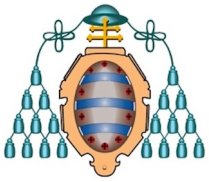

Universidad de Oviedo

Departamento de Filosofía

Tesis doctoral

EL AUTOENGAÑO:
PROBLEMAS CONCEPTUALES

Examen crítico de algunas teorías recientes

Víctor Manuel Santamaría Navarro

2009

„Die Wirklichkeit mancher *inneren* Lüge, welche die Menschen sich zu Schulden kommen lassen, zu beweisen, ist leicht, aber ihre Möglichkeit zu erklären, scheint doch schwerer zu sein“

[Es fácil probar la realidad de algunas mentiras *internas* de las que los hombres se culpabilizan; sin embargo, explicar su posibilidad parece más difícil.]

[Kant, Ak VI, 430]

„In jedem ernsteren philosophischen Problem reicht die Unsicherheit bis an die Wurzeln hinab.

Man muß immer darauf gefaßt sein, etwas *ganz* Neues zu lernen“

[En todo problema filosófico serio, la incertidumbre se extiende hasta las raíces.

Se debe estar siempre preparado para aprender algo *totalmente* nuevo.]

[Wittgenstein, *BF* I-15]

ÍNDICE

INTRODUCCIÓN

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.

Otra mujer comienza a tener extraños síntomas que le hacen acudir al médico. Siempre ha confiado en la medicina occidental, y además acude a un prestigioso hospital en el que una reputada especialista la examina. Una vez allí se somete a diversas pruebas que confirman que padece una grave enfermedad en fase terminal; sus fuerzas van disminuyendo día a día y el médico le confirma que el avance es irreversible; sin embargo, la mujer cree que su enfermedad es pasajera.

Un joven lleva años enamorado de una muchacha que reside en su mismo barrio; él le ha declarado su amor en numerosas ocasiones, a lo que

ella ha respondido siempre con el mayor tacto pero la mayor claridad que le ha sido posible que no le corresponde en ese sentimiento. Hace un par de meses ella comenzó una relación con otro chico que le gustaba hacía casi un año. Sin embargo, nuestro joven cree que en realidad ella está enamorada de él, y que sólo lo está poniendo a prueba.

Casos como éstos 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? ¿Y desde el moral? Por último, ¿es en todo caso, siquiera posible?

La controversia generada en los últimos años no ha conseguido que los distintos autores se hayan puesto de acuerdo, no sólo en la propia explicación del fenómeno en sí, sino tampoco acerca de qué es aquello que debe ser considerado como caso genuino de autoengaño.

El debate y la necesidad de ofrecer una explicación al supuesto problema es producto de la combinación de dos hechos: por un lado, el autoengaño se nos aparece como un fenómeno real, empírico y totalmente cotidiano, pero por otro, filósofos y psicólogos no son capaces de ofrecer una única explicación satisfactoria acerca de los mecanismos que intervienen en este fenómeno que dé cuenta además de la irracionalidad que parece implicar. Como era de esperar, tampoco los experimentos empíricos que se han ensayado al respecto han zanjado el debate, arrojando muy poca luz a la discusión conceptual.

Por supuesto, ha de evitarse que la discusión se convierta en una lucha vacía en torno a "palabras" en la que lo único que separe una teoría de las otras sea el uso convencional y arbitrario de distintas etiquetas para describir lo mismo. Nuestro propósito es examinar las más importantes teorías, clasificarlas y valorar sus aportaciones, para dirimir si es posible o no que un fenómeno como el autoengaño se produzca y, si puede producirse, qué mecanismos intervendrían, de qué modo y cómo se lleva a cabo el proceso sin dar lugar a un estado totalmente irracional. Ha de señalarse que el problema puede ser atacado de diversas maneras. Un modo de hacerlo sería comenzar ofreciendo mi propia visión del asunto y a partir de ahí evaluar tanto casos particulares típicos que ejemplifiquen el problema como las distintas interpretaciones que han ofrecido para esta cuestión los autores más influyentes. Otro modo de acercarse al tema sería exponiendo directamente casos prototípicos para tratar de extraer los rasgos principales y a partir de ahí examinar las dificultades.

Sin embargo, dado que parece que el consenso entre teóricos es difícil y que sus posturas se hayan muy enfrentadas, parece claro que no es posible fijar ni las hipótesis de partida, ni describir algún caso prototípico, pues diferirán en la interpretación de sus rasgos principales. No queda otra salida, en esta situación, que partir de la definición ofrecida por el diccionario, lo cual parece, en todo caso, el modo más neutro de acercarse al tema; a partir de ahí trataré de afinar más la caracterización diferenciándolo de otros problemas similares o cercanos.

Falsedad, mentira, engaño

La primera dificultad que nos encontramos, es que el *Diccionario de la Real Academia de la Lengua Española* no contiene una entrada para "autoengaño". La entrada más cercana conceptualmente es la de "engaño", que es definido como la "acción y efecto de engañar". Ya a primera vista, el problema que esto suscita es la interpretación del prefijo "auto", que puede ser interpretado como "a sí mismo", "para con uno mismo" o "respecto de uno mismo".

