AUTOENGAÑO Y EVIDENCIA*

GUSTAVO FERNÁNDEZ ACEVEDO
Universidad Nacional de Mar del Plata
Departamento de Filosofia
Mar del Plata
Argentina
gustavofernandezacevedo@gmail.com

Article info
CDD: 100
Received: 02.09.2018; Accepted: 13.09.2018
DOI: http://dx.doi.org/10.1590/0100-6045.2018.V41N3.GA

Keywords:
Self-deception
Evidence
Beliefs

Palabras claves:
Autoengaño
Evidencia
Creencias

Abstract: A condition usually considered necessary for self-deception is that the belief the agent acquires or sustains is not supported by the evidence available to him. However, this formulation is too wide and requires higher accuracy. This article presents a characterization of that condition that intends to overcome the objections that affect the available formulations.

Resumen: Una condición usualmente considerada necesaria para el autoengaño consiste en que la creencia que se adquiere o mantiene no debe ser sostenida por la evidencia a disposición del

* Deseo agradecer las útiles observaciones del árbitro de *Manuscrito*, que me permitieron mejorar varios aspectos de este artículo. Identifico tales observaciones en notas al pie.

agente. Sin embargo, esta formulación general de la condición es demasiado amplia y requiere de mayor precisión. En el presente artículo se presenta una caracterización de tal condición que intenta superar las objeciones que pueden elevarse contra las formulaciones existentes.

I

El fenómeno del autoengaño, del cual se ha ocupado tradicionalmente la filosofía (Davidson, 1982, 1985; Demos, 1961; Fingarette, 1969; Mele, 1997, 2001; Oksenberg-Rorty, 1988; Pears, 1984) y, más recientemente, diversas disciplinas científicas (Hirstein, 2005; Metcalfe, 1998; Taylor y Brown, 1988; Trivers, 2000, 2011; von Hippel y Trivers, 2011), se ha revelado como notoriamente elusivo y controvertido, a punto tal que hasta fechas relativamente recientes persistían los cuestionamientos respecto de su propia existencia (Borge, 2003). Los debates acerca de él no han limitado a aspectos específicos de su naturaleza (por ejemplo, si su formación requiere de una intención del agente)[1], sino que su misma definición continúa siendo objeto de polémicas (Deweese-Boyd, 2016).

Pese a las perennes controversias acerca de la caracterización del autoengaño, existe un consenso respecto de la clase de hechos en la cual está incluido, esto es, en la de aquellos fenómenos conocidos como "irracionalidad motivada", categoría que incluye, entre sus integrantes más importantes, a la *akrasia* y al pensamiento desiderativo. Al

[1] Cabe aclarar aquí que, aunque a lo largo de este artículo emplearemos la palabra "agente", su uso no implica la adhesión a una concepción intencionalista del autoengaño; en particular, no implica la aceptación de la tesis según la cual quien se autoengaña ejecuta un plan deliberado con el fin de inducir en sí mismo una creencia falsa.

igual que este último, y a diferencia de la primera, el autoengaño constituye una forma de irracionalidad *teórica*, es decir, una clase de distorsión en los procedimientos racionales de adopción (o mantenimiento) de creencias que conduce a aceptar una creencia que no debería sostenerse en vista de la evidencia disponible. Ahora bien, los fenómenos incluidos en la categoría de irracionalidad motivada, en particular aquellos que constituyen formas de irracionalidad teórica, no constituyen meros errores debido a factores como fallas en el razonamiento, cansancio o insuficiente comprensión; se supone que la adopción o mantenimiento de la creencia está fuertemente influida por procesos no cognitivos, como los deseos y los temores, que sesgan la racionalidad teórica.

Sobre la base de lo anterior, es posible sugerir algunos rasgos que parecen caracterizar al fenómeno que nos ocupa; 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 evidencia contraria a ella,
c. motivados por procesos mentales no cognitivos (por ejemplo, deseos o emociones) que favorecen su adquisición y/o retención.

Ahora bien, todas y cada una de las notas de esta caracterización intuitiva pueden ser cuestionadas o, como mínimo, matizadas o debilitadas. En primer lugar, podría señalarse que estas 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: al igual que el engaño interpersonal, requiere de una intención para su ejecución (Pears, 1984; Davidson, 1985). También se ha sostenido 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.[2] Por otra parte, el primer punto de la caracterización supone 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*], una disposición a afirmar una proposición con sinceridad, pero que carece de conexiones profundas con la acción, rasgo que sí caracteriza a las creencias. 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

[2] Lynch (2010), observa que hay dos casos distintos en los que es posible hablar de autoengaño, casos que deberían ser distinguidos: "estar autoengañado *en* creer que *p*", y "haberse engañado a uno mismo *para* creer que *p*". 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 que crea* que esa proposición es verdadera. Sin desestimar la posibilidad de que Lynch esté en lo correcto, en este trabajo nos limitaremos a examinar los casos típicos de autoengaño, esto es, aquellos casos en los cuales la creencia que el agente favorece de modo autoengañoso es falsa.

lo generan. Entre los principales candidatos se encuentran el deseo (Pears, 1984; Davidson, 1985), la ansiedad (Barnes, 1999) y las emociones (Lazar, 1999). También hay modelos explicativos que no recurren a un solo factor no cognitivo desencadenante del autoengaño, sino a varios de ellos (Mele, 2001).

Los esfuerzos en la caracterización del autoengaño, por otra parte, no sólo han tenido como objetivo una definición precisa de este fenómeno, sino también el logro de una distinción entre él y otras formas de irracionalidad motivada, como el pensamiento desiderativo. Diversos autores, como veremos, han hecho intentos de diferenciar el autoengaño de otros fenómenos emparentados con él, aunque no siempre resultan claras las motivaciones que justifican tales tentativas.

En este trabajo quiero ocuparme de uno de los problemas que plantea la caracterización del autoengaño, esto es, el referente al requisito de evidencia. Este requisito suele enunciarse de una forma simplificada, como vimos, de la siguiente manera: quien se autoengaña debe sustentar su creencia en presencia de evidencia contraria a ésta. Sin embargo, como puede verse mediante un somero examen, tal formulación deja indefinidas muchas dimensiones importantes del requisito. Entre ellas, ¿debe poseer efectivamente el agente la evidencia contraria a su creencia, o basta con que esta evidencia esté a su disposición y pueda obtenerla si desea hacerlo? ¿Cuál es el grado de fortaleza que debe tener la evidencia contraria a la creencia que el agente favorece?[3] ¿Debe el requisito ser formulado de

[3] Si bien no es muy frecuente que se haga referencia al grado de fortaleza que debe tener la evidencia contraria a la creencia sostenida por el agente, se ha sugerido alguna especificación de esta dimensión. Deweese-Boyd (2008), por ejemplo, caracteriza al autoengaño como la adquisición y mantenimiento de una creencia falsa frente a fuerte evidencia en contrario, estado que es causado por deseos o emociones que sesgan el propio manejo de la

idéntica manera para la adquisición autoengañosa de una creencia y para su mantenimiento? En particular, quiero exponer algunos argumentos en favor de la tesis según la cual, para que sea posible hacer una atribución de autoengaño, es necesario que el agente esté en posesión no sólo de la evidencia favorable a la creencia que sostiene, sino también de la evidencia contraria a tal creencia y que, considerada globalmente, debería conducir al rechazo de ésta. Trataré de mostrar que, en ausencia de esta condición, la categoría de autoengaño se torna inaceptablemente difusa y se dificulta la distinción entre el autoengaño y otras formas de irracionalidad motivada que conviene diferenciar.

