teorema
Vol. XXVI/3, 2007, pp. 5-17
ISSN: 0210-1602

Introducción*

Víctor M. Santamaría Navarro

> Siempre que haya de dibujarse el retrato de un hombre, destacándose aquello que en él es más humano, sea noble o innoble, seguramente deberíamos colocar muy en primer plano la enorme capacidad humana para el autoengaño.
>
> HERBERT FINGARETTE

Un hombre que ha confiado durante años en su esposa comienza a tener motivos para sospechar de ella. La mujer llega a casa más tarde de lo usual después del trabajo, dos o tres veces por semana sale por la noche y se niega a dar ninguna explicación, se preocupa más por su imagen e incluso se muestra más distante con su familia. Además, un buen amigo le hace al marido la confidencia de que la ha visto en cierto local en compañía de otro individuo. Sin embargo, el marido continúa creyendo que su mujer le es fiel. Casos como éste se esgrimen para apoyar la existencia de comportamientos, aparentemente irracionales, en los que los sujetos mantienen ciertas creencias en contra de la mayor parte de la evidencia de la que disponen.

Durante los últimos 50 años, el autoengaño ha planteado a filósofos y psicólogos problemas de difícil solución y ha dado lugar a un gran número de libros y artículos que tratan de dar cuenta de un fenómeno aparentemente tan común como paradójico. Pero, ¿qué queremos decir cuando afirmamos que alguien se autoengaña? ¿Es posible que alguien lo haga de modo voluntario? ¿Es necesario que sea consciente del proceso? ¿Hemos de exigir que el sujeto mantenga creencias contradictorias? ¿Es esto aceptable desde el punto de vista epistémico o moral? Y en todo caso, ¿es siquiera posible?

En primer lugar, hemos de distinguir el autoengaño de otros fenómenos similares como la debilidad de la voluntad, la ceguera intelectual, el pensamiento desiderativo o la disonancia cognitiva. Mientras que la *debilidad de la voluntad* consiste en realizar un curso de acción libre, consciente y voluntariamente, aun cuando el sujeto crea que dispone de otro mejor, el *pensamiento desiderativo* estriba en la formación de creencias con arreglo a un deseo en ausencia de evidencia significativa. Por su parte, la *ceguera intelectual* nos

presenta a un individuo cuyo razonamiento se ha nublado por la fuerza de un deseo o emoción, y la *disonancia cognitiva* consiste en la tensión psicológica debida a una incongruencia dentro del sistema de creencias, ideas y valores de un sujeto, que éste tiende a reducir mediante diversas estrategias, entre las que se encuentra el autoengaño.

El mayor obstáculo que surge al enfrentarse a los problemas planteados por el autoengaño ha sido generado por los intentos de interpretarlo como un caso especial de engaño, esto es, bajo el modelo del "engaño a otro". Tales intentos dan lugar a dos paradojas, conocidas como la paradoja estática o doxástica y la dinámica o de la estrategia. La primera consiste en el hecho de que el mismo sujeto habría de mantener a la vez dos creencias cuyo contenido proposicional es contradictorio. Como engañador, el sujeto en cuestión tiene una creencia verdadera que se habría de ocultar a sí mismo para mantener, como engañado, una creencia falsa (y contradictoria con la anterior). Hay varias fórmulas que expresan esto, unas más fuertes que otras, dependiendo de que el sujeto crea que *p* y *no* crea que *p*, crea que *p* y *no-p* o crea que *p* y crea que *no-p*. La primera de ellas es una contradicción lógica, y por ello ha de rechazarse; la segunda supone que el sujeto cree una contradicción manifiesta, lo cual hace también excesivamente problemática esta interpretación. Sin embargo, la tercera, aunque paradójica, es quizás aceptable bajo el supuesto de que el sujeto mantiene ambas creencias sin ser consciente de ello. Esta es la estrategia que siguen autores como Donald Davidson (1982; 1985), David Pears (1982; 1991), Amélie O. Rorty (1972; 1983) y un buen número de autores partidarios de la hipótesis de la "división" de la mente. La segunda paradoja, la dinámica, se produce por el hecho de que el sujeto habría de diseñar *conscientemente* una estrategia y llevarla a cabo en tanto que engañador, siendo *inconsciente* a la vez de todos sus detalles y ejecución en tanto que engañado. De nuevo, la división de la mente ha sido esgrimida como hipótesis heurística por los autores anteriormente citados con el objeto de dar una explicación racional del fenómeno, ya que tal "división" mantendría separadas por un lado la evidencia dolorosa junto a la creencia evidencial —basada en esa evidencia— de que *p* es el caso, y por otro el deseo de que *no-p* sea el caso junto a la creencia contraevidencial.