Según el *Diccionario de la Real Academia de la Lengua Española* "engañar" es "inducir a otro a tener por cierto lo que no lo es, valiéndose de palabras o de obras aparentes y fingidas". Por tanto el engaño supone un *proceso* mediante el cual el engañador *conscientemente diseña* una estrategia más o menos compleja mediante la cual *induce al error* a otro sujeto. Así pues, el engaño consiste en la inducción a error de forma *consciente y voluntaria*. Pero entonces, surge el primer problema, pues parece difícil que el autoengaño

pueda ser visto como un caso particular de engaño. Efectivamente, si la persona que engaña y la que es engañada es una y la misma, parece que una conducta "aparente y fingida" dirigida a uno mismo resultaría de lo más inútil para obtener el resultado deseado, a saber, inducirnos a error, pues no se nos ocultarían ni el fin de nuestra propia estrategia ni los detalles de la misma.

Y sin embargo, la experiencia diaria nos ofrece casos difícilmente explicables si no acudimos a la noción de "autoengaño". La gente parece que "se miente a sí misma" y "se engaña a sí misma". Es interesante y absolutamente necesario detenerse aquí para distinguir "mentir" de "engañar", pues no son lo mismo. Según el *Diccionario de la Real Academia de la Lengua Española,* "mentir" es "decir lo contrario de lo que se sabe, cree o piensa". Mentir sería también un fenómeno intencional, pero consistiría —como dice Donald Davidson—, en algo así como ser insincero con respecto a la representación de las propias creencias [cf. Davidson (1985), p. 207]. Por tanto, aunque mentir y engañar son dos fenómenos intencionales[1], la mentira es un intento de ocultar lo que uno

[1] Léase, que conllevan intención. En algunas ocasiones algunos autores hablan de "intencionalidad" en el (auto)engaño, lo cual puede llevar a equívoco. Éste es un rasgo crucial. De hecho, por esta razón es necesario distinguir también el decir cosas falsas del mentir, pues a veces se dice "me has engañado" aun cuando se sabe que no había intención alguna por parte de quien indujo al error de que su interlocutor se formase una idea equivocada de lo que era el caso, esto es, de engaño. Sin embargo, hemos de considerar esto como casos de aplicación extralimitada guiados, a buen seguro, por el hecho de que también se produce esa inducción al error. Una vez más insisto en que, salvo que se indique lo contrario, hago uso del concepto «intencionalidad» de modo que tiene que ver exclusivamente con intenciones, y no con la teoría de Franz Brentano, según la cual todo estado

sabe, cree o piensa, mientras engañar es el intento de inducir [a otro] al error con respecto a lo que es el caso.

La razón por la que muchas veces aparecen casi como sinónimos es que el engaño se sirve en muchas ocasiones de la mentira para conseguir su objetivo, aunque esto no siempre sea así. En realidad para que estemos ante una mentira sólo se exige que el sujeto sea insincero con respecto a lo que sabe, cree o piensa y no exigimos induzca a error a su interlocutor; por tanto, no exigimos que se produzca un engaño. Puede haber mentira sin que haya engaño, y nadie dudaría en decirle a un amigo *"¡me has mentido!"*, aun cuando no hubiese sido engañado. Lo que sucede es que el objetivo de la mentira suele ser el engaño, y la mentira puede ser considerada como exitosa (o no) en función de si consigue (o no) su objetivo de inducir al error.

Por otro lado, mentir no es el único modo de inducir al error a otro individuo, pues a veces la estrategia que diseña quien engaña puede pasar por decir lo que en realidad sabe, cree o piensa, si cree que su interlocutor no va a creerle y va a retorcer lo que dice por desconfianza acerca de su credibilidad. En este caso el sujeto no mentiría, pero sí estaría *tratando* de que su interlocutor se formase una idea errónea, y por tanto de engañarle.

Por último hemos de distinguir la "mentira" y el "engaño" de la proferencia de falsedades. Decir algo falso no implica ni voluntad, ni consciencia, ni insinceridad, ni proyecto alguno; simplemente aquello que

mental tiene la propiedad de apuntar o referirse a algo externo. El uso que hago es, por lo demás, común en toda la literatura relacionada con el autoengaño.

decimos no es lo que verdaderamente es el caso. Tenemos por tanto tres fenómenos distintos, aunque cercanos, que se presentan a veces valiéndose unos de otros. Cuando mentimos, decimos algo en lo que no creemos, pero esto no implica que digamos algo falso, pues nuestras creencias pueden estar felizmente equivocadas. Supongamos que Pedro cree que tiene un examen el próximo miércoles; entonces Juan, un compañero que no es de su agrado, se acerca a preguntarle por la fecha de dicho examen, a lo que Pedro responde diciendo algo que cree falso, esto es, mintiendo: "el examen es el viernes". Supongamos además que Pedro estaba equivocado, y que verdaderamente el examen es el viernes. Pedro ha mentido, pero no ha dicho nada falso. Por tanto, mentir no implica decir cosas falsas, sino únicamente ser insincero (generalmente con la intención de engañar).

