria. En segundo lugar, la cualidad ilusoria de las autopercepciones positivas se pone de manifiesto en investigaciones que comparan las autoevaluaciones con

[99] El riesgo de hablar de una creencia que se aleja de la realidad, señalan Taylor y Brown, es caer en los debates filosóficos relativos a cómo es posible conocer la realidad. Consideran que es posible, en cierto grado, escapar a este riesgo por medio de las definiciones operacionales ofrecidas por la psicología social.

juicios hechos por observadores: las autoevaluaciones fueron significativamente más positivas que las puntuaciones de los observadores. Por último, existe un grupo de individuos que aceptan los aspectos buenos y malos de sí mismos, tal como muchas concepciones de la salud mental aseveran: las personas con baja autoestima, y/o moderadamente deprimidas, son más equilibradas en sus autopercepciones.

Con respecto al segundo tipo de ilusiones, si bien muchos teóricos han mantenido que un sentido de control personal es esencial para el autoconcepto y la autoestima, la evidencia sugiere que las creencias de las personas en el control personal son a veces mayores que lo que puede ser justificado. Esta evidencia incluye el sesgo por el cual las personas, en estudios controlados que adoptaron el formato de juegos, actúan a menudo como si tuvieran control en situaciones que están en realidad determinadas por el azar. Además, existe una amplia bibliografía sobre estimación de covariación que indica que las personas sobreestiman sustancialmente su grado de control sobre sucesos fuertemente determinados por el azar. Por último, entre las personas moderada y severamente deprimidas se observa una menor vulnerabilidad a la ilusión de control.

Por último, con respecto al optimismo no realista, Taylor y Brown señalan que el optimismo penetra los pensamientos de las personas acerca del futuro; muchas personas creen que el presente es mejor que el pasado y que el futuro será todavía mejor. No obstante, se debe probar que este optimismo es efectivamente irrealista. La primera evidencia de esta naturaleza irrealista proviene de la comparación de los juicios de los propios sujetos con juicios de terceros: si bien la visión del futuro cálida y generosa que los individuos mantienen respecto de sí mismos se extiende a todas las personas, es decididamente más marcada para ellos mismos. Conversamente, cuando se interroga a los sujetos acerca de las probabilidades de experimentar una amplia gama de sucesos negativos, muchas personas creen que es mucho menos probable que ellos mismos lo experimenten, a la vez que creen que esta probabilidad es mayor para otras personas. Por último, las predicciones de los sujetos sobre lo que ocurrirá respecto de una amplia variedad de tareas corresponden cercanamente a lo que quisieran que ocurriera o a lo que es socialmente deseable, más que a lo que es objetivamente probable. En contraste con las concepciones extremadamente positivas del futuro desplegadas por los individuos normales, las personas levemente deprimidas y aquellas con baja autoestima parecieron sostener evaluaciones más equilibradas de sus probables circunstancias futuras.

Los hallazgos anteriores relativos a personas normales y a personas depresivas y/o que padecen de baja autoestima, en síntesis, parecen ser inconsistentes con la noción de que el autoconocimiento preciso es el sello de la salud mental.

Si bien Taylor y Brown consideran que los hallazgos anteriores son sólidos, admiten que una cosa es decir que las ilusiones positivas existen en las personas normales, y otra distinta identificar el modo en que contribuyen con la salud mental. Hacer esto último requiere del establecimiento de criterios de salud mental, para luego determinar si las consecuencias de las ilusiones positivas precedentes se ajustan a esos criterios. El dilema que surge inmediatamente es que muchas definiciones formales de salud mental incorporan las autopercepciones precisas como un criterio. Los elementos comunes que Taylor y Brown encuentran en distintas definiciones de salud mental son: felicidad o satisfacción, la habilidad para ocuparse de otros y la capacidad para trabajar productiva y creativamente. Con respecto al primero, señalan que existe una mayor probabilidad de que las personas que informan las características mencionadas (autoconfianza, sensación de control, optimismo respecto del futuro) sean más felices que aquellas que no las poseen. Esta asociación entre las ilusiones y el humor positivo parece ser consistente, pero, admiten, es ampliamente correlacional y no causal. Sin embargo, se ha informado de alguna evidencia de que las ilusiones influencian directamente el humor. Si bien algunas de estas investigaciones no descartan la posibilidad de que el humor positivo cause las ilusiones, esto es, que las variables estén recíprocamente relacionadas, proveen evidencia de que las ilusiones promueven la felicidad. Con respecto al segundo elemento, la habilidad para ocuparse de otros, existe evidencia de que las ilusiones positivas están asociadas con ciertos aspectos del vínculo social. Las ilusiones pueden afectar la habilidad para ocuparse de otros por medio de su capacidad para crear un humor positivo, lo que incrementa la probabilidad de ayudar, de iniciar conversaciones, de expresar acuerdos y evaluaciones positivas en general y de reducir el uso de estrategias contenciosas. Por último, respecto de la capacidad para trabajar productiva y creativamente, la influencia de las ilusiones puede producirse en dos sentidos: facilitar el funcionamiento intelectualmente creativo en sí mismo e incrementar la motivación, la persistencia y el rendimiento.

Los análisis previos, señalan Taylor y Brown, presentan varios dilemas teóricos y prácticos. El primero refiere a la forma de reconciliar el punto de vista tradicional de la salud mental con los hallazgos de la Psicología cognitiva y social que han reseñado. El

segundo, a la manera en que las ilusiones pueden ser mantenidas y, más importante, pueden ser funcionales ante la presencia de evidencia realista y a menudo contradictoria proveniente desde el entorno.

Un punto de partida útil en el intento de reconciliar las perspectivas tradicionales con los hallazgos que examinan consiste en analizar las debilidades potenciales en los métodos de recolección de datos de la literatura relevante en Psicología clínica y Psicología social, de la cual derivan las respectivas concepciones. Históricamente, observan, las construcciones clínicas de la salud mental han estado dominadas por la terapia y la investigación con personas anormales. Dado que un atributo de muchas personas psicológicamente perturbadas es una inhabilidad para monitorear la realidad de manera efectiva, es posible que el retrato del individuo saludable sea el de aquel que mantiene un estrecho contacto con la realidad; así, pueden haber pasado inadvertidas desviaciones más sutiles de los estándares objetivamente precisos en las percepciones y cogniciones. Por otra parte, una concepción de la salud mental derivada solamente de la investigación sobre la cognición social puede ser sesgada de un modo tal que revele un énfasis excesivo en las ilusiones. ¿Cuál es la forma de reconciliar estos puntos de vista? En primer lugar, señalan Taylor y Brown, un cierto grado de contacto con la realidad parece ser esencial para desempeñar las tareas de la vida diaria. Por el otro lado, es evidente que, cuando los errores y los sesgos ocurren, no están distribuidos uniformemente. Se desvían de manera consistente en una dirección positiva, hacia el automejoramiento del yo y del mundo en el cual uno debe funcionar. La clave para la integración de las dos concepciones de la salud mental, entonces, yace en la comprensión de las circunstancias bajo las cuales las ilusiones positivas pueden ser más útiles. Estas circunstancias parecen ser aquellas que resultan adversas, esto es, aquellas condiciones en las cuales podría esperarse que se produjeran depresión o carencia de motivación. Bajo tales circunstancias, la creencia en uno mismo como un actor competente, eficaz, que se comporta en el mundo con un sentido del futuro generalmente positivo, puede ser especialmente útil para superar contratiempos, posibles golpes a la autoestima y la erosión potencial a la propia concepción del futuro. Esto involucra una serie de filtros sociales y cognitivos en el manejo de la retroalimentación negativa, que incluyen, entre otros, la construcción social de la retroalimentación social, sesgos en la codificación, interpretación y recuperación de la información, deriva cognitiva, y el reconocimiento de áreas de incompetencia.

Taylor y Brown no dejan de poner de manifiesto las limitaciones de su estudio y las precauciones que requiere su análisis. Entre estas se cuentan la debilidad de algunas

conexiones que requieren investigación empírica posterior, especialmente en lo que refiere a los nexos entre ilusiones y afecto positivo. En segundo lugar, la naturaleza correlacional y no causal de la evidencia relativa a ciertas conexiones fácticas. Tercero, la ausencia en sus análisis de un criterio común de salud mental, esto es, la capacidad para el crecimiento personal y el cambio. Cuarto, la naturaleza experimental de mucha de la evidencia, expuesta a varios sesgos potenciales, como la tendencia a extraer a las personas de sus entornos acostumbrados, someterlos a estímulos no familiares y a obtener conclusiones generales acerca de la conducta humana que pueden ser en parte una respuesta a la novedad. Por último, el interrogante referente a si las ilusiones positivas son siempre adaptativas. De hecho, pueden verse riesgos inherentes a cada una de las ilusiones descriptas. Es importante recordar que las evaluaciones de las personas son sólo un aspecto de los juicios acerca de la situación, y puede haber información inherente no relacionada con el yo que contrarreste los efectos de las ilusiones y lleve la gente a corregir su conducta. El precedente argumento no pretende sugerir que las ilusiones positivas carezcan de pasivos; de hecho, reconocen, tienen muchos. Sin embargo, consideran que no se deberían extraer conclusiones apresuradas relativas a las cargas negativas de las ilusiones positivas sin una evaluación de las posibles fuerzas que pueden ayudar a contrarrestar tales consecuencias.

Como es fácil de suponer, habida cuenta de la naturaleza controvertida de la mayoría de las tesis relativas al autoengaño, las afirmaciones de Taylor y sus colaboradores fueron seguidas por una serie de cuestionamientos de distintas clases. Ciertos autores[100] han alertado acerca de los peligros que las "ilusiones positivas" pueden acarrear en el plano individual, pero también en el colectivo. Otros (Mertz, 2004), en una línea más cercana a los cuestionamientos relativos a la naturaleza moral del autoengaño, han señalado algunos de los riesgos que involucra, como el daño emocional y el perjuicio para las relaciones. En términos generales, las críticas contra las tesis de Taylor y sus colaboradores pueden dividirse en dos grupos: las que han tratado de probar que el autoengaño no contribuye efectivamente con nuestra salud mental, y aquellas que han intentado mostrar cómo la felicidad lograda por medio del autoengaño no es realmente digna de ese nombre. Tales objeciones han provenido tanto desde la Psicología como desde la Filosofía. Dentro de la primera, Robins y Beer (2001), han señalado que las creencias automejoradoras pueden ser adaptativas en el corto plazo, pero no a largo plazo. Especialmente citados e influyentes han sido los estudios de Colvin y otros (1994, 1995), quienes han objetado las conclusiones

[100] Cfr. al respecto las observaciones de Goleman (1989) expuestas en el apartado 4 de este capítulo.

de Taylor y sus colaboradores, señalando lo que consideran deficiencias conceptuales y empíricas que impiden que tales conclusiones sean adecuadamente sustentadas. La presunta conexión entre la muy influyente corriente de la psicología positiva, por un lado, y las "ilusiones positivas" y el autoengaño, por el otro, también ha sido desestimada por los fundadores de aquel enfoque.[101] Así como hemos comentado al inicio de este capítulo las observaciones de algunos filósofos según las cuales en algunos casos el autoengaño podría evitar al individuo sufrimientos innecesarios, otros autores (Bhadwar, 2008; Van Leeuwen, 2009; Martin, 2012) han señalado los riesgos de suponer que el autoengaño podría constituir la base para el logro de una felicidad genuina. En lo que sigue nos concentraremos en los comentarios críticos de estos autores, especialmente en los del tercero.

Martin (2012) considera tres preguntas acerca del modo en que el autoengaño y la felicidad están conectados: primero, si es posible que estemos autoengañados respecto de nuestra propia felicidad; segundo, si el autoengaño favorece la felicidad o más bien es una amenaza para ella; tercero, de qué modo el encontrarse "felizmente autoengañado" interactúa con otros aspectos de una vida buena. Los psicólogos han documentado ampliamente, señala, que cuando perseguimos la felicidad nos encontramos bajo una miríada de ilusiones específicas, y es plausible suponer que el autoengaño se encuentra involucrado. Esto, agrega, no es sorprendente, sino que refleja el sentido común. Su posición es más favorable a la influencia positiva que, dentro de ciertos límites, pueda tener el autoengaño sobre la felicidad que lo que encontramos en Van Leeeuwen y Badhwar;[102] no obstante, los argumentos de Taylor le parecen susceptibles de crítica. En su opinión, Taylor se desliza de modo inadecuado desde una definición inicial de ilusiones como falsas creencias a una mucho más amplia que las caracteriza como creencias que carecen de apoyo fáctico o que requieren una percepción optimista de los hechos. Esta noción, señala Martin,

[101] Compárese lo expresado en la nota 99 en el presente capítulo con las siguientes observaciones: "cualesquiera sean los orígenes personales de nuestra convicción de que ha llegado el momento para la psicología positiva, nuestro mensaje es recordar a nuestro campo que la psicología no es sólo el estudio de la patología, la debilidad y el daño; es también el estudio de la fortaleza y la virtud. (…) La psicología no es precisamente una rama de la medicina concerniente a la enfermedad o la salud; es mucho más amplia. Se ocupa del trabajo, la educación, el *insight*, el amor, el crecimiento y el juego. Y en esta búsqueda de lo que es mejor, la psicología positiva no descansa sobre *el pensamiento desiderativo, la fe, el autoengaño* (…); trata de adaptar lo mejor del método científico a los problemas singulares que presenta la conducta humana para aquellos que desean comprenderla en toda su complejidad" (Seligman & Csikszentmihalyi, 2000, p. 7. Cursivas nuestras).

[102] En su opinión, es verdad que, tal como afirman el sentido común y los psicólogos, el autoengaño es, dentro de ciertos límites, un amigo de la felicidad, en la medida en que promueva el placer, el significado y el amor integral en nuestras vidas. No obstante, no deja de señalar que el autoengaño puede también socavar nuestro sentido de lo que es significativo y, de este modo, diluir la felicidad así como debilitar la aspiración a llevar una buena vida.

combina creencias optimistas de dos clases muy distintas: 1) creencias falsas y sesgadas contrarias a la evidencia disponible (lo que incluye al autoengaño), y 2) creencias no probadas, que incluyen la fe y la esperanza, que van más allá de la evidencia pero que podrían resultar verdaderas. Las últimas, advierte, no son ilusiones en ningún sentido usual; las creencias verdaderas no son ilusiones. Taylor, sospecha Martin, ha tomado prestado de modo acrítico el uso que Freud hace del término "ilusión" en *El porvenir de una ilusión*. Allí, Freud señala que lo característico de las ilusiones es que derivan de deseos humanos, lo que no implica que sean necesariamente falsas, es decir, irrealizables o contradictorias con la realidad. Este uso no ortodoxo, señala Martin, fusiona autoengaño con fe y esperanzas razonables. Por supuesto, observa, la esperanza y la fe, que van más allá de la evidencia, son vitales para la felicidad de la mayoría de nosotros, pero no siempre están enraizadas en ilusiones o autoengaño. Más aun, en respuesta a la objeción de que las ilusiones pueden causar un gran daño y ser una señal de un trastorno mental, Taylor señala que las "ilusiones positivas" son sólo ilusiones suaves, no patológicas, que contrastan con las ilusiones "extremas" de la paranoia o la megalomanía. Martin observa que esta respuesta incurre entonces en cierta circularidad: "suave" implica salud mental.

Si Martin muestra una posición más conciliadora respecto de los potenciales beneficios del autoengaño para nuestra felicidad, Badhwar (2007) es mucho más escéptica respecto de ellos; argumenta que, entendido de modo apropiado y contrariamente a lo que proponen Taylor y Brown, el realismo acerca de nosotros mismos y nuestras circunstancias es bueno para nuestras vidas.[103] Considera que, para la atención que han generado sus afirmaciones, la evidencia que presentan Taylor y Brown en su favor es sorprendentemente débil, y está plagada de problemas conceptuales y lógicos. No obstante esta última afirmación, Badhwar señala también debilidades en el presunto apoyo empírico y en la interpretación de éste que hacen Taylor y Brown. Para comenzar, la principal fuente de evidencia acerca de la ubicuidad de las ilusiones positivas en el artículo pionero de Taylor y Brown proviene de investigaciones con estudiantes universitarios, si bien, con el objetivo de ampliar el sustento de sus afirmaciones, citan también estudios llevados a cabo con pacientes terminales en su respuesta a Colvin y Block. Respecto de las primeras, Badhwar advierte que los estudiantes universitarios no son representativos de la población norteamericana en general, hacia la cual Taylor y Brown extrapolan los hallazgos obtenidos.

[103] Respecto de la concepción de felicidad de Taylor y sus colaboradores, Badhwar (al igual que Van Leeuwen) observa que, al igual que la de otros psicólogos contemporáneos, es puramente subjetiva: no hace mención a bienes *objetivos* como componentes esenciales de una felicidad digna de ese nombre.

Dada su juventud e inexperiencia, observa, debería esperarse que los estudiantes sean particularmente susceptibles a las ilusiones de control y a un optimismo exagerado. Pero aun si se concede, prosigue, que la mayoría de las personas se consideran a sí mismas como más felices y superiores en sus habilidades, logros, grado de control y perspectivas futuras que los demás, no se sigue, como sostienen Taylor y Brown, que se encuentren en un estado de ilusión. Supóngase, dice Badhwar, que el 60% de las personas piensan que son más felices, mejores y que poseen un futuro más brillante que el 60%. Podría ocurrir que sólo el 20% de ellas esté equivocado, porque resulta posible para el 40% del total encontrarse en ventaja en esos aspectos respecto del 60% restante. Si esto fuese así, sólo el 20% se encontraría bajo el efecto de ilusiones positivas. Así, en ausencia de un número que los respondentes tengan en mente cuando se evalúan comparativamente con la "mayoría de las personas", Taylor y Brown no pueden concluir, a partir de los datos, que la mayoría de las personas se encuentran sistemáticamente, aunque de modo suave, bajo la influencia de ilusiones.

La observación cotidiana, prosigue Badhwar, también sugiere que la mayoría de las personas no es globalmente positiva de un modo no realista acerca de sí mismas, ni abiertamente optimista acerca del futuro la mayor parte del tiempo. La mayoría de nosotros, señala, somos realistas en algunos dominios de nuestras vidas, optimistas o pesimistas de modo no realista en otros, y ni realistas ni irrealistas de modo consistente en el resto.

Pero las afirmaciones de Taylor y Brown, prosigue Badhwar, pueden ser objeto de más cuestionamientos. Supóngase que la mayoría de las personas se encuentra sistemáticamente bajo la influencia de ilusiones acerca de sí mismas. ¿Sobre qué base, pregunta, Taylor y Brown afirman que están mentalmente sanos o son subjetivamente felices? Una razón es simplemente que la mayoría de las personas no pueden estar mentalmente enfermos o ser infelices. No obstante, esta conclusión parece apresurada. Existen grados de salud mental y felicidad subjetiva, así como hay grados de salud física y de virtud. Las personas que no están deprimidas o psicóticas pueden igualmente estar insatisfechas o neuróticas en variados grados. La proliferación de terapeutas, consejeros, gurúes y libros de autoayuda es un indicador, señala Badhwar, de que la insatisfacción y la neurosis gozan de buena salud. Pero concédase de todos modos, continúa, que la mayoría de las personas goza de tan buena salud y es tan subjetivamente feliz como resulte posible. ¿Qué elementos de juicio tienen Taylor y Brown para afirmar que eso es así a causa de sus ilusiones? Toman como base para afirmar esto el hecho de que los individuos más realistas

tienden a ser depresivos. Pero, como Colvin y Block mostraron, señala Badhwar, la evidencia respecto de estos "realistas depresivos" es contradictoria o ambigua. De hecho, en su respuesta a estos autores, Taylor y Brown reconocieron que varios estudios indicaban que los depresivos, en vez de ser realistas, se hallaban bajo el efecto de un sesgo negativo hacia sí mismos, y que los "realistas depresivos" podrían no existir. Esto último tiene consecuencias graves para de Taylor y Brown, ya que torna vacuas sus tesis. Su afirmación principal, recuerda Badhwar, es comparativa: sostiene que las ilusiones suaves son más conducentes a la felicidad subjetiva y a la salud mental que el realismo. De este modo, se hace necesario contar con un grupo de personas realistas infelices o al menos felices en una medida inferior para compararlos con los optimistas que se encuentran bajo la influencia de las ilusiones. Taylor y Brown sostuvieron que, aun cuando las personas depresivas no fueran realistas, su tesis original se mantendría sin cambios. Sin embargo, concluye Badhwar, si los pesimistas no fueran realistas, no existirían casos de depresivos realistas con los cuales comparar a los optimistas y, en consecuencia, no habría evidencia para la tesis de que las ilusiones positivas acerca de uno mismo son más conducentes a la felicidad subjetiva y a la salud mental que el realismo.