En general, estos enfoques que interpretan el autoengaño sobre el modelo del engaño a otro son conocidos como "intencionalistas", pues requieren que el sujeto tenga la intención de alcanzar una creencia contraevidencial. De estos enfoques, sólo algunos de ellos postulan una partición de la mente o una teoría de sistemas con subagentes. Kent Bach (1981), William J. Talbott (1995) o José Luis Bermúdez (2000) son ejemplos de autores intencionalistas que buscan explicaciones alternativas a la teoría de sistemas sin renunciar al carácter intencional del autoengaño.

Sin embargo en los últimos años han ganado peso las explicaciones "no intencionalistas" que se alejan del modelo de engaño interpersonal, y descri-

ben el fenómeno como un proceso motivacional, evitando así las paradojas ya mencionadas. El más influyente de estos enfoques es el que ofrece Alfred Mele (1997; 2001; 2003), para quien la intención de adquirir una creencia contraevidencial por parte del sujeto, aunque puede ser el motor del autoengaño en algunos casos atípicos, no es condición necesaria en la mayoría de los casos habituales. De hecho, Mele establece las siguientes condiciones suficientes: 1) la creencia que el sujeto adquiere es falsa, 2) el sujeto sesga la evidencia relevante o aparentemente relevante para la verdad de *p*, 3) el sesgo es una causa no desviada en la adquisición de *p* por parte del sujeto, y 4) el cuerpo de datos que éste posee ofrece mayor apoyo para la verdad de *no-p* que para la verdad de *p*. Las condiciones de Mele *no* incluyen voluntariedad alguna en el proceso, ni que el sujeto mantenga a la vez dos creencias contradictorias; por esta razón se la considera como una interpretación *deflacionaria* del autoengaño. Mele da cuenta de la adulteración de la evidencia acudiendo a la teoría Friedrich-Trope-Liberman (FTL), resultante de la combinación de la teoría de Friedrich con la tesis de Trope-Liberman. La teoría de Friedrich [*Primary error detection and minimization* (PEDMIN)] afirma que al evaluar hipótesis el sujeto está mucho más ocupado en tratar de minimizar la cantidad de errores que en buscar la verdad, mientras la tesis de Trope-Liberman afirma que al evaluar la evidencia establecemos un umbral de confianza que indica la cantidad de evidencia que el sujeto necesita para aceptar o rechazar una creencia. Por supuesto, el umbral de la creencia que se adapta a sus deseos es mucho más bajo que el que es necesario superar para forjar la creencia que nos resulta dolorosa. El sujeto sesgaría la evidencia de diversos modos, entre los que se encuentran la mala interpretación negativa, la mala interpretación positiva, la focalización/atención selectiva y el acopio selectivo de pruebas.

Otras explicaciones conceptuales en términos no intencionales en las que no podemos entrar aquí son las de Robert Audi (1982), o la tesis de la reducción de ansiedad de Annette Barnes (1990) y Mark Johnston (1988). Merece la pena sin embargo señalar que la teoría de la represión de Sigmund Freud (1915), así como su modelo de lo consciente/inconsciente de la primera tópica (1900) y del Ello/Yo/Súper-Yo de la segunda (1923), ha sido recogida en mayor o menor medida por muchos autores tanto de la perspectiva intencionalista (heredera suya es la teoría particionista) como de la no-intencionalista. Efectivamente, algunos han visto en los mecanismos de adulteración de la evidencia procesos preconscientes o incluso inconscientes de represión por parte del sujeto; tal es el caso de Edward Erwin (1988) o Herbert Fingarette (1969; 1998). No hará falta recordar que la posibilidad de una explicación teórica del engaño a uno mismo recurriendo a la división de la psique que el psicoanálisis freudiano postulaba, fue objeto de dura crítica por parte de Jean Paul Sartre (1943) en su análisis de la *mala fe*.

Desde la biología social algunos autores como Robert Trivers advierten del peligro de caer en la tentación de volver a la psicología de hace un siglo, e interpretar el autoengaño como un mecanismo de defensa de mi frágil ego o una protección de mi débil psique esgrimiendo hipótesis psicoanalíticas o algunas otras de carácter psedocientífico, pues eso nos impedirá ver la característica agresiva del autoengaño. En opinión de Trivers resulta más acertado dar cuenta del autoengaño desde un punto de vista científico aludiendo a su presunta funcionalidad y carácter adaptativo, pues un sujeto que es capaz de autoengañarse será más hábil en la tarea de ocultar información y engañar a los demás, lo cual sería ventajoso para su supervivencia [Trivers (1985; 2000)].

No han faltado, por supuesto, quienes han manifestado una postura escéptica o eliminativista respecto al fenómeno, explicándolo en términos de otras actitudes más o menos cínicas, o fenómenos menos paradójicos, como R.M. Haight (1980) y David Kipp (1981), o incluso categorizándolo como una mera construcción social, como hace Kenneth J. Gergen (1985).