El trabajo estará estructurado de la siguiente forma. En la segunda sección presentaré algunos intentos de distinguir el autoengaño de otros fenómenos de irracionalidad motivada sobre la base de la diferente posición del agente respecto de la evidencia contraria a su creencia, y describiré algunas de las réplicas a estos intentos. En la tercera sección examinaré las posiciones precedentemente descriptas y defenderé la tesis según la cual la condición de evidencia resulta esencial para distinguir el autoengaño de otras formas de irracionalidad motivada, tanto por razones teóricas como por razones prácticas. En la cuarta sección sugeriré un criterio de evidencia para la adquisición autoengañosa de creencias que evite algunas de las dificultades que presentan los criterios existentes, por una parte, y los problemas que trae el considerarlo una condición meramente suficiente para el autoengaño, por la

evidencia pertinente para la evaluación de la creencia. Observemos que no sólo estipula que la evidencia contraria a la creencia preferida por el agente debe ser sólida, sino que también distingue entre la aceptación inicial de una creencia y su mantenimiento posterior.

otra.[4] Por último, en la quinta sección examinaré si la condición de evidencia sugerida resulta aplicable al

[4] Al examinar el modo en que debe concebirse la condición de evidencia en el autoengaño no es posible evitar mencionar, aunque sea de modo tangencial, el problema relativo a los muy distintos grados de complejidad que puede tener la evidencia pertinente para la evaluación de una creencia. Por mencionar un ejemplo entre muchos posibles, Richard Holton (2000), propone como caso de autoengaño el de Jean Marie, un racista que está convencido de la superioridad de los blancos sobre los negros y los árabes. Este ejemplo, así como otros ejemplos análogos (el del fanático religioso o el del adherente acérrimo a ideas o regímenes políticos), plantean espinosas cuestiones relativas a la racionalidad en la formación de creencias que refieren a hechos, por así decirlo, muy lejanos a nuestra experiencia inmediata. Entre las diversas dificultades que esta clase de ejemplos plantea podemos mencionar las siguientes. En primer lugar, no se trata de un conocimiento sobre estados o sucesos singulares; se trata de afirmaciones sobre regularidades que exceden ampliamente el dominio de los hechos que nos incumben directamente (si bien esta es una categoría vaga, sin duda tiene casos claros de aplicación, como los referidos a nuestro estado de salud o a la fidelidad de nuestra pareja). En segundo lugar, se trata de creencias acerca de las cuales resulta mucho más difícil tener evidencia directa (como sí se puede, por ejemplo, sobre la infidelidad de la pareja). La mayor parte de la evidencia pertinente para la formación de creencias que se refieren a colectivos (como los grupos étnicos) será una evidencia mediatizada (por ejemplo, por terceros y/o por medios de comunicación). Esto hace mucho más difícil establecer, en un sentido normativo, cuál es la evidencia pertinente a disposición del agente a partir de la cual puede sustentarse la creencia rechazada por el agente. Tercero, tales ejemplos comparten el atributo de que muy a menudo son compartidos por colectivos más o menos extensos; no se trata de meros procesos de autoengaño individual. Escapa a los propósitos de este trabajo determinar si tales casos constituyen ejemplos de autoengaño o a alguna categoría más amplia de irracionalidad colectiva. A los fines de la presente argumentación,

mantenimiento autoengañoso de creencias o si se requiere alguna modificación adicional.[5]

II

Algunos filósofos han sostenido que la posesión de la evidencia contraria a la creencia favorecida por el agente es un requisito necesario para la definición del autoengaño. Cito al respecto un ejemplo ilustrativo:

> [U]n agente está en un estado de autoengaño si y sólo si sostiene una creencia que a) es contraria a lo que sus normas epistémicas, *en conjunción con la evidencia disponible*, usualmente dictarían y b) un deseo por la obtención de cierto estado de cosas, o por poseer cierta creencia, hacen una diferencia causal para que la creencia sea sustentada en un modo epistémicamente ilegítimo (Van Leeuwen, 2007, p. 332. Cursivas mías).

Obsérvese, no obstante, que el hecho de que de la evidencia esté disponible no es equivalente a la *posesión* de la evidencia por parte del agente (éste podría evitar de modo más o menos activo tomar contacto con ella), ni a su

---

nos limitaremos a analizar el caso de la evidencia relativa a hechos cercanos a la experiencia directa del agente.

[5] Si bien podría razonablemente afirmarse que todo autoengaño tiene un componente social, ya que muy a menudo requiere de colaboración interpersonal (Ruddick, 1988), y esto vale también, sin duda, a la consideración de las dimensiones de la condición de evidencia, a los fines del análisis que desarrollaremos omitiremos, por razones de simplicidad, el examen de tales factores interpersonales o sociales.

inteligibilidad (esto es, que la información en cuestión sea susceptible de ser reconocida por parte del agente como evidencia pertinente para la evaluación de la creencia), ni tampoco a la *competencia* del agente (esto es, que éste sea capaz de reconocer que un fragmento de información constituye evidencia en favor o en contra de una determinada creencia). Sin embargo, a veces sí se especifica la condición de evidencia de modo tal que se limita de manera importante la ambigüedad de las implicaciones de este requisito. La definición de Davidson proporciona un buen ejemplo al respecto:

> 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 (Davidson, 1985, pp. 111-112).

Puede advertirse que la caracterización de Davidson, a diferencia de la precedente, supone no sólo que el agente es conciente de la existencia de evidencia contraria a su creencia, sino que la interpreta correctamente como evidencia que sustenta la creencia opuesta a la que finalmente adopta.

Como anticipé, la especificación de los requisitos del autoengaño ha supuesto también la posibilidad de distinguir este fenómeno de otras clases de pensamiento supuestamente irracional, como 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*.[6] En esta línea de pensamiento, Szabados (1974) considera que la diferencia crucial entre el autoengaño y el pensamiento desiderativo reside en que, mientras que en el primero la evidencia es contraria a la creencia que se adopta, esto no ocurre en el pensamiento desiderativo. A diferencia del autoengañado, quien sostiene un pensamiento desiderativo no pervierte los procedimientos por medio de los cuales establecemos la verdad o falsedad de nuestras creencias. Si tal persona es confrontada con evidencia incompatible con su creencia será capaz de reconocer, aunque tal vez con reticencia, su carácter contrario a aquello que cree. Sin embargo, agrega Szabados, si procede a resistir "por medio de tácticas ingeniosas" las implicaciones de la evidencia presentada sentiremos que se encuentra autoengañada. Bach (1981), por su parte, distingue el autoengaño del pensamiento desiderativo y de

[6] Se ha señalado que el pensamiento desiderativo no opera de un modo tan simple como esta 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.

otra clase de irracionalidad motivada, la ceguera intelectual. A diferencia del autoengaño, en los casos de pensamiento desiderativo no existe ningún razonamiento o esbozo de él. Quien piensa desiderativamente imagina algún estado de cosas agradable y supone que eso sucederá; no intenta justificar su creencia, tal vez por resultarle suficiente la ausencia de evidencia en favor o en contra de ella. En caso de que el agente sea consciente de la presencia de evidencia en contrario, prosigue, y de la necesidad de lidiar con ella, nos encontraríamos ante un caso especial de autoengaño. Tampoco es el autoengaño, advierte Bach, un caso de ceguera intelectual. Este fenómeno consiste en la incapacidad de advertir la dirección hacia la que apunta la evidencia; por el contrario, quien se autoengaña percibe correctamente tal dirección, al menos inicialmente.