Si bien las críticas de Badhwar parecen sólidas, las objeciones más agudas a las tesis de Taylor y sus colaboradores provienen, a nuestro entender, de un filósofo ya conocido, Neil Van Leeuwen, en un artículo titulado "Self-deception won"t make you happy" (2009). Van Leeuwen comienza por señalar que el trabajo de Taylor no es explícitamente normativo; no obstante, implica que muchas de las ilusiones que examina deben ser fomentadas, y es justamente la cuestión normativa la que desea examinar. Formulada de modo interrogativo, plantea lo siguiente: si deseamos la felicidad, ¿es consistente con la racionalidad práctica la prosecución de políticas de autoengaño? Específicamente, Van Leeuwen quiere refutar el siguiente razonamiento: la posesión de creencias positivas acerca de uno mismo conduce a la felicidad; esto ocurre, como supuestamente muestran los trabajos de Taylor, aun cuando tales creencias se aparten de la verdad debido a su carácter excesivamente halagador para el sí mismo. Para lograr la felicidad, por lo tanto, es una buena idea perseguir una política de autoengaño con respecto al yo y a la propia situación, tornándolos mejores que lo que admitiría una apreciación honesta. Van Leeuwen afirma, por el contrario, que la honestidad hacia uno mismo es una política superior que el autoengaño si se desea lograr la felicidad.

El argumento de Van Leeuwen tiene tres aspectos: primero, el trabajo de Taylor no puede ser usado apropiadamente para apoyar el razonamiento en cuestión; segundo, hay varios argumentos para sustentar la conclusión de que una política de autoengaño no produce un tipo de felicidad a la que denomina "digna de elección" [*choiceworthy*]; tercero, sostiene que un camino efectivo hacia la felicidad yace en el ejercicio de la imaginación honrada, lo que involucra honestidad con uno mismo acerca de las propias habilidades y las ofertas del entorno, e imaginación positiva acerca de lo que es posible hacer con ellas.

Cualquier intento de examinar las relaciones del autoengaño con la felicidad requiere de un mínimo análisis de este último concepto. Van Leeuwen considera que existen dos dimensiones distintas (aunque no netamente separables) constitutivas de la felicidad: la existencia de sentimientos positivos y la posesión de bienes externos genuinos. Combinadas, estas dimensiones originan cuatro tipos distintos de estado, de las cuales se ocupa especialmente de dos: la felicidad digna de elección, que implica la existencia de ambos componentes, y la que denomina "felicidad Matrix", que sólo requiere de la posesión de sentimientos positivos, pero en la que se carece de bienes externos valiosos. Van Leeuwen considera que la única clase de "felicidad" que la persecución de una política de autoengaño es probable que promueva es la felicidad Matrix. Ahora bien, esta clase de felicidad puede dividirse en dos subtipos. Un subtipo puede conducir a la felicidad digna de elección, dado que los sentimientos positivos pueden originar acciones que conduzcan a bienes externos genuinos. No obstante, otro subtipo simplemente sienta las bases para la derrota y la desilusión. Van Leeuwen advierte que es necesario recordar que está hablando de la *mejor* política para el logro de felicidad digna de elección; en consecuencia, un ejemplo de autoengaño que conduzca a la felicidad, digna de elección o de otra clase, no refutará el argumento de que no es instrumentalmente racional autoengañarse para lograr la felicidad.

El modo en el que una persona debe engañarse a sí misma con el fin de alcanzar la felicidad debe implicar que desea creer que ciertas proposiciones negativas son falsas y ciertas proposiciones positivas son verdaderas. Llevar adelante una política de autoengaño, entonces, requiere poseer creencias que se encuentren en tensión con las creencias deseadas. Aquí existe una diferencia importante con la mayoría de las ilusiones positivas que Taylor y sus colaboradores discuten. "Sesgo de autoinflación" [*self-inflation bias*] refiere a una tendencia general a formar creencias acerca de uno mismo que son más halagadoras de lo que la realidad justifica, pero no necesariamente controvertidas por otra información que

el agente posee. Por el contrario, un estado de autoengaño requiere de la existencia de una creencia que es contraria tanto a la evidencia que el agente posee como a sus normas epistémicas, y se trata de un estado en el cual la motivación está causalmente implicada en su producción. La política de autoengaño para el logro de la felicidad involucrará entonces, una conciencia de las áreas de la propia vida que se perciben como deficitarias o acerca de las cuales se siente inseguridad, una selección de creencias que promoverán la felicidad, y un intento de procesar la información que parezca confirmar las creencias favorables a la felicidad e ignorar las contrarias.

Perseguir una política de autoengaño significará, en síntesis, intentar formar creencias que son contrarias a la información negativa acerca del sí mismo que *ya se posee,* y es esta política lo que Van Leeuwen tratará de mostrar que no conduce a la felicidad.

Van Leeuwen señala que hay varias razones por las cuales la afirmación de que el autoengaño puede contribuir con la felicidad digna de elección es problemática.

En primer lugar, y dicho en sus propios términos, observa que Taylor sostiene que las ilusiones positivas, tales como la creencia de que uno es más atractivo que lo que realmente es el caso, contribuyen al logro de felicidad digna de elección. Así, Taylor sostendría que las ilusiones positivas conducen al logro de la felicidad subjetiva (en forma de afecto y creencias positivas), y ésta, a su vez, al logro de bienes externos. Aun cuando los nexos entre unos y otros estados sean correctos, ¿muestra esto que las ilusiones positivas generan bienes externos genuinos? A menudo las "ilusiones" a las que Taylor se refiere resultan ser correctas. Parece haber entonces una confusión conceptual al llamarlas "ilusiones", dado que las ilusiones son cosas que no resultan ser verdaderas. No obstante, hay una versión más matizada de la concepción que Taylor sugiere. Esta versión consistiría en afirmar que poseer una estimación elevada de un modo no realista de la probabilidad de éxito eleva la probabilidad de éxito (si bien no al nivel de la evaluación), y merece en consecuencia el nombre de "ilusión". Esta versión del argumento, observa Van Leeuwen, no sufre de la confusión conceptual de la versión original, pero no encuentra evidencia en el trabajo de Taylor que la sostenga. Es verdad, señala, que los individuos operan con más de una reserva de representaciones y pueden cambiarlas dependiendo de la situación en que se encuentren. Sin embargo, sería un error la tendencia a rotular como "ilusión" a cada representación que no se ajusta enteramente a la realidad.

En segundo lugar, sin duda las personas tienen muchas ilusiones acerca de sí mismas y de sus vidas. Ahora bien, la afirmación de Taylor de que las ilusiones conducen a

resultados positivos futuros fusiona la posesión de creencias ilusorias acerca del sí mismo con la posesión de fantasías orientadas hacia el futuro. Las representaciones del futuro no necesitan ser ilusorias, aun cuando es improbable que sus contenidos resulten finalmente ser verdaderos. Simplemente pueden ser fantasías positivas orientadas hacia el futuro. La unificación bajo el mismo rótulo de "ilusiones positivas" tanto de las creencias infladas acerca de uno mismo como de las fantasías respecto de resultados futuros es perjudicial para la comprensión de estos fenómenos.

Por último, no parece haber buenas razones para pensar que poseer creencias tales como que uno es mejor conductor o más inteligente que otras personas nos hará más felices. Es posible que esas creencias estén *correlacionadas* con afecto positivo, ¿pero hay razones para pensar que tales creencias *causan* afecto positivo, de modo tal de dar sentido a una política de inculcarlas vía autoengaño? Muchas personas saben que tienen una inteligencia superior al promedio, y pese a eso están deprimidas. Van Leeuwen no se muestra convencido de que tales creencias, ilusorias o no, realmente contribuyan causalmente al afecto positivo más que de un modo temporario. Conversamente, muchas personas felices saben que no son especialmente inteligentes, que no son los mejores conductores y que sus hijos pueden incluso ser inferiores al promedio. Claramente, dice, esas creencias no son necesarias para la felicidad. Van Leeuwen considera que lo que ocurre en los trabajos empíricos que Taylor cita es que las personas no deprimidas tienen simplemente una tendencia sistemática a sobreestimar varios aspectos de valor personal. No obstante, una interpretación probable de este hallazgo es que el afecto positivo causa la sobreestimación (y no viceversa). Si las creencias concernientes a rasgos positivos causaran el afecto positivo, simplemente no se esperaría ver tantas personas inteligentes, apuestas o exitosas que son infelices; sin embargo, resulta que hay muchas de esas personas.

Van Leeuwen sugiere que lo que sí resulta necesario para la felicidad es la creencia en el propio valor como ser humano. Esta creencia es una clase diferente de creencia de aquellas con contenidos como "soy más popular que el promedio". Es incluso problemático llamar creencia a la conciencia del propio valor como ser humano; ciertamente no se trata de una creencia con contenido fáctico. Van Leeuwen sugiere referirse a ella simplemente como sentido del propio valor, y señala que no hay nada ilusorio en un ser humano normal que posea un sentido de su propio valor.

Además de que la evidencia empírica que, *prima facie*, parece favorecer una política de autoengaño para la consecución de la felicidad no apoya realmente esa política, Van Leeuwen ofrece argumentos para rechazarla por completo.

Poseer creencias verdaderas posibilita a las personas lograr objetivos. Por una parte, es necesario conocer las posibles opciones antes de elegir entre ellas. Además, tener creencias acerca de los resultados de las opciones de elección guiará la elección. Si alguna de las creencias mencionadas en este esquema es falsa (ya sea acerca de las opciones o de los resultados), señala Van Leeuwen, entonces probablemente la persona terminará insatisfecha. En consecuencia, las falsas creencias tienden hacia la insatisfacción. Pero, como sabemos, el autoengaño conduce a falsas creencias; luego, el autoengaño tiende hacia la insatisfacción. En conclusión, el autoengaño aleja de la felicidad.

El defensor del autoengaño podría hacer notar, en primer lugar, que hay algunas cosas en la vida sobre las cuales uno no tiene control, pero que afectan a la felicidad (inteligencia o apariencia, por ejemplo). Entonces la receta general para que el autoengaño promueva la felicidad, prosigue el argumento, involucra no autoengañarse acerca de cosas que uno puede controlar, sino acerca de cosas que uno no puede controlar.

Tres puntos deben ponerse de manifiesto, observa Van Leeuwen. En primer lugar, esta posibilidad nos sitúa fuera del territorio de la competencia por la felicidad digna de elección, de modo que el defensor del autoengaño que adopte esta línea de defensa ha resuelto optar por la felicidad Matrix (al menos en un área determinada de la vida). En segundo lugar, aun si uno no puede cambiar los componentes innatos de la inteligencia o el aspecto, tener creencias falsas sobre esas cuestiones puede aun así socavar la satisfacción del deseo; a menudo conocer las propias limitaciones es esencial para determinar las elecciones óptimas que las compensen. Tercero, el autoengaño también disminuye la propia habilidad para saber *cuáles* son las situaciones sobre las cuales uno tiene control y puede hacer mejoras, y cuáles no. Los razonamientos basados en creencias verdaderas, entonces, son necesarios para satisfacer los deseos. El autoengaño es contrario a las creencias verdaderas y, en consecuencia, tiende a alejar de la satisfacción del deseo y la felicidad. Los argumentos en su favor ignoran el daño que produce a la cognición fidedigna en cuestiones de importancia práctica, y el hecho de que es contrario a la felicidad digna de elección, ya que socava el logro de bienes externos genuinos.

Existe una tensión curiosa, observa Van Leeuwen, en la posición del defensor del autoengaño. Por un lado, defiende al autoengaño como un medio para la felicidad,

sosteniendo que engañarnos a nosotros mismos en la dirección de un afecto positivo proporcionará un panorama brillante, lo que a su vez influenciará positivamente el mundo que nos rodea. Por otro lado, parte de la estrategia para defender el autoengaño ha sido la admisión de que a veces somos impotentes para cambiar la situación y, dada esta condición, podemos convencernos a nosotros mismos de que la situación es mejor de lo que realmente es. Como hemos visto, Van Leeuwen ha intentado arrojar dudas sobre ambas afirmaciones. Sin embargo, el punto que quiere tratar es que ambas defensas del autoengaño *no pueden requerir el mismo autoengaño*. La defensa "el afecto positivo cambia el futuro" presupone que el agente tiene algún tipo de control sobre el entorno que lo rodea respecto de los asuntos sobre los cuales se autoengaña. La defensa "no es posible cambiar la situación" presupone que el agente no puede cambiar el entorno acerca del cual se autoengaña. La pregunta, entonces, es esta: ¿en qué tipo de situación es el autoengaño conducente a la felicidad: en las situaciones en las cuales hay control sobre el entorno o situaciones en las cuales no lo hay? El defensor del autoengaño dirá: "en ambas". Pero la cuestión, observa Van Leeuwen, es si existe alguna incoherencia en esta respuesta; y considera que sí la hay. La respuesta "ambos" está comprometida no sólo con el autoengaño en los dos tipos de situación, sino también con la confusión relativa a la situación en que se encuentra el agente. Pero entonces, ¿cuáles deberían ser los contenidos de su autoengaño? ¿Que la mala situación es buena? ¿O que es posible cambiarla? Quien se encuentra comprometido con la utilidad del autoengaño en *ambos* tipos de situaciones socava su habilidad para decir en qué tipo de situación se encuentra y, en consecuencia, que tipo de contenidos de autoengaño debe perseguir. Por consiguiente, observa Van Leeuwen, la política del autoengaño para la consecución de la felicidad parece tener una oportunidad de resultar mejor que la auto honestidad sólo en situaciones que son malas y en las cuales uno no puede alterar las circunstancias, y aun en ese caso el resultado de la política no es positivo. Pero el problema es que la política de autoengaño socava la propia habilidad para decir si se está en el tipo de situación para la cual el autoengaño podría ser apropiado. Entonces, si se persigue tal política, es posible hacerlo incluso en situaciones en las cuales hay disponibles opciones mucho mejores y racionalmente defendibles. Van Leeuwen considera, en resumen, que perseguir una política de autoengaño para la consecución de la felicidad es una mala política.

Las objeciones de Badhwar y Van Leeuwen, así como las críticas formuladas por Colvin y Block, arrojan fundadas dudas acerca de la real contribución de las ilusiones

positivas (y, *a fortiori*, del autoengaño) para nuestro bienestar y felicidad. Si bien la evidencia empírica puede ser favorable en alguna medida a la posición de Taylor (aunque, como hemos visto, no está exenta de controversias), los nexos conceptuales y, en especial, el escaso análisis de nociones fundamentales, como la de felicidad, debilitan fuertemente la pretensión de que las ilusiones positivas constituyen una vía para la mejora de la salud mental y un mayor desarrollo de la felicidad personal. Sin embargo, no se debería pensar que los trabajos de Taylor y sus colaboradores constituyen meramente una desviación no significativa con respecto a las concepciones tradicionales. Posiblemente han tenido el mérito, a nuestro modo de ver, de mostrar que la extendida creencia en que una percepción correcta de nosotros mismos y del mundo que nos rodea es invariablemente ventajosa admite algunas desviaciones. Si bien, como hemos visto, las observaciones de diversos filósofos respecto de la posible contribución del autoengaño a nuestro bienestar anteceden largamente las investigaciones empíricas sobre el tema, no debería negarse a Taylor y sus colaboradores el reconocimiento por el esfuerzo de aportar evidencia fáctica favorable a esa tesis.

3. Autoengaño y creencias religiosas

En buena parte de los análisis previos se ha hablado genéricamente de la creencia falsa, adquirida o mantenida, como producto del autoengaño. Como hemos visto en el capítulo II, hablar de tal producto no debe hacer olvidar que no es adecuado considerarlo meramente una creencia aislada, sino que es más plausible concebirlo como la cúspide de una jerarquía de creencias coherentes, mutuamente sustentantes y similarmente sesgadas. En otros términos, la creencia falsa producto del autoengaño forma parte de un sistema que, pese a la irracionalidad del proceso, posee una lógica interna indudable. Ahora bien, admitir que las creencias falsas pueden constituir un sistema coherente y autosustentante hace posible plantear un nuevo interrogante, que examinaremos brevemente en este apartado: ¿puede el autoengaño jugar un rol de importancia en sistemas de creencias compartidos socialmente; en particular, juega un rol en la adquisición o mantenimiento de las creencias religiosas?[104] Esta cuestión, en consecuencia, implica preguntarse si el

[104] Sería ingenuo pensar que un fenómeno que posee la ubicuidad del autoengaño sólo puede tener lugar en sistemas de creencias no caracterizados por el empleo de procedimientos racionales para la determinación de qué creencias deben ser aceptadas. En consecuencia, también es posible preguntarse si el autoengaño puede tener lugar en el ámbito de la ciencia. Esto es, dando por supuesto que la ciencia (o, mejor dicho, los científicos) *cree* que determinadas proposiciones son verdaderas o aproximadamente verdaderas, ¿es posible el

autoengaño puede desempeñar un papel más amplio en nuestra vida que el meramente individual, pregunta que tendrá conexión con los temas examinados en el siguiente apartado, referente a la posibilidad del autoengaño colectivo.

No sólo los filósofos se han ocupado de la pregunta relativa a si las creencias religiosas involucran alguna clase de proceso de irracionalidad motivada; algunos psicólogos han reflexionado también sobre el tema. Triandis (2013) sostiene que el autoengaño está implicado en diversos debates contemporáneos (como los relativos al calentamiento global, los excesos en el consumo, el aborto y la homosexualidad), debates que incluyen los relativos a la religión. Distingue dos clases de religiones: "externas" e "internas"; las primeras aceptan la existencia de entidades sobrenaturales, mientras que las segundas (como el Budismo en su manifestación original) se interesan por sucesos internos, como el logro de "iluminación" [*enlightenment*]. Mientras que las externas, sostiene, constituyen casos "clásicos" de autoengaño, las internas, por el contrario, lo evitan. Las religiones externas se basan en los deseos humanos de contar con entidades que nos protejan y nos guíen; en consecuencia, se inventarán dioses de acuerdo con estos deseos y necesidades. En sus propios términos: "Dios es un autoengaño maravilloso y cognitivamente simple. Se ajusta a nuestras esperanzas, necesidad y deseos de poseer una entidad poderosa que nos ayudará a ganar nuestras batallas" (p. 1072). La postulación de

---

autoengaño dentro de las comunidades científicas? Hay autores que han sostenido que los científicos son especialmente vulnerables a los procesos que conducen al autoengaño. Este es el caso de William Broad y Nicholas Wade (1982), quienes pasan revista a un considerable número de casos de fraude y autoengaño en la historia de la ciencia, que incluye los casos de los errores en la confirmación de la teoría de Copérnico por parte de Hooke y Flamsteed, el "descubrimiento" de los rayos N por el físico francés Blondlot, el "Hombre de Piltdown", las investigaciones sobre lenguaje animal (incluyendo al caso pionero de Clever Hans), el Sudario de Turín y otros; se mencionan también las investigaciones de Rosenthal sobre efectos de expectativa del investigador y el maestro. Según estos autores, el riguroso entrenamiento en observación objetiva que caracteriza a los científicos es a menudo una defensa endeble contra el deseo de obtener un resultado determinado; las expectativas de un experimentador acerca de aquello que verá darán forma reiteradamente a los datos que registra, en detrimento de la verdad. Agregan que el moldeamiento inconsciente de los resultados puede ocurrir en modos numerosos y sutiles, y que no se trata de un fenómeno exclusivamente individual, sino que a veces una comunidad entera de investigadores es víctima de una ilusión compartida. La expectativa, observan Broad y Wade, "conduce al autoengaño, y el autoengaño conduce a la propensión a ser engañado por otros. Los grandes engaños científicos (…) demuestran los extremos de credulidad a los cuales grandes científicos pueden ser conducidos por su deseo de creer. De hecho, los magos profesionales afirman que los científicos, a causa de su confianza en su propia objetividad, son más fáciles de engañar que otras personas" (p. 42). El estudio de Broad y Wade no profundiza en los mecanismos explicativos psicológicos o sociales que hacen posible las formas de autoengaño individual o colectivo en la ciencia (como el caso de Blondlot) más que en la observación de que los científicos, paradójicamente, son presa más fácil que otras personas. Sería muy poco plausible afirmar que la ciencia está exenta de engaño y autoengaño, como lo prueban los casos mencionados por estos autores y otros no mencionados; al fin y al cabo, la ciencia es desarrollada por seres humanos y no por ángeles o máquinas. Sin embargo, si la ciencia y los científicos fueran tan severamente vulnerables a los sesgos como sostienen los autores, es sumamente improbable que se hubiera llegado a los avances que hoy podemos observar. *Mutatis mutandis*, se podría aplicar a las afirmaciones de Broad y Wade lo que se ha observado sobre las concepciones relativistas e irracionalistas respecto de la ciencia: "si la comunidad científica es una mafia, ¿por qué fue la comunidad científica, y no la mafia, la que puso al hombre en la Luna?" (Comesaña, 2011, p. 220).

tales entidades sobrenaturales, en síntesis, constituye en sí misma un autoengaño. A diferencia de aquellas, observa, Buda aspiraba a percibir el mundo de un modo realista, sin que los deseos interfirieran en nuestra percepción de él; en consecuencia, sin autoengaño. Las religiones internas, en consecuencia, no son en sí mismas autoengañosas, a menos que incorporen lo que denomina "autoengaños culturales" (por ejemplo, la incorporación de entidades que se ajustan a los esquemas y prejuicios de la sociedad en la cual se difunde la doctrina original). Las consideraciones de Triandis adolecen, como hemos visto previamente, de una deficiencia usual en los estudios psicológicos sobre el autoengaño: la ausencia de una definición mínimamente estricta del fenómeno que permita diferenciarlo de otras formas de irracionalidad motivada, ya sea individual o colectiva. Entre otras muchas cuestiones que quedan sin responder, no es posible saber si se trata de casos de autoengaño o meramente pensamiento desiderativo. Como examinaremos a continuación, la determinación del presunto carácter autoengañoso de las creencias religiosas requerirá un análisis mucho más sutil.