Las contribuciones que recoge este número monográfico de **teorema** analizan el fenómeno del autoengaño desde distintos puntos de vista. A continuación se hace una breve sinopsis de las mismas.

M. ROSARIO HERNÁNDEZ BORGES hace una exposición de las dos principales líneas de explicación del fenómeno del autoengaño, a saber, las intencionalistas y las no intencionalistas. Toma a Donald Davidson como paradigma del enfoque intencionalista —según el cual la intención del sujeto de engañarse ha de jugar un papel central y necesario— rechazando posteriormente esta postura por las paradojas a las que nos conduce. La propuesta de Mele, aunque sortea las paradojas estática y dinámica de las explicaciones intencionalistas, se encuentra con lo que Bermúdez denomina el "problema selectivo": hay muchas cosas desagradables que me inclinarían a sesgar la evidencia y autoengañarme, para las que además tengo un umbral bajo de rechazo y alto de aceptación, y sin embargo, hablamos de autoengaño sólo cuando me engaño con respecto a *algunas* de ellas. Pero, ¿por qué me engaño con respecto a unas y no a otras? Para Rosario Hernández, que abraza la explicación no intencionalista de Mele, esto plantea un serio problema que hay que solventar haciendo una nueva caracterización del modo en que la motivación ejerce su rol: hay que explicar por qué en unos casos aparece esa fuerza motivacional y no lo hace en otros. La razón habría de buscarse, según la autora, en las falsas creencias que uno tiene acerca de sí mismo, en cuya formación tienen un papel fundamental la emoción y el afecto.

Por su parte, VASCO CORREIA ataca también la postura intencionalista, ya que, además de convertir al sujeto en un agente acrático en el plano epistemológico, da lugar a cuatro nuevas paradojas además de las anteriormente señaladas. Su postura supone una defensa de las caracterizaciones no inten-

cionales y, más concretamente, de aquellas que dan cuenta del fenómeno del autoengaño como un estado irracional provocado por las interferencias que las emociones provocan en el proceso cognitivo de formación de creencias. Esto no obsta para que al sujeto que se autoengaña se le pueda exigir responsabilidad, puesto que siempre tiene cierto control sobre los procesos que desencadenan las emociones.

Frente a estas defensas de la teoría deflacionaria, ANNA NICHOLSON cree que, aunque es necesaria la exigencia de que la intención tenga un papel motor en el autogaño, tal circunstancia no ha de forzarnos a abandonar la estructura doxástica de la explicación de Mele. Nicholson, no obstante, realiza varias críticas. En primer lugar, Mele no diferencia convincentemente el autoengaño del pensamiento desiderativo; en segundo lugar, la idea de que el sujeto *intencionalmente* sesga la evidencia dado su deseo de que *p* sea al caso, no se diferencia funcionalmente de la postura según la cual el sujeto *intencionalmente* busca abrazar la falsa creencia de que *p* es el caso; pero además, Mele se ve enredado en otro problema: lo que Nicholson llama el "problema de la clasificación", según el cual, cuando sesgamos la evidencia, los estímulos que rechazamos son necesariamente procesados en algún nivel y, además, los mecanismos de control mental generan efectos no deseados, como, por ejemplo, los "procesos irónicos", en virtud de los cuales aquello que pretendemos rechazar queda fijado en nuestra mente con mayor fuerza. La propuesta de Nicholson consiste en mantener la estructura del proceso de formación de creencias que propone Mele y, al mismo tiempo, introducir en el modelo explicativo la intención del sujeto de autoengañarse, ya que ésta otorgaría el combustible necesario para dar cuenta de la fuerza motivacional que el deseo por sí solo no es capaz de explicar.

Según DON BERKICH, el comportamiento acrático de los sujetos —esto es, la realización de una acción de modo voluntario pese a que uno es consciente de que puede llevar a cabo otro curso de acción mejor— parece indicar tanto la existencia de autoengaño como, quizá, de irracionalidad. Generalmente se han excluido de la discusión sobre la *akrasia* las situaciones en las que el deseo que el sujeto tiene de que algo sea el caso le nubla el juicio y oculta la mejor elección (ceguera intelectual), así como aquellas en las que, aunque el sujeto es consciente de cuál ha de ser el curso de acción que ha de seguir, no es capaz de hacerlo. Quien actúa de este modo lo hace, según Davidson, guiado por el llamado *Principio de Medea.* Al haber conciencia pero falta de voluntad, este tipo de *akrasia* supone una pérdida de control que la asemeja en este aspecto a los casos de adicción. Berkich propone un enfoque global según el cual el término "*akrasia*" denote toda "falta de control" por parte del sujeto; en este sentido, la *akrasia* estricta sería aquella en la que el sujeto, conociendo cuál es el mejor curso de acción, deseando autocontrolarse y poseyendo la capaci-

dad de ejercer ese autocontrol, libre y voluntariamente no realiza aquello que su mejor juicio le dicta. Según Davidson, hay tres principios que nos parecen autoevidentes pero que en conjunción resultan aparentemente contradictorios, a saber, P1: Si un agente desea hacer *x* más de lo que desea hacer *y*, y puede hacer libre y voluntariamente *x* o *y*, entonces, hará intencionalmente *x*; P2: Si un agente valora más hacer *x* que hacer *y*, entonces desea hacer *x* más de lo desea que hacer *y*; y P3: Hay acciones acráticas o incontinentes.