Según estas caracterizaciones, en síntesis, el autoengaño parece requerir del agente la comprensión de que la evidencia es contraria a la creencia que favorece, y en esto se diferencia el autoengaño del pensamiento desiderativo y de otras formas de irracionalidad motivada.

Ahora bien, no todos los filósofos que se han ocupado del problema están de acuerdo con la tesis de que el autoengaño debe requerir necesariamente que el agente sea conciente de que la evidencia es contraria a la creencia preferida. Un ejemplo especialmente destacado de este desacuerdo es el proporcionado por la perspectiva de Alfred Mele. Mele (2001) 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 desviante de que *S* adquiera la falsa creencia de que *p*.
4. El cuerpo de datos poseídos por *S* en ese momento provee mayor garantía para ¬ *p* que para *p* (p. 50-51).

Mele cuestiona la objeción de algunos autores respecto de que la condición 4 es demasiado débil. Según esta objeción, la condición 4 debe formularse de modo tal que permita atribuir al sujeto un reconocimiento de que su evidencia proporciona mayores garantías para ¬*p* que para *p*. La objeción se basa en que, si es el caso que nos engañamos a nosotros mismos al creer que *p*, debemos ser concientes de que la evidencia que poseemos favorece a ¬*p*, y es esta conciencia lo que explica nuestro tratamiento tendencioso de los datos. Sin esa conciencia, se argumenta, no habría razones para tratar los datos de modo motivacionalmente sesgado, ya que éstos no serían percibidos como amenazantes y, en consecuencia, no nos comprometeríamos con una cognición motivacionalmente desviada. Mele considera que una exigencia tal tiende a estar ligada a una concepción intencionalista del autoengaño: el agente se fija una meta, determina los medios para promover su logro y procede en consecuencia. Sin embargo, objeta, este modelo establece exigencias excesivas sobre quienes se autoengañan. La cognición fría o no motivada no es explicada sobre la base de la acción intencional, y la motivación puede poner en marcha y sostener el funcionamiento de mecanismos que sesgan de manera "fría" los datos, sin que seamos concientes o lleguemos a creer que la evidencia que poseemos favorece a una determinada proposición por sobre otra. Más aun, Mele observa explícitamente que esta cuarta condición no debe ser considerada una condición necesaria para que sea posible hablar de autoengaño respecto de *p* (algo que sí

ocurre, según él, con la condición 1). Considera que, en algunos casos de recolección motivacionalmente sesgada de la evidencia, las personas pueden llegar a creer en una proposición falsa aun cuando la proposición contradictoria con ella está mucho mejor sustentada en la evidencia que podrían fácilmente obtener; debido a la selectividad en el proceso de recolección, observa, la evidencia que efectivamente poseen favorece a la creencia falsa y no a la verdadera. No obstante, en su opinión esas personas son evaluadas naturalmente, en igualdad de circunstancias, como autoengañadas.[7] La ignorancia, como sostiene en *Irrationality* (1987), no excluye al autoengaño; por el contrario, hay casos en los que parece contribuir con él.

Así como la ignorancia de la evidencia no elimina la posibilidad de atribuir autoengaño, Mele tampoco considera que el autoengaño pueda ser nítidamente diferenciado del pensamiento desiderativo sobre la base de este requisito. En *Irrationality* discute brevemente el intento de Szabados de diferenciar ambos fenómenos. Señala que, si hay alguna diferencia entre pensamiento desiderativo y autoengaño, esta puede radicar simplemente en que el primero constituye un género denotado por el término "autoengaño". Observa que, 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 apropiada e influenciada por el deseo, quien se autoengaña carece de buenas bases para rechazar la proposición que sostiene autoengañosamente. Si bien, agrega, 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

[7] Mele se ha mantenido en una defensa consistente de esta posición. Cfr. al respecto Mele (2012).

"pensamiento desiderativo", habláramos de "falsa creencia desiderativa", expresión que le parece correcta si nos basamos en el análisis que hace Szabados.

III

Lo expuesto, en síntesis, parece fijar de manera bastante nítida dos posiciones. La primera, sostenida por varios filósofos, tiende a considerar a la condición de posesión de evidencia por parte del agente (y, en el caso de Davidson, la comprensión de éste de que la evidencia apunta en la dirección contraria a la creencia preferida) como requisito necesario del autoengaño y también como criterio para distinguirlo de otras formas de irracionalidad, como el pensamiento desiderativo. Mele, por su parte, niega claramente la necesidad del requisito y no considera que pueda establecerse sobre su base una distinción clara con el pensamiento desiderativo. Creo que Mele está en lo correcto al señalar que la posibilidad de una falta de captación de la evidencia a causa de un sesgo en su recolección difumina en alguna medida la distinción entre los fenómenos en cuestión. Sin embargo, hay varias razones por las cuales considero que la condición de evidencia debe, en primer lugar, ser mantenida como requisito necesario, y no sólo como una de las condiciones conjuntamente suficientes, del autoengaño y, en segundo lugar, que esto requiere de una formulación más fuerte que la mera disponibilidad de evidencia contraria a la creencia preferida.

En primer lugar, no parece forzoso conceder que las personas que adquieren una falsa creencia motivada a causa de una recolección sesgada de evidencia sean evaluadas "naturalmente" como autoengañadas, tal como sostiene Mele. Si "naturalmente" hace referencia a los usos comunes del término en el lenguaje ordinario, parece dudoso que tal referencia pueda ser de mucha ayuda en la discusión de

distinciones como la que nos ocupa. En el caso de que una persona acepte una determinada proposición sin percibir, debido a un deseo o temor, que una parte importante de la evidencia es contraria a esa proposición, entiendo que, así como podríamos decir que se autoengaña, con igual derecho podríamos afirmar que se trata de una persona "cegada por el deseo" o "cegada por la pasión", distinguiendo este fenómeno del caso en el cual la persona percibe claramente la evidencia contraria a su creencia y la interpreta de manera irracional y distorsionada, de modo de sustentar la creencia predilecta. No se ganaría demasiado, a nuestro entender, si se adoptaran expresiones como "autoengaño ciego" o "autoengaño infundado" para designar a esta clase de casos; esta alternativa no sería más que un subterfugio lingüístico que demostraría la necesidad de distinguir entre fenómenos de distinta clase, aunque emparentados.[8]

---

[8] El árbitro ha señalado que, contrariamente a lo sostenido en esta sección, la intuición de Mele según la cual aquellos casos en los cuales el agente cree que p en vez de $\neg p$ debido a una recolección sesgada de evidencia son casos de autoengaño es bastante sólida. Considera que tales casos son lo suficientemente similares a los ejemplos paradigmáticos de autoengaño como para descartar que Mele esté empleando el término "autoengaño" con un significado diferente. Sugiere, además, que la manera adecuada de descartar la intuición de Mele es o bien a través de un contraejemplo (esto es, un caso de adopción irracional de una creencia por medio de procedimientos de recolección sesgada que no despierte la intuición de autoengaño defendida por Mele), o bien por medio de argumentar que incluso los casos de recolección sesgada de evidencia son casos en los que el agente evalúa de modo motivacionalmente sesgado la evidencia que posee, con la diferencia de que en tales casos la evidencia no es de primer orden, sino de segundo. La evidencia de segundo orden consiste en evidencia acerca de la existencia de evidencia de primer orden respecto de una proposición o de la importancia de tal evidencia. Sobre la base de esta distinción, el caso propuesto

Por otra parte, el rechazo del carácter necesario del requisito de evidencia tiene consecuencias negativas que no conviene soslayar. Notemos en primer término que la categoría de autoengaño resultante de las condiciones suficientes de Mele sin el requisito de evidencia parece redundar en una clase abarcativa casi tan amplia como es "irracionalidad motivada", y ser tan vaga e incluyente como la de "ilusiones" propuesta por algunos psicólogos (Taylor y Brown, 1988). Admite desde casos en los cuales el sujeto que se autoengaña ignora de manera involuntaria la evidencia contraria a su creencia hasta los (raros) casos de autoengaño intencional y deliberado, pasando por las situaciones en las cuales el sujeto percibe la existencia de evidencia extremadamente sólida en contra de su creencia preferida. No parece imposible afirmar que todos estos son casos de autoengaño, pero sí parece haber diferencias importantes entre ellos que, creo, son significativas, y que se traducen en otras diferencias de importancia, tanto desde el punto de vista filosófico como desde la perspectiva psicológica.