Los interrogantes relativos a la medida en que la adopción de creencias religiosas requiere alguna clase de maniobra o estrategia contraria a la racionalidad teórica tienen una larga tradición en la historia del pensamiento; en particular, también la tiene la idea de que las religiones constituyen sistemas de creencias que distorsionan de modo sistemático la percepción correcta de la realidad.[105] Como hemos señalado al examinar la relación del voluntarismo doxástico con el autoengaño,[106] los argumentos de Pascal constituyen un ejemplo paradigmático en ese sentido.[107] No resulta obvio, sin embargo, que sea posible adoptar voluntariamente un sistema de creencias incompatible en principio con las mejores

[105] A Marx (recordemos aquí se célebre afirmación según la cual "la religión es el opio del pueblo" y a Nietzsche, en *El Anticristo*, se sumó posteriormente Freud. En *El porvenir de una ilusión* (1927), señala que las creencias religiosas no son decantaciones de la experiencia ni resultados de una actividad intelectual, sino *ilusiones*, cumplimientos de los más antiguos, intensos y urgentes deseos de nuestra especie: "Una ilusión no es lo mismo que un error; tampoco es necesariamente un error (…) Lo característico de la ilusión es que siempre deriva de deseos humanos; en este aspecto se aproxima a la idea delirante de la psiquiatría, si bien tampoco se identifica con ella, aun si prescindimos del complejo edificio de la idea delirante. Destacamos como lo esencial en esta última su contradicción con la realidad efectiva; en cambio, la ilusión no necesariamente es falsa, vale decir, irrealizable o contradictoria con la realidad (…) *Por lo tanto, llamamos ilusión a una creencia cuando en su motivación esfuerza sobre todo el cumplimiento de deseo; y en esto prescindimos de su nexo con la realidad efectiva, tal como la ilusión misma renuncia a sus testimonios* (p. 22. Cursivas nuestras). Así, la existencia de una Providencia divina bondadosa, señala Freud, permite calmar las angustias que se originan en los peligros de la vida, y satisfacer las demandas de justicia a través de un orden ético en el universo, demandas rara vez realizadas en la cultura humana; también, agrega, la existencia de una vida posterior a la terrenal genera las condiciones para el cumplimientos de tales deseos ancestrales. Todo esto, observa, implica un enorme alivio psíquico para el individuo, a quien se le quitan de encima los conflictos alrededor del complejo paterno no totalmente superados originados en la infancia. La descripción de Freud, cabe señalar, coincide claramente lo que se designa ampliamente como pensamiento desiderativo, esto es, una creencia orientada claramente por el deseo de cierto estado de cosas, no necesariamente falsa, pero sí huérfana de evidencia favorable.

[106] Capítulo II, § 2.

[107] Para un análisis acerca de la posibilidad de que los argumentos de Pascal impliquen el autoengaño, *cfr.* Jones (1998).

pruebas a disposición del agente. En este sentido, un caso notable de conflicto entre creencias religiosas y creencias científicas, que parecería constituir *prima facie* un ejemplo extremo de voluntarismo doxástico, es narrado por Richard Dawkins en su libro *The God's Delusion* (2006). Pese a que este episodio no constituye en absoluto un caso típico de autoengaño, sin dudas su análisis es sumamente interesante en este contexto. Dawkins describe allí la penosa historia del geólogo Kurt Wise, responsable del Centro para la Investigación de los Orígenes del Bryan College[108] en Dayton, Tennessee. Wise era un investigador joven y sumamente calificado, señala Dawkins, que podría haber desarrollado una brillante carrera científica. Sin embargo, sugiere, el conflicto entre su formación científica y su temprana educación en ideas religiosas fundamentalistas constituyó una fuente de sufrimiento crecientemente intolerable; la incompatibilidad entre una doctrina que lo obligaba a creer, entre otras cosas, en una Tierra de menos de diez mil años de antigüedad, por una parte, y la solidez empírica de las ciencias geológicas, por la otra, lo llevaron a un estado de desesperación que no podía sostenerse en el tiempo. Dawkins cita aquí al propio Wise, quien describe cómo dedicó esfuerzos sistemáticos a examinar la Biblia y a eliminar de ella todo lo que fuera incompatible con una visión científica del mundo. El resultado de este ensayo fue que, a su pesar, casi nada quedaba en pie del texto revelado. La solución que encontró para resolver este conflicto fue igualmente radical: optó por el texto bíblico y renunció a Teoría de la Evolución y a la ciencia. Dawkins remarca lo patética que, a su modo de ver, resulta la elección de Wise: todo lo que tenía que hacer era renunciar a la Biblia o bien interpretarla simbólica o alegóricamente, como hacen los teólogos. La diferencia que el presenta el caso de Wise con el de otros fundamentalistas, prosigue, es que éste es dolorosamente honesto. Sin embargo, muestra lo que usualmente ocurre de modo encubierto con los fundamentalistas cuando encuentran evidencia científica que contradice sus creencias: si todas las pruebas del mundo se inclinaran en contra del creacionismo, observa Wise, él sería el primero en aceptarlo, pero aún así seguiría adhiriendo al creacionismo, ya que esto es lo que indica la palabra de Dios.

Dawkins no examina la posibilidad de que el caso de Wise constituya un ejemplo notorio de autoengaño; en su libro, de hecho, el término "autoengaño" aparece sólo una vez a lo largo de la obra, para hacer referencia a la teoría de Robert Trivers sobre este fenómeno. Dawkins deja pasar, a nuestro modo de ver, una interesante oportunidad para analizar la teoría sobre la creencia religiosa como autoengaño. ¿Creía *realmente* Wise en la

[108] Llamado así en honor de William Jennings Bryan, el fiscal del proceso contra el maestro de ciencias John Scopes, llevado a juicio por enseñar a sus alumnos la Teoría de la Evolución. Este proceso, conocido como "Juicio del Mono", tuvo lugar en Dayton en 1925.

verdad del creacionismo, aun cuando supiera que la evidencia científica en contrario resulta abrumadora? La situación hipotética que plantea el mismo Wise, esto es, el caso en el que aun cuando toda la evidencia del universo se volcara contra el creacionismo, él seguiría creyendo en su verdad, ¿es psicológicamente posible? Parece razonable pensar que no: o bien Wise no cree *realmente* en la verdad de la evidencia científica, o bien su creencia en la verdad del creacionismo es un artificio. Quizás sería posible pensar que Wise no cree realmente en la verdad del creacionismo, sino que meramente *cree* que cree en ella; esto es, *cree* que efectivamente acepta de buena fe la doctrina creacionista, pero de hecho no lo hace, simplemente se equivoca al sostener que lo hace. Esta posibilidad parece relacionarse con un enfoque que hemos mencionado en el capítulo I, según el cual es posible explicar ciertos aspectos intrigantes del autoengaño por medio de la estrategia de negar que el producto de este fenómeno sea una creencia; en vez de ella, el resultado del autoengaño sería una *manifestación* [*avowal*] (Audi, 1982; Rey, 1988). Para ambos autores una manifestación es básicamente una disposición a afirmar una proposición con "sinceridad"; sin embargo, tal proposición carece de conexiones profundas con la acción. Entonces, si quien se autoengaña meramente manifiesta que *p* encontrándose en el estado de autoengaño de que *p*, no sería necesario atribuir al agente tanto la creencia de que *p* como la creencia de que no *p*, con lo cual se evitaría el problema de explicar la coexistencia de creencias contradictorias. Cuando un agente está autoengañado respecto de que *p*, tanto Audi como Rey sostienen que el agente posee una creencia (que Rey llama "creencia central") de que no *p*. Como ha observado Van Leeuwen (2007) la disociación entre creencias y acciones es un componente esencial del enfoque de la manifestación; de otro modo resultaría completamente oscuro en qué respecto se supone que una manifestación es distinta de una creencia genuina. Para Rey, podría decirse que, *ceteris paribus*, una persona cree manifiestamente que *p* si afirmara sincera y decididamente que *p* si se le preguntara al respecto. Creencias manifestadas, para este autor, son aquellas que se atribuyen sobre la base de la conducta verbal, mientras que creencias centrales son aquellas que se atribuyen sobre la base de las acciones. Audi, a su vez, advierte que el hecho que S esté autoengañado con respecto a *p* no implica su creencia –incluso consciente- de que *p*; lo que su posición requiere respecto de la actitud positiva de S hacia *p* es que S esté dispuesto sinceramente a manifestarlo.[109] Si se aplica esta perspectiva al caso de Wise, se diría que éste no *cree* realmente en la verdad del creacionismo, sino que sólo *manifiesta sinceramente* su creencia en

[109] Como otros debates filosóficos acerca del autoengaño, la viabilidad del enfoque de la manifestación permanece como una cuestión abierta. Para un examen acerca de si el resultado del autoengaño puede ser una manifestación, y no una creencia, cfr. Van Leeuwen (2007).

él. Sin embargo, este enfoque no parece ser suficiente para explicar el caso. En efecto, Wise, a diferencia de supuestos agentes que *manifiestan*, pero no *creen*, parece efectivamente comprometido desde la acción con su credo religioso; su función como director de un centro de enseñanza en doctrinas fundamentalistas desmiente una adhesión meramente verbal a tales creencias. En consecuencia, la disociación entre la manifestación y la acción que requiere de modo esencial el enfoque defendido por Rey y Audi parece estar claramente ausente en este caso.

El caso de Wise, que parece refractario a explicaciones simples, trae también a colación la cuestión relativa a la complejidad de la evidencia relativa a ciertos casos de autoengaño y a la dificultad para determinar su verdad. [110] Wise, como hemos visto, sostiene que hay razones para aceptar la doctrina creacionista respecto de la edad de la Tierra. Más allá de que esta supuesta evidencia con toda seguridad está groseramente sobreestimada, no hay duda de que la complejidad de la evidencia científica relativa a esa y a otras cuestiones conexas (así como en los casos de autoengaño en cuestiones ideológicas y políticas) es inconmensurablemente mayor que en los ejemplos típicos que pueblan la bibliografía filosófica (por ejemplo, la madre que descree de que su hijo use drogas pese a las abundantes pruebas en contrario).[111] El ítem relativo a la naturaleza de la evidencia y a la verdad de las proposiciones que constituyen las doctrinas religiosas, de hecho, será un aspecto importante en algunos exámenes acerca de la posibilidad de que las creencias religiosas sean un producto directo del autoengaño, como veremos a continuación.

J. Räikkä (2007) señala que la pregunta referente a la relación entre creencias religiosas y autoengaño ha recibido escasa atención, en comparación con otros debates filosóficos sobre el problema. Su objetivo es analizar si las creencias religiosas son o pueden ser autoengañosas en el sentido ordinario del término, y observa que hay tres respuestas a la pregunta anterior. Puede afirmarse que las creencias religiosas a) siempre, b) en ocasiones, o c) nunca son autoengañosas. Räikkä defiende la tesis de que no puede decirse de las creencias religiosas típicas que son autoengañosas en el sentido común del término. En consecuencia, argumenta en contra de la afirmación de que las creencias religiosas son siempre autoengañosas (las religiones son formas de autoengaño colectivo) y también contra la concepción moderada de que algunas creencias religiosas están basadas en el autoengaño, mientras que otras no lo están.

---

[110] No carece de interés analizar el caso de Wise bajo los términos de la teoría de la disonancia cognitiva, y en particular examinar los vínculos de este fenómeno con el autoengaño. Para un estudio acerca de estas relaciones cfr. Scott-Kakures (2009).

[111] Este aspecto volverá a tener relevancia al momento de examinar los presuntos casos de autoengaño colectivo, como veremos en el apartado siguiente de este capítulo.

Una creencia religiosa típica es que de una u otra forma la vida continúa después de la muerte, afirmación incluida en muchas religiones. Räikkä pasa revista a las distintas concepciones sobre esta otra vida en distintas religiones, antes de plantear por qué alguien podría sostener que tal creencia está basada en el autoengaño. Observa que no es necesario tener una actitud hostil hacia la religión para sostener que la creencia en la vida después de la muerte está muy probablemente basada en el autoengaño, lo que se debe a que parece satisfacer dos criterios básicos para las creencias autoengañosas. Cuando una persona se engaña a sí misma, 1) cree algo que desea que sea verdadero, y 2) el objeto de la creencia no es apoyado apropiadamente por la evidencia. Es quizás imposible, observa, demostrar que la vida no continúa después de la muerte; sin embargo, hay ciertamente fuertes razones para pensar que lo anterior es verdadero. La continuidad de la vida después de la muerte es una idea problemática, tanto conceptual como físicamente. Sin embargo, observa, las personas tienen fuertes motivos para sostener la creencia en cuestión. Una persona normal no desea morir, y vivir después de la propia muerte es ser, en algún sentido, inmortal. También es reconfortante pensar en que amantes, familiares o amigos fallecidos continúan viviendo en alguna parte, y quizás sea posible encontrarse con ellos nuevamente. Por último, es muy importante tener en mente, señala Räikkä, que de acuerdo con muchas religiones tal continuidad implica una corrección a todas las injusticias de este mundo. Por estas razones, considera que la afirmación de que la vida después de la muerte está basada en el autoengaño tiene cierta plausibilidad.

Pese a lo anterior, Räikkä considera que hay razones para sostener que quienes creen en la vida después de la muerte *no* se autoengañan. Lo que es común a todos los casos de autoengaño, observa, es que una persona cree contra la evidencia de un modo que otras personas encontrarían sorprendentemente extraño si tuvieran la misma evidencia y no se autoengañasen. Ahora bien, un observador ideal no podría decir que una persona que cree en la vida después de la muerte está equivocada del mismo modo en que lo están quienes se autoengañan (por ejemplo, cuando una madre cree, contra la evidencia, que su hijo no usa drogas). Resulta conceptualmente necesario para el autoengaño que las creencias autoengañosas sean falsas;[112] no hay pruebas, sin embargo, de que la creencia en cuestión sea falsa. Räikkä agrega que, por supuesto, si no hubiera vida después de la muerte, entonces las personas que creen en ella estarían en un error. Pero es imposible decir si hay o no vida después de la muerte; en consecuencia, no podemos saber si las

[112] Conviene recordar que esta afirmación, aunque plausible, no goza de unanimidad.

creencias religiosas son a veces autoengañosas.[113] Si bien la creencia en la vida después de la muerte es problemática, no puede decirse de ella que sea falsa como sí puede decirse de la creencia de la madre engañada respecto de su hijo. Lo mismo puede decirse, señala, de otras creencias a las que llamamos "religiosas" (por ejemplo, la creencia de que Dios existe).[114] Sin duda, reconoce, sería difícil dar criterios específicos para las creencias que cuentan como "religiosas", pero aun así hay casos claros, tales como la creencia en Dios y la creencia en la vida después de la muerte.

Räikkä examina la objeción según la cual la afirmación de que quienes creen en la vida después de la muerte no pueden autoengañarse debido a la carencia de pruebas en contra de su creencia está basada en el supuesto de que es conceptualmente necesario para el autoengaño que las creencias autoengañosas sean falsas.[115] Pero, advierte la objeción, esta última afirmación sobre la necesidad conceptual es incorrecta. La objeción posee cierta plausibilidad, admite Räikkä, dado que es claro que un procesamiento incorrecto de la información (un tratamiento motivacionalmente sesgado de los datos) puede conducir a una creencia verdadera; si bien esto puede ocurrir, es cuestionable que represente autoengaño, más que irracionalidad. Mele, observa, probablemente está en lo correcto cuando argumenta que desde un punto de vista puramente léxico el autoengaño debe incluir una creencia falsa. Ciertamente, agrega, sería raro culpar a una persona de autoengaño cuando ella está absolutamente en lo correcto respecto de lo que piensa.

Una segunda objeción enfatiza el elemento subjetivo del autoengaño. Räikkä presenta el caso de una madre atea que, al perder a su hijo en un accidente, súbitamente pasa a creer en la vida después de la muerte. En este caso, observa, existe la tentación de considerarla autoengañada, ya que la única razón por la cual ella cambia su creencia parece estar relacionada con sus deseos y temores; ella conocía bien los problemas de la creencia en la vida después de la muerte. La objeción, observa Räikkä, descansa en una percepción correcta. Sin embargo, señala, es útil notar que la madre no se autoengaña en un sentido normal. Compara este caso con el caso en el cual a la madre se le informa que su hijo ha muerto en un accidente, y pese a la evidencia a favor, rechaza creerlo, engañándose a sí

[113] Manuel Comesaña (comunicación personal) me ha señalado que es como mínimo plausible la idea de que la suspensión del juicio es racional cuando hay equivalencia de razones. Si esto fuese correcto, en ausencia de razones en favor de la existencia de vida después de la muerte no deberíamos creer en ella; en todo caso, la carga de la prueba recae en quien sostiene que tal cosa existe.

[114] Räikkä tiene razón en señalar que las creencias religiosas (así como, podríamos agregar, con diferencias de grado, las creencias relativas a cuestiones más allá del alcance de nuestro conocimiento directo, como las cuestiones históricas) presentan diferencias importantes respecto de las creencias involucradas en los ejemplos usuales de autoengaño. Pero es dudoso que estas diferencias sean tan grandes como para anular toda posible acusación de autoengaño.

[115] No todos los autores están de acuerdo con esta especificación. Véanse los argumentos de Lynch en este mismo apartado.

misma. La diferencia entre los dos casos reside en que sólo en el segundo caso la persona se engaña a sí misma en el sentido normal, al interpretar los datos de un modo motivacionalmente sesgado y aceptar una creencia que es claramente falsa. Respecto del primer caso podemos hablar de autoengaño sólo en un sentido extendido, y deberíamos ser cuidadosos en no confundir ambos.[116]

Räikkä se plantea por último si es trivial afirmar que típicamente las creencias religiosas no son nunca autoengañosas en el sentido ordinario. Esto podría ser el caso si las creencias religiosas no fueran creencias en absoluto. Discutiblemente, si S realmente cree que *p*, entonces debe tener que creer que hay evidencia adecuada para *p*. Sin embargo, una persona religiosa puede decir muy sinceramente que cree que Dios existe, pero que no está justificada en creer que Dios existe. Esto sugiere que (algunas) creencias religiosas no son realmente creencias, y que en consecuencia no pueden ser creencias autoengañosas. Este argumento sugiere preguntas interesantes, observa Räikkä, (por ejemplo, como definir "creencia"), pero añade que aquí es suficiente puntualizar que la conclusión del argumento es consistente con su posición.

Una primera réplica obvia al argumento de Räikkä consiste en señalar que, aun cuando sea correcto que el caso de la creencia en la vida después de la muerte constituya un ejemplo innegable de creencia religiosa, pero no un caso de autoengaño, es altamente probable que existan casos de creencias que claramente califican como religiosas (por ejemplo, la creencia en la existencia de ciertos fenómenos sobrenaturales, como los milagros, que desafían todo lo que se sabe en la ciencia moderna) y cuya aceptación requeriría del autoengaño; en consecuencia, no podría afirmarse, como pretende Räikkä, que las creencias religiosas *nunca* son sostenidas por medio del autoengaño. Sin embargo, perseguir esta línea argumentativa conduciría sin duda a la intrincada cuestión relativa a cuáles son las creencias religiosas, lo que excedería con mucho los objetivos de este apartado. Ahora bien, es posible que no haga falta profundizar en la línea anterior para mostrar que Räikkä no consigue probar adecuadamente su tesis.