Tras varios intentos de formalizar simbólicamente estos tres principios, Berkich no logra dar cuenta de la presunta inconsistencia. Según él, esto sucede porque hacemos afirmaciones sobre acciones en un contexto de creencia, para dar luego un salto a otro contexto de creencia diferente en el que hacemos otras afirmaciones sobre esas mismas acciones. En realidad, en los casos de *akrasia* estricta, hay un fallo en la identificación de las propias creencias. Esto apunta a que la autoridad de primera persona sobre nuestras propias creencias no es absoluta. Sin embargo, el autor se guarda de distanciarse de la postura de Ryle (1949), según la cual no tenemos acceso privilegiado a los contenidos de nuestra propia mente. Siguiendo a Davidson afirma que, aunque tenemos autoridad de primera persona sobre nuestras creencias, el error es posible y, por tanto, la auto-atribución de creencias no es incorregible.

THOMAS STURM se sitúa en una perspectiva, si no escéptica, sí prudente con respecto a la afirmación de la existencia del fenómeno del autoengaño. Ni da por supuesta su existencia, ni niega la posibilidad de que se trate de una mera atribución por parte de los observadores externos. El mayor problema es que mientras los estudios empíricos acerca del autoengaño realizados por psicólogos presuponen muy a menudo nociones distintas, e incluso incompatibles, del fenómeno, los enfoques conceptuales de los filósofos o bien carecen de base empírica o bien toman dichos estudios de los psicólogos sin reparar en que hay conceptos problemáticos que se aceptan sin ponerse en cuestión. Esto conduce tanto a caracterizaciones confusas de la racionalidad, como a modelos ontológicos enfrentados del yo —el homuncular, el eliminativista y el reduccionista—. Además, el autoengaño no supone necesariamente un engaño acerca del yo, esto es, sobre uno mismo —aunque sea lo más común—; en primer lugar, porque no se distinguiría entre autoengaño y simple ignorancia o error sobre uno mismo y, en segundo lugar, porque hay muchos casos en los que a primera vista atribuimos autoengaño, y sin embargo el contenido de la creencia errónea no versa acerca del individuo en tanto que objeto, sino de un amigo o de su esposa.

Según Sturm, los estudios empíricos de Ruben C. Gur y Harold A. Sackheim quizá constituyan la primera tentativa de estudio empírico del autoengaño tomando como punto de partida un enfoque conceptual (en este caso concreto, adoptan el ofrecido por Raphael Demos). Sin embargo, estos experimentos tienen que hacer frente a varias críticas acerca de la fiabilidad de sus

resultados y además exigen la intervención de la voluntad del sujeto, cuando parece claro que hay casos de creencias adulteradas motivacional pero no intencionalmente.

El mayor esfuerzo por incorporar al estudio conceptual los resultados de las investigaciones de la psicología está representado —a juicio de Sturm— por Alfred Mele, a cuyo enfoque plantea tres objeciones: que hay casos de autoengaño que carecen de motivación, que hay creencias adulteradas motivacionalmente que no suponen casos de autoengaño y que no existe una única teoría de la racionalidad desde la que explicar el problema. Su propuesta es que toda teoría debe enfrentarse al test de la realidad, por lo que se necesitan estudios empíricos más serios donde, además, los casos sometidos a valoración *no* puedan ser interpretados desde ninguna teoría alternativa como conductas racionales; pero estas investigaciones empíricas requerirían a su vez un estudio riguroso del concepto normativo de racionalidad.