---

por Mele podría ser plausiblemente redescripto del siguiente modo: el agente tiene evidencia acerca de la existencia de evidencia contraria a *p* (evidencia de segundo orden respecto de *p*) y evalúa de manera motivacionalmente sesgada esa evidencia de segundo orden, esto es, ignora o es indiferente a evidencia respecto de la existencia de evidencia contraria a *p* debido a sesgos motivacionales. Si esto fuese correcto, señala el árbitro, incluso los casos de recolección sesgada de evidencia serían casos en los cuales el agente evalúa de modo motivacionalmente sesgado evidencia que posee, con la diferencia que en tales casos la evidencia sería indirecta, de segundo orden, sobre *p*, en lugar de evidencia directa respecto de *p*. Encuentro plausible esta sugerencia del árbitro (de hecho más plausible que el proponer un contraejemplo convincente) como respuesta a la objeción por él mismo planteada.

La primera diferencia filosófica se relaciona con la atribución de racionalidad, y puede ser explicada mediante un análisis de dos de los casos mencionados que, si se adopta el criterio de Mele, caen dentro de la categoría de autoengaño. Estos casos son, en primer lugar, el autoengaño en el cual el agente no percibe, a causa de procedimientos sesgados en la recolección de la información, la evidencia contraria a su creencia, esto es, el caso de autoengaño que ya hemos descripto; en segundo lugar, el autoengaño que, empleando un término utilizado por Mele, llamaremos "extremo". Estos casos son aquellos en los cuales la evidencia contraria a la creencia favorecida por el agente es notoriamente sólida y el agente es consciente de la dirección hacia la que apunta.[9] Ahora bien, si consideramos que el autoengaño es un fenómeno básicamente irracional, parecería que esta forma de autoengaño constituye una variante máximamente irracional; esto es, el agente enfrenta evidencia extremadamente fuerte contraria a la creencia que favorece, lo que necesariamente conduce a un tratamiento sesgado igualmente extremo de ésta. En este aspecto, entonces, esta variante de autoengaño difiere cuantitativamente, pero en gran medida, de la modalidad en la cual el agente ignora la evidencia existente en contra de su creencia. Sin embargo, parece poco plausible la suposición de que tal tratamiento no genera ninguna consecuencia en absoluto en el funcionamiento cognitivo del agente; por el contrario, estos casos parecen especialmente aptos para la producción de un estado que, según algunos autores, es característico del autoengaño. Me refiero aquí a la existencia de una *tensión psíquica peculiar* que acompañaría al estado de autoengaño y que se manifestaría, por ejemplo, por dudas recurrentes acerca de la creencia en cuestión o por un estado de

[9] Cfr. Oksenberg-Rorty (1988) y Mele (2001) para ejemplos de esta forma de autoengaño.

conflicto cognitivo producido por la coexistencia de la sospecha de que *p* y la creencia de que ¬*p*.[10] Aunque este estado no sea un componente necesario del autoengaño[11] y, en consecuencia, pueda no estar presente en todos los casos, es plausible suponer (aunque esto debe ser objeto de investigación empírica) que su probabilidad de ocurrencia será muy desigual en los distintos casos de autoengaño. Por ejemplo, parece probable que no esté presente en los casos de autoengaño en los cuales el agente, a causa de la actuación de mecanismos de recolección sesgada de la evidencia, tiende a percibir sólo la evidencia favorable a su creencia y no la contraria; por el contrario, es muy probable que esté presente en los casos de autoengaño extremo, en los cuales el agente debe lidiar con evidencia muy sólida y consistente en contra de su creencia.

Un último argumento en favor de la conveniencia de distinguir los casos de irracionalidad motivada debidos a la selección sesgada de la evidencia de aquellos casos en los cuales el agente está en posesión de tal evidencia está basado en consideraciones prácticas más que teóricas. Estas consideraciones se relacionan con la posibilidad de formular una atribución de responsabilidad moral por el autoengaño y, en particular, por los grados de responsabilidad moral que tal estado implica para quien se encuentra en él. Tradicionalmente se ha sostenido que el autoengaño implica un deterioro moral en quien se encuentra en tal estado (Butler, 1726; Sartre, 1943; Demos, 1961; Baron, 1988) y, además, que la persona autoengañada es responsable por su estado (Sartre, 1943; Demos, 1961). Si bien más recientemente (Levy, 2004), y sobre la base

[10] Cfr. Audi (1997) respecto de la necesidad de incluir tal tensión como un componente fundamental del autoengaño.

[11] Cfr. Mele (2001) para un argumento en contra de la necesidad de este requisito.

proporcionada por los denominados modelos deflacionistas del autoengaño, se ha puesto en tela de juicio la responsabilidad moral por tal estado, sigue siendo mayoritaria la opinión tradicional respecto de este último punto. Si se admite la premisa de que quien se autoengaña es responsable por su estado, entonces puede examinarse la posibilidad de que distintas formas de irracionalidad motiva impliquen distintos grados de responsabilidad moral. *Prima facie*, entiendo que no atribuiríamos la misma responsabilidad moral a quien omite evidencia por la acción de sesgos en la recolección de la evidencia que el que interpreta de manera extremadamente distorsionada la información que tiene a su disposición.

Lo anterior puede, no obstante, parecer insuficiente para fundamentar la tesis que sostiene una diferencia significativa entre autoengaño y otras formas de irracionalidad motivada (especialmente pensamiento desiderativo) sobre la base del requisito de evidencia. Una posible objeción se basaría en las siguientes consideraciones.[12] Mele mantiene que fenómenos como el autoengaño y el pensamiento desiderativo son especies de un mismo género; esto no le impide distinguirlos en un sentido interesante, lo que incluye una distinción en los mismos términos que aquí se intenta aplicar. Este autor podría sostener que sería posible clasificar ambos fenómenos en la misma categoría en virtud de satisfacer el resto de las condiciones; esto es, en los dos casos se formaría una creencia falsa sobre la base del tratamiento motivacionalmente sesgado de la información, ya sea ésta poseída efectivamente o fácilmente accesible en el contexto. Al mismo tiempo, los fenómenos podrían ser diferenciados en tanto la posesión o no de la evidencia genera una fenomenología diferente (la tensión psíquica que acompañaría al autoengaño) y los diferentes juicios de

[12] Debo esta consideración al árbitro de *Manuscrito*.

responsabilidad moral y práctica. Mele, finaliza la objeción, no sostiene que los fenómenos son iguales, sino solamente que son casos de autoengaño, pero esto es compatible con la existencia de diferencias (la posesión o no de la evidencia) que explicarían las diferencias morales y fenomenológicas.