Podría parecer *prima facie* que el hecho de que Pedro no diga algo falso es la razón de que no haya engaño (pues Pablo, persuadido por la mentira de Pedro, acude a su examen el viernes). Sin embargo, en un caso extraño pero no imposible, un sujeto puede decir algo verdadero y engañar a otro individuo: supone con acierto que su interlocutor no le creerá y retorcerá su argumento; anticipándose a esta situación, el sujeto es sincero, pero con la intención clara de llevar a error al otro sujeto. Ocasionalmente puede tener éxito en lo que se propone.

Finalmente, desde luego puede decirse algo falso con intención de engañar sin que se produzca engaño, a saber, aquel caso en el que, por la razón que sea, el interlocutor no cree al sujeto y no es inducido al error.

Podemos por tanto, combinando todas las posibilidades, extraer todas las situaciones posibles. Así, un sujeto puede:

A) *Decir cosas falsas sin saber que lo son y sin intención de inducir a error.* Esto es un caso común que no representa mentira (pues aunque el sujeto diga algo falso, no es consciente de ello sino que dice lo que piensa). Dado que no hay intención, no puede ser tampoco un caso de engaño aun cuando induzca a error.

B) *Tratar de engañar sin mentir.* Esto sucede cuando el sujeto dice algo que *cree que es cierto (pudiendo ser efectivamente verdad o no)*, pero además tiene un proyecto de engaño, pues cree que su interlocutor no va a creerle y va a retorcer lo que dice. En este caso pueden darse varias situaciones:

B.1) Si el interlocutor retuerce lo que dice el sujeto porque no cree lo que se le dice y aquello que le dice el sujeto es algo verdadero, habrá caído en la estratagema diseñada por el sujeto y, al retorcer algo verdadero, habrá sido inducido al error, esto es, engañado.

B.2) Si el interlocutor retuerce lo que dice el sujeto porque no cree lo que le dice y aquello que dice el sujeto es algo falso, pese a caer en la estrategia del sujeto no habrá sido inducido a algo erróneo, y por tanto no será engañado[2].

[2] Que no ha sido engañado pese a caer en la trampa del sujeto es claro, pues ante este tipo de situaciones podemos imaginarnos al interlocutor diciendo: "Has *tratado* de engañarme", lo cual sólo puede decirlo porque presupone intención de engaño por parte del individuo y porque el intento no ha tenido éxito.

B.3) Si el interlocutor no retuerce lo que dice el sujeto porque [equivocadamente] confía en él y aquello que le dice es algo verdadero, ni habrá caído en la estrategia diseñada por el sujeto ni habrá sido conducido al error.

B.4) Si el interlocutor no retuerce lo que dice el sujeto porque [equivocadamente] confía en él y aquello que le dice es algo falso, pese a no caer en la estrategia del sujeto, ha sido conducido al error por confiar en él y creer algo falso.[3]

C) *Tratar de engañar mintiendo.* El sujeto miente, es decir, dice que es el caso aquello que cree o piensa que *no* es el caso *(pudiendo ser lo que dice efectivamente falso o no)*, con la intención de engañar a su interlocutor.

[3] Aquí, como ocurrirá en el caso C.3), puede debatirse si esto puede de llamarse "engaño", pues pese a que hay inducción al error, la intención del sujeto era la de engañarle mediante esa estrategia que diseña (la de esperar que el sujeto retuerza sus argumentos y así se equivoque) y de hecho esa estrategia fracasa. Este asunto está relacionado con el problema de las *cadenas causales desviadas*. Supongamos que deseo disparar a una diana y dar en el centro; como consecuencia de mi deseo, *intencionalmente* apunto y disparo. Por desgracia, mis cálculos resultan a la postre completamente equivocados y la bala sale totalmente desviada. Sin embargo, de modo sorprendente golpea en una piedra, sale de ahí rebotada a la diana, y alcanza finalmente el centro. Sin duda efectúe intencionalmente mi disparo, y tenía la intención de dar la diana. Como resultado de mi intención he efectuado un disparo que ha alcanzado la diana, pero ¿he dado en el centro de la diana a propósito, intencionalmente? Parece que no.

En nuestro caso concreto, B.4, sin duda tenía la intención de engañar y como resultado de mi acción, he engañado al sujeto. Pero no ha sido de un modo directo; alguien podría estar dispuesto a defender que pese a que la intención primera del sujeto fracasa, quizá hemos de preguntarnos si la verdadera y más general intención del sujeto no era simplemente la de "inducir al error". Discutiremos este asunto más adelante.

C.1) Si por estar equivocado, con su mentira dice algo cierto y el interlocutor cree la mentira, *no* se habrá producido el engaño, en tanto que no ha sido inducido al error.[4]

C.2) Si por el contrario el contenido de la mentira consiste en algo realmente falso y el interlocutor cree la mentira, entonces habrá sido inducido al error, y por tanto, engañado. Este es el caso fetén de engaño con éxito.