Además de las posturas filosóficas clásicas de, entre otros muchos, Robert Audi, Amélie O. Rorty, Donald Davidson o Alfred Mele, una de las posibles explicaciones conceptuales del fenómeno del autoengaño se apoya en las ideas propuestas por Freud sobre la existencia, tanto de varios sistemas discretos dentro de la mente individual, como de un mecanismo que reprimiría aquellos elementos que pusieran en peligro la unidad o integridad de la psique. En esta línea, ANTONI GOMILA argumenta que, aunque la tradición anglosajona se ha preocupado especialmente por los problemas generados por el holismo en la atribución de creencias en el caso del autoengaño, las posturas divisionistas tipo Davidson constituirían el correlato filosófico de la teoría psicoanalítica. Parece que muchas de las posibles explicaciones conceptuales del fenómeno descansarían en la plausibilidad de la represión, término que ha sido muy criticado y manipulado en exceso. Ahora bien, según Gomila, los nuevos experimentos en psicología social sobre memoria autobiográfica, recuperación de situaciones traumáticas olvidadas, olvido intencional o dirigido y evitación activa, demuestran empíricamente la existencia de la represión, dando por tanto nueva legitimidad a la teoría freudiana y colocando en una situación insostenible a quienes niegan la propia posibilidad del fenómeno. Sin embargo, sería necesario rechazar los enfoques divisionistas y la metapsicología de subsistemas, pues ninguno de ellos puede evitar caer en la falacia homuncular, precisamente por la mala comprensión que tienen del proceso de la represión. Más bien habría que entender ésta como una interacción de los procesos inconscientes, automáticos y asociativos que tienen un papel motivacional o causal, con el nivel de intervención personal, es decir, la consciencia.

ALAN THOMAS se posiciona en defensa de los ataques que ha recibido la teoría de la represión de Freud por parte de Jean Paul Sartre. Efectivamente, la teoría freudiana de la división de la mente y la represión parece salvar

las paradojas que encierra el autoengaño, dando la oportunidad de comprender cómo el agente puede mantener fuera de la consciencia pensamientos que desempeñan un papel causal en su acción. No obstante, la crítica de Sartre a la teoría freudiana de la represión era ésta: en el sujeto ha de haber alguna instancia censora que, a la vez, sea consciente de los pensamientos que *ha* de reprimir y no sea consciente de ellos, *ex hipothesi*, en tanto que reprimidos. Parece, por tanto, que la teoría freudiana caería en la falacia homuncular, pues sólo ha reubicado el problema, aplicando propiedades de la mente a partes de la mente. Sin embargo, según Thomas, la confusión de Sartre reside precisamente en concebir el "censor" interno como una representación microcósmica del control de los pensamientos que el sujeto muestra como macrocosmos. Esto demostraría que es la exposición sartriana, y no la freudiana, la que es regresiva. El error de Sartre descansa en la creencia de que tener un estado mental consciente significa ser consciente de estar en ese estado mental, esto es, que el control de los estados mentales supone una metarrepresentación de los mismos. Pero es esta idea sartriana lo que convertiría el mero acto de pensar en un regreso infinito. De hecho, la represión consiste simplemente en el proceso de clasificar los pensamientos, y no requiere reflexividad. Ejercitamos nuestra racionalidad y tenemos estados mentales intencionales sin necesidad de representárnoslos a nosotros mismos o a otros.

Otro aspecto fundamental en el estudio del autoengaño supone analizar en qué condiciones atribuimos este estado mental a un sujeto. Los artículos de Charles Hermes y Fernando Martínez Manrique investigan este asunto. CHARLES HERMES examina las cláusulas del test cognitivo entre pares que Mele propuso en 2001 y que modificó posteriormente. Según este test, para poder atribuir autoengaño a un sujeto *S* que cree que *p* es el caso, la mayoría de sujetos de un grupo de control debería creer en la falsedad de *p* ante el mismo conjunto de evidencia del que dispone *S*. Mele añade posteriormente que los pares no deben disponer de más tiempo al evaluar ese conjunto de evidencia [Mele (2003)]. Parece que si los pares disponen de la misma evidencia y no poseen ningún deseo de que *p* o *no-p* sean el caso —es decir, son neutrales— la adopción de una creencia final distinta ha de deberse a la fuerza motivacional del deseo que posee el sujeto de estudio, tal como quiere Mele. Sin embargo, Hermes señala que el test tiene en esencia dos puntos débiles: por un lado, no es sensible históricamente, esto es, si el sujeto y los pares adoptan finalmente la misma creencia, no podríamos atribuirle autoengaño, aun cuando durante el proceso de acopio de pruebas ambas partes pudieran haber diferido en sus creencias; y por otro lado no soporta contraejemplos. Para Hermes, la adulteración de la evidencia es una propiedad *disposicional*, por lo que el hecho de que un sujeto no haya adulterado cierta evidencia, no indica que su deseo no posea fuerza motivacional para sesgar la evidencia dolorosa, en caso de que sea necesario en el futuro.