Reconozco, en primer lugar, que las diferencias entre ambos fenómenos no son nada fáciles de establecer, y no supongo que mi planteo esté libre de objeciones. Sin embargo, los casos de irracionalidad motivada que pueden plausiblemente considerarse pensamiento desiderativo no quedan abarcados completamente por el análisis de Mele. Es innegable que, en principio, pueden existir casos de irracionalidad motivada a partir de una selección sesgada de la evidencia que conduzca al agente a adoptar una creencia que no debería adoptar; en tales casos la evidencia, si hubiera sido recolectada apropiadamente, favorecería $\neg p$ en vez de p. Aun cuando se admita que los anteriores constituyen casos de pensamiento desiderativo, es perfectamente posible que muchos casos de este fenómeno no se ajusten a la caracterización anterior. Esto es, resulta posible pensar en casos de pensamiento desiderativo en los que no haya recolección sesgada de la evidencia; el agente simplemente adopta la creencia sobre la base de su deseo, aun cuando la inferencia no resulte tan simple como "deseo que p, por lo tanto p".[13] La carencia de evidencia podría deberse entonces a la recolección sesgada de la evidencia, pero también a que factores motivacionales conducen a la aceptación de la creencia sin búsqueda alguna de datos que la sustenten. De hecho, esta alternativa es más compatible con las caracterizaciones del pensamiento desiderativo que se describieron en la segunda sección. No veo un modo plausible de considerar a este tipo de casos como una variante de autoengaño. Podría objetarse que tales casos no

[13] Véase el comentario al respecto en la nota 6.

son posibles y que siempre debe haber alguna clase de evidencia, pero esta afirmación no parece ser susceptible de ser resuelta sobre bases puramente conceptuales y, además, creo que la carga de la prueba recae sobre quien sostenga que tales casos no pueden existir. Resultaría posible, en consecuencia, reservar la expresión "pensamiento desiderativo" exclusivamente para los casos en los que la carencia de evidencia es debida sólo a que el agente no busca evidencia pertinente para determinar la adopción o rechazo de la creencia, y la adopta a partir de su deseo de que sea verdadera, y considerar como casos de autoengaño no típicos a aquellos en los cuales la adopción de la creencia tiene lugar a partir de una recolección sesgada de la evidencia a disposición del agente. De este modo, entiendo, puede mantenerse la distinción entre pensamiento desiderativo y autoengaño sobre la base del requisito de evidencia.

Una última objeción a la distinción entre autoengaño y pensamiento desiderativo basada en el requisito de evidencia que aquí defendemos estaría basada en la siguiente posibilidad. Podría ocurrir que hubiera casos de autoengaño en los que la evidencia en su conjunto no favoreciera a *p* ni a ¬*p*, esto es, que apoyara la suspensión del juicio respecto de *p*. Si, en tal caso, un agente poseyera evidencia que determinara que sería racional suspender el juicio pero aun así creyera que *p* influido por sus deseos, ansiedades o emociones, parecería estar autoengañándose. La existencia de tales casos podría impactar de manera directa en la formulación del requisito de evidencia, ya que en lugar de requerir que la evidencia total apoye la negación de la creencia sostenida por el agente, debería pedirse meramente que la evidencia no apoye la creencia sostenida de modo irracional por el agente. Ahora bien, una vez que el requisito de evidencia fuese relajado de ese modo, es posible que la distinción entre autoengaño y pensamiento

desiderativo se viera afectada.[14] Sin duda esta alternativa perfectamente posible, y coincido en que tal modificación impactaría de modo directo en la distinción entre pensamiento desiderativo y autoengaño. Sin embargo, creo que la afirmación de que el agente parecería estar autoengañándose merece un análisis más detallado. Es indudable que la adopción, basada en deseos o emociones, de una creencia que no es confirmada ni disconfirmada por la evidencia global que se posee constituiría un caso de irracionalidad motivada. No obstante, entiendo que tal caso podría considerarse, con toda justicia, como un caso muy cercano al pensamiento desiderativo. Como hemos visto, las caracterizaciones típicas del pensamiento desiderativo ponen énfasis en la ausencia de evidencia favorable o desfavorable a la creencia que se adopta sobre la base de un deseo. El caso propuesto diferiría en que sí existiría evidencia pertinente para la evaluación de la creencia que finalmente se adopta, pero esta evidencia no apoyaría en mayor medida a esta creencia que a su negación. La equivalencia de razones, entiendo, sería el rasgo central en común entre ambos casos, que podrían ser considerados variantes del pensamiento desiderativo. Si esta interpretación alternativa fuese plausible, la formulación del requisito de evidencia que aquí propondremos no se vería afectara por la existencia de tales casos.

[14] Debo esta consideración al árbitro de *Manuscrito*.

IV

Parece haber razones bastante buenas para pensar, en resumen, que hay diferencias importantes entre los ejemplos de irracionalidad motivada debido a procesos sesgados de selección de evidencia y casos como los de autoengaño extremo como para admitir la conveniencia de no incluirlas en una misma categoría de fenómenos. Creo que nada de lo expuesto constituye un argumento concluyente en contra de la negativa a considerar necesario el requisito de evidencia; sin embargo, entiendo que lo anterior indica que la clasificación resultante es peor de lo que podría ser si se aceptara alguna restricción subordinada al empleo de ese criterio. Si se admite que es importante mantener, por las razones expuestas, la necesidad del requisito de evidencia y, sobre la base de éste, una posible distinción entre el autoengaño y otras formas de irracionalidad motivada, resulta necesario determinar, o al menos esbozar, qué formulación de esta condición puede cumplir esta función de manera satisfactoria.

En primer lugar, y contra quienes sostienen que la condición 4 debe ser fortalecida, entiendo que no puede exigirse que el sujeto interprete correctamente que la evidencia que posee no sustenta la creencia que él preferiría adoptar. Aun cuando se considere que una interpretación correcta de la evidencia no debe estar atada a un modelo intencionalista, parece una restricción excesiva el considerar como autoengaño sólo a aquellos casos en los cuales el agente está en posesión de la evidencia contraria a su creencia y evalúa correctamente la dirección en la cual apunta (suponiendo que tales casos fueran posibles). Entiendo que el requisito sólo debe exigir, en consecuencia, que el agente esté en posesión de la evidencia contraria a la creencia que finalmente adopta. Sin embargo, a la posesión de la evidencia debe adicionarse, a mi modo de ver, el requerimiento que el agente evalúe correctamente que la

información con la que cuenta constituye evidencia pertinente para la evaluación de su creencia. Consideremos al respecto el siguiente ejemplo. W es un médico competente y bien formado, que ha estado experimentando desde hace bastante tiempo síntomas cada vez más claramente compatibles con la enfermedad de Huntington, que incluyen alteraciones en los movimientos (movimientos faciales anómalos, marcha inestable, movimientos espasmódicos rápidos y súbitos en distintas partes del cuerpo), deterioro del habla, alteraciones en la conducta social y en el estado anímico (comportamientos antisociales, irritabilidad, malhumor), alucinaciones, desorientación y pérdida de la memoria y la capacidad de discernimiento. Por otra parte, posee información proveniente de sus familiares que sostiene que su padre, con quien ha perdido contacto muchos años atrás, sufrió la enfermedad, lo cual, en caso de ser cierto, implicaría que la probabilidad de que él la padeciera fuese ciertamente elevada. W no obstante, niega vehementemente ante sus íntimos la posibilidad de padecer ese mal. Desestima la información proveniente de sus familiares y conjetura explicaciones que, si bien no completamente inverosímiles, son claramente menos plausibles que la alternativa más indeseable, como reacciones atípicas al estrés generado por su trabajo; asimismo, rechaza de modo tajante la posibilidad de consultar a un colega especializado en neurología, e intenta tenazmente que su comportamiento cotidiano se vea alterado lo menos posible. W *no niega* que los síntomas que ha estado experimentando constituyen evidencia pertinente para la evaluación de su creencia: niega que constituyan pruebas que apunten en una dirección nítidamente definida, en este caso, hacia la atribución de una enfermedad neurológica. No parece existir ninguna razón de peso para negar que casos como el descripto constituyan ejemplos de autoengaño; más aun, podrían ser considerados, dado el

tratamiento claramente sesgado y motivado de los datos disponibles, como casos paradigmáticos de ese fenómeno.