Lynch (2010) señala que el argumento de Räikkä está basado en una cuestión conceptual acerca del autoengaño, esto es, que el autoengaño debe involucrar una creencia

[116] Räikkä reconoce que no ha proporcionado una definición específica del autoengaño, y que ésa es una tarea por derecho propio. Para sus propósitos presentes, observa, es suficiente caracterizar el autoengaño de una manera general. "Quienes se autoengañan tienen éxito en interpretar la evidencia de modo tal que les hace posible adoptar proposiciones que desean aceptar. Debido a razones puramente léxicas, el autoengaño debe incluir una creencia que sea falsa, si bien debe admitirse que el tratamiento motivacionalmente sesgado de los datos puede en ocasiones conducir a una creencia que sea verdadera. Las dudas y sospechas molestas son comunes en el autoengaño, pero no son conceptualmente necesarios para él (p. 523)."

falsa, e intenta mostrar que el argumento fracasa una vez que se advierte la distinción entre estar autoengañado *en* creer que *p* y engañarse a uno mismo *para* creer que *p*. Denomina "Condición de Falsa Creencia" a la condición enunciada por Räikkä por la cual quienes creen en la vida después de la muerte no pueden ser acusados de autoengaño (debido a que no es posible probar la falsedad de esa creencia).

Lynch observa que inicialmente Räikkä da la impresión de construir un argumento que pruebe que las creencias religiosas no pueden ser clasificadas como autoengañosas, pero que, finalmente, concluye con una afirmación mucho más débil. Los observadores objetivos, según el argumento de Räikkä, mayoritariamente concluirían que la evidencia disponible no justifica concluyentemente ni la creencia en la vida después de la muerte ni la creencia contraria. Pero las personas religiosas creen en esa vida. Entonces, al menos un buen número de ellos probablemente crea de un modo injustificado, debido a que su deseo sesga su juicio. Sin embargo, por lo que sabemos, podría no existir la vida después de la muerte. Si este fuese el caso, señala Lynch, entonces la Condición de Falsa Creencia sería satisfecha, y estas personas estarían autoengañadas después de todo (aunque esto sería desconocido para nosotros). Por lo tanto, dice Lynch citando a Räikkä, "la conclusión debe ser que *no podemos saber* si las creencias religiosas son a veces autoengañosas" (p. 1074. Cursivas del autor) . Y esta afirmación más débil, observa Lynch, parece ser todo lo que Räikkä está autorizado para concluir a partir de sus supuestos. En otras palabras, no es que la Condición de Falsa Creencia no sea satisfecha en tales casos, sino que no sabemos (y no podemos saber) si lo es; no podemos saber si quienes creen en la vida después de la muerte se autoengañan.

Más aún, Lynch señala que, sin disputar la visión general del autoengaño de este autor, considera que la Condición de Falsa Creencia lo deja expuesto a objeciones de mayor peso. Pese a la defensa de Räikkä de la posición según la cual el autoengaño debe incluir una creencia falsa, descuida el hecho de que hay otra construcción gramatical para atribuir autoengaño que no acarrea esta implicación de falsa creencia. Esta construcción tiene la forma "*S* se engaña a sí mismo para creer que *p*". "Estar autoengañado *en* creer que *p*", y "haberse engañado a uno mismo *para* creer que *p*" deberían ser distinguidas, observa Lynch, y sugiere que si bien es contradictorio decir que *S* está autoengañado al creer algo que es verdadero, no es contradictorio decir que *S* se engaña a sí mismo para creer algo que es verdadero. La idea anterior puede ser apoyada, observa Lynch, por los casos de engaño interpersonal. Supóngase que yo sé que *p* es verdadero y sé que X sólo creerá que es verdadero si escucha que Z considera que lo es, ya que Z es la única persona en la que X

confía. En tal caso yo puedo mentirle a X diciéndole que he estado en contacto con Z, y que él ha dicho que *p* es verdadera. En esta situación X no está engañado al creer que *p*, ya que *p* es verdadera. Sin embargo, podemos decir que X fue engañado por mí *para creer* que esa proposición es verdadera.

La cuestión crucial acerca del autoengaño, entonces, concluye Lynch, parece yacer en creer contra buena evidencia a causa de un deseo o temor, y que la creencia sea verdadera o falsa no parece ser una consideración crítica. Por lo tanto, si bien el creyente en la vida después de la muerte puede no estar autoengañado en creer esto (aunque no podemos estar seguros) bien puede haberse engañado a sí mismo para creer en la vida después de la muerte, aun cuando haya una.

Los argumentos de Lynch parecen sólidos. Sin embargo, hay razones adicionales para cuestionar la defensa del carácter no autoengañoso de las creencias religiosas. Räikkä sostiene que, como no podemos saber si la creencia en la vida después de la muerte es falsa, entonces no es posible predicar autoengaño. Sin duda (y el hecho de que él mismo admita que la idea es problemática, tanto desde el punto de vista conceptual como físico, pero esto no es suficiente para concluir que es falsa) está adoptando aquí, de modo implícito, un sentido fuerte de la noción de conocimiento proposicional; esto es, para predicar autoengaño deberíamos poseer pruebas concluyentes de que la proposición "existe vida después de la muerte" es falsa. Si no fuese así, las razones invocadas por Räikkä para descreer en la vida después de la muerte serían suficientes para satisfacer una de las condiciones necesarias para afirmar que quienes creen en ella se han engañado a sí mismos. Si bien parece muy plausible la idea de que es más difícil obtener pruebas aceptables de algunas creencias, más que de otras, ¿tenemos en algún caso pruebas *concluyentes* de que una creencia es falsa? Si no fuese así, el argumento de Räikkä tendría la incómoda consecuencia de que nunca podríamos predicar autoengaño; en consecuencia, el argumento de Räikkä probaría más de lo que él desea, ya que deja en el mismo barco a todos los casos de autoengaño, y no sólo a los presuntos casos de autoengaño relativos a las creencias religiosas. Y si nunca fuese posible decir que alguien está autoengañado, entonces el término carecería por completo de casos seguros de aplicación, lo que lo tornaría completamente inservible. Räikkä observa lo siguiente: "Lo que tienen en común todos los casos de autoengaño es que una persona cree contra la evidencia de un modo que otras personas encontrarían sorprendentemente extraño si tuvieran la misma evidencia en su poder y no estuviesen también en un estado de autoengaño (…) ¿Qué diría un observador

objetivo acerca de la creencia en la vida después de la muerte? ¿Diría el observador que la creencia es falsa como lo es el pensamiento desiderativo de una madre cuando cree contra obvia evidencia que su hijo no consume drogas? Por supuesto que no. Un observador ideal apuntaría a que hay mucha evidencia que muestra que la creencia en la vida después de la muerte es problemática, pero también a que *no hay evidencia concluyente* que muestre que la creencia es falsa" (p. 520. Cursivas nuestras). Notemos que Räikkä no sostiene que las creencias religiosas requieren, para ser aceptadas, de una clase de evidencia cualitativamente diferente de la que requieren otras creencias. Lo expuesto no alcanza, en consecuencia, para asignar un carácter especial a las creencias religiosas, esto es, para poder sostener que las creencias religiosas presentan diferencias de clase, y no de grado, con otras creencias.

Räikkä tiene razón en señalar que las creencias religiosas presentan diferencias importantes respecto de las creencias involucradas en los ejemplos usuales de autoengaño. Pero resulta al menos dudoso que estas diferencias sean tan grandes como para anular toda posible acusación de autoengaño o, si se acepta el argumento de Räikkä, de ninguna creencia podemos decir que ha sido adquirida o mantenida por medio del autoengaño. Y, si esto último resulta inadmisible, parece inevitable concluir que las creencias religiosas (o al menos un subconjunto de ellas) tienen un lugar bien ganado dentro del dominio de la irracionalidad motivada.

4. *¿Autoengaño colectivo?*

> Su mente se deslizó por el laberíntico mundo del *doblepensar*. Saber y no saber, hallarse consciente de lo que es realmente verdad mientras se dicen mentiras cuidadosamente elaboradas, sostener simultáneamente dos opiniones sabiendo que son contradictorias y creer sin embargo en ambas; emplear la lógica contra la lógica, repudiar la moralidad mientras se recurre a ella, creer que la democracia es imposible y que el Partido es el guardián de la democracia; olvidar cuanto fuera necesario olvidar y, no obstante, recurrir a ello, volverlo a traer a la memoria en cuanto se necesitara y luego olvidarlo de nuevo; y sobre todo, aplicar el mismo proceso al procedimiento mismo. Esta era la más refinada sutileza del sistema: inducir conscientemente a la inconciencia, y luego hacerse inconsciente para no reconocer que se había realizado un acto de autosugestión (G. Orwell, *1984*, pp. 43-44).

El fragmento de *1984* con el que se inicia este apartado ilustra de manera notable, a nuestro entender, la complejidad de los procesos de distorsión de las creencias cuando en

su formación intervienen no sólo mecanismos psicológicos individuales, sino también dinamismos sociales. No parece tratarse simplemente de casos de distorsión de creencias asimilables o reductibles a procesos de autoengaño individual; parecería que estamos ante procesos mucho más complejos en los cuales los sesgos cognitivos, motivaciones y emociones característicos del autoengaño individual juegan un rol importante, pero en modo alguno suficiente, para explicar tales fenómenos cuando tienen lugar en sistemas constituidos por seres humanos, ya sean estos pequeños (como la familia) o muy grandes (como sociedades enteras).

La idea de que existen mecanismos colectivos de distorsión de creencias, esto es, procesos que actúan de modo tal que las creencias resultantes no reflejan los hechos, en particular los hechos sociales, de manera fidedigna, se inserta en una larga tradición en la historia de las ideas. No sería arriesgado mencionar aquí la teoría de la ideología de Marx como una de las más influyentes perspectivas en esta línea de pensamiento. Por el contrario, otros pensadores que pueden considerarse precursores respecto del problema del autoengaño sólo han considerado de manera tangencial su dimensión social, como es el caso de Sartre.[117]

En este apartado examinaré brevemente, en primer lugar, el concepto marxiano[118] de ideología, tal como aparece en su trabajo *La ideología alemana* de 1845, en el cual este fenómeno es presentado como una forma sistemáticamente sesgada de concebir la realidad. Luego de esta breve revisión, voy a exponer someramente dos análisis de la teoría marxiana de la ideología que buscan establecer el posible rol del autoengaño en la conformación y mantenimiento de la ideología, debidos a J. Elster (1986) y a A. Wood (1988).[119] Luego revisaré algunas formulaciones recientes acerca de los procesos de distorsión colectiva de creencias, y sostendré que los fenómenos que suelen agruparse bajo el rótulo "autoengaño colectivo" tienen una complejidad mucho mayor que los procesos de autoengaño individual, y que tal rótulo induce a pensar en una similitud entre ambos que resulta ser sólo superficial.

---

[117] En referencia a su conocido ejemplo del camarero que cumple con celo excesivo su rol laboral, Sartre observa que este intento de mala fe está apoyado por el público en general, el cual demanda de todas las personas al servicio de la industria que renuncien a su estatus como humanos autónomos y se agoten completamente a sí mismos en el servicio de su función social.

[118] Seguiré aquí el uso del término "marxiano" para hacer referencia a la obra del propio Marx y "marxista" para designar a la doctrina de los autores que adhirieron y continuaron la obra de Marx.

[119] La elección del estos autores y textos no es casual. Elster es uno de los más importantes representantes de la corriente conocida como marxismo analítico, y también un importante estudioso de los problemas de la racionalidad; como tal, se ha ocupado en distintos trabajos sobre el problema del autoengaño (en su libro *Sour Grapes* y en la compilación *The Multiple Self*, así como en varios artículos). Respecto de Wood, además de ser un autor que ha escrito varios trabajos sobre Marx (incluyendo un libro), su ensayo se encuentra incluido en un libro titulado *Perspectives on Self-Deception*, en especial en una sección titulada "The social dimension of self-deception", ubicación que no deja dudas acerca de los propósitos del texto y el interés para nuestro análisis.

Aunque ocasionalmente se emplea la expresión "falsa conciencia" para hacer referencia a una concepción profunda y sistemáticamente distorsionada de la realidad se ha observado que, contra lo que a veces se cree, Marx no empleó esa expresión.[120] El término que sí ocupa un lugar central en la teorización de Marx es el de *ideología*, que es el que examinaremos.[121]

En *La ideología alemana* Marx señala que, al contrario de lo que ocurre en la filosofía alemana (en particular en la obra de aquellos a quienes denomina "jóvenes hegelianos"), su propio examen no parte de lo que los hombres dicen, se representan o se imaginan, ni tampoco del hombre pensado, representado o imaginado, para llegar luego al hombre de carne y hueso. El punto de partida es el hombre que realmente actúa; de allí, de su proceso de vida real, se expone el desarrollo de los reflejos ideológicos y de los ecos de este proceso. De este modo, la moral, la religión, la metafísica y cualquier otra ideología y las formas de conciencia que a ellos correspondan pierden la apariencia de su propia sustantividad, característica de la filosofía alemana. No tienen su propia historia ni su propio desarrollo, sino que los hombres que desarrollan su producción material y su trato material cambian también, al cambiar esta realidad, su pensamiento y los productos de su pensamiento. Aquí Marx formula su famosa frase: "No es la conciencia la que determina la vida, sino la vida la que determina la conciencia".

Una vez afirmada esta relación entre la producción de las condiciones materiales de existencia y la producción de las formaciones ideológicas (religión, filosofía, moral), Marx observa que las ideas de la clase dominante en una sociedad determinada son las ideas dominantes en cada época. Esto es, la clase que ejerce el poder material es simultáneamente la clase que sustenta el poder espiritual dominante. La posesión por parte de esta clase de los medios para la producción material involucra también la posesión de los medios para la

[120] Según Little (2007), la expresión "falsa conciencia" fue introducida de manera sistemática en la teoría marxista a través del escrito del filósofo Georg Lukacs "Class Consciousness", en su libro *History and Class Consciousness. Studies in Marxist Dialectic*. Sin embargo, antes que Lukacs, Engels empleó esta expresión: "La ideología es un proceso realizado conscientemente por el así llamado pensador, en efecto, pero con una conciencia falsa; por ello su carácter ideológico no se manifiesta inmediatamente, sino a través de un esfuerzo analítico y en el umbral de una nueva conyuntura histórica que permite comprender la naturaleza ilusoria del universo mental del período precedente" (carta de Engels a Mehring de 14 de junio de 1893).

[121] El concepto de "ideología" no es el único concepto de la obra de Marx relacionado con una visión sistemáticamente sesgada de la realidad. Little (2007), señala que otros conceptos cercanamente relacionados son los de "fetichismo de la mercancía" y "mistificación", mientras que Wood (1988), menciona "ilusión social" y, al igual que el anterior, "fetichismo de la mercancía". Por otra parte, *La ideología alemana* no es el único texto de Marx que se ocupa el problema de la ideología, pero sin duda es un texto de primera importancia y es más que suficiente para nuestro propósito de mostrar algunos antecedentes del problema que nos ocupa.

producción espiritual. En opinión de Marx, esto redunda en que quienes no poseen los medios necesarios para la producción espiritual se sometan a las ideas de la clase dominante. Tales ideas no son más que la expresión ideal de las relaciones materiales dominantes. Los miembros de la clase dominante son conscientes de su predominio y, en consecuencia, es esperable que extiendan su dominación a todo el ámbito de una época histórica; de este modo, también lo harán como pensadores y como quienes regulan la producción y distribución de las ideas de su tiempo, que no son más que las ideas dominantes generadas por ellos.

La división del trabajo, señala Marx, se extiende también en el interior de la clase dominante como división del trabajo material y espiritual. Una parte de esta clase es la que proporciona sus pensadores, esto es, los ideólogos que generan para esta clase las ilusiones acerca de sí misma; otra parte adopta ante esas ideas una actitud en general pasiva y receptiva, lo que no se debe más que al hecho de que sus miembros, al ser integrantes activos de esta, disponen de poco tiempo para la formación de ilusiones acerca de sí mismos. Marx observa también que esta división al interior de la clase dominante es susceptible de originar cierta hostilidad entre ambas partes. No obstante, esta hostilidad se extingue en la medida en que surja cualquier conflicto práctico capaz de generar peligro para la clase misma; en esa circunstancia, además, desaparece cualquier apariencia de que las ideas dominantes no pertenecen a la clase dominante y que poseen un poder propio.

Si se olvida, sostiene Marx, la relación inseparable de las ideas con las condiciones de producción y los productores de esas ideas, se llegará a creer que, por ejemplo, en la época en la que dominó la aristocracia dominaron las ideas del honor, la lealtad y otras, y que cuando el predominio correspondió a la burguesía imperaron los ideales de la igualdad, la libertad y otras. Si bien así es como se imagina las cosas, por lo general, la clase dominante, esta concepción de la historia se encuentra necesariamente con el imperio de ideas cada vez más abstractas. Cada clase que llega a dominar se encuentra obligada, para lograr sus fines, a representar su propio interés como el interés general de la sociedad, esto es, a presentar sus ideas como las únicas racionales y provistas de vigencia absoluta.

Toda la apariencia de que la dominación de una clase dada es sólo el predominio de ciertas ideas desaparece, observa Marx, cuando la dominación de clases deja de ser la forma característica de organización de la sociedad; en consecuencia, ya no resulta necesario que un interés particular sea presentado como el interés de la sociedad en su totalidad.

En el capítulo 9 de su libro *An Introduction to Karl Marx* (1986), Jon Elster analiza el concepto de ideología sobre la base de distintos textos marxianos, no sólo *La ideología alemana*, sino también en *Theories of Surplus-Value*, y "Contribution to the critique of Hegel's Philosophy of Law: Introduction". En particular, nos interesará de este capítulo considerar las explicaciones que Elster proporciona del surgimiento e implantación colectiva de las formaciones ideológicas; en este examen encontraremos antecedentes de análisis recientes del autoengaño.

Elster señala, respecto de las fuerzas que moldean y mantienen al pensamiento ideológico, que la respuesta estándar del marxismo oficial es el *interés*; más específicamente, el interés de la clase dominante. La cuestión central (no resuelta, dice, por el marxismo o el propio Marx) es el *cómo*, la manera en que el interés de la clase dominante moldea las concepciones de otros miembros de la sociedad. La idea según la cual los dominadores y explotadores moldean el mundo de los oprimidos por una manipulación consciente y cínica es demasiado simplista; el cinismo de los dominadores conduciría al cinismo, no a la creencia, en los dominados. Conversamente, observa Elster, el adoctrinamiento exitoso requiere que los dominadores crean en aquello que predican. No hace falta decir, agrega, que el mero hecho de que la clase dominante se beneficie de las ilusiones de otros sujetos no prueba que sea causalmente responsable por ello.

Marx no siempre adhirió a su respuesta oficial, observa Elster, sino que también sugirió que las ideologías pueden surgir o enraizarse espontáneamente en las mentes de los sujetos, sin asistencia alguna de otros. Las creencias ideológicas compartidas surgirían entonces en dos modos: pueden emerger simultánea y espontáneamente en las mentes de muchas personas, quienes son expuestas a similares influencias externas y sujetas a procesos psicológicos similares. También pueden surgir en la mente de una persona y diseminarse a otras personas, quienes por alguna razón están dispuestas a aceptarlas.

Señala Elster que hay dos clases de actitudes que están sujetas a sesgos ideológicos: afectivas y cognitivas, o "calientes" y "frías". Lo que las personas valoran para sí mismas, lo que creen que es moralmente requerido para ellos mismos o para otros, cómo piensan que los bienes sociales deben ser distribuidos, son cuestiones que involucran directamente sus pasiones. Lo que ellos creen con respecto a cuestiones particulares de hecho y a conexiones causales generales no son cuestiones que en sí mismas involucren sus pasiones, excepto posiblemente su pasión por la verdad. Los sesgos que forman las actitudes ideológicas también pueden ser en sí mismos afectivos o cognitivos, calientes o fríos. En consecuencia, es posible distinguir entre cuatro tipos de actitudes ideológicas.

En primer lugar, las actitudes afectivas pueden ser formadas por procesos afectivamente sesgados. Esto ocurre en el caso descripto en la fábula de la zorra y las uvas: los agentes ajustarán sus aspiraciones a lo que parece factible, y evitarán vivir con la tensión y frustración causada por el deseo de lo inalcanzable. Segundo, las motivaciones calientes pueden ser formadas por factores cognitivos fríos. En tercer lugar, las actitudes cognitivas son a menudo formadas por procesos motivacionales, como ocurre con fenómenos como el pensamiento desiderativo, el autoengaño y otros. Por último, la cognición puede estar sujeta a distorsiones específicamente cognitivas, como cuando las personas tienen mucha confianza a partir de muestras pequeñas u otros modos de ignorancia de los principios básicos de inferencia estadística.