Para FERNANDO MARTÍNEZ MANRIQUE el autoengaño, fenómeno muy común, no es una anomalía, sino el resultado de un sistema funcional y adaptativo en la protección del yo y la regulación de metas. En su contribución, establece dos criterios para distinguir el autoengaño de otros fenómenos en cierto modo similares, como el engaño, la fabulación y la confabulación: *sinceridad* y *accesibilidad* a la evidencia. Para dar cuenta del fenómeno del autoengaño, Manrique propone un mecanismo de filtración de evidencia: el *supresor*. Tal mecanismo no supone problemas para la primera de las teorías de lectura de la mente que Manrique señala, la "Teoría Teoría". Sin embargo, sí supone un escollo para la "Teoría de la simulación", pues la simulación perfecta arruina la posibilidad de la atribución de autoengaño. Según Manrique, es cierto que raramente la simulación es perfecta, pero parece extraño que la simulación requiera, para que tenga éxito en su propósito, el que sea defectuosa. Otro problema que Manrique señala es que las teorías de la simulación son predictivas, no explicativas, y raramente predecimos el autoengaño. No decimos "*S* se autoengañará", sino "*S* está —o estuvo— autoengañado". Estos problemas de las teorías de la simulación no deben llevarnos, sin embargo, a su abandono, aun cuando debamos limitar su aplicación en la lectura de la mente y sus procesos.

La pregunta por la ética en la formación de creencias fue ya debatida por William Clifford (1877) y William James (1897). Mientras para Clifford creer en ausencia de evidencia suficiente es moralmente condenable siempre y en todo lugar, James otorga —en ciertos casos— una oportunidad a la libertad de creer; para James, algunos tipos de creencia no requieren evidencia, porque no es posible encontrarla. Pero esto no puede impedirnos creer, sino que la decisión de creer —o de no creer— debe ser tomada voluntariamente. Esto no implica tampoco que se pueda creer "por las buenas" cualquier cosa, sino que a veces sería necesario "dar un rodeo". Bernard Williams (1973) atacó duramente estas conclusiones argumentando que el hecho de que la creencia aspire a la verdad la convierte en producto de un proceso incontrolable por nuestra voluntad. Creer a voluntad es una imposibilidad conceptual.

Precisamente en esta línea de debate pueden situarse las preguntas por el autoengaño, no sólo desde un punto de vista conceptual, sino moral. Las tesis de Williams parecen negar que nos sea posible creer algo de modo intencional, voluntario. Pero, aun admitiendo que esto pudiera ser el caso, nos resta dirimir la cuestión de si es moral o no. Clifford no negaba la posibilidad de creer a voluntad, pero sí lo condenaba moralmente; James sin embargo parece que dejaba a cada cual la decisión de creer —al menos en ciertas cuestiones.

En la literatura más reciente, ya específicamente sobre el autoengaño, las cuestiones morales que más han preocupado giran en torno a asuntos como la responsabilidad del sujeto que se autoengaña, si el autoengaño es algo

que incluye intencionalidad, si es al menos susceptible de cierto control por parte del sujeto o si, en todo caso, implica o no un caso de negligencia moral y/o epistémica. Según Neil Levy (2004), han de cumplirse dos condiciones para hablar de responsabilidad en el autogaño: 1) que el sujeto crea que se trata de una cuestión importante —moral o de otro tipo—, y 2) que tenga alguna duda acerca de la verdad de la creencia que abraza.

Según JESÚS COLL, quien se autoengaña es responsable en la medida en que podría haberlo evitado si hubiese tomado la precaución y la molestia necesarias. Sin embargo, aunque desde los enfoques intencionalistas parece que la intención de engañarse sería suficiente para exigirle responsabilidad al sujeto, casi todos los intencionalistas defienden en última instancia alguna modalidad de partición de la mente o la existencia de diferentes agentes y subagentes mentales, de modo que la intención de engañarse se hallaría escondida y se tornaría inaccesible para el sujeto, con lo que se desencadenaría un proceso incontrolable. Las propuestas no-intencionalistas como la de Mele parecen exculpar al sujeto en tanto que no tiene intención y se nos aparece como una víctima involuntaria de procesos y motivaciones que le llevan a sesgar la evidencia. Por ello habría que recuperar, según Coll, algún elemento de las teorías intencionales; en un primer momento parecería que ese elemento es la *intención*, pero si el sujeto conoce la intención, la tarea de engañarse resulta imposible; si se la oculta de algún modo y no la conoce, carecemos de control sobre el proceso. Coll le reprocha a Mele el hecho de que parezca olvidar que el problema del autoengaño no comporta sólo explicar un estado, sino un proceso, y que en éste se revela la tensión existente en el sujeto entre la dirección en la que apunta la evidencia de la que dispone y el estado de cosas que desea que sea el caso; la formación de creencias orientadas por el deseo en ausencia de evidencia tendría que denominarse pensamiento desiderativo, no autoengaño. El autoengaño requiere, pues, que el sujeto disponga de evidencia y que, a causa de su deseo de que la realidad sea de otro modo, se activen —activación de la que el sujeto es conocedor— algunos mecanismos no intencionales, innatos o adquiridos socialmente, formándose una creencia contraevidencial. Por ello, la responsabilidad que puede exigírsele al agente consiste en el control de estos procesos epistémicos. De este modo, Coll da cuenta de las dos cláusulas que exige Levy para poder hablar de responsabilidad en un sujeto que se autoengaña.