Ahora bien, es posible que un agente esté en posesión de cierto tipo de información que constituye evidencia favorable o contraria a una creencia, pero no ser capaz de interpretar no sólo la dirección de la evidencia, sino que tal información constituye evidencia pertinente para la evaluación de una creencia. Consideremos la siguiente alternativa al caso precedente. W' ha estado experimentando desde hace tiempo síntomas notoriamente similares a los descriptos en el caso de W y, al igual que éste, posee información relativa a que su padre ha sufrido la enfermedad de Huntington. Sin embargo, W' carece de formación médica, ignora por completo los signos y síntomas característicos de ese mal y desconoce que su grado de heredabilidad es elevado. ¿Podría afirmarse que W' se encuentra autoengañado respecto de la posibilidad de sufrir la enfermedad de Huntington? Entiendo que la respuesta es claramente negativa: W', a diferencia de W, no malinterpreta, minimiza o busca explicaciones poco plausibles para la información que posee; simplemente es incapaz de interpretarla de un modo tal que la vincule de un modo significativo con la proposición relativa a padecer esa enfermedad. Tampoco es el caso, como debe resultar visible, que W' *carezca* de información adecuada para formarse alguna creencia relativa a su estado de salud. Por supuesto, lo anterior no impide que W' interprete de modo sesgado y poco realista los síntomas que experimenta y niegue de modo autoengañoso la posibilidad de padecer *alguna* enfermedad.

Sobre la base de lo expuesto, propongo la siguiente formulación del requisito de evidencia para el autoengaño: a) el agente debe estar en posesión del cuerpo de datos pertinente para la evaluación de su creencia; b) este cuerpo de datos provee mayor justificación para $\neg p$ que para p; y c) el agente debe ser capaz de interpretar estos datos como

pertinentes para la evaluación de su creencia. Mientras que la condición b) no parece requerir de mayor interpretación o justificación (si no se la incluyera, mal podría hablarse de autoengaño), si parece conveniente alguna especificación sobre las restantes. En primer lugar, no debe suponerse que la condición a) requiere que el agente esté en posesión de todos los datos pertinentes para la evaluación de su creencia (exigencia que tornaría prácticamente imposible cualquier atribución de autoengaño); debe entenderse en el sentido más moderado según el cual el agente debe poseer un cuerpo de datos pertinentes a su creencia que determinen que es más racional rechazarla que aceptarla. La condición c), como adelantamos, no implica que el agente deba ser capaz de interpretar correctamente que los datos que posee son contrarios a la creencia que favorece, sino que sólo requiere que comprenda inicialmente que tales datos son pertinentes para evaluarla racionalmente; los mecanismos cognitivos y motivacionales o emocionales implicados en el autoengaño derivarán en una interpretación de estos datos que favorecerá el rechazo de la creencia mejor sustentada por esa evidencia y la aceptación de una creencia incompatible con ella. a), b) y c), conjuntamente, constituyen un requisito necesario para una caracterización del autoengaño que aspire a cierta plausibilidad inicial.

Cabe señalar, por último, que la condición c) aquí propuesta podría ser considerada redundante.[15] Para algunos epistemólogos de peso, la evidencia es tal en tanto pueda jugar un rol en la justificación epistémica. De este modo, si un sujeto S no tiene ningún tipo de conciencia o aprehensión de la conexión entre un fragmento de información i y una proposición *p*, i no puede ser considerada como evidencia que S posee para *p*. Si esta posición fuese correcta, concluye, la condición a), que requiere la posesión de evidencia, ya implicaría la condición

[15] Debo esta observación el árbitro de *Manuscrito*.

c). Creo que una respuesta posible a esta observación se basa en el perjuicio potencial de su mantenimiento o eliminación para la propuesta aquí defendida. Si la posición defendida por tales epistemólogos fuese correcta (cuestión sobre la cual no estamos en condiciones de pronunciarnos aquí) entonces la condición c) podría ser eliminada sin pérdida. Sin embargo, si la posición arriba descripta resultara incorrecta, la eliminación de la condición c) redundaría en que la formulación del requisito de evidencia aquí sugerida resultaría irremediablemente incompleta y, en tanto tal, inadecuada. Entiendo, en conclusión, que consideraciones de prudencia y racionalidad recomiendan mantener la condición c) de modo explícito.

V

Si el examen de la condición de evidencia para la adquisición autoengañosa de creencias ha generado, como vimos, un debate de cierta relevancia, el análisis de la misma condición para el mantenimiento autoengañoso de creencias ha sido virtualmente inexistente. Sin embargo, esta omisión no implica que el análisis de este problema sea irrelevante. Veremos enseguida que hay buenas razones para pensar que el mantenimiento (autoengañoso o no) de creencias implica la actuación de procesos cognitivos distintos de los que intervienen en su adquisición. Como anticipamos en la sección introductoria, en esta última parte intentaremos determinar si la condición de evidencia presentada en la sección precedente es válida también para el mantenimiento autoengañoso de creencias. La condición previamente formulada es adecuada, a nuestro modo de ver, como requisito para la adquisición autoengañosa de creencias falsas, pero no necesariamente para su mantenimiento. Esto se debe a que parece plausible suponer que, una vez que el agente ha adquirido de modo

autoengañoso una creencia falsa (o al menos no sustentada por la evidencia), el proceso continuará con la activación de "filtros" de distintas clases (perceptivos, mnémicos o cognitivos) que limiten la entrada de evidencia contraria a la creencia adquirida. La anterior suposición resulta apoyada por algunos teóricos del autoengaño. Van Leeuwen (2008), observa que uno de los rasgos de la cognición humana normal que puede contribuir con el autoengaño es la propiedad denominada "inercia de la red de creencias". Como sabemos, las creencias típicamente se encuentran en relación con una red de creencias en la cual se insertan. Este conjunto o red manifiesta inercia, esto es, no cambia globalmente con facilidad debido a la existencia de hechos que son anómalos desde la perspectiva de una red en particular. Tal rasgo es ampliamente ventajoso para la coherencia de nuestro sistema de creencias, ya que sin él experimentaría una revolución con cada descubrimiento de hechos anómalos y se encontraría en un estado de flujo cognitivo permanente. La existencia de esta inercia, considera Van Leeuwen, explica en cierta medida por qué el sesgo de confirmación (que consiste en la tendencia a buscar casos confirmatorios de aquello en lo que ya creemos, y no casos contrarios) puede ser una ventaja, ya que nos protege contra la revolución cognitiva constante. Sin embargo, agrega, la inercia de la red también puede facilitar el autoengaño. En particular, puede tornar más fácil mantener creencias bajo la influencia de un deseo, aun cuando la evidencia disponible resulte convincente en favor de la creencia opuesta.

