EL AUTOENGAÑO DESENMASCARADO

Alfred R. Mele

CÁTEDRA

Alfred R. Mele

El autoengaño desenmascarado

Traducción de Víctor Manuel Santamaría Navarro

Índice

A mi padre, Al

Nota a la traducción española (2016)

Me complace mucho que se publique ahora una traducción española de *Self-Deception Unmasked*. Cuando escribí el libro, el punto de vista deflacionario sobre el autoengaño que yo defendía era muy radical. Ahora leo algunas veces que es la «concepción estándar». En cualquier caso, es, según mi opinión, un punto de vista muy atractivo. Después de publicar el libro recibí muy pocas invitaciones para escribir artículos sobre el autoengaño. Como respuesta, he publicado artículos sobre las relaciones entre el autoengaño y fenómenos tales como las creencias de tipo conspirativo e ilusorio, y he explorado con mayor profundidad que en el libro el lugar de las emociones en el autoengaño. Los lectores de este libro a los que les resulte posible leer filosofía en inglés, pueden estar interesados en mis siguientes artículos: «Delusional Confabulations and Self-Deception», en W. Hirstein (ed.), *Confabulation: Views from Neuroscience, Psychiatry, Psychology, and Philosophy,* Oxford University Press, 2009; «Self-Deception and Delusions», *European Journal of Analytic Philosophy,* 2006; «Emotion and Desire in Self-Deception», en A. Hatzimoysis (ed.), *Philosophy and the Emotions,* Cambridge University Press, 2003; «When Are We Self-Deceived?», *Humana.Mente – Journal of Philosophical Studies,* 2012; «Approaching Self-Deception: How Robert Audi and I Part Company», *Consciousness and Cognition,* 2010; «Have I Unmasked Self-Deception or Am I Self-Deceived?», en C. Martin (ed.), *The Philosophy of Deception,* Oxford University Press, 2009.

Alfred Mele
Abril de 2016

Prefacio

Lo que me motivó a aceptar la amable invitación de Harry Frankfurt a que enviara un manuscrito para esta serie fue la oportunidad de presentar y defender de modo sistemático una postura respecto al autoengaño que ha evolucionado parcialmente a partir de mis primeras tentativas de arrojar luz sobre el fenómeno. Aunque me baso en trabajos publicados con anterioridad, aquí se ofrece una defensa mucho más robusta de mis tesis centrales acerca del tipo de autoengaño que ha recibido mayor atención en la literatura filosófica y psicológica, y se muestra una nueva e importante dimensión que se beneficia del trabajo empírico reciente en la evaluación de hipótesis. La postura general sobre el autoengaño que se presenta aquí es también considerablemente más comprehensiva que la que había sido capaz de esgrimir en mis incursiones en el asunto desperdigadas a lo largo de los años.

Me he pasado más tiempo del que deseo reconocer reflexionando sobre el autoengaño. Mi primer abordaje del asunto (Mele, 1982), un breve comentario sobre un provocativo artículo de Robert Audi (1982), contiene algunas de las semillas de la postura que se presenta en este libro. Allí mi tesis central era que «no hay una analogía estrecha entre el autoengaño y el engaño *intencionado* a otro» y que «cuando se rompe nuestra adhesión a esta analogía, hay bastante menos motivación para postular creencias [verdaderas] inconscientes en los casos cotidianos de autoengaño» (Mele, 1982, pág. 164). En Mele (1983), ofrezco una defensa mucho más rigurosa de este punto, una exposición de las características y condiciones conjuntamente suficientes del autoengaño, y una solución de la paradoja «estáti-

ca» del autoengaño que se subraya en el capítulo 1. Ese artículo fue la base del capítulo 9 de *Irrationality* (Mele, 1987a). Hay otro capítulo de *Irrationality* que se enfrenta al autoengaño: en el capítulo 10, apoyado en ciertos escritos de psicología social, ofrezco una solución a la paradoja sobre la dinámica del autoengaño. Estos dos capítulos son en parte la base de mis artículos «Two Paradoxes of Self-Deception» (1998b) y «Real Self-Deception» (1997a). El primero se escribió como ponencia invitada para un congreso interdisciplinar sobre el autoengaño organizado por Jean-Pierre Dupuy en 1993 en la Universidad de Stanford y fue publicada posteriormente en un volumen que contiene varios de los artículos que se presentaron allí. Debido a la naturaleza de la invitación, no tuve reparo en tomar ideas de *Irrationality,* pero las extendí en dos direcciones. Empleé nueva literatura empírica que apoyaba una hipótesis central que se adelantaba en *Irrationality* sobre la creencia motivacionalmente sesgada y, a petición de Dupuy, examiné un ejemplo literario de autoengaño (mi discusión del capítulo 3 sobre el cuento de Isaac Bashevis Singer «Gimpel el tonto» deriva en parte de ese artículo). Tiempo después del congreso en Stanford me invitaron a enviar a *Behavioral and Brain Sciences* un artículo que serviría de diana a los críticos. Ese artículo (Mele, 1997a) es sucesor del texto del congreso, con mayor alcance y más empírico. Se benefició de dos tandas de corrección realizadas mediante quince informes de los revisores.