C.3) Si con su mentira dice algo que es en realidad verdadero pero el interlocutor no le cree y retuerce el argumento, habrá inducido, pese a que esa no era la forma en la que pensaba hacerlo, al sujeto a error.[5]

C.4) Si su mentira dice algo que en realidad es ciertamente falso pero el interlocutor, por desconfianza, no le cree no habrá producido el engaño, en tanto que no habrá ni caída en la estrategia ni inducción al error.

La mentira es, por tanto, una de las herramientas utilizadas para el engaño, pero no la única. De hecho, aunque el engaño a uno mismo resulta abiertamente paradójico, podemos encontrar casos en los que el sujeto

[4] De modo similar a los casos anteriores, es dudoso si podemos hablar de engaño cuando falta alguno de sus rasgos característicos. ¿Es la mera caída en la estrategia un rasgo de engaño o es necesaria la inducción de una creencia falsa? Sin duda hemos hecho caer en nuestra trampa al sujeto, pero no le hemos inducido al error.

[5] *Vid.* nota 3.

puede mentirse a sí mismo sin que esto haya de ser considerado como problemático. Esto nos puede resultar extraño en principio, y la causa de la extrañeza reside en que generalmente el objetivo de la mentira es que aquél a quien va dirigida no se percate de la insinceridad implícita en la mentira. Esto, a primera vista no puede ocurrir con uno mismo, por lo que cuando nos mentimos a nosotros mismos parece que sólo puede tratarse de una especie de "juego", un modo de tratar de animarnos y espolearnos para superarnos ante algún reto, adversidad o dificultad, como quien se dice, sabiendo que no es cierto, "voy a aprobar este examen" o "vamos, que *tú* puedes" (nótese que muchas veces hay este juego de desdoblamiento en el que el sujeto parece decirle a otro las cosas).

Por tanto la mentira, del mismo modo que el engaño, parece inútil y estéril —salvo la utilidad secundaria a la que hacíamos referencia— en tanto que dirigida al propio sujeto. De hecho, esta incapacidad para ocultarnos nuestras intenciones, estrategias e insinceridad (común a la mentira y al engaño dirigidos contra uno mismo[6]) es una de las razones por las que la propia posibilidad de existencia del autoengaño, definido en términos de engaño dirigido a uno mismo, se nos muestra como problemática.

[6] El caso que propone Alfred Mele [Mele (2001), p. 16] de un individuo sumamente olvidadizo que, consciente de su carencia memorística, escribe algo falso en su diario con la intención de resultar engañado cuando semanas más tarde vuelva a leerlo nos parece —como el propio Mele reconoce— sumamente extraño y marginal (en caso de tener verdaderamente éxito) y que, de cualquier manera, dista mucho de poder ser considerado como caso prototípico de autoengaño.

Las Paradojas: engaño interpersonal vs. autoengaño

Como hemos dicho, la interpretación del autoengaño bajo el modelo de engaño a otro, es decir, bajo el modelo de engaño interpersonal, da lugar a paradojas de difícil solución. Concretamente produce dos paradojas, una *estática* (*static puzzle*) consistente en que el sujeto ha de creer al mismo tiempo que *p* y *no-p* son el caso y otra *dinámica* (*dynamic puzzle*), según la cual para llegar a este estado ha de diseñar una estrategia mediante la que, a pesar de mantener la creencia —apoyada en la evidencia— de que *p* es el caso, se convence en contra de toda esa evidencia de que *no-p* es el caso.

La primera de ellas se basa en que en el engaño interpresonal o engaño a secas, el sujeto trata de que su interlocutor se forme una creencia errónea y contradictoria con la suya. Si el engaño es efectivo, engañado y engañador mantendrán a la vez dos creencias contradictorias; esto, obviamente no supone problema alguno, pues la dualidad de conciencias permite que ambas creencias no entren en contacto. Sin embargo, al tratar de interpretar el autoengaño bajo el modelo del engaño interpersonal habríamos de exigir que el sujeto mantenga al mismo tiempo dos creencias contradictorias, en tanto que engañado y engañador.

La segunda tiene que ver con el plan trazado para lograr el engaño; nuevamente el engaño interpersonal no supone obstáculo alguno para que un sujeto diseñe estrategias y lleve a cabo éstas a escondidas del engañado. El problema aparece cuando el mismo sujeto ha de diseñar un plan para engañarse a sí mismo. Ha de ser tan sutil e ingenioso que el sujeto esté al

corriente de todos y cada uno de sus detalles en tanto que engañador, pero permanezca en la más absoluta ignorancia de ellos en tanto que engañado.

Como veremos más adelante, la mayoría de los defensores del autoengaño bajo el modelo interpersonal se ven forzados a suponer intenciones inconscientes y una mente dividida cuyas partes presentan un funcionamiento al menos relativamente autónomo (donde no saben las unas de las otras) a fin de reintroducir la dualidad que posibilita la coexistencia de creencias u otras cogniciones contradictorias sin que surja tal contradicción. Los autores que por el contrario rechazan este modelo interpersonal, se ven obligados a renunciar al papel de la intención en el proceso de autoengaño, dando lugar a teorías *deflacionarias*.