De estos mecanismos, señala Elster, todos menos el segundo tienen alguna importancia en la teoría de la ideología de Marx. El primero subyace a la a menudo citada afirmación de Marx de que "la religión es el opio del pueblo". El tercero opera en la selección de concepciones del mundo: entre las muchas perspectivas diferentes de la causación económica y social, cada grupo o clase seleccionará uno que parezca justificar su consideración especial por sus intereses. El último es importante cuando Marx sugiere que la *posición de clase* más que el *interés de clase* es la fuente del pensamiento ideológico. En consecuencia, Elster considera que los estudios reales de Marx sobre el pensamiento ideológico difieren de su "teoría oficial" –esto es, que las ideas dominantes son aquellas ideas que sirven al interés de la clase dominante- en dos sentidos. En primer lugar, cuando refiere al interés como una explicación de la ideología, es a menudo en un modo causal más que funcional. En vez de apuntar a las consecuencias de una cierta creencia con respecto a ciertos intereses, señala Elster, Marx cita el interés como la causa de la creencia. No puede concluirse que las creencias generadas por el interés servirán al interés del creyente, agrega, debido a que "las creencias formadas por la pasión sirven a la pasión malamente", o que servirán al interés de la clase dominante dado que algunas de las creencias de esa clase pueden en sí mismas ser formadas por el interés. En segundo lugar, tanto la posición de clase como el interés de clase entran en la explicación del pensamiento ideológico.[122] Tales ilusiones basadas en la clase no tenderán a servir al interés de los miembros de la clase o de la clase dominante si sus miembros son también víctimas de este mecanismo.

Elster señala que la concepción según la cual todo fenómeno superestructural (como la ideología) tiende a estabilizar la estructura económica sirviendo a los intereses de la clase dominante es errónea, así como la idea de que la superestructura puede ser

[122] En *Sour Grapes* Elster define la ideología como "un conjunto de creencias o valores que pueden ser explicados a través de la posición o el interés (no cognitivo) de algún grupo social" (p. 141).

explicada de esta manera. Aun cuando ciertas creencias sirvan a los intereses de la clase dominante, esto no necesita ser parte de la explicación; la explicación puede ser hallada eventualmente en los intereses y las necesidades de los sujetos.

En su estudio de 1988, Wood procede a una revisión de los aspectos más destacados de la teoría de la ideología de Marx y, a continuación, plantea explícitamente la relación entre la ideología y el autoengaño. Por "autoengaño" Wood entiende una clase determinada de irracionalidad motivada, algo que pertenece al mismo género que la *akrasia*. Distingue ambas clases de irracionalidad, no obstante, en que en el autoengaño nos vemos forzados a explicar la irracionalidad en términos de la suposición de que la mente del sujeto se encuentra de algún modo "dividida"; los motivos y los mecanismos que producen la irracionalidad se encuentran excluidos de la percatación consciente del sujeto, y tal exclusión es motivada.

La ideología, observa Wood, es cualquier forma de conciencia que distorsiona o falsea la percepción de la realidad de las personas, y cuya prominencia social es explicada por su carácter funcional para el modo de producción predominante o para la promoción de los intereses de una clase social. En la medida en que la ideología distorsiona la realidad, agrega, y especialmente en la medida en que tiende a ocultar su propia influencia distorsionante, la ideología puede ser considerada una forma de engaño. No obstante, el autoengaño parece tener lugar en ciertos casos de ideología. Dado que las ideologías de la clase dominante típicamente representan a los miembros de esa clase bajo una perspectiva favorable, es relativamente fácil observar cómo tal sistema de creencias responderá a los deseos de tal clase, y cómo los miembros de esa clase lo sostendrán de un modo que puede ser considerado como autoengañoso. Asimismo, observa Wood, Marx puso de manifiesto a menudo que las ideologías son enseñadas a la clase oprimida por los representantes pagos de la clase dominante (sacerdotes, periodistas, académicos y pedagogos) y sirven a esta clase al engañar a los oprimidos acerca de su condición. En tales casos los oprimidos no parecen víctimas de *auto*engaño, sino de una distorsión que les es impuesta. Reducir el análisis a esta visión, no obstante, ignoraría el hecho de que los oprimidos pueden lograr una suerte de confort en la creencia de que sus sufrimientos son inevitables o merecidos, y que podría resultar muy angustiante advertir que sus sufrimientos no son inevitables ni inalterables, especialmente si modificarlos puede ser percibido como difícil, costoso y riesgoso. Las ideologías (incluyendo aquellas que engañan a los oprimidos) frecuentemente resultan funcionales debido a que proveen el confort y consuelo que las personas desean; las

personas, en síntesis, están sujetas a condiciones que requieren de ilusiones. En consecuencia, las ilusiones de las personas operan a menudo por medio del autoengaño.

Pese a todo lo anterior, Wood no considera probable que el autoengaño desempeñe un rol en la mayoría de las ideologías, ya que si bien provee un mecanismo que puede prestar un servicio a la ideología, es dudoso que resulte necesario para ella. En el autoengaño, la toma de conciencia psíquicamente perturbadora está peligrosamente cercana, y el psiquismo del individuo debe adoptar procedimientos considerablemente drásticos para evitar el peligro. Un orden social cuya ideología funcional dependiera de mecanismos de autoengaño de sus miembros individuales, observa Wood, resultaría mucho menos seguro que otro que encontrara diferentes maneras de inducir en ellos las ilusiones necesarias. Por otra parte, señala Wood, hay varias consideraciones que muestran que las sociedades no necesitan depender del autoengaño. Los distintos órdenes sociales, sus perspectivas de cambio y las necesidades, capacidades e intereses de sus miembros son materias complejas. La verdad de las teorías acerca de ellas, sobre todo cuando se trata de sociedades en proceso de cambio, es difícil de establecer y probablemente inestable. En tal situación muy probablemente varios conjuntos de ideas favorables a la estabilidad social o a los intereses de una clase dada estarán disponibles, y tenderán a ser socialmente predominantes si los mecanismos sociales que regulan la producción y difusión de las ideas los favorecen.

Cada sociedad, observa Wood, requiere de algunos mecanismos para seleccionar cuales ideas van a formar parte de la ortodoxia pedagógica y científica, y cuales temas y teorías recibirán los mayores recursos de investigación. Estos mecanismos típicamente implicarán elecciones entre teorías rivales, elecciones que son realizadas por personas (políticos, burócratas, editores, quienes financian la investigación, o simplemente consumidores llanos) que no poseen tantos conocimientos especializados como aquellos que producen tales teorías. Es natural que las elecciones de tales personas, prosigue, sean influenciadas por sus convicciones, sostenidas honestamente y sin autoengaño, las cuales armonizan con sus intereses o prejuicios de clase. En consecuencia, las teorías que preferirán tenderán a armonizar con sus intereses o prejuicios de clase. Wood no niega que el autoengaño pueda estar a menudo involucrado en la producción de teorías, o en la selección hecha por aquellos a quienes el orden social posibilita la elección, o en ambas partes a la vez. Sin embargo, considera que no es difícil de entender como las ideas prevalecientes tenderán a estar en armonía con los intereses de clase y las necesidades del

modo de producción, aun cuando el autoengaño no constituya en absoluto un factor en juego.

La concepción de que la ideología opera a través del autoengaño, concluye Wood, conduce naturalmente al supuesto de que cuando las personas son "honestas", no experimentan tensión psíquica y no necesitan invertir la energía psíquica característica del autoengaño, la ideología no puede operar. Este supuesto, sin embargo, es uno de los errores que da a la ideología el manto de invisibilidad que requiere para actuar. La noción de que la ideología opera a través del autoengaño, finaliza este autor, es en sí misma un fragmento de ideología.

De lo expuesto se desprende que, para las dos perspectivas descriptas, el autoengaño tiene algún rol en la producción de ilusiones colectivas. Si bien tal rol no es central para ninguna de ellas, para Wood parece ser más marginal que para Elster. Parte de esta diferencia está relacionada directamente con la diferente caracterización que se hace de aquel fenómeno: la tensión psíquica que Wood identifica como un componente fundamental del autoengaño no aparece en Elster, lo que lo convierte en un mecanismo menos confiable para asegurar la estabilidad de las producciones ideológicas. En lo que sigue defenderemos una tesis que tiene cierta similitud con lo anterior, esto es, que el autoengaño colectivo no puede concebirse o explicarse como una simple suma de autoengaños individuales meramente fortalecidos por procesos sociales. Para esto no hay necesidad de adherir a la concepción materialista histórica de la ideología, y resulta posible explorar la posibilidad de formas colectivas de distorsión de creencias cuya interpretación no se base en sus premisas, cosa que haremos a continuación.

La distinción entre autoengaño individual y autoengaño colectivo ha recibido una atención considerablemente menor, en términos comparativos, que otros aspectos del fenómeno del autoengaño. Podría afirmarse que todo autoengaño tiene un componente social, ya que muy a menudo requiere colaboración social (Ruddick, 1988); en este sentido, la distinción entre ambos tipos de autoengaño puede parecer diluida en alguna medida. Sin embargo, no puede haber dudas de que existen diferencias importantes entre los casos autoengaño individual en los cuales sólo el agente que se autoengaña sustenta la creencia falsa, en contra incluso de la opinión de personas cercanas a él y en quienes confía, y los casos en los cuales son grupos, organizaciones o comunidades enteras las que sustentan la

creencia falsa. Para este tipo de fenómenos parecería razonable reservar la expresión "autoengaño colectivo".

No resultan infrecuentes en distintas disciplinas las referencias a las distorsiones en la percepción de la realidad que aquejan a totalidades mayores que el individuo: instituciones, organizaciones, comunidades o sociedades enteras. Suele hablarse de falta de memoria colectiva, de negación de realidades que son evidentes para observadores externos, de minimización de hechos de gravedad institucional o social. El sociólogo Stanley Cohen, en su libro *States of Denial* (2001), enumera una serie de expresiones que, en su opinión, expresan los comportamientos de personas y colectivos cuando son confrontados con información demasiado perturbadora, amenazante o anómala para ser completamente aceptada; en estos casos, señala, la información es en alguna medida reprimida, rechazada, ignorada o reinterpretada:[123] "enterrar la cabeza en la arena"; "vio lo que quería ver", "la ignorancia es felicidad"; "sólo oyó lo que quería oír"; "vivir una mentira"; "conspiración de silencio"; "no tiene nada que ver conmigo"; "no hay que hacer olas"; "no hay nada que pueda hacer acerca de eso"; "no puedo creer que eso esté sucediendo"; "no quiero saber/escuchar/ver nada más"; "la sociedad íntegra se encontraba en un estado de profunda negación"; "eso no puede ocurrirle a personas como nosotros";[124] "ignorancia deliberada"; "miró para otro lado"; "no lo admitió, ni siquiera ante sí mismo"; "no lavar la ropa sucia en público"; "no ocurrió bajo mi vista"; "debí haberlo sabido desde el principio".

En los últimos años la distorsión colectiva de las creencias ha sido mencionada en diversos escritos académicos que emplean el rótulo "autoengaño colectivo" (Dupuy, 2004; Güth & Kliemt, 2004; Runciman, 2008; Michel & Newen, 2010). Pero pese a este uso relativamente extendido, no son tan habituales los intentos de caracterizar con precisión el fenómeno al que se designa con esa expresión. *Prima facie*, con ella parecería hacerse referencia a un proceso de distorsión de las creencias similar al autoengaño individual, pero en un contexto social. Esto es, se trataría de un proceso en el cual cada agente es víctima de sus propias motivaciones y sesgos, que lo conducen a adoptar creencias distorsionadas de la realidad, proceso reforzado por la presencia de un contexto social que lo favorece y fortalece. Si bien lo anterior indicaría que tal expresión es ampliamente aceptada y su

[123] Traduzco aquí sólo algunas de las expresiones que menciona Cohen.

[124] No carece de interés señalar que algunas de estas conductas de distorsión colectiva de creencias, como la que precede a esta nota, quizás puedan estar relacionadas con sesgos ampliamente estudiados por la psicología social y que se producen con independencia del autoengaño, en particular, la llamada "creencia en un mundo justo". Este fenómeno consiste en culpar a las víctimas de sucesos negativos (como accidentes y crímenes), en un intento de apoyar la presunción de que a quien sostiene esa creencia no le ocurrirían tales desgracias. Cfr. al respecto van den Bos & Maas (2009), para un estudio reciente sobre este fenómeno.

significado es inequívoco, esto dista de ser claro cuando se analizan algunos de esos textos; no es inusual que se hable simplemente de "autoengaño colectivo" sin tratar de establecer de qué tipo de estado se está hablando. No resulta trivial, en consecuencia, la pregunta acerca de la naturaleza del autoengaño colectivo. Adelantaré aquí una respuesta a este interrogante: no existe algo como "autoengaño colectivo", si con tal expresión se hace referencia a un estado que consiste en la simple suma de autoengaños individuales sostenidos por el contexto social y relativos a una misma creencia falsa. La negación de la existencia del autoengaño colectivo, así entendido, no implica negar el hecho de que es posible que en ocasiones los grupos, las organizaciones e incluso sociedades enteras tengan percepciones severamente distorsionadas de sí mismas y del entorno que las rodea, y que tales percepciones pueden conducir a consecuencias desastrosas para sí mismas y para ese entorno. En lo que sigue sostendré que:

1. Lo que suele denominarse "autoengaño colectivo" no es simplemente una colección de autoengaños individuales reforzados por procesos de interacción social. Es un proceso en el cual la dimensión social tiene una relevancia explicativa como mínimo igual a la dimensión psicológica.
2. Los procesos de distorsión colectiva de la realidad incluyen interacciones sociales en diversas dimensiones, que involucran procesos de influencia en escala tanto micro y macrosocial como en dirección ascendente y descendente.
3. La expresión "autoengaño colectivo" tiene el efecto de generar la creencia en la existencia de un fenómeno social análogo al autoengaño individual, con el que en realidad tiene sólo una similitud superficial y engañosa.

Un ejemplo convincente de autoengaño colectivo podría constituir un primer paso en la dirección de confirmar o desestimar la existencia de tal fenómeno y, en caso de que esta se admita, determinar qué es lo que lo diferencia del autoengaño individual. Goleman (1989) describe un presunto caso de tal fenómeno.[125] Este autor considera que la acción de los mecanismos de producción de ilusiones (en el sentido dado por Taylor y Brown, a quienes cita) que actúan en el individuo y en pequeños grupos pueden también ser observados dentro de la sociedad como un todo. Cita al respecto un ejemplo al que considera excelente: la denominada "teoría de la percepción", doctrina enunciada por

[125] Conviene aclarar que la preocupación principal de Goleman no es el análisis detallado del supuesto fenómeno del autoengaño colectivo, sino las consecuencias colectivas del autoengaño, tanto individual como social.

estrategas del Pentágono en la década del '70. Los supuestos que subyacen a esta teoría son, simplificadamente, los siguientes:

a. En una era de capacidad termonuclear excesiva, ningún sistema nuevo de armas produce una diferencia real en el equilibrio del poder militar.
b. Sin embargo, si se logra que un nuevo sistema de armas parezca tener importancia militar, entonces tendrá importancia psicológica y, por lo tanto, política.
c. En consecuencia, los planificadores del Pentágono deberían actuar como si los nuevos sistemas de armas importaran militarmente, con lo que importarán psicológicamente.

La farsa anterior puede persistir, observa Goleman, a causa del rol que juega en las ilusiones en las que el colectivo mayor parece desear creer. Menciona al respecto una encuesta nacional realizada en 1986 en los Estados Unidos, la cual encontró que cerca del 90% de los encuestados estaba de acuerdo con la afirmación de que una guerra nuclear no puede ser ganada, mientras que a la vez el 70% estaba de acuerdo con la afirmación de que su país debía construir armas nuevas y mejores. Sostener esas creencias contradictorias,[126] observa Goleman, implica una ilusión de invulnerabilidad; es el sentir que otros se encuentran más en riesgo que nosotros mismos, que algo puede hacerse para protegernos incluso de la más abrumadora de las amenazas, la guerra nuclear. Es en el nivel colectivo, señala, en el que las ilusiones positivas descriptas por Taylor pierden su utilidad. Pueden tener el efecto de construir un "capullo" psicológico, un sentimiento de bienestar personal a expensas de una perspectiva clara de las amenazas que enfrentamos nosotros y nuestro planeta como un todo.

Independientemente del presunto bienestar psicológico o función defensiva presente en el caso particular, el ejemplo propuesto por Goleman no parece adecuado para elucidar la naturaleza del supuesto autoengaño colectivo (lo cual, además, no es su objetivo). Más allá del empleo de expresiones como "autoengaño masivo" (para hacer referencia a la falta de conciencia de nuestra especie de los peligros inminentes que nos acechan), o "autoengaño colectivo", no está claro que los fenómenos a los que se hace referencia en el trabajo merezcan realmente el rótulo de "autoengaño". Como se desprende bastante claramente del ejemplo, estamos en presencia de un fenómeno al que sería mejor calificar como engaño-autoengaño: existe un agente social interesado en la aceptación social de ciertas creencias falsas, y un conjunto de agentes sociales dispuestos a adoptar tales

[126] En realidad no lo son desde el punto de vista lógico, y posiblemente no sean siquiera incompatibles expresadas del modo en que aparecen; pueden revelar, no obstante, flaquezas en las concepciones individuales acerca del tema.

creencias. La creencia falsa sostenida socialmente no se origina en cada uno de estos agentes; surge como tentativa de engaño, intento en el que se cuenta con la colaboración de sus víctimas potenciales. Quizás corresponda aplicar, en este caso, la observación de Paul Ekman al describir la interacción entre Hitler y Chamberlain en el contexto de sus conversaciones previas a la Segunda Guerra Mundial: uno engañaba deliberadamente, y el otro deseaba ser engañado.

Si bien el ejemplo aportado por Goleman no contribuye de manera apreciable a una mejor comprensión del autoengaño colectivo, es posible encontrar un intento de caracterizar con más precisión este fenómeno y distinguirlo del autoengaño individual en Deweese-Boyd (2008). Este autor parte de la observación de que el autoengaño colectivo ha recibido escasa atención filosófica directa en comparación con la variante individual, e intenta lograr una caracterización "de trabajo" de la primera variante (ya que, al igual que el de Goleman, no es el objetivo principal del texto). Sobre la base de la adopción de una perspectiva "deflacionista" respecto del autoengaño individual,[127] describe al autoengaño como la adquisición y mantenimiento de una falsa creencia frente a fuertes pruebas en contra, causadas por deseos o emociones que sesgan el propio manejo de los elementos de juicio disponibles y pertinentes respecto de la creencia.

Ahora bien, Deweese-Boyd reconoce una ambigüedad relativa a la expresión "autoengaño colectivo": puede referir simplemente a una colección de individuos autoengañados de manera similar, o a una creencia autoengañosa sostenida por un colectivo tomado como un todo. Los especialistas en gnoseología social, señala, clarifican esta ambigüedad en la atribución de actitudes grupales por medio de la distinción entre comprensiones sumatorias y no sumatorias de las actitudes de los grupos. En una comprensión sumatoria, sólo los individuos que integran un grupo son objetos apropiados para la atribución de creencias, deseos y otros estados análogos. De acuerdo con esto, un grupo sostiene la creencia de que *p* sólo en el caso de que todos o la mayoría de sus miembros crean que *p*. En una comprensión no sumatoria, los grupos existen por su propio derecho por encima y más allá de los individuos que los componen y, como tales, son en sí mismos objetos apropiados para la atribución de propiedades psicológicas. Según sea la estructura del grupo, las creencias de los miembros o de subgrupos pueden diferir de las del grupo. En este sentido no sumatorio, el hecho de que todos o la mayoría de los miembros del grupo sostengan la creencia de que *p* no constituye una condición ni

[127] Véase al respecto el capítulo III, § 3.

necesaria ni suficiente para que el grupo sostenga la creencia de que *p*. Sobre la base de esta distinción, "autoengaño colectivo" puede ser entendido en un sentido sumatorio o no sumatorio. En un sentido sumatorio, "autoengaño colectivo" referirá a un yo colectivo y no simplemente a la suma de los individuos que componen una totalidad mayor. En este caso, tanto la creencia en cuestión como el autoengaño son comprendidos como atributos del colectivo y no primariamente de los individuos que lo integran.[128] En el sentido sumatorio, por otro lado (que es aquel en el que se concentrará el autor), "autoengaño colectivo" hará referencia a un grupo de individuos que comparten similares sesgos motivados y, como consecuencia de ello, forman y mantienen la misma creencia frente a fuertes pruebas en contrario.