Para este libro he tomado libremente ideas de mi trabajo publicado. Parte del capítulo 1 deriva de tres fuentes: «Real Self-Deception», *Behavioral and Brain Sciences,* 20, págs. 91-102 (Mele, 1997a); mi respuesta a los comentarios en *BBS,* «Understanding and Explaining Real Self-Deception», *Behavioral and Brain Sciences,* 20, págs. 127-134 (Mele, 1997b); y «Motivated Belief and Agency», *Philosophical Psychology,* 11, págs. 353-369 (Mele, 1998a). Parte del capítulo 2 deriva de Mele (1997a), Mele (1998a), y otro

artículo: «Twisted Self-Deception», *Philosophical Psychology,* 12, págs. 117-137 (Mele, 1999a)[1]. Parte del capítulo 3 deriva de los mismos tres artículos del capítulo 1 y una fuente adicional: «Two Paradoxes of Self-Deception», en J.-P. Dupuy (ed.), *Self-Deception and Paradoxes of Rationality,* Stanford, CSLI, págs. 37-58 (Mele, 1998b). Parte del capítulo 4 deriva de las tres mismas fuentes del capítulo 1. El capítulo 5 se basa en Mele (1999a).

Naturalmente, he aprendido de los autores que publicaron respuestas a mi trabajo previo sobre el autoengaño. Le estoy agradecido a Robert Audi, Kent Bach, Jim Friedrich, Rainer Reisenzein y Bill Talbott por las valiosas charlas informales e indicaciones por escrito. También me resultaron útiles los consejos de dos revisores anónimos.

Concluí la revisión final de este libro mientras disfrutaba en 1999-2000 de una beca del Fondo Nacional para las Humanidades *(National Endowment for the Humanities,* NEH) para Profesores Universitarios. (La beca apoyó el trabajo sobre otro libro —cuyo título tentativo es *Motivation and Agency*— que está bien encaminado en el momento en el que aparece este libro). Buena parte de la revisión se hizo mientras era profesor visitante del Programa de Filosofía en la Universidad Nacional Australiana (ANU) de junio a agosto de 1999. Le estoy agradecido al NEH y a la ANU por su apoyo, y al Davidson College por el año sabático 1999-2000.

---

1 El editor de *Philosophical Psychology,* donde apareció Mele, 1998a y 1999a, es Taylor & Francis Ltd./Carfax/Routledge (http://taylorandfrancisgroup.com).

Capítulo 1

Introducción: enfoques, paradojas, sesgos y agencia

«Una encuesta a profesores universitarios mostró que el 94% creía que era mejor en su trabajo que el promedio de sus colegas» (Gilovich, 1991, pág. 77). ¿Son los profesores universitarios excepcionalmente adeptos al autoengaño? Quizá no. «Una encuesta a un millón de estudiantes universitarios de último curso mostró que [...] *todos* los estudiantes creían que estaban por encima de la media» con respecto a su «capacidad para relacionarse con los demás [...] y el 25% pensaba que estaba en el 1% superior» *(ibíd.)*. Se podría pensar que quienes contestaban a las encuestas no eran completamente sinceros en sus respuestas. Entonces, de nuevo, ¿cuántos profesores universitarios conoce usted que *no* se crean mejores en su trabajo que el promedio de sus colegas?

Datos como éstos sugieren que a veces nos engañamos. Y esto plantea interesantes preguntas. ¿Cómo nos engañamos a nosotros mismos? ¿Por qué nos engañamos? ¿Qué es engañarse a uno mismo? ¿Es siquiera posible el autoengaño? Estas preguntas son las que orientan la exposición en este libro.