Por su parte, IAN DEWEESE-BOYD cree que ninguna de las condiciones propuestas por Levy es necesaria —aunque serían suficientes— para atribuirle responsabilidad al sujeto que se autoengaña. La falta de estas condiciones no implica la ausencia de "*control de orientación*" sobre los desencadenantes de los procesos de adulteración, los deseos, la emoción y los sentimientos. Siguiendo a Ronald Milo, DeWeese-Boyd afirma que el sujeto será responsa-

ble no sólo de no ejecutar la capacidad que controla la orientación de las creencias respecto de la motivación y la emoción, sino de no *cultivar* dicha capacidad. El sujeto tiene además el deber de conocer las consecuencias morales de sus acciones, y el escrutinio de la evidencia que sirve de base para la formación de creencias que guiarán nuestras acciones ha de ser minucioso si estas acciones pueden comportar daños morales. El sujeto que se autoengaña es responsable de sus creencias cuando éstas pueden dar lugar a un perjuicio moral, y es moralmente culpable cuando su negligencia epistémica facilita el daño moral, dependiendo su grado de culpabilidad de la seriedad del mal moral ocasionado y del esfuerzo que le hubiese sido necesario para evitar caer en el autoengaño.

SAMI PIHLSTRÖM ofrece un enfoque novedoso en el tratamiento del autoengaño cuando propone que, además de los casos prototípicos, hay otro tipo de autoengaño: el *autoengaño transcendental*. Este nuevo enfoque enriquecerá las caracterizaciones tradicionales del autoengaño, a la vez que el concepto de *conocimiento transcendental*. Pese a que alguien podría tratar de dar cuenta del autoengaño de modo que el yo empírico engañase al yo transcendental o viceversa, este enfoque parece poco prometedor, ya que no se trata de dos yoes, sino de un solo yo bajo dos aspectos. El autoengaño transcendental sería un engaño acerca del yo, en el sentido de que el contenido de la creencia falsa que adquirimos versa sobre la esencia de nuestro yo y los límites de nuestra condición humana. El autoengaño sería producto más bien de los errores metafísicos generados por los extravíos de la ilusión transcendental, como ocurre, por ejemplo, en los "paralogismos" kantianos.

Hay dos tipos de autoengaño transcendental: el metafísico y el ético. El *autoengaño transcendental metafísico* tiene a su vez dos variedades: el inflacionario y el deflacionario. El de tipo inflacionario aparece cuando, a partir de la unidad de la apercepción transcendental infiero, empujado por una tendencia ilegítima de la ilusión transcendental, la permanencia o unidad real absoluta de una entidad metafísica: el alma. Mientras las ilusiones transcendentales son una tendencia inevitable para el hombre, estas inferencias dialécticas metafísicas sí pueden —y deben— evitarse. El tipo deflacionario se produce por la tendencia a mantener una visión exclusivamente científica del mundo, en la que el yo no tiene ninguna realidad. Esta clase de autoengaño cristaliza, entre otras, en las tesis defendidas por Daniel Dennett, para quien el yo no existe como una unidad y es, en realidad, una historia ficticia. El problema de esta postura es que no sólo niega la unidad del yo empírico kantiano, sino que rechaza el yo transcendental como condición de posibilidad de toda experiencia subjetiva por ser una entidad misteriosa y fantasmagórica, poco científica. Esto da lugar a una visión deflacionaria de la verdadera entidad del yo.

El *autoengaño transcendental ético* consiste en el olvido de que la condición humana está estructurada por la responsabilidad moral y, por tanto, por una ineliminable potencialidad de culpabilidad moral. La culpabilidad —en tanto que elemento *potencial* al menos— es condición de posibilidad de nuestra relación con otros individuos, así como del uso de conceptos y juicios morales. El yo moral *transcendentalmente culpable* no es tanto un objeto del mundo como un punto de vista del mundo: supone ver el mundo bajo una luz ética. Cuando nos ocultamos esto, actuamos de un modo autoengañoso.

Departamento de Filosofía
Universidad de Oviedo
Campus del Milán, E-33080, Oviedo
E-mail: vmsantamaria@hotmail.com

NOTAS

* La realización de este trabajo ha sido posible gracias a la subvención por parte del Ministerio de Educación y Ciencia a través de una beca de investigación del Plan Nacional de Formación de Profesorado Universitario, ref. AP2003-4762, y a la financiación ofrecida por el Proyecto de Investigación ref. HUM2007-65921, concedido asimismo por el Ministerio de Educación y Ciencia.