Así pues, dado el escaso avance logrado al abordar el problema mediante la definición ofrecida por el diccionario, ensayaré una primera definición tosca a partir de los rasgos más generales del fenómeno, en la que podrían reconocerse la gran mayoría de teorías.

En este primer acercamiento tentativo, el autoengaño quedaría definido como la *autoinducción de una creencia falsa*. No exigiremos, al menos por el momento, ni intención ni consciencia de engañarse y en un caso típico, el elemento que desencadena este proceso es el dolor, displacer o angustia que producen una creencia, evidencia, situación o pensamiento previos. En primer lugar, ha de señalarse que quien se autoengaña tiene en ese displacer un *motivo* para ello. La gente no trata de autoengañarse con respecto a todo, ni respecto de todas las cosas que no se adecúan a sus deseos, de las que le son dolorosas, o de algunas cosas de modo aleatorio,

sino que parece tratarse de casos muy concretos. El sujeto, en la mayoría de los casos considerados como ejemplos de autoengaño, trata de evitar seguir manteniendo una creencia que le está produciendo un dolor o malestar en cierto modo insoportable[7]. Concretamente, pese a que su evidencia apunta de un modo más claro, esto es, apoya de modo más fuerte la verdad de que, por ejemplo, *p* era el caso, el sujeto acaba manteniendo —de un modo que trataremos de explicar— la creencia de que *no-p* es el caso. Por tanto, es la tendencia a evitar el dolor que la verdad de *p* le produciría junto con el deseo de que *no-p* sea el caso, lo que conduce al sujeto a creer aquello que, al menos en principio, estaba mucho menos apoyado por la evidencia disponible. Esta propensión a formarse una creencia en la dirección a la que apunta el deseo acerca el autoengaño a otro fenómeno conocido en la literatura filosófica como "pensamiento desiderativo", aunque conviene distinguir ambos fenómenos, pues presentan importante diferencias. Veremos esto más adelante.

AUTOENGAÑO DIRECTO Y AUTOENGAÑO RETORCIDO

Un aspecto sorprendente a primera vista y relevante del autoengaño es que no siempre está dirigido hacia lo que el sujeto desea y por ello las instancias de autoengaño pueden ser clasificadas, siguiendo a Mele, en dos grandes grupos: autoengaño directo (*straight self-deception*) y autoengaño retorcido (*twisted self-deception*). El propio Mele explica que, aunque *bent* ("torcido" o "indirecto") hubiese sido el antónimo más claro de *straight*, el

[7] Decimos "en la mayoría", porque veremos a continuación que hay distintos tipos de autoengaño.

calificativo de "retorcido" se debe a cuestiones estilísticas y ni es peyorativo ni señala una patología en el sujeto. [Mele (2001), pp. 4-5; cap. 5]

El "autoengaño directo" sería aquel en el que el sujeto abraza una creencia en contra de la evidencia y cuyo contenido coincide con el contenido de su deseo, como la madre que, pese a que la evidencia de la que dispone favorece o incluso apoya concluyentemente la creencia de que su hijo toma drogas, su deseo de que esto no sea el caso hace que rechace esta evidencia y crea efectivamente que no las toma.

Por el contrario, el "autoengaño retorcido" sería aquel en el que el sujeto, en contra de la evidencia que posee y le lleva a creer que *p* es el caso, pasa a creer, en contra además de su deseo de que efectivamente *p* sea el caso, que *no-p* lo es. Un ejemplo clásico de esto que se repite en la literatura acerca del asunto que nos ocupa sería el que describe al esposo celoso e inseguro que, pese a la muy débil evidencia que puede aportar de que su esposa le es infiel y pese a su deseo de que ésta no lo sea efectivamente, cree finalmente que es engañado por ella. Otro caso típico sería el del paciente que pese a la poca evidencia que apoya el diagnóstico de que padece una grave enfermedad y pese su deseo de estar sano, cree que los médicos no son capaces de detectar algo serio que padece. Es discutible si estos casos de autoengaño retorcido no cruzarían ya la línea de lo patológico, constituyendo un caso de celos enfermizos o patológicos el primero y un cuadro de hipocondría severa el segundo.

Hemos podido ver que parece que hay un fenómeno que nos ofrece problemas si queremos entender cómo ciertos sujeto pueden abrazar