Con el enfoque deflacionista del autoengaño en mente, Deweese-Boyd sugiere que la expresión "autoengaño colectivo" hace referencia a un grupo de individuos que comparten sesgos similares y como consecuencia forman y mantienen la misma creencia falsa en presencia de pruebas en contra, poseídas o fácilmente disponibles para cada miembro del colectivo. Lo que distingue al autoengaño colectivo del individual, a su modo de ver, es simplemente el contexto social, esto es, que ocurre dentro de un grupo que comparte las actitudes hacia la creencia falsa y la creencia falsa en sí misma. Agrega que el autoengaño dentro de una entidad colectiva es más fácil de promover y del que es más difícil escapar, ya que es instigado por los esfuerzos autoengañosos de otros miembros del grupo. Si bien Deweese-Boyd señala que casi todos los casos de autoengaño tienen un componente social (debido a que son consciente o inconscientemente sustentados por las personas cercanas a nosotros), el caso del autoengaño colectivo es diferente, ya que la dimensión social se ubica en primer plano: cada miembro de la totalidad ayuda inadvertidamente a sostener la creencia autoengañosa de otros miembros del grupo.

Un enfoque bastante diferente de los anteriores es el que puede encontrarse en algunos estudios sociológicos relativos a los procesos colectivos de distorsión de las creencias. Describiré brevemente aquí los análisis de Cohen (2001) y Zerubavel (2006).

[128] La caracterización no sumatoria puede generar la conocida sospecha acerca de la posibilidad de atribuir a entidades colectivas estados típicamente considerados propios de entidades individuales. Por ejemplo, puede parecer dudoso que pueda darse a la oración "el grupo mantenía una creencia autoengañosa" un sentido razonable mientras no sea condición de verdad necesaria de tal afirmación que todos o la mayoría de sus miembros sostengan tal creencia. No obstante, el análisis detallado de esta posibilidad llevaría inevitablemente a adentrarse en los complejos problemas relativos a la ontología de los fenómenos sociales, en especial, al debate que enfrenta a los partidarios del individualismo (ontológico o metodológico) con los defensores del holismo o colectivismo, lo cual está por completo fuera de los alcances de este trabajo. La existencia de este debate basta, a mi modo de ver, para hacer innecesaria una argumentación adicional que dé sustento a las dudas acerca de tal concepción del autoengaño colectivo.

Zerubavel (2006) recurre a dos categorías distintas, aunque relacionadas, para hacer referencia a los procesos colectivos de distorsión de las creencias: "negación" y "conspiración de silencio". Ambas coinciden en un rasgo fundamental: no son productos exclusivamente individuales, sino que son el resultado de un esfuerzo colectivo. La categoría de "negación" no es entendida entonces en su sentido tradicional de fenómeno intrapsíquico, sino como un proceso tanto individual como social. En las conspiraciones de silencio, un grupo de personas acuerda tácitamente exteriorizar ignorancia acerca de algo de lo que cada uno es personalmente consciente; tales conspiraciones presuponen la negación, en un proceso en el que al menos dos individuos colaboran en la evitación del reconocimiento de algún hecho. Tal proceso no implica una actitud pasiva por parte de los agentes involucrados; por el contrario, involucra un esfuerzo deliberado para abstenerse de tomar conocimiento de algo. Individuos, familias, organizaciones y sociedades enteras pueden ser sujetos activos en estos fenómenos. Múltiples causas pueden actuar, solas o combinadas, para producir la negación y el silencio, como el miedo, la vergüenza y el dolor.[129]

Cohen (2001), a su vez, emplea el término "negación" de una manera abarcativa; no hace referencia únicamente a un mecanismo psicológico fijo, ni a un proceso social universal. En este sentido, la negación, ya sea como proceso individual o como proceso social excede ampliamente al autoengaño como proceso intrapsíquico. Cohen distingue una serie de categorías de negación, según sea su status psicológico (consciente o inconsciente), su contenido (literal, interpretativa o implicativa),[130] su relación con el individuo o el colectivo (personal, oficial o cultural), su ubicación temporal (histórica o contemporánea), su relación con el agente (víctimas, perpetradores o espectadores), y su espacio y lugar (en el territorio propio o en el de otros). Un somero examen de la distinción mediante el criterio consciente-inconsciente que emplea este autor ayuda a comprender el muy diferente sentido del término "negación" con respecto a "autoengaño". Cohen observa que las declaraciones de negación son afirmaciones de que algo no ha ocurrido, no existe, no es verdadero o no se sabe nada acerca de él. Las posibilidades relativas al valor de verdad de tales aseveraciones son tres: o bien las aseveraciones son de hecho verdaderas, justificadas y correctas, o bien se trata de intentos intencionales y deliberados de engañar, o bien

[129] El autor menciona esta combinación de factores en casos como el del Holocausto.

[130] En la negación literal o factual, es el hecho o conocimiento del hecho lo que es negado. En la interpretativa, los hechos puros (algo que ha ocurrido), no son negados; más bien, se les asigna un significado diferente de aquel que parece claro para otros. En la implicativa, no se hace un intento de negar los hechos o su interpretación convencional; lo que se niega o minimiza son las implicaciones psicológicas, políticas o morales que se siguen convencionalmente.

(posibilidad más compleja e intrigante) tal negación no es un intento de decir la verdad ni un intento deliberado de engañar o mentir. En este caso, el estatus del "conocimiento" acerca de la verdad de la aseveración no es completamente claro. Parecería, observa Cohen, que hay estados de la mente, o incluso de culturas enteras, en las cuales sabemos y no sabemos al mismo tiempo.

Las explicaciones psicológicas de tales estados de negación, observa Cohen, van desde la perspectiva psicoanalítica hasta las teorías provistas por la psicología cognitiva. Independientemente de cual sea la explicación correcta, señala que los ecos políticos de los estados mentales descriptos por esas teorías pueden ser encontrados en la negación masiva característica de estados represivos, racistas y colonialistas. En tales casos, los grupos dominantes parecen inexplicablemente capaces de esconder o ignorar las injusticias y el sufrimiento que los rodean. En sociedades más democráticas, las personas esconden ciertas situaciones no mediante la coerción sino mediante el hábito cultural de hacer la vista gorda de los recordatorios visibles de la existencia de personas sin hogar, las privaciones, la pobreza y la decadencia urbana. No sólo las familias, sino las burocracias gubernamentales, los partidos políticos, las asociaciones profesionales, las religiones, el ejército y la policía tienen sus propias formas de encubrimiento y mentira.

Las consideraciones del precedente apartado sobre los procesos sociales de distorsión de las creencias permiten, a mi modo de ver, agregar algún sustento al escepticismo expresado inicialmente con respecto al autoengaño colectivo como suma de autoengaños individuales sostenidos por la interacción social. Varios factores sugieren que la expresión "autoengaño colectivo" sólo tiene una semejanza superficial con el autoengaño individual.

En primer lugar, al analizar la posibilidad de autoengaño colectivo, parece pertinente tener en cuenta la relevancia social, política, económica o cultural de las creencias que pueden ser objeto de distorsión. Así como los casos de autoengaño individual no acontecen en áreas irrelevantes de la vida de una persona, sino en áreas de importancia vital (carecemos de motivos para autoengañarnos respecto de cuestiones triviales de nuestra vida), es muy plausible suponer que los casos de distorsión social de las creencias también ocurrirán en áreas de importancia colectiva. Sin duda, algunos de los ejemplos paradigmáticos de negación o conspiraciones de silencio sugeridos por los estudios sociológicos antes mencionados avalan esta presunción: el genocidio armenio, el Holocausto, las matanzas tribales en Ruanda o la limpieza étnica en la ex Yugoslavia. En

tales casos sin duda habrá un interés estatal o institucional en la difusión o mantenimiento de la creencia falsa. Diversos mecanismos "descendentes" (esto es, desde el estado u organización hacia el individuo) se pondrán en marcha para asegurar la difusión y consolidación del sistema de falsas creencias. En su examen de lo que denomina "negación oficial" (variante iniciada, estructurada y sostenida por los recursos masivos del estado moderno), Cohen señala que existirán diversos mecanismos apropiados para la presentación interesada de los hechos. La modalidad de estos mecanismos dependerá, al menos en parte, de la naturaleza del colectivo en cuestión. Así como la coerción será seguramente uno de los medios privilegiados empleados por estados u organizaciones autoritarias, los estados y organizaciones más democráticas apelarán a mecanismos más sutiles de manipulación y distorsión: filtraciones tendenciosas a la prensa, empleo de publicistas encargados de presentar interpretaciones favorables de los hechos, preocupación selectiva acerca de víctimas apropiadas o negaciones interpretativas acerca de asuntos de política exterior. En cualquier caso, la dinámica del supuesto autoengaño colectivo, en casos paradigmáticos como los mencionados, se asemejará más al modelo engaño-autoengaño que a un proceso de origen individual y ascendente (esto es, un autoengaño individual que llega a ser colectivo).

En segundo lugar, es razonable resaltar el rol de los múltiples mecanismos interpersonales de nivel horizontal, que se suman a los mecanismos psicológicos intrapsíquicos. Como se ha observado en algunos estudios sociológicos respecto del conocimiento de la realidad de la vida cotidiana,[131] el diálogo (común, no destinado específicamente a la reafirmación de la realidad) contribuye de manera decisiva al mantenimiento de tal conocimiento; si ciertas creencias sobre esta realidad son falsas, este mecanismo contribuirá a su preservación. Pero también habrá otra clase de procesos más específicos por los cuales se mantendrá la "realidad" de la creencia falsa. La crítica explícita, el ridículo, la desautorización de la duda o la amenaza velada serán sin duda mecanismos de mantenimiento más activos que el diálogo que da por supuesto tal creencia falsa. Cuál es el peso de los mecanismos intrapsíquicos y cuál el de los procesos interpersonales en la producción de creencias colectivas falsas es algo que, entiendo, debe ser materia de investigación empírica. No obstante, y en conjunción con el punto anterior, parece al menos dudoso que el rol de los mecanismos interpersonales se limite al mantenimiento de la creencia falsa generada individualmente; parece una alternativa mucho más plausible el

[131] Por ejemplo, Berger y Luckmann, 1966.

pensar que tales mecanismos no actúan meramente en el mantenimiento de tal creencia, sino también en su generación y aceptación.

Un tercer y último aspecto en el cual el presunto autoengaño colectivo difiere del autoengaño individual es debido tanto a la naturaleza de los hechos que son objeto de creencias falsas como a las pruebas a disposición de los agentes sociales, variables ambas que posibilitan o no una atribución de autoengaño. En los casos típicos de autoengaño individual la creencia falsa refiere a hechos relativamente "simples" (infidelidad de la pareja, estados de salud o enfermedad, las propias habilidades o capacidades). Por el contrario, los hechos que típicamente serán objeto de supuesto autoengaño colectivo serán hechos comparativamente "complejos", de mucha más difícil caracterización e interpretación (episodios de conmoción política, casos de corrupción gubernamental, crisis económicas). Aun cuando aspectos importantes de esos hechos no sean de naturaleza controvertida, en muchos casos sí lo serán sus interpretaciones o implicaciones y los modos en que son explicados. A esto se suma que la naturaleza de los elementos de juicio, en los casos de presunto autoengaño colectivo, será muy diferente a los casos de autoengaño individual.[132] Las pruebas, en los casos típicos de autoengaño individual, son, al menos parcialmente, pruebas directas y, o bien son poseídas por el agente, o bien están fácilmente disponibles para él. Por el contrario, en muchos de los casos de presunto autoengaño colectivo, como se puede inferir del ejemplo propuesto por Goleman, la información que el agente posee es *mediada* por un número mucho mayor de agentes sociales (medios masivos de comunicación, funcionarios estatales, redes sociales, etc.). Por otra parte,[133] el autoengaño no es incompatible con la ignorancia; el desconocimiento o la incomprensión sobre aspectos esenciales de los hechos en cuestión, cuando estos hechos tienen amplias implicaciones políticas, sociales o culturales será la regla, más que la excepción, en los procesos sociales de distorsión de creencias.

Lo anterior no implica, en absoluto, la afirmación de que sólo existen interpretaciones diversas sobre los hechos y no es posible predicar verdad o falsedad de ninguna de ellas, extremo sólo admisible para un posmodernista o construccionista radical;[134] por el contrario, entiendo que sí es posible tener representaciones distorsionadas sobre cualquier aspecto de la realidad, y que la contracara de esta afirmación es que es posible tener representaciones correctas, o verdaderas, de esa realidad. Tampoco implica

[132] No debe ser casual que cuando Cohen hace referencia a la negación, por parte de los pobladores de la zona, del conocimiento de la existencia de un campo de concentración en Mauthausen, el ejemplo remita a algo de lo cual se podía llegar a tener pruebas relativamente directas.
[133] Como ha señalado Mele (1987).
[134] Cfr. al respecto el § 2.1. del capítulo III.

que sea imposible que puedan existir procesos de distorsión generalizada de creencias que surjan, por así decirlo, "de abajo hacia arriba", esto es, de los individuos hacia el colectivo. Lo que sí implica es que la formación de creencias colectivas sesgadas implicará, en muchos casos, procesos de muy superior complejidad a los estudiados en los casos de autoengaño individual, diferencia que debilita considerablemente, a mi entender, las analogías que parecen desprenderse del uso de la expresión "autoengaño colectivo".[135]

[135] Lo expuesto tiene, a mi modo de ver, consecuencias no menores para el problema de la responsabilidad moral por el autoengaño. Como hemos visto, la progresiva puesta en tela de juicio de los enfoques intencionalistas y el surgimiento de modelos explicativos que prescinden de la intención del agente no parece haber disminuido demasiado la influencia de la posición según la cual tenemos responsabilidad moral por nuestro autoengaño. Sin embargo, parece haber razones atendibles para pensar que una atribución de responsabilidad resulta mucho menos plausible en modelos que eliminen la intención del agente (Levy, 2004). Si a los mecanismos psicológicos que inducen a sesgos en la formación individual de creencias se suman mecanismos sociales similarmente sutiles y complejos (cuando no una manipulación deliberada y planificada por terceros) que inducen a la formación de creencias falsas, parecería que la posibilidad de atribuir responsabilidad moral individual por el mantenimiento de creencias colectivamente distorsionadas resulta aun más difícil.

El porvenir de un problema

El filósofo John Austin describió la transición por la cual un problema filosófico llega a ser un problema científico de un modo que ha devenido clásico y profusamente citado: "En la historia de las indagaciones humanas la filosofía ocupa el lugar de un sol central originario, seminal y tumultuoso. De tanto en tanto ese sol arroja algún trozo de sí mismo que adquiere el *status* de una ciencia, de un planeta frío y bien regulado, que progresa sin pausa hacia un distante estado final" (1961, pp. 179-80). Creo que el problema del autoengaño ilustra muy bien en la actualidad el proceso al que Austin hace referencia. El estudio de este fenómeno ha dejado de ser competencia casi exclusivamente filosófica, para convertirse en un abanico de interrogantes examinados por distintas disciplinas científicas. Sin embargo, esta transición no parece haber finalizado, y tampoco está lejos de ser problemática. Esto se debe al menos a dos razones; las consideraciones que haremos en primer término son específicas para la Psicología pero, *mutatis mutandis*, pueden aplicarse a otras disciplinas.[136]

En primer lugar, como hemos señalado a lo largo del libro, diversas posiciones sobre el autoengaño parten de formas muy discutibles de comprender teóricamente este fenómeno, lo que redunda no sólo en falsas soluciones a algunos problemas, sino que limita la utilidad de investigaciones empíricas potencialmente relevantes. Recordemos brevemente dos de ellas.

Hemos visto en el capítulo II que Bandura (1991) ha negado de manera tajante la existencia de este fenómeno. Basa su posición escéptica en la conocida tesis de la imposibilidad de ser simultáneamente quien engaña y quien es engañado; es lógicamente imposible, señala, engañarse a uno mismo para creer algo mientras simultáneamente se sabe que es falso. Los intentos de resolver esta paradoja mediante la apelación a partes inconcientes de la mente, agrega, han tenido escaso éxito: esta estrategia, más que explicar el autoengaño, lo aniquila. Estas concepciones del yo dividido fracasan al tener que explicar como un yo conciente puede mentir a un yo inconciente; el yo engañador tiene que ser conciente de lo que el yo engañado cree para saber como fraguar el engaño. Por otra parte, en el capítulo III hemos reseñado una serie de investigaciones empíricas tendientes a

[136] Cfr. Sturm (2007) para un análisis del problema del autoengaño tanto desde la Filosofía como desde la Psicología.

probar la existencia del autoengaño, entre las cuales ocupan un lugar especialmente destacado las de Gur y Sackheim (1979). En este intento, estos autores proponen los siguientes criterios como necesarios y suficientes para el autoengaño: "1. el individuo sostiene dos creencias contradictorias (que *p* y que no *p*). 2. Esas dos creencias contradictorias son sostenidas simultáneamente. 3. El individuo no es conciente de que sostiene una de esas creencias. 4. El acto que determina cual creencia es objeto de la conciencia y cual no es un acto motivado" (p. 149).[137] Sobre la base de esta definición, consideran que sus estudios prueban la existencia del fenómeno en cuestión.

Los estudios mencionados son susceptibles de una misma clase de objeción: el partir de caracterizaciones como mínimo discutibles del fenómeno del autoengaño. Esto se debe a que comparten un mismo supuesto cuestionable: el de que el autoengaño debe implicar esencialmente la coexistencia de creencias contradictorias. La adopción de este supuesto conduce, a su vez, a dos alternativas igualmente innecesarias: o bien, como es el caso de Gur y Sackheim, a proponer partes inconcientes de la mente de una forma *ad hoc*, o bien, como es el caso de Bandura, a negar la existencia del fenómeno. Si, como vimos en el capítulo II a partir del desarrollo de las posiciones deflacionistas, el supuesto de existencia de creencias contradictorias es abandonado, desaparece tanto la premisa principal para el argumento escéptico de Bandura como la base conceptual para el experimento de Gur y Sackheim.

Encontramos, entonces, que el descuido de análisis conceptuales detallados (muchos de los cuales provienen de la filosofía) conduce a consecuencias teóricas y empíricas inadecuadas o insuficientemente fundadas.[138] Ahora bien, como adelantamos, existe una segunda razón para considerar la posible transición del problema como un proceso complejo e inacabado. Esta segunda razón se basa en el hecho de que parecería que nos encontramos lejos de una teoría explicativa unificada del fenómeno lo que, a su vez, nos conduce a nuevos interrogantes filosóficos. Una enumeración comprehensiva, aunque no necesariamente exhaustiva (Fernández Acevedo, 2014), de las explicaciones del autoengaño incluye:

---

[137] Podemos recordar también, respecto de esta crítica, a la posición de Paulhus (2007), quien define al autoengaño como "el acto de mentirse a uno mismo". Refiere a esta conducta como casos en los cuales las "personas aparentemente creen algo que saben que es falso". Aclara que esta conducta no incluye exageración, falsificación o simple mentira; el autoengaño es algo más profundo y complicado, incluso paradójico. Su explicación requiere el reconocimiento de la existencia de partes inconcientes de la mente. Sólo en tales partes inconcientes puede un conflicto emocional influenciar realmente la conducta de un individuo y, pese a eso, ser inaccesible.

[138] Cabe recordar también aquí, aunque en referencia a un concepto diferente, los cuestionamientos a la concepción puramente subjetiva de la felicidad sostenida por Taylor y Brown.

a. Explicaciones neuropsicológicas (Ramachandran, 1996; Hirstein, 2000). El común denominador de este tipo de explicaciones es la comprensión de la patología neurológica, expresada en déficits como la anosognosia o la confabulación, como la base para la explicación del funcionamiento mental en personas en quienes pueden observarse procesos de autoengaño no patológico.
b. Explicaciones evolucionistas (Van Leeuwen, 2007; von Hippel y Trivers, 2011). Estas explicaciones se caracterizan o bien por postular una función evolutiva para el autoengaño (esto es, una contribución a la aptitud inclusiva), o bien por concebir este fenómeno como un subproducto estructural (*spandrel*) de sistemas seleccionados por su contribución a la aptitud.
c. Explicaciones "disolucionistas" (Lewis, 1996; Clegg y Moissinac, 2005). Para estas perspectivas, basadas en filosofías construccionistas y relativistas, el autoengaño no existe dentro del yo individual del así llamado "autoengañado". Existe sólo en la mente (más precisamente en las narrativas) del observador, quien supone acceder a información más completa o correcta o a una representación "verdadera" del mundo. El autoengaño es un fenómeno cultural, no natural.
d. Explicaciones cognitivo-motivacionales (Lazar, 1999; Mele, 1997, 2001). Para esta perspectiva el autoengaño es el resultado de una compleja combinación de procesos sesgados de procesamiento cognitivo y emotivo-motivacionales, combinación en la que los deseos del agente suelen jugar un rol decisivo en la producción del fenómeno.
e. Explicaciones psicológico-sociales (Cohen, 2001; Zeruvabel, 2006). Para esta perspectiva, el autoengaño (o, como lo denominan, "negación" o "conspiración de silencio") es el resultado de múltiples procesos y mecanismos tanto individuales como sociales, no un proceso unitario en el interior de un individuo. El autoengaño se descompone en una multiplicidad de categorías, discriminadas a través de criterios tales como su status psicológico, su relación con el individuo o el colectivo y su ubicación temporal, entre otros.