Algunos teóricos entienden que el autoengaño es en buena medida isomorfo al engaño interpersonal estereotípico. Esta visión, que ha generado rompecabezas o «paradojas» tan debatidos, está en la base de influyentes trabajos sobre el autoengaño no sólo en filosofía, sino también en psicología, psiquiatría y biología[2]. A la vez que trato de resolver las paradojas más importantes, defiendo que in-

tentar comprender el autoengaño bajo el modelo del engaño interpersonal estereotípico es fundamentalmente un error. La postura que se defiende aquí respecto al autoengaño es *deflacionaria*. Si estoy en lo cierto, el autoengaño no es ni irresolublemente paradójico ni misterioso, y se puede explicar sin acudir a exotismos mentales. Aunque un teórico cuyo interés en el autoengaño se limitase a los límites externos de lógica o a la posibilidad conceptual podría ver esto como despojar al tema de la intriga conceptual, la principal fuente de un interés más amplio y perdurable en el autoengaño es la preocupación por comprender y explicar el comportamiento de los seres humanos reales.

1. Un primer vistazo

El autoengaño se presenta aparentemente bajo dos formas, «directo» y «retorcido». Los casos directos de autoengaño han recibido mayor atención en los trabajos filosóficos y empíricos. En estos casos, la gente se autoengaña al creer algo que quieren que sea cierto —por ejemplo, que no están gravemente enfermos, que sus hijos no toman drogas o que un ser querido es inocente de una acusación criminal—. En los casos retorcidos, la gente se autoengaña al creer algo que quieren que sea falso (y tampoco quieren que sea verdad). Por ejemplo, un marido inseguro y celoso puede creer que su mujer está teniendo una aventura a pesar de que sólo posee evidencias endebles de esa proposición y a pesar de que no desea que sea el caso que tenga tal aventura[3]. Si algún autoengaño es retorcido en este sentido, hay al menos una afirmación relativamente común sobre el autoengaño que es falsa —la afirmación de que, que *S* se autoengañe sobre *p,* requiere que *S* desee que *p*[4] —. Además, el autoengaño retorcido incluso amenaza aparentemente la afirmación mucho más modesta de que todo autoengaño es motivado o tiene un componente motiva-

cional[5]. Aunque el antónimo más obvio de «directo» *(straight)* es «torcido» *(bent)*, prefiero «retorcido» *(twisted)* por cuestiones de estilo. No uso el término de modo peyorativo ni considero que el autoengaño retorcido sea esencialmente patológico.

En los capítulos 2 y 3 ofrezco una explicación de la naturaleza y etiología del autoengaño directo cotidiano y resuelvo algunas paradojas comunes sobre el autoengaño. En el capítulo 4, someto a revisión y rechazo algunos intentos de demostración empírica de autoengaño «estricto» en el que quien se autoengaña cree una proposición, *p*, al tiempo que cree también su negación, ~*p*. En el capítulo 5, desarrollo un par de enfoques para explicar el autoengaño retorcido —un enfoque centrado en la motivación y un enfoque híbrido en el que toman parte tanto la motivación como la emoción— con el fin de desplegar nuestros recursos para explorar y explicar el autoengaño retorcido y mostrar que estos prometedores enfoques son consistentes con mi postura sobre el autoengaño directo.

2. Tres enfoques para caracterizar el autoengaño y un par de paradojas

Al definir el autoengaño se pueden distinguir tres enfoques habituales: el *léxico*, en el que el teórico comienza con una definición de «engañar» o «engaño» valiéndose de un diccionario o del uso cotidiano como guía, y luego lo emplea como modelo para definir el autoengaño; el *basado en ejemplos*, en el que se hace un escrutinio de ejemplos representativos de autoengaño y se intenta identificar las características esenciales que tienen en común; y el *guiado por una teoría*, en el que la búsqueda de una definición está orientada por una teoría de sentido común respecto a la etiología y naturaleza del autoengaño. También son habituales los enfoques híbridos de los tres.

El enfoque léxico puede parecer el más prudente. Quienes emplean el enfoque basado en ejemplos corren el riesgo de considerar un rango de casos demasiado reducido. El enfoque guiado por una teoría descansa, en sus manifestaciones más típicas, en hipótesis explicativas de sentido común que podrían estar desencaminadas: incluso si la gente corriente suele acertar cuando identifica casos de autoengaño, podría ser poco fiable al diagnosticar qué ocurre en ellos. En sus versiones más prístinas, el enfoque léxico descansa fundamentalmente en la definición de «engañar» que ofrece un diccionario. ¿Y qué podría constituir una mejor fuente de definiciones que un diccionario?

Las cosas no son tan sencillas, sin embargo. Hay sentidos más débiles y más fuertes de «engañar» tanto en el diccionario como en el lenguaje cotidiano. Los lexicalistas necesitan un sentido de «engañar» que sea apropiado para el autoengaño. ¿Sobre qué base van a identificar ese sentido? ¿Al final han de volver a casos representativos de autoengaño o teorías de sentido común acerca de lo que ocurre en los casos de autoengaño?