creencias contradictorias (incluso con el agravante de que parece que una es sostenida por la otra) a la vez que seguimos considerándoles como seres racionales. Al acercarnos al problema aparecen paradojas y se nos presentan muchas preguntas acerca de la naturaleza del autoengaño: ¿Qué es el autoengaño? ¿Es posible? ¿Cómo es posible? ¿Es racional?... y junto al autoengaño, hemos visto algunas nociones que se entrecruzan con nuestro tema principal, —haciéndolo aun más confuso—, pero que son totalmente necesarias para ubicar con exactitud nuestro tema central, y que en cualquier caso nos van a ser totalmente imprescindibles para exponer y comprender algunas posturas que veremos a continuación. Lo que sí parece claro es que el asunto no va a ser cerrado fácilmente, como sugiere un hecho muy curioso con el que he topado al indagar sobre la literatura disponible: Tras leer varios artículos y libros, un artículo llamó mi atención: "Self-Deception Needs No Explaining". Ahora parecía que, después disputas inacabadas durante décadas, según Herbert Fingarette el autoengaño no necesitaba de ninguna explicación especial... Lo cierto es que al ver que precisaba de doce páginas para explicar que el autoengaño no necesitaba explicación, sospeché algo... al leer el artículo se iba confirmando poco a poco que Fingarette trataba de explicar el fenómeno —y no era mucho más convincente que otros— para acabar por percatarme de otro detalle: El artículo tiene fecha de 1998, y el señor Fingarette había publicado un libro en 1969 titulado *Self-Deception*. Resulta curioso que algo que no necesita explicación le lleve a uno al menos 30 años de su vida.

Mi propósito en este trabajo no es, obviamente, discutir minuciosamente todo lo que se ha dicho acerca del autoengaño; ni siquiera podría establecer un catálogo que contuviese la mayor parte de los argumentos que se han esgrimido en los diversos intentos por esclarecer la etiología y dinámica de este paradójico fenómeno. Esta tarea se antoja, si no ya imposible, sí tediosa y engorrosa en su manejo, ya que la cantidad de artículos que han sido publicados durante los últimos 50 años es ingente. Por ello, mi tarea se centrará en agrupar y evaluar críticamente los principales aportes en distintos apartados.

No obstante, una tarea previa se nos impone: he de tratar de deslindar el autoengaño de otros fenómenos similares, con el objeto de aislar el problema y evitar así equívocos y discusiones vanas. Por esta razón, voy a dedicar la primera parte del trabajo a analizar distintos fenómenos cercanos al autoengaño. Esta labor no sólo constituirá un ejercicio negativo que sirva para aclarar qué no es autoengaño, sino que en el análisis crítico del material iré esculpiendo las principales características del autoengaño.

Posteriormente, en la segunda parte analizaré algunas propuestas del psicoanálisis, ya que esta teoría nos brindó las herramientas teóricas que han permitido posteriormente a muchos autores desbrozar el terreno para tratar con un fenómeno de carácter cuasi-contradictorio. Esa línea supuso un estímulo para la mayor parte de las explicaciones alternativas posteriores; aun cuando renieguen de la distinción freudiana Yo/Ello/Súper-Yo, suelen incorporar alguna de sus ideas, como la

escisión de la mente, los subsistemas, la atención selectiva, la represión o el inconsciente. Veremos en esta parte también las críticas sartrianas y los primeros planteamientos ya directamente sobre el autoengaño.

En la tercera parte entraremos en el asunto de modo más directo. Trataré de conjugar una exposición en orden cronológico con una de tipo sistemático; aunque soy consciente de las limitaciones que puede suponer esta elección, me ha parecido preferible a una exposición meramente histórica, que sería más fiel en cuanto a la visualización de la línea de nuevas hipótesis, su autoría, así como las consecuentes deudas, pero daría lugar a un batiburrillo enormemente confuso; por otro lado, una exposición sistemática tiene una clara ventaja clasificatoria pero no está libre de dificultades al perder la línea histórica y no dejar claro a quién está respondiendo cada autor o de dónde saca tal o cual idea. He intentado, en la medida de mis capacidades, subsanar algunas de esas deficiencias.

Analizo las propuestas de los dos grandes grupos de teorías, intencionalistas y no-intencionalistas, a través de distintos autores y problemas en cada uno los bloques, subrayando sus puntos de encuentro y desencuentro. A continuación, presento otras hipótesis explicativas más escépticas, que abarcan desde las explicaciones sociologistas hasta las eliminativistas.

Dedico la cuarta parte a los aportes que se han realizado desde las ciencias. Concretamente, algunos experimentos que ha llevado a cabo la psicología social para intentar comprobar la existencia de casos empíricos de autoengaño, y algunos apuntes desde la teoría biológica de la evolución.

En la quinta parte ofrezco ya mi propia visión del asunto. En primer lugar, pongo sobre el tapete algunas consideraciones de partida acerca de la naturaleza de la creencia, la evidencia y la formación de creencias necesarias para la discusión. Posteriomente analizo en qué consistiría el autoengaño, como *proceso* y como *estado*, y someto a crítica las distintas propuestas que se expusieron en la segunda, tercera y cuarta parte. A continuación examino las condiciones bajo las cuales atribuimos a autoengaño y, finalmente, se exponen las conclusiones.

PRIMERA PARTE

I - PROBLEMAS AFINES

I.1 - Ceguera intelectual

La ceguera intelectual (*intellectual blindness*) tiene lugar cuando un sujeto, por el fuerte deseo de que *p* sea el caso, no "ve" la evidencia que desacredita *p*, quedando así la evidencia total adulterada y favoreciendo con ello finalmente la formación de la creencia de que *p* es el caso. Aunque el sujeto se forma una creencia en virtud de una evidencia sesgada, no constituiría un caso genuino de autoengaño porque ni hay intención de que así ocurra, ni el sujeto es tampoco consciente de que hay otras alternativas igualmente o más apoyadas por la evidencia. Como veremos más adelante al caracterizar y distinguir la debilidad de la voluntad del autoengaño, Aristóteles definió de tal modo la debilidad de la voluntad, que el sujeto parecía estar cegado por el deseo o la pasión, y por tanto, a nuestro juicio padecería más bien ceguera intelectual [*EN*, 1147a 15-25].