Si bien lo expuesto en el párrafo precedente podría hacer pensar que el mantenimiento autoengañoso de una creencia depende simplemente de mecanismos cognitivos autónomos, esto es, el agente simplemente mantiene la creencia en cuestión sin mayor esfuerzo cognitivo, es posible pensar en situaciones de autoengaño que no se ajustan a una descripción tan simple. Podemos distinguir

como mínimo dos tipos de mantenimiento autoengañoso de una creencia: a) aquellos casos en los cuales se adquiere de modo autoengañoso una creencia falsa, que se mantiene como tal; b) aquellos casos en los cuales se adquiere una creencia verdadera, que se convierte en falsa debido a cambios en el mundo externo, pero pasa a ser mantenida de modo autoengañoso. Examinaremos brevemente ambas posibilidades.

El tipo a), pese a su aparente simplicidad, admite al menos dos variantes, a_1) y a_2). a_1) puede ser descripta del siguiente modo. Un agente ha adquirido de modo autoengañoso la creencia falsa de que *p*; no se han producido modificaciones en la evidencia que posee, de modo que no tiene motivos o razones para buscar o idear justificaciones adicionales para la creencia que privilegia. En este caso, propiedades de nuestro sistema cognitivo, como la inercia de la red de creencias, facilitarán el mantenimiento de la creencia adoptada de manera autoengañosa. a_2) La segunda puede ser ilustrada mediante una extensión del caso de W descripto en la sección previa. Como dijimos, W posee la creencia de que no sufre de la enfermedad de Huntington, y ha mantenido una cerrada defensa de esa creencia, pese a la abundante evidencia en contrario de que dispone. Ahora bien, de modo inesperado surge nueva evidencia que indica de modo mucho más convincente la falsedad de su creencia: la historia clínica de su padre, encontrada por azar en el archivo del hospital en el que W trabaja, muestra de modo inequívoco que sufrió la enfermedad de Huntington; esta información, sumada a los síntomas que experimenta, tornan mucho menos plausible aún su creencia de que no padece la enfermedad. Pese a esta nueva evidencia W se mantiene incólume en su creencia de que él no sufre ese mal: cuestiona la confiabilidad de los registros del hospital, alega que es posible que se haya producido una confusión de identidad, desconfía de los procedimientos diagnósticos empleados en

el pasado, etc. Sin embargo, para cualquier observador externo su creencia se ha tornado absolutamente insostenible sobre la base de la evidencia de que dispone.

El tipo b) puede ser ilustrado mediante el siguiente caso. X está convencido de la fidelidad de su cónyuge, quien ha sido durante muchos años una esposa devota y leal; X jamás ha tenido motivos para desconfiar de ella. Sin embargo, desde hace algunos meses la esposa de X muestra comportamientos que, para un observador competente e imparcial[16] constituirían evidencia inequívoca de un cambio importante respecto de su historia previa. Llega mucho más

[16] Mele (2007) ha sugerido un criterio, al que denomina "test del observador imparcial", para evaluar el nivel de sesgo motivacional o emocional adecuado para que una persona adquiera de modo autoengañoso una creencia. Dado el hecho de que S adquiere la creencia de que *p* y D es el conjunto de datos relevantes fácilmente disponibles para S durante el proceso de adquisición de la creencia, si D estuviera igualmente disponible para los pares cognitivos imparciales de S, y estos pares reflexionaran sobre tal evidencia tanto como S y (como mínimo) luego de una reflexión moderada, el número de los que concluirían que *p* es falsa superaría significativamente el número de los que considerarían que es verdadera. Por "pares cognitivos" Mele entiende personas que son muy similares en educación e inteligencia a la persona que es testeada; los pares cognitivos que comparten ciertos deseos pertinentes con el sujeto que es testeado podrían a menudo adquirir la misma conclusión injustificada que el sujeto testeado, dados los mismos datos. Sin embargo, agrega Mele, para los propósitos del análisis que está llevando a cabo los pares cognitivos son observadores imparciales. Un requisito mínimo de imparcialidad en este contexto es que ninguno de los pares comparta el deseo de que *p* ni posea tampoco el deseo de que ¬*p*; otro requisito plausible consiste en que ninguno prefiera evitar uno de los siguientes errores por sobre el otro: creer falsamente que *p* o creer falsamente que ¬*p*; por último, un tercer requisito consiste en que no se posea un interés emocional en la verdad o en la falsedad de *p*.

tarde de su trabajo sin explicaciones convincentes; dedica mucho tiempo a revisar su teléfono móvil, que además oculta a la vista de su esposo; cuida su aspecto físico de un modo en que no lo ha hecho con anterioridad; viaja frecuentemente alegando razones de trabajo, cuando no hay obligación aparente. X, interrogado por un amigo acerca de esas conductas, acepta que resulta razonable tenerlas en cuenta para la evaluación de la fidelidad (o infidelidad) de su esposa. Sin embargo, niega obstinadamente que sean evidencia de infidelidad; por el contrario, idea explicaciones rebuscadas y extravagantes incompatibles con la infidelidad de su cónyuge. Al igual que en el caso de W descripto en la sección anterior, X *no niega* que las conductas que ha observado constituyen evidencia pertinente para la evaluación de su creencia: niega que constituyan pruebas que apunten en una dirección definida, en este caso, hacia la atribución de infidelidad.

Ahora bien, los casos a_1), a_2) y b) difieren en un aspecto fundamental. Mientras que en a_1 no surge evidencia adicional que eventualmente fuerce al agente a estrategias tendientes a mantener el estado de autoengaño, tanto en a_2) como en b) tal situación es lo que de hecho acaece. ¿Sería posible que, tanto en a_2) como en b) el agente simplemente omitiera la consideración de la evidencia contraria a su creencia debido a la acción de mecanismos sesgados de selección de evidencia? No es posible desestimar esta posibilidad; no obstante, esto no cambiaría el estatus de autoengañados de W y X: simplemente diríamos que se ha añadido un mecanismo adicional de mantenimiento de la creencia. Sin embargo, tanto W como X no sólo están en posesión sino que comprenden perfectamente que la evidencia nueva de que disponen es pertinente para la evaluación de su creencia y para determinar si es racional seguir sosteniéndola (en el caso de W) o sustituirla por una incompatible con ella (en el caso de X).

Los casos anteriores (excepto en a_1, en el que no surge nueva evidencia que fuerce al empleo de estrategias adicionales de mantenimiento del autoengaño), por lo tanto, se asemejan en que los agentes evalúan correctamente que la evidencia nueva de que disponen es pertinente para la evaluación de su creencia; sin embargo, la interpretación de tal evidencia no los conduce al cuestionamiento de las creencias previas, sino a su mantenimiento. Parece plausible concluir, en consecuencia, que la condición de evidencia propuesta en la sección previa es válida tanto para la adquisición autoengañosa de creencias como para su mantenimiento.