REFERENCIAS BIBLIOGRÁFICAS

AUDI, R. (1982), "Self-Deception, Action and Will", *Erkenntnis*, 18, pp. 133-158.
BACH, K. (1981), "An Analysis of Self-Deception", en *Philosophy and Phenomenological Research*, 41 (3), pp. 351-370.
BARNES, A. (1990), "When Do We Deceive Others?", *Analysis*, 50 (3), pp. 197-202.
BERMÚDEZ, J.L. (2000), "Self-deception, intentions and contradictory beliefs", *Analysis*, 60, pp. 309-319.
CLIFFORD, W.K. (1877), "La ética de la creencia", en *La voluntad de creer. Un debate sobre la ética de la creencia*, Madrid, Tecnos, 2003, pp. 91-134.
DAVIDSON, D. (1982), "Paradoxes of Irrationality", en *Problems of Rationality*, Oxford, Clarendon Press, 2004, pp. 169-187.
— (1985), "Deception and Division", en *Problems of Rationality*, Oxford, Clarendon Press, 2004, pp. 199-212.
ERWIN, E. (1988), "Psychoanalysis and self-deception", en B.P. McLaughlin & A.O. Rorty (eds.), *Perspectives on self-deception,* Berkeley, University of California Press, pp. 228-245.
FESTINGER, L. (1957), *Teoría de la disonancia cognoscitiva*, Madrid, Instituto de Estudios Políticos, 1975.
FINGARETTE, H. (1969), *Self-Deception*, Londres, Routledge and Kegan Paul.
— (1998), "Self-Deception Needs No Explaining", *The Philosophical Quarterly*, 48 (192), pp. 289-301.

FREUD, S. (1900), "La interpretación de los sueños", en Freud, S. (1886/1939), *Obras completas*, Madrid, Biblioteca Nueva, 9 tomos, 1974. Tomo II, pp. 343-720.
— (1915), "La represión", en Freud, S. (1886/1939), *Obras completas*, Madrid, Biblioteca Nueva, 9 tomos, 1974. Tomo I, pp. 1045-1051.
— (1923), "El 'Yo' y el 'Ello'", en Freud, S. (1886/1939), *Obras completas*, Madrid, Biblioteca Nueva, 9 tomos, 1974. Tomo VII, pp. 2701-2728.
GERGEN, K.J. (1985), "The Ethnopsychology of Self-Deception", en Mike W. Martin (ed.), *Self-Deception and Self-Understanding*, Lawrence, Kansas University Press.
HAIGHT, M.R. (1980), *A Study of Self-Deception*, Sussex, The Harvester Press.
JAMES, W. (1897), "La voluntad de creer", en *La voluntad de creer. Un debate sobre la ética de la creencia,* Madrid, Tecnos, (2003), pp. 135-180.
JOHNSTON, M. (1988), "Self-Deception and the Nature of Mind", en B.P. McLaughlin and A.O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley, University of California Press, pp. 63-91.
KIPP, D. (1980), "On Self-Deception", *The Philosophical Quarterly*, 30, pp. 305-317.
LEVY, N. (2004), "Self-Deception and Moral Responsibility", *Ratio (new series)*, XVII, pp. 294-311.
MELE, A.R. (1987), "Real Self-Deception", *Behavioral and Brain Sciences*, 20, pp. 91-136.
— (2001), *Self-Deception Unmasked*, Cambridge, Harvard University Press.
— (2003), "Emotion and Desire in Self-Deception", en Hatzimoysis, A. (ed.) (2003), *Philosophy and the Emotions,* Cambridge University Press, pp. 163-179.
PEARS, D. (1982), "The Goals and Strategies Of Self-Deception", en Elster, J. (ed.) *The Multiple Self*, Cambridge, Cambridge University Press, 1986, pp. 59-77.
— (1991), "Self-Deceptive Belief-Formation", *Synthese*, 89, pp. 392-405.
RORTY, A.O. (1972), "Belief and Self- Deception", *Inquiry*, 15, pp. 387-410.
— (1983), "Akratic Believers", *American Philosophical Quarterly*, 20, pp. 175-183.
RYLE, G. (1949), *El concepto de lo mental*, Paidós, Barcelona, 2005.
SARTRE, J.P. (1943), *El ser y la nada*, Losada, Buenos Aires, 2005.
TALBOTT, W.J. (1995), "Intentional Self-Deception in a Single Coherent Self", *Philosophy and Phenomenological Research*, 55 (1), pp. 27-74.
TRIVERS, R. (1985), "Deceit and Self-Deception", en *Social Evolution*, Menlo Park, CA, Benjamin/Cummings, pp. 395-420.
— (2000), "Elements of a Scientific Theory of Self-Deception", *Annals of the New York Academy of Sciences*, 907, pp. 114-131.
VALDÉS VILLANUEVA, L.M. (2003), "El derecho de creer", en *La voluntad de creer. Un debate sobre la ética de la creencia*, Tecnos, Madrid, 2003, pp. 9-90.
WILLIAMS, B. (1973), "Deciding to Believe", en *Problems of the Self*, Cambridge, Cambridge University Press, pp. 136-151.