Es cierto también que generalmente la causa de la ceguera intelectual es el propio deseo de que *p* sea el caso. En otras ocasiones, la causa se halla en alguna hipótesis o creencia previa que conforma la evidencia que va

recogiendo el sujeto, quedando cegado por su propia teoría. En cualquier caso, al no haber ni consciencia, ni intención de alcanzar una creencia errónea, ni siquiera formación de una creencia contra el peso de la evidencia que efectivamente maneja el sujeto, no podríamos hablar de autoengaño. Esto sin embargo, ha de matizarse del modo siguiente: aun cuando el sujeto que se halle cegado transitoriamente no presente intención ni voluntad en ese momento, no es totalmente claro que no sea responsable de su acción, ni que su ceguera no responda a un proyecto de autoengaño de mayor envergadura. Efectivamente el sujeto podría haber tomado voluntaria y conscientemente, poco a poco, el hábito de obviar cierta evidencia desagradable, lo cual daría lugar a posteriores situaciones de ceguera evaluativa, quizá inconsciente.

No negamos, así pues, que la práctica del autoengaño pudiera incluir cierto tipo de ceguera. Sin embargo, los casos de ceguera a los que nos referimos mediante la caracterización "ceguera intelectual" serían, más bien, aquellos en los que es el mero deseo, rabia, dolor, etc. lo que nubla el juicio del individuo. Por tanto, pese a que quien está cegado por su pasión puede hacer juicios apresurados y adquirir creencias erróneas en la mayor parte de las ocasiones —aunque no siempre—, y que la causa que produce este error no es externa, sino que proviene de una pasión del propio sujeto, no se trata de un engaño autoprovocado porque *no es un engaño;* aunque la causa es interna, no es ni consciente ni voluntaria. Se trata más de un *error evaluativo* causado por el propio deseo que de un engaño. En los casos de autoengaño no hay un mero error evaluativo: el sujeto no se

equivoca el evaluar, sino que *se engaña* a sí mismo [cf. Bach (1981), 1981), pp. 351-352; Davidson (1985), p. 104].

I.2 - Pensamiento desiderativo

El pensamiento desiderativo (*wishful thinking*) es descrito generalmente como aquella situación en la que un sujeto, ante la *ausencia de evidencia significativa* que le conduzca a formase la creencia o bien de que *p* es el caso o bien de que *no-p* lo es —este aspecto es crucial— y en presencia del deseo de que *p* sea el caso, finalmente se forma la creencia de que *p* es el caso.

Aunque este matiz no es necesario por el momento, me gustaría decir aquí que el pensamiento desiderativo así definido está, en mi opinión, incorrectamente descrito. La razón está en que en ausencia de evidencia significativa el sujeto no puede tener una creencia sino una —mayor o menor— esperanza, de que la realidad sea tal y como él desea, esto es, sólo puede albergar una esperanza guiada por un deseo. Esta esperanza, que como toda esperanza comparte con la creencia la propiedad de ser disposicionalmente una guía para la acción[8], puede hacer al sujeto más dispuesto a buscar evidencia a favor de aquello que desea que sea el caso, pero no puede dar lugar a una creencia, pues esta demanda evidencia que la apoye, si no concluyentemente, al menos de un modo fuerte. De hecho, me parece que una muy buena traducción de "whisful thinking" sería

[8] Esta característica de poder funcionar como guías para la acción común a creencias y esperanzas, quizá confunda a más de uno, que tal vez ven creencias donde no hay más que esperanzas.

"hacerse ilusiones". Uno no tiene creencia alguna cuando se hace ilusiones; quizá tiene algún indicio, pero son pocos como para que tenga una creencia. Lo que tiene es una esperanza o una ilusión de que algo sea el caso, y por supuesto, las ilusiones son motor de la acción en no pocas ocasiones.

Por tanto, el autoengaño y el pensamiento desiderativo se asemejan en que el sujeto tiene una cierta disposición a buscar de algún modo evidencia favorable a una situación concreta —o a anegar la evidencia contraria— (sesgando la evidencia si es preciso), pero se diferencian en dos cosas:

En primer lugar, el pensamiento desiderativo, aunque manifieste esa búsqueda de evidencia, sólo puede surgir en ausencia de evidencia significativa. En caso de que el sujeto disponga de evidencia significativa las alternativas son dos: o el sujeto se acomoda a aquello que más apoya su evidencia (el caso de creencia normal, no problemático) o el sujeto cree aquello que está menos apoyado por su evidencia, es decir, presenta aquello que Davidson ha llamado "debilidad de la justificación" (*weakness of the warrant*)[9] y, por tanto, se autoengaña.