Podría existir, no obstante, una controversia acerca de qué significa poseer evidencia, sobre la que es necesario fijar posición.[17] Existen diferentes posiciones sobre esta cuestión, que van desde quienes consideran que la evidencia poseída abarca las creencias olvidadas y difícilmente recordables hasta aquellos que admiten incluso las creencias ocurrentes, de las que se tiene conciencia en el momento, pasando por aquellos que incluyen sólo creencias de fácil acceso. El siguiente caso resulta útil para plantear la cuestión: un agente adquiere de forma autoengañosa una creencia de que *p*, encontrándose en posesión de la evidencia E. Con el paso del tiempo va olvidando progresivamente dicha evidencia. Es posible distinguir aquí dos posibilidades: la primera, en la que olvida por completo toda la evidencia, de modo que resultaría muy difícil o imposible recordarla; la segunda, en la cual olvida la evidencia pero podría recordarla con cierta facilidad si reflexionara sobre la cuestión. Estas dos posibilidades suscitan la pregunta relativa a en cuál de los dos casos es autoengañoso el mantenimiento de la creencia y esto, a su vez, implica tomar posición respecto de los requisitos para poseer evidencia. Aquellos que consideren que el agente

[17] Debo esta observación al árbitro de *Manuscrito*.

solo posee la evidencia fácilmente accesible mediante la reflexión dirán en principio que la primera posibilidad no constituye un caso de mantenimiento autoengañoso, dado que no se satisface el requisito de evidencia, esto es, el agente no posee evidencia contraria a su creencia. Por el contrario, quienes admitan la evidencia completamente olvidada como parte de la evidencia que posee el agente considerarán que la primera posibilidad sí constituye un caso de mantenimiento autoengañoso. Consideraciones similares pueden hacerse, *mutatis mutandis*, acerca de la segunda posibilidad. Excede sin duda las posibilidades de este trabajo fijar posición sobre la cuestión general relativa a qué implica poseer evidencia. Sin embargo, creo que el mantenimiento autoengañoso de una creencia requiere, como mínimo, que el agente pueda recuperar con cierta facilidad la evidencia que ya posee contraria a la creencia que mantiene de modo irracional. El caso en el cual el agente olvida por completo la evidencia, de modo que su recuperación resultaría muy difícil o imposible, no parece constituir un caso de mantenimiento autoengañoso a mi modo de ver. El agente, en tal caso, sólo estaría en posesión de la evidencia favorable a su creencia, pero no de aquella evidencia en contra. Tal vez se podría plantear, no obstante, la siguiente objeción a lo anterior: podría tratarse de un caso de mantenimiento autoengañoso en caso de que el agente, habiendo olvidado toda la evidencia contraria a su creencia, *no poseyera* evidencia favorable a esta. Creo, no obstante, y en línea con lo señalado en la nota 8, que este caso constituiría un ejemplo de pensamiento desiderativo, pero no de autoengaño. Se trataría, a mi modo de ver, de un caso distinto de equivalencia de razones al sugerido en esa nota, pero que se asemejaría más al pensamiento desiderativo que al autoengaño.

La condición de evidencia sugerida permite, a mi modo de ver, distinguir de manera más precisa el autoengaño de otras formas de irracionalidad motivada. Si bien es

innegable que pueden existir casos dudosos y no siempre puede resultar segura una atribución de autoengaño, el rechazo al carácter necesario de la condición, a la posesión de la evidencia por parte del agente y al reconocimiento de éste como tal redunda, entiendo, en una indiferenciación de este fenómeno de otros emparentados con él que, tanto por razones teóricas como prácticas, conviene distinguir.

BIBLIOGRAFÍA

AUDI, R. "Self-Deception, Action, and Will". *Erkenntnis* 18, pp. 133–58, 1982.

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

BACH, K. "An Analysis of Self-Deception", *Philosophy and Phenomenological Research*, *41*, 3, pp. 351-370, 1981.

BARNES, A. *Seeing through self-deception*, New York: Cambridge University Press, 1997.

BARON, M. "What is Wrong with Self-Deception?", en Brian McLaughlin & Amelie Oksenberg-Rorty (eds.), 1988.

BERMÚDEZ, J. L. "Self-deception, intentions and contradictory beliefs", *Analysis 60*, 4, pp. 309-19, 2000.

BORGE, S. "The Myth of Self-Deception", *The Southern Journal of Philosophy*. *41*, 1, pp. 1–28, 2003.

BUTLER, J. (1726): Sermon X. Upon Self-Deceit. Extraído el 24/07/2010 de http://anglicanhistory.org/butler/rolls/10.html

CORREIA, V. "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

DAVIDSON, D. "Engaño y división", en D. Davidson, *Mente, mundo y acción*. Barcelona, Paidós, 1985.

DEMOS, R. "Lying to Oneself", *Journal of Philosophy*, *57*, pp. 588–95, 1960.

DEWEESE-BOYD, I. "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

_____ "Self-Deception". *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.). Extraído el 22/2/2017 de https://plato.stanford.edu/archives/win2016/entries/self-deception

FINGARETTE, H. *Self-deception*, London: Routledge & Kegan Paul, 1969.

HIRSTEIN, W. *Brain Fiction. Self-Deception and the Riddle of Confabulation*, Cambridge: The MIT Press, 2005.

HOLTON, R. "What is the Role of the Self in Self-Deception?", *Proceedings of the Aristotelian Society*, 101, pp. 53-69, 2000.

LAZAR, A. "Deceiving Oneself or Self-Deceived? On the Formation of Beliefs 'Under the Influence'", *Mind 108*, 430, pp. 265-290, 1999.

LEVY, N. "Self -Deception and Moral Responsibility", *Ratio (new series)*, *XVII*, pp. 294-311, 2004.

LYNCH, K. "Self-Deception, Religious Belief, and the False Belief Condition", *The Heythrop Journal*, *LI*, pp. 1073–1074, 2010.

MELE, A. *Irrationality. An Essay on Akrasia, Self-Deception, and Self-Control*, New York-Oxford: Oxford University Press, 2010.

_____ *Self-deception Unmasked*, Princeton: Princeton University Press, 2001.

_____ "Self-Deception and Three Psychiatric Delusions", en Mark Timmons, John Greco & Alfred R. Mele (eds.), *Rationality and the Good*, Oxford, Oxford University Press, 2007.

_____ "When Are We Self-Deceived?", en P. Pedrini (ed.), *Philosophy of Self-deception. HumanaMente*, 20, pp. 1-15, 2012.

METCALFE, J. "Cognitive Optimism: Self-Deception or Memory-Based Processing Heuristics?", *Personality and Social Psychology Review; 2*, pp. 100-110, 1998.

OKSENBERG-RORTY, A. "The Deceptive Self: Liars, Layers and Lairs", en Brian McLaughlin & Amelie Oksenberg-Rorty (eds.), *Perspectives on Self-Deception*. Berkeley, University of California Press, 1988.

PEARS, D. *Motivated Irrationality*, New York: Oxford University Press, 1984.

REY, G. "Toward a computational account of Akrasia and self-deception", en A. O. Rorty & B. P. McLaughlin (eds.), 1988.

RUDDICK, W. "Social Self-Deceptions", en Brian McLaughlin & Amelie Oksenberg-Rorty (eds.) *Perspectives on Self-Deception*, 1988.

SARTRE, J.-P. *El Ser y la Nada*, Buenos Aires: Losada, 1943.

SZABADOS, B. "Wishful Thinking and Self-Deception", *Analysis*, 33, 6, pp. 201-205, 1973.

TAYLOR, S. & J. BROWN "Illusion and Well-Being: A Social Psychological Perspective on Mental Health", *Psychological Bulletin 103*, 2, pp. 193-210, 1988.

TRIVERS, R. "Deceit and Self-Deception", en Peter M. Kappeler & Joan B. Silk (eds.), *Mind the Gap. Tracing the Origins of Human Universals*. Springer, pp. 373-393, 2010.

_____ *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*, New York: Basic Books, 2011.

VAN LEEUWEN, D. S. N. "The Spandrels of Self-Deception: Prospects for a Biological Theory of a

Mental Phenomenon", *Philosophical Psychology 20*, 3, pp. 329–348, 2007.

_____ "Finite rational self-deceivers", *Philosophical Studies 139*, pp. 191–208, 2008.

VON HIPPEL, W. & R. TRIVERS "The evolution and psychology of self-deception", *Behavioral and Brain Sciences 34*, pp. 1–56, 2011.