Como adelantamos, esta proliferación explicativa trae aparejados nuevos interrogantes filosóficos. En particular, cabe preguntar si la coexistencia de explicaciones alternativas y potencialmente competitivas entre sí constituye un progreso cognoscitivo, o en realidad se trata de un estado intelectualmente indeseable; como un filósofo ha señalado, "demasiadas explicaciones [de un mismo fenómeno] pueden ser una fuente de incoherencia

en vez de incrementar la coherencia" (Kim, 1989, p. 258). Nada indica que tales explicaciones sean esencialmente incompatibles y es perfectamente posible que se encuentren formas de conciliarlas (o, también, que algunas de ellas sean simplemente eliminadas). Sin embargo, esto no ha ocurrido hasta el momento, por lo que los interrogantes filosóficos al respecto mantienen plenamente su vigencia.

Si bien la reflexión acerca de los derroteros futuros en el estudio del problema no puede superar en cierta medida el plano de la especulación, quizás sea posible avanzar en algunas consideraciones guiadas por la prudencia. En primer lugar, es posible que el autoengaño deje de ser considerado un fenómeno único, para pasar a ser considerado una familia de fenómenos cercanamente relacionados. Indicios de esta tendencia pueden encontrarse en las distinciones realizadas por algunos estudiosos del problema, quienes identifican diversas variantes, en ocasiones muy diferentes entre sí, de autoengaño. En segundo lugar es probable, también, que se profundice el fenómeno de proliferación explicativa observado en los últimos años, esto es, que cada vez más disciplinas científicas realicen aportes para nuestra comprensión del fenómeno. Con esto es posible que estemos cada vez más lejos de una teoría unificada acerca de este fenómeno. En consecuencia, los interrogantes filosóficos relativos a la proliferación explicativa no sólo no desaparecerán, sino que se fortalecerán. Por otra parte, aun cuando la ciencia se apropie cada vez más del estudio del autoengaño, parece improbable que esta apropiación sea completa. La responsabilidad moral por el autoengaño, por ejemplo, o su contribución a la felicidad, parecen ser territorios fértiles para la reflexión filosófica, a menos que se crea (suposición poco plausible hasta el momento) que estos temas *también* pasarán a ser competencia exclusiva de la ciencia.

Pase lo que pase con el estudio del autoengaño, y pese a las persistentes (y quizás perennes) controversias alrededor de él, una idea parece gozar de bastante consenso: este fenómeno está profundamente enraizado en nuestra naturaleza. Para afirmar esto no hace falta limitarse a las agudas observaciones de filósofos y escritores a lo largo de los siglos. Cualesquiera que sean sus orígenes evolutivos, sin duda nuestros sistemas perceptivos, cognitivos y afectivos se encuentran estructurados e interrelacionados de un modo tal que hacen posible, o incluso facilitan, el autoengaño, así como otras formas de irracionalidad motivada. Quizás podamos incrementar, tanto de modo individual como colectivo, nuestra capacidad para ser conscientes de su actuación y minimizarla, pero parece improbable que alguna vez seremos capaces de eliminarlo por completo. Creo que seríamos prudentes si tuviéramos siempre presente esta posibilidad.

Referencias bibliográficas

Adler, Jonathan E. (2002). *Belief's Own Ethics*. Cambridge, The MIT Press.

Asociación Psiquiátrica Americana (1994). *Manual Diagnóstico y Estadístico de los Trastornos Mentales*, *Cuarta Edición*. Barcelona, Masson.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. Washington, American Psychiatric Publishing.

Atran, Scott, (2005), Adaptationism for Human Cognition: Strong, Spurious or Weak? *Mind & Language*, 20. 39-67.

Audi, Robert (1982). Self-Deception, Action, and Will. *Erkenntnis* 18: 133–58.

Audi, Robert (1997). Self-deception vs. self-caused deception: A comment on Professor Mele. *Behavioral and Brain Sciences* 20 (1). 104.

Austin, John (1961). *Philosophical Papers*, editado por G. J. Warnock y J. O. Urmson, Oxford, Clarendon Press. Citado por Genaro Carrió y Eduardo Rabossi, "La filosofía de John L. Austin", en John Austin, *Cómo hacer cosas con palabras*, Barcelona, Paidós, 1990, p. 27.

Bach, Kent (1981). An Analysis of Self Deception. *Philosophy and Phenomenological Research*, *41*, 3. 351-370.

Badhwar, Neera (2008). Is Realism bad for You? A Realistic Response. *The Journal of Philosophy*, *CV*, 2. 85-107.

Bandura, Albert (1991). Social cognitive theory of moral thought and action. En W. M. Kurtines & J. L. Gewirtz (eds.), *Handbook of moral behavior and development. Vol. 1*. Hillsdale, NJ, Erlbaum. 45-103.

Bandura, Albert (2011). Self-deception: A paradox revisited. *Behavioral and Brain Sciences*, 34, 16-17.

Barkow, Jerome, Leda Cosmides & John Tooby (Eds.), *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*. New York; Oxford University Press.

Barnes, Annette (1997). *Seeing through self-deception*. New York, Cambridge University Press.

Baron, Marcia (1988). What is Wrong with Self-Deception? En Brian McLaughlin & Amelie Oksenberg-Rorty (eds.).

Baumeister, Roy, Karen Dale y Kristin Sommer (1998). Freudian Defense Mechanisms and Empirical Findings in Modern Social Psychology: Reaction Formation, Projection, Displacement, Undoing, Isolation, Sublimation and Denial. *Journal of Personality 66*, 6. 1081-1124.

Bayne, Tim (2008). Delusion and Self-Deception: Mapping the Terrain. En T. Bayne and J. Fernandez (eds.) *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation*. Hove, East Sussex, Psychology Press.

Beck, Aaron, Fred Wright, Corey Newman y Bruce Liese (1993). Terapia cognitiva de las drogodependencias. Barcelona, Paidós.

Bennett, Jonathan (1990). Why Is Belief Involuntary? *Analysis*, *50*, 2. 87-107.

Berger, Peter y Thomas Luckmann (1966). *La construcción social de la realidad*. Buenos Aires, Amorrortu.

Bermúdez, José Luis (2000). Self Deception, intentions and contradictory beliefs. *Analysis 60*, 4.

Boag, Simon (2007). Realism, Self-Deception and the Logical Paradox of Repression. *Theory & Psychology*. 17(3): 421–447.

Bok, Sissela (1980). The self deceived. *Social Science Information 19*. 923-935.

Booth, Anthony (2007). Doxastic voluntarism and self-deception. *Disputatio*, II, 22. 115-130.

Borge, Steffen (2003). The Myth of Self-Deception. *The Southern Journal of Philosophy*. *41*, 1. 1–28.

Bortolotti, Lisa & Rochelle E. Cox (2009). 'Faultless' ignorance: Strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition 18*. 952–965.
Broad, William & Nicholas Wade (1982). Self-Deception and Gullibility. En William Broad & Nicholas Wade, *Betrayers of the Truth*. New York, Simon and Schuster.
Buss, David (1992). Mate Preference Mechanism: Consequences for Partner Choice and Intrasexual Competition. En Barkow, J., Cosmides, L., Tooby, J. (Eds.).
Buss, David (1995). Evolutionary Psychology: A New Paradigm for Psychological Science. *Psychological Inquiry*, 6, 1-30.
Buss, David, Martie Haselton, Todd Shackelford, April Bleske & Jerome Wakefield (1998). Adaptations, Exaptations, and Spandrels. *American Psychologist*, 54, 533-548.
Butler, Joseph (1726). Sermon X. Upon Self-Deceit. Extraído el 24/07/2010 de http://anglicanhistory.org/butler/rolls/10.html
Byrne, Christopher & Jeffrey Kurland (2001). Self-deception in an Evolutionary Game. *Journal of Theoretical Biology*. 212, 457-480.
Canfield, John & Patrick McNally (1961). Paradoxes of Self-Deception. *Analysis 21*, 6. 140-144.
Caporael, Linda (2001). Evolutionary Psychology: Toward a Unifying Theory and a Hybrid Science. *Annual Review of Psychology*.
Catalano, Joseph (1990). Successfully Lying to Oneself: A Sartrean Perspective. *Philosophy and Phenomenological Research*, *50*, 4. 673-693.
Champlin, T. S. (1977). Self-deception: A reflexive dilemma. *Philosophy*, *52*. 281-299.
Clegg, Joshua W. & Luke Moissinac (2005). A relational theory of self-deception. *New Ideas in Psychology*, 23. 96–110.
Cohen, Stanley (2001). *States of Denial. Knowing about Atrocities and Suffering*. Oxford, Blackwell.
Colvin, Randall, Jack Block & David Funder (1995). Overly Positive Self-Evaluations and Personality: Negative Implications for Mental Health. *Journal of Personality and Social Psychology*, *68*, 6. 1152-1162.
Colvin, Randall & Jack Block (1994). Do Positive Illusions Foster Mental Health? An Examination of the Taylor and Brown Formulation. *Psychological Bulletin*, *116*, 1. 3-20.
Comesaña, Manuel (2011). ¿En qué sentido es racional la ciencia?, en Ana Rosa Pérez Ransanz y Ambrosio Velasco Gómez (coord.), *Racionalidad en ciencia y tecnología. Nuevas perspectivas iberoamericanas*. México, UNAM.
Cook, J. Thomas (1987). Deciding to Believe Without Self-Deception. *The Journal of Philosophy*, *84*, 8. 441-446.
Correia, Vasco (s/f). Sour Illusions. What is adaptive about misbelief. Extraído el 17/7/2012 de http://fcsh-unl.academia.edu/VascoCorreia/Papers/581527/Sour_illusions_What_is_adaptive_about_illusional_beliefs
Correia, Vasco (2007). Une conception émotionnaliste de la self-deception. Teorema, XXVI, 3.
Cosmides, Leda & John Tooby (1992). Cognitive Adaptations for Social Exchange. En Barkow, J., Cosmides, Leda, & John Tooby (Eds.).
Cosmides, Leda & John Tooby (1997). Evolutionary Psychology: A Primer. Extraído el 25/2/2003 de http://www.psych.ucsb.edu/research/cep/primer.html.
Cosmides, Leda & John Tooby (2000). Evolutionary Psychology and the Emotions. En Lewis, M., Haviland-Jones, J.M. (Eds.), *Handbook of Emotions*, 2nd edition, NY, Guilford.
Daly, Martin & Margo Wilson (1988). *Homicidio*. Buenos Aires, Fondo de Cultura Económica.

Darwall, Stephen (1988). Self-Deception, Autonomy and Moral Constitution. En Brian McLaughlin & Amelie Oksernberg-Rorty (eds.).
Davidson, Donald (1963). Acciones, razones y causas. En A. White (comp.) (1976), *La Filosofía de la acción*. Madrid, Fondo de Cultura Económica.
Davidson, Donald (1970). ¿Cómo es posible la debilidad de la voluntad? En *Ensayos sobre acciones y sucesos*. México-Barcelona, Instituto de Investigaciones Filosóficas. UNAM-Crítica.
Davidson, Donald (1982). Paradoxes of irrationality. En Richard Wollheim & James Hopkins (eds.).
Davidson, Donald (1985) Engaño y división. En *Mente, mundo y acción*. Barcelona, Paidós.
Davies, Martin (2008). Delusion and Motivationally Biased Belief: Self-Deception in the Two-Factor Framework. En T. Bayne and J. Fernandez (eds.) *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation*. Hove, East Sussex, Psychology Press.
Dawkins, Richard (1976). *El gen egoísta*. Barcelona, Salvat.
Dawkins, Richard (2006). *The God Delusion*. London, Bantam Press.
Demos, Raphael (1960). Lying to Oneself. *Journal of Philosophy*, *57*. 588–95.
Deweese-Boyd, Ian (2007). Self-deception, Culpability and Control. *Teorema,* XXVI/3. 161-176.
Deweese-Boyd, Ian (2008). Collective self-deception, collective injustice: Consumption, sustainability and responsibility. Extraído el 27/10/10 de http://www.colorado.edu/philosophy/center/rome/papers/DeWeese-boyd_CollectiveSelfDeception_CollectiveInjustice.pdf
Deweese-Boyd, Ian (2012). Self Deception. Stanford Encyclopedia of Philosophy. Extraído el 10/06/10 de. http://plato.stanford.edu/entries/self-deception/
Dupuy, Jean-Pierre (1995). Not to Know What One Knows: Some Paradoxes of Self-Deception. *Diogenes*, *43*, 169. 53-68.
Dupuy, Jean-Pierre (2004). Intersubjectivity and Embodiment. *Journal of Bioeconomics 6*, 3. 275-294.
Durrant, Richard & B. J. Ellis (2003). Evolutionary Psychology. En Gallagher, M., R.J. Nelson, R. J. (Eds.), *Comprehensive Handbook of Psychology*, *Vol. 3: Biological Psychology*, New York, Wiley & Sons.
Dyke, Daniel (1614). *The Mystery of Selfe-Deceiuing*: *or, a Discourse and Discovery of the Deceitfulnesse of Mansheart*. London, Griffin and Mab. Extraído el 10/04/2012 de https://books.google.com.ar/books?id=T-EUAAAAQAAJ&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
Egan, Andy (2008). Imagination, Delusion, and Self-Deception. En T. Bayne and J. Fernandez (eds.) *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation*. Hove, East Sussex, Psychology Press.
Ekman, Paul (1997). Deception, Lying and Demeanor. En Diane F. Halpern & Alexander Voiskounsky (eds.), *States of Mind. American and Post-Soviet Perspectives on Contemporary Issues in Psychology*. New York, Oxford University Press.
Elga, Adam (2005). On Overrating Oneself. . . and Knowing it. *Philosophical Studies*, *123*, 1-2. 115-124.
Elster, Jon (1979). *Ulysses and the Sirens*. New York, Cambridge University Press.
Elster, Jon (1986). *An Introduction to Karl Marx*. Cambridge, Cambridge University Press.
Elster, Jon (2007). *Explaining Social Behavior. More Nuts and Bolts for Social Sciences*. Cambridge, Cambridge University Press.

Erez, Amir, Diane E. Johnson & Timothy A. Judge (1995). Self-Deception as a Mediator of the Relationship Between Dispositions and Subjective Well-Being. *Personality and Individual Differences. 19*. 5. 597-612.

Fairbanks, Rick (1999). The Availability of Self-Deception. *Philosophical Investigations 22*, 4. 335-340.

Fernández, Jordi (2013). Self-deception and self-knowledge. *Philosophical Studies*, *162*. 379–400

Fernández Acevedo, Gustavo (2008). Psicología evolucionista: un difícil equilibrio entre naturalismo, no reduccionismo y dualismo. En E. Kronmüller y C. Cornejo (comps.), *Ciencias de la Mente: Aproximaciones desde Latinoamérica*, Santiago de Chile, JCSaez Editor. 265-294.

Fernández Acevedo, Gustavo (2011). ¿Cómo debe entenderse la condición de evidencia en el autoengaño? En José M. Gil y Gastón Gil (eds.), *Análisis epistemológico II*. Mar del Plata, Ed. Martín.

Fernández Acevedo, Gustavo (2014). El pluralismo explicativo en Psicología. Un examen del caso de las teorías psicológicas sobre el autoengaño. En Ana Talak (comp.) *Problemas actuales de las explicaciones en psicología*. Buenos Aires, Prometeo Libros.

Fernández Acevedo, Gustavo (2015). Autoengaño, sistemas de creencias y errores en el autoconocimiento. *Areté. Revista de Filosofía de la Pontificia Universidad Católica de Perú*. Vol. XXVII, Nº 1. 69-85.

Fingarette, Herbert (1969). *Self deception*. London, Routledge & Kegan Paul.

Fodor, Jerry (1998). The Trouble with Psychological Darwinism. *London Review of Books*, 20.

Forrester, Mary (2002). Self-Deception and Valuing Truth. *American Philosophical Quarterly 39*, 1. 31-47.

Frenkel-Brunswick, Else (1939). Mechanisms of Self-Deception. *Journal of Social Psychology*, 10.

Freud, Sigmund. La negación (1925, vol. XIX); El porvenir de una ilusión (1927, vol. XXI); La escisión del yo en el proceso defensivo (1938, vol. XXIII, p. 77); Esquema del Psicoanálisis (1938, vol. XXIII, p. 41). En *Obras completas*. Buenos Aires, Amorrortu.

Funkhouser, Eric (2003). Willing Belief and the Norm of Truth. *Philosophical Studies*, *115*. 179–195.

Funkhouser, Eric (2005). Do the Self Deceived Get what they Want? *Pacific Philosophical Quarterly. 86*. 295 –312.

Funkhouser, Eric (2009). Self-Deception and the Limits of Folk Psychology. *Social Theory and Practice. 35*, 1. 2-13.

Gardiner, Patrick (1970). Error, Faith and Self Deception. *Proceedings of the Aristotelian Society*, 70.

Garssen, Bert (2007). Repression: Finding Our Way in the Maze of Concepts. *Journal of Behavioral Medicine 30*, 6. 471–481.

Gendler, Tamar Szabó (2007). Self-Deception as Pretense. *Philosophical Perspectives, 21*, *Philosophy of Mind*. 231-258.

Gergen, Kenneth (1985). The ethnopsychology of self-deception. In M. W. Martin (Ed.), *Self-deception and selfunderstanding* (pp. 228–243). Lawrence, KS: University Press of Kansas.

Godfrey-Smith, Peter (2001). Three Kinds of Adaptationism. En Orzack, S. H. & Sober, E., (Eds.), *Adaptationism and Optimality*. Cambridge, Cambridge University Press.

Goleman, Daniel (1989). What is Negative About Positive Illusions? When Benefits for the Individual Harm the Collective. *Journal of Social and Clinical Psychology*, *8*. 190–197.

Gomila, Antoni (2007). El retorno de la represión. *Teorema* XXVI/3. 97-111.

Gould, Stephen Jay (1997). Evolution: The Pleasures of Pluralism. *New York Review of Books*, June 26.

Gould, Stephen Jay & Richard Lewontin (1979). The spandrels of San Marco and the panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London*, Series B, 205. 581-598.

Greenwald, Anthony (1996). Self-Knowledge and Self-Deception: Further Consideration. En Michael S. Myslobodsky, (ed.), *The Mythomanias: The Nature of Deception and Self-deception*. Hillsdale NJ: Lawrence Erlbaum.51-71.

Gudjonsson, Gisli H. & Jon Fridrik Sigurdsson (2004). The Relationship of Suggestibility and Compliance with Self-Deception and Other-Deception. *Psychology, Crime & Law*, *10* (4). 447-453.

Gur, Ruben & H. Sackeim (1979). Self-Deception: A Concept in Search of a Phenomenon. *Journal of Personality and Social Psychology* 37:147–69.

Güth, Werner & Hartmut Kliemt (2004). Perfect or Bounded Rationality? Some Facts, Speculations and Proposals. *Analyse & Kritik* 26. 364–381.

Guttenplan, Samuel (1994). Self Deception. En Samuel Guttenplan (ed.). *A Companion to the Philosophy of Mind.* Oxford, Blackwell.

Hales, Steven (1994). Self-Deception and Belief Attribution. *Synthese 101*. 273-289.

Hamlyn, D. & H. O. Mounce (1971). Self Deception. *Proceedings of the Aristotelian Society*, supl., 45. 45-72.

Hanson, Craig (2009). *Thinking about Addiction. Hyperbolic Discounting and Responsible Agency*. Amsterdam - New York, Rodopi.

Harré, Rom (1988). The Social Context of Self.Deception. En Brian McLaughlin & Amelie Oksenberg-Rorty (eds.)

Hartung, John (1988). Deceiving Down. Conjectures on the Management of Subordinate Status. En Joan S. Lockard and Delroy L. Paulhus (eds.) *Self-Deception: An Adaptive Mechanism*. Englewood Cliffs, NJ, Prentice-Hall.