En segundo lugar, la otra diferencia importante es que mientras en el caso del pensamiento desiderativo la creencia final siempre está orientada hacia el deseo del sujeto, en el caso del autoengaño sin embargo parece que hay ejemplos en los que el sujeto busca evidencia y abraza una

[9] En qué consiste este fenómeno lo explicaremos más adelante (§ I.6).

creencia que no sólo no es apoyada por la evidencia de la que dispone, sino que además es contraria al deseo del sujeto. Es cierto que este otro tipo de autoengaño (que vimos ejemplificado previamente en el esposo de celos enfermizos y en el hipocondríaco) es mucho menos común y que generalmente el autoengañado busca evitar el dolor o malestar refugiándose en aquello que desea que sea el caso, pero el simple hecho de que se produzca este otro tipo de autoengaño ya sería razón suficiente para distinguir el autoengaño del "pensamiento desiderativo", que obviamente siempre orienta las creencias en la misma dirección que los deseos.

De este modo, el pensamiento desiderativo se distinguiría del autoengaño en tanto que requiere la existencia evidencia significativa a favor tanto de *p* como de *no-p,* y se separa de la ceguera intelectual en tanto que el sujeto sí es capaz de ver la evidencia contraria y dañina.

I.3 - Falsa conciencia

Otro concepto que suele ser asociado al de autoengaño es el de falsa conciencia. La expresión "falsa conciencia" (*falsches Bewusstsein*), en muchas ocasiones atribuida a Karl Marx, fue introducida —aunque no definida— en realidad por Friedrich Engels en una carta a Franz Mehring[10] escrita en

[10] Engels, Friedrich (1893), 'Carta a Franz Mehring', 14 de julio de 1893, en K. Marx y F. Engels, Correspondencia, Buenos Aires, Editorial Cartago, 1973, pp. 406-408; *Obras escogidas*, tomo III, pp. 522-527].

1893, es decir, 10 años después de la muerte del propio Marx. En dicha carta, Engels afirma:

> La ideología es un proceso que se opera por el llamado pensador conscientemente, en efecto, pero con una conciencia falsa. Las verdaderas fuerzas propulsoras que lo mueven, permanecen ignoradas para él; de otro modo, no sería tal proceso ideológico. Se imaginan, pues, fuerzas propulsoras falsas o aparentes [Engels (1893), tomo III, p. 523].

En este texto queda claro que Engels considera que la falsa consciencia es la característica principal de toda ideología, aunque no se trata de un mero error en la concepción de algo, no es una "conciencia equivocada". Lo que cabe preguntarse es si la falsedad de las ideologías supone o no autoengaño.

Con respecto al concepto de "ideología", ni podemos ni es pertinente detenerse aquí en un análisis pormenorizado de un concepto que por tantos avatares ha pasado desde que Destutt de Tracy lo introdujera en sus famosos *Éléments D'Idéologie* (1801-1815) con la pretensión de establecer una ciencia de las ideas. Como es bien sabido, posteriormente Napoleón y sus partidarios le otorgaron un sentido peyorativo debido a que Destutt y los "ideólogos" rechazaban la monarquía, promovían una forma de republicanismo que simpatizaba con el modelo político americano, apoyaban el *laissez-faire* y defendían el uso de la razón.

Más tarde, Marx le otorga un matiz más sociológico que psicológico o subjetivo, concibiendo la ideología como el modo en que los individuos de un determinado grupo captan la realidad social. En algunas ocasiones,

aunque los sujetos de tal grupo no son conscientes de ello, el origen y naturaleza de las relaciones sociales que mantienen les son desconocidas, lo cual produce una alienación. Marx expuso en el capítulo dedicado al fetichismo de la mercancía[11] el fenómeno por el cual, al mantener los individuos siempre sus relaciones con otros individuos a través de mercancías, las relaciones entre individuos se perciben como relaciones entre mercancías (incluyendo el dinero y la fuerza de trabajo del propio individuo entre tales mercancías). Así, las relaciones entre los individuos quedarían enmascaradas bajo las relaciones entre mercancías. En efecto, el modo de producción capitalista daría lugar a que, por ejemplo, el valor de la propia mercancía no dependiese ni del trabajo ni de la voluntad de quien la produce, sino de las necesidades de los consumidores, es decir, de la demanda. Los sujetos están por tanto alienados, ya que ni perciben las verdaderas relaciones que mantienen con otros individuos, ni el valor de lo que producen queda fijado ni por su valor de uso ni por el valor de la fuerza de trabajo del productor; los sujetos reemplazan sin apercibirse de ello la maraña social real por una relación fantasmagórica y metafísica. Este tipo de conciencia deformada es precisamente una forma de lo que Engels denomina "falsa conciencia".

Por supuesto hay ideologías de diversa índole pero, según Engels, siempre hay un denominador común: el sujeto tiene una concepción

[11] Marx, Karl (1867), 'Der Fetischcharakter der Ware und sein Geheimnis', en *Das Kapital*, Hamburg, Meissner. Reeditado en Karl Marx y Friedrich