El enfoque léxico es el preferido por los teóricos que niegan que el autoengaño sea posible (por ejemplo, Gergen, 1985; Haight, 1980; Kipp, 1980). Hay un par de supuestos léxicos habituales:

1. Por definición, la persona *A* engaña a la persona *B* (donde *B* puede ser o no ser la misma persona que *A)* para que crea que *p* sólo si *A* sabe, o al menos cree sinceramente, que ~*p* y causa que *B* crea que *p*.
2. Por definición, el engaño es una actividad intencionada: el engaño sin intención es conceptualmente imposible.

Cada supuesto va asociado a una paradoja familiar sobre el autoengaño.

Si es cierto el supuesto 1, entonces engañarse de tal modo que llegue a creer que *p* exige que uno sepa, o al menos crea cierto, que ~*p* y que uno mismo se cause la creencia de que *p*. Como mínimo, uno empieza creyendo que ~*p* y luego de algún modo se lleva a sí mismo a creer que *p*. Algunos teóricos interpretan que esto implica que, en algún momento, quienes se autoengañan creen tanto que *p* como que ~*p* (por ejemplo, Kipp, 1980, pág. 309). Y afirman que éste no es un estado mental posible: la propia naturaleza de la creencia impide que uno crea a la vez que *p* es verdad y que *p* es falso[6]. Por tanto, estamos ante una paradoja del autoengaño *estática:* el autoengaño, de acuerdo con esta visión, requiere estar en un *estado mental* imposible.

De hecho, el supuesto 1 no implica que en todos los casos de engaño haya un momento en el que la persona que engaña crea que ~*p* y la persona engañada crea que *p*. En algunos casos de engaño interpersonal, *A* ha dejado de creer que ~*p* en el momento en el que causa que *B* crea que *p*. Imagínese que el medio por el que *A* intenta engañar a alguien sea una carta. En su carta, *A* trata de engañar a *B* para que crea que *p* mintiéndole: *p* es falso y su afirmación de *p* en la carta constituye una mentira. Cuando envía la carta, *A* está seguro de que ~*p,* pero cuando *B* recibe la carta, *A* ya cree que *p*. Si la mentira de *A* tiene éxito, *A engaña a B* para que crea que *p* de un modo tal, que respalda el supuesto 1. Pero no hay un momento en el que *A* crea que ~*p* y *B* crea que *p* (véase Sorensen, 1985).

Un teórico inclinado a creer que en «el concepto de engaño» hay base para la afirmación de que quienes se autoengañan creen que *p* y creen que ~*p* simultáneamente, no debe desanimarse por la observación anterior. Bien puede ser cierto que en los casos estereotípicos de engaño interpersonal haya un momento en el que *A* cree que ~*p* y *B* cree que *p*. Y un teórico puede defender aún que el au-

toengaño sólo se entiende adecuadamente bajo el modelo de engaño interpersonal estereotípico.

La afirmación de que el autoengaño ha de entenderse bajo el modelo mencionado produce una paradoja más respecto al estado de autoengaño. En los casos estereotípicos de engaño interpersonal hay un momento en el que quien engaña *no* tiene la creencia de que *p* y la persona engañada tiene la creencia de que *p*. Si el autoengaño es estrictamente análogo al engaño interpersonal estereotípico, hay un momento en el que quien se autoengaña tiene la creencia de que *p* y no tiene la creencia de que *p*, una condición desconcertante[7], ciertamente.

El supuesto 2 genera una paradoja *dinámica*, una paradoja sobre la dinámica del autoengaño. Por un lado, resulta difícil imaginar cómo podría una persona engañar a otra para que crea que *p* si la última sabe exactamente qué trama la primera, y es difícil ver cómo podría resultar más fácil el truco si quien aspira a engañar y la pretendida víctima son la misma persona. Por otro lado, normalmente el hecho de que quien engaña tenga o ejecute intencionadamente una estrategia para engañar facilita el engaño. Si, con el fin de evitar que resulten boicoteados los propios esfuerzos de autoengaño, uno no ha de ejecutar intencionadamente ninguna estrategia para engañarse a sí mismo, ¿cómo podría tener éxito? El reto reside en explicar cómo es que el autoengaño es en general un proceso psicológicamente posible. Si quienes se autoengañan se engañan intencionadamente a sí mismos, uno se pregunta qué impide que la intención rectora socave su propio funcionamiento efectivo. Y si el autoengaño no es intencionado, ¿qué motiva y dirige los procesos del autoengaño?[8].

Un teórico que considere que el autoengaño es un fenómeno genuino puede tratar de resolver las paradojas dejando los supuestos 1 y 2 sin tocar. Un camino alternativo consiste en socavar esos supuestos y mostrar la relevancia