Hirstein, William (2000). Self-Deception and Confabulation. *Philosophy of Science. 67*. 418-429.

Hirstein, William (2005). *Brain Fiction. Self-Deception and the Riddle of Confabulation*. Cambridge, The MIT Press.

Holton, Richard (2000). What is the Role of the Self in Self-Deception? Meeting of the Aristotelian Society, held in Senate House, University of London, on Monday, 6th November.

Huddleston, Andrew (2011). Naughty beliefs. *Philosophical Studies*, *160*, 2. 209-222.

Jamner, Larry & Gary Schwartz (1986). Self-Deception Predicts Self-Report and Endurance of Pain. *Psychosomatic Medicine 48*, 3/4. 211-223.

Johnson, Dominic, Richard Wrangham & Stephen Rosen (2002). Is military incompetence adaptive? An empirical test with risk-taking behaviour in modern warfare. *Evolution and Human Behavior 23*. 245–264.

Johnston, Mark (1995). Self-deception and the nature of mind. En C. Macdonald (ed.), *Philosophy of Psychology: Debates on Psychological Explanation*. Cambridge: Blackwell.

Jones, Ward E. (1998). Religious Conversion, Self-Deception, and Pascal's Wager. *Journal of the History of Philosophy*, *36*, 2, pp. 167-188.

Kennair, Leif (2002). Evolutionary Psychology: an Emerging Integrative Perspective Within the Science and Practice of Psychology. *Human Nature Review*, 2, 17-61.

Kim, Jaegwon (1989). Mecanismo, propósito y exclusión explicativa. *Análisis Filosófico*, X (1), 15-47.

Kinghorn, Kevin (2007). Spiritual Blindness, Self-deception and Morally Culpable Nonbelief. *Heythrop Journal* , *XLVIII*. 527–545.

Kipp, David (1980). On Self-Deception. *Philosophical Quarterly 30*. 305–17.

Knight, Martha (1988). Cognitive and Motivational Bases of Self-Deception: Commentary on Mele's *Irrationality*. *Philosophical Psychology*, 1. 179–88.

Kurzban, Robert & C. Athena Aktipis (2007). Modularity and the social mind: Are psychologists too selfish? *Personality and Social Psychology Review*, *20*. 1–19.
Lazar, Ariela (1999). Deceiving Oneself or Self-Deceived? On the Formation of Beliefs 'Under the Influence'. *Mind 108*, 430. 265-290.
Lee, Sunhee & Howard Klein (2002). Relationships Between Conscientiousness, Self-Efficacy, Self-Deception, and Learning Over Time. *Journal of Applied Psychology, 87*, 6. 1175–1182
Levy, Neil (2003). Self Deception and Responsibility for Addiction. *Journal of Applied Philosophy*, *20*, 2. 133-142.
Levy, Neil (2004). Self Deception and Moral Responsibility. *Ratio (new series)*, *XVII*. 294-311.
Lewis, Brian (1996). Self-deception: A postmodern reflection. *Journal of Theoretical and Philosophical Psychology*, *16*, 1. 49-66.
Little, Daniel (2007). False Consciousness. En William A. Darity (ed.), *International Encyclopedia of the Social Sciences, second edition*. New York, Macmillan.
Livneh, Hanoch (2009a). Denial of Chronic Illness and Disability. Part I. Theoretical, Functional, and Dynamic Perspectives. *Rehabilitation Counseling Bulletin 52*, 4. 225-236.
Livneh, Hanoch (2009b). Denial of Chronic Illness and Disability. Part II. Research Findings, Measurement Considerations, and Clinical Aspects. *Rehabilitation Counseling Bulletin 53*, 1. 44–55.
Lowe, E. Jonathan (2000). *Filosofía de la mente*. Barcelona, Idea Books.
Lynch, Kevin (2009). Prospects for an Intentionalist Theory of Self-Deception. *Abstracta, 5,* 2. 126-138.
Lynch, Kevin (2010). Self-Deception, Religious Belief, and the False Belief Condition. *The Heythrop Journal*, *LI*. 1073–1074.
Lynch, Kevin (2013). Self-Deception and Stubborn Belief. *Erkenntnis*, *78*, 6. 1337-1345.
Martin, Mike (1979). Self-Deception, Self-Pretence, and Emotional Detachment. *Mind 88*. 441-446.
Martin, Mike (2012). *Happiness and the Good Life*. New York, Oxford University Press.
Martínez Manrique, Fernando (2007). Attributions of Self-Deception. *Teorema*, *XXVI, 3*. 131-143.
Marx, Karl y Friedrich Engels (1846). *La ideología alemana*. En K. Marx y F. Engels, *Obras escogidas, Tomo I*. Moscú, Progreso.
McKay, Ryan T. & Daniel C. Dennett (2009). The evolution of misbelief. *Behavioral and Brain Sciences 32*. 493–561.
McLaughlin, Brian & Amelie Oksenberg-Rorty (eds.) (1988). *Perspectives on Self-Deception*. Berkeley, University of California Press.
Mealey, Linda (1995). The Sociobiology of Sociopathy: An Integrated Evolutionary Model. *Behavioral and Brain Sciences*, 18, 523-541.
Mele, Alfred (1983). Self-Deception. *The Philosophical Quarterly, 33*, 133. 365-377.
Mele, Alfred (1987). *Irrationality. An Essay on Akrasia, Self-Deception, and Self-Control*. New York-Oxford, Oxford University Press.
Mele, Alfred (1997). Real Self Deception. *Behavioral and Brain Sciences* 20 (1): 91-136.
Mele, Alfred (2001). *Self-deception Unmasked*. Princeton, Princeton University Press.
Mele, Alfred (2003). Emotion and Desire in Self-Deception. En Anthony Hatzimoysis (ed), *Philosophy and the Emotions*. Cambridge, Cambridge University Press. 163-179.
Mele, Alfred (2004). Motivated Irrationality. En Alfred Mele & Piers Rawling (eds.) *The Oxford Handbook of Rationality*. Oxford, Oxford University Press.
Mele, Alfred (2006). Self-Deception and Delusions. *EUJAP*. 2. 1. 109-124.

Mele, Alfred (2007). Self-Deception and Three Psychiatric Delusions. En Mark Timmons, John Greco & Alfred R. Mele (eds.), *Rationality and the Good*. Oxford, Oxford University Press.
Mercier, Hugo (2001). Self-deception: Adaptation or by-product? *Behavioral and Brain Sciences 34*. 35.
Mertz Hsieh, Diana (2004). False Excuses and Moral Growth. 6th International Carnegie Mellon-University of Pittsburgh Graduate Philosophy Conference. 20 March 2004. Extraído el 23/07/2010 de http://www.dianahsieh.com/docs/feamg.pdf
Metcalfe, Janet (1998). Cognitive Optimism: Self-Deception or Memory-Based Processing Heuristics? *Personality and Social Psychology Review; 2*. 100-110.
Metzger, Lawrence (1988). *From Denial to Recovery*. San Francisco, Josey Bass.
Michel, Christoph & Albert Newen (2010). Self-deception as pseudo-rational regulation of belief. *Consciousness and Cognition 19*, 3. 731–744.
Miller, William y Stephen Rollnick (1991). *La entrevista motivacional*. Buenos Aires, Paidós.
Moomal, Zubair & Stephanus Petrus Henzi (2000). The Evolutionary Psychology of Deception and Self-Deception. *South African Journal of Psychology*, *30*, 3.
Nelkin, Dana K. (2002). Self-Deception, Motivation, and the Desire to Believe. *Pacific Philosophical Quarterly 83*. 384–406.
Nesse, Randolph (1998). Emotional disorders in evolutionary perspective. *British Journal of Medical Psychology*, 71, 397-415.
Nesse, Randolph & Alan T. Lloyd (1992). The Evolution of Psychodynamic Mechanisms. En Barkow, J., L. Cosmides, & J. Tooby, (eds.) (1992), *The Adapted Mind. Evolutionary Psychology and the generation of Culture*. New York, Oxford University Press.
Nicholson, Anna (2007). Cognitive Bias, Intentionality and Self-Deception. *Teorema* XXVI/3. 45-58.
Noordhof, Paul (2003). Self-Deception, Interpretation and Consciousness. *Philosophy and Phenomenological Research. LXVII*, 1. 75-100.
Noordhof, Paul (2009). The Essential Instability of Self-Deception. *Social Theory and Practice*, *35*, 1. 45-71.
Norem, Julie K. (2002). Defensive self-deception and social adaptation among optimists. *Journal of Research in Personality 36*. 549–555.
Norris, Amanda, Larry Powell & Mark Hickson, III (2007).The Relationship Between Self-Deception and Extrinsic-Personal Religiosity. *Human Communication*, *10*, 3. 373 – 380.
Oksenberg-Rorty, Amelie (1985). The Deceptive Self: Liars and Layers. *Analyse & Kritik 7*. 141-161.
Oksenberg-Rorty, Amelie (1988). The Deceptive Self: Liars, Layers and Lairs. En Brian McLaughlin & Amelie Oksenberg-Rorty (eds.).
Oksenberg-Rorty, Amélie (1994). User-Friendly Self-Deception. *Philosophy*, *69*. 268. 211-228.
Orwell, George (1948). *1984*. Barcelona, Destino.
Palmer, Anthony (1979). Characterising Self-deception. *Mind, New Series. 88*. 349. 45-58.
Paluch, Stanley (1967). Self-Deception. *Inquiry 10*. 268–78.
Panksepp, Jaak & Jules Panksepp (2000). The Seven Sins of Evolutionary Psychology. *Evolution and Cognition*, 6, 108-131.
Pataki, Tamas (1997). Self-Deception and Wish-Fulfilment. *Philosophia*, 25, 1-4. 297-322.
Paulhus, Delroy (2007). Self Deception. En Roy Baumeister & Kathleen Vohs (eds.). *Encyclopedia of Social Psychology*. Los Angeles, Sage.
Pears, David (1982). Motivated irrationality, Freudian theory and cognitive dissonance. En Richard Wollheim & James Hopkins (eds.).
Pears, David (1984), *Motivated Irrationality*. New York, Oxford University Press.
Pears, David (1991). Self-Deceptive Belief-Formation. *Synthese* 89. 393–405.

Pedrini, Patrizia (2005). Self-Deception: What is it to Blame After All? *Annali del Dipartimento di Filosofia*, *XI*. 147-179.
Pihlström, Sami (2007).Transcendental Self-Deception. *Teorema* XXVI/3. 177-189.
Pinker, Steven (1997). *Cómo funciona la mente*. Barcelona, Destino.
Pinker, Steven & Paul Bloom (1992). Natural Language and Natural Selection. En Barkow, J., Cosmides, L., Tooby, J. (Eds.).
Platek, Steven M. & Todd Shackleford (eds.) (2009). *Foundations in Evolutionary Cognitive Neuroscience*. New York, Cambridge University Press.
Quattrone, George & Amos Tversky (1985). Self-deception and the voter's illusion. En Jon Elster (ed.) (1985).
Radden, Jennifer (2011). *On Delusion*. New York, Routledge.
Räikkä, Juha (2007). Self-Deception and Religious Beliefs. *The Heythrop Journal XLVIII*. 513–526.
Ramachandran, V. S. (1996). The Evolutionary Biology of Self-Deception, Laughter, Dreaming and Depression: Some Clues for Anosognosia. *Medical Hypothesis 47*. 347-362.
Ramachandran, V. S. & Sandra Blakeslee (1998). *Phantoms in the Brain. Probing the Mysteries of the Human Mind*. New York, Harper Collins.
Rey, Georges (1988). Toward a computational account of Akrasia and self-deception. En A. O. Rorty & B. P. McLaughlin (eds.).
Robins, Richard & Beer, Jennifer (2001). Positive Illusions About the Self: Short-Term Benefits and Long-Term Costs. *Journal of Personality and Social* Psychology, *80*, 2. 340-352.
Ruddick, William (1988). Social Self-Deceptions. En Brian McLaughlin & Amelie Oksenberg-Rorty (eds.) *Perspectives on Self-Deception*.
Runciman, David (2008). *Political Hypocrisy. The Mask of Power from Hobbes to Orwell and Beyond*. Princeton and Oxford, Princeton University Press.
Ryan, Sharon (2003). Doxastic compatibilism and the ethics of belief. *Philosophical Studies, 114*. 47–79.
Sackeim, Harold & Ruben Gur (1985). Voice Recognition and the Ontological Status of Self-Deception. *Journal of Personality and Social Psychology* 48:1365–68.
Sartre, Jean-Paul (1943). *El Ser y la Nada*. Buenos Aires, Losada.
Saunders, John (1975). The Paradox of Self-Deception. *Philosophy and Phenomenological Research, 35*, 4. 559-570.
Scott-Kakures, Dion (1996). Self-Deception and Internal Irrationality. *Philosophy and Phenomenological Research, 56*, 1. 31-56.
Scott-Kakures, Dion (2000). Motivated Believing: Wishful and Unwelcome. *Nous 34*, 3. 348–375.
Scott-Kakures, Dion (2001). High anxiety: Barnes on what moves the unwelcome believer. *Philosophical Psychology*, *14*, 3. 313-326.
Scott-Kakures, Dion (2002). At "Permanent Risk": Reasoning Self-Knowledge in Self-Deception. *Philosophy and Phenomenological Research*, *LXV*, 3. 576-603.
Scott-Kakures, Dion (2009). Unsettling Questions: Cognitive Dissonance in Self-Deception. *Social Theory and Practice, 35*, 1. 73-106.
Searle, John (2001). *Rationality in Action*. Cambridge, The MIT Press.
Seligman, Martin E. P. & Mihaly Csikszentmihalyi (2000). Positive Psychology. An Introduction. *American Psychologist*, *55*, 1. 5-14.
Shean, Glenn (1993). Delusions, Self-Deception, and Intentionality. *Journal of Humanistic Psychology*, *33*. 45-66.
Siegler, Frederick (1963). Self-deception. *Australasian Journal of Philosophy, 41*, 1. 29-43.
Sorensen, Roy (1985). Self-Deception and Scattered Events. *Mind 94*. 64–69.

Starek, Joanna & Caroline Keating (1991). Self-Deception and Its Relationship to Success in Competition. *Basic and Applied Social Psychology*, *12*, 2. 145-155.

Steup, Matthias (2008). Doxastic freedom. *Synthese*, 161. 375–392.

Stevens, Sean, Kevin Guise, William Christiana, Monisha Kumar & Julian P. Keenan (2006). Deception, Evolution and the Brain. En Steven Platek, Julian P. Keenan & Todd Shackleford (eds.) *Evolutionary Cognitive Neuroscience*. Cambridge, The MIT Press.

Sturm, Thomas (2007). The Self Between Philosophy and Psychology: The Case of Self-Deception. En Mitchell G. Ash & Thomas Sturm (eds.), *Psychology's Territories. Historical and Contemporary Perspectives from Different Disciplines*. New Jersey, Lawrence Erlbaum.

Surbey, Michele (2011). Adaptive significance of low levels of self-deception and cooperation in depression. *Evolution and Human Behavior 32*. 29–40.

Surbey, Michele & Jeffrey McNally (1997). Self-Deception as a Mediator of Cooperation and Defection in Varying Social Contexts Described in the Iterated Prisoner's Dilemma. *Evolution and Human Behavior* 18. 417-435.

Symons, Donald (1992). On the Use and Misuse of Darwinism in the Study of Human Behavior. En Barkow, J., Cosmides, L., J. Tooby, J. (Eds.).

Szabados, Bela (1973). Wishful Thinking and Self-Deception. *Analysis 33*, 6. 201-205.

Szabados, Bela (1974). Self-Deception. *Canadian Journal of Philosophy* 4. 51–68.

Talbott, William (1995). Intentional Self-Deception in a Single Coherent Self. *Philosophy and Phenomenological Research 55*, 1. 27-74.

Tattersall, Ian (2001). Evolution, Genes, and Behavior. *Zygon*, 36, 657-666.

Taylor, Shelley & Jonathon Brown (1988). Illusion and Well-Being: A Social Psychological Perspective on Mental Health. *Psychological Bulletin 103*, 2. 193-210.

Taylor, Shelley & Jonathon Brown (1994). Positive Illusions and Well-Being Revisited Separating Fact From Fiction. *Psychological Bulletin*, *116*, 1. 21-27.

Taylor, Shelley & Peter Gollwitzer (1995). Effects of Mindset on Positive Illusions. *Journal of Personality and Social Psychology, 69*, 2. 213-226.

Taylor, Shelley, Margaret Kemeny, Geoffrey Reed, Julienne Bower, and Tara Gruenewald (2000). Psychological Resources, Positive Illusions, and Health. *American Psychologist*, *55*, 1. 99-109.

Tenbrunsel, Ann & David M. Messick (2004). Ethical Fading: The Role of Self-Deception in Unethical Behavior. *Social Justice Research, 17*, 2.

Thagard, Paul (2011). Critical Thinking and Informal Logic: Neuropsychological Perspectives. *Informal Logic* 31, 3. 152-170.

Tooby, John & Leda Cosmides (1992). The Psychological Foundations of Culture. En Barkow, J., Cosmides, L., Tooby, J. (Eds.).

Triandis, Harry (2013). Self-deception: An Introduction. *Acta de Investigación Psicológica*, 3 (2). 1069-1078.

Trivers, Robert & Huey P. Newton (1982). The crash of Flight 90: Doomed by self-deception? *Science Digest* (Nov.): 66–67, 111. Reimpreso en Robert Trivers (2002). *Natural Selection and Social Theory: Selected Papers of Robert Trivers*. New York, Oxford University Press.

Trivers, Robert (2000). The Elements of a Scientific Theory of Self-Deception. *Annals of the New York Academy of Sciences* 907. 114-131.

Trivers, Robert (2002). Self-deception in service of deceit. En Robert Trivers, *Natural Selection and Social Theory: Selected Papers of Robert Trivers*. New York, Oxford University Press.

Trivers, Robert (2010). Deceit and Self-Deception. En Peter M. Kappeler & Joan B. Silk (eds.), *Mind the Gap. Tracing the Origins of Human Universals*. Springer. 373-393.

Trivers, Robert (2011). *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. New York, Basic Books.

Twerski, Abraham (1997). *Addictive Thinking: Understanding Self-deception*. Hazelden Publishing.
Vaillant, George (2000). Adaptive Mental Mechanisms. Their Role in a Positive Psychology. *American Psychologist*, *55*, 1. 89-98.
van den Bos, Kees & Marjolein Maas (2009). On the Psychology of the Belief in a Just World: Exploring Experiential and Rationalistic Paths to Victim Blaming. *Personality and Social Psychology Bulletin 35*, 12. 1567-1578.
Van Leeuwen, D. S. Neil (2007a). The Product of Self-Deception. *Erkenntnis 67*. 419–437.
Van Leeuwen, D. S. Neil (2007b). The Spandrels of Self-Deception: Prospects for a Biological Theory of a Mental Phenomenon. *Philosophical Psychology 20*, 3. 329–348.
Van Leeuwen, D. S. Neil (2008). Finite rational self-deceivers. *Philosophical Studies 139*.191–208.
Van Leeuwen, D. S. Neil (2009). Self-Deception Won't Make You Happy. *Social Theory and Practice*, *35*, 1. 107-132.
Van Leeuwen, D. S. Neil (2013). Self-deception. Extraído el 14/06/13 de http://www.academia.edu/1808355/Self-Deception
von Hippel, William & Robert Trivers (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences 34*. 1–56.
Westland, Sandra & Pnina Shinebourne (2009). Self-deception and the therapist: An interpretative phenomenological analysis of the experiences and understandings of therapists working with clients they describe as self-deceptive. *Psychology and Psychotherapy: Theory, Research and Practice*, *82*. 385–401.
Williams, Bernard (1973). Deciding to believe. En *Problems of the Self. Philosophical Papers 1956-1972*. Cambridge, Cambridge University Press.
Wilson, Catherine (1980). Self-deception and Psychological Realism. *Philosophical Investigations*, *3*, 4. 47–60.
Wilson, Edward (1975). *Sociobiología. La nueva síntesis*. Madrid, Omega.
Wollheim, Richard & James Hopkins (eds.) (1982). *Philosophical essays on Freud*. New York, Cambridge University Press.
Wood, Allen (1988). Ideology, False Consciousness and Social Illusion. En Brian McLaughlin & Amelie Oksenberg-Rorty (eds.).
Wrangham, Richard (1999). Is Military Incompetence Adaptive? *Evolution and Human Behavior 20*. 3–17.
Zerubavel, Eviatar (2006). *The Elephant in the Room. Silence and Denial in Everyday Life*. New York, Oxford University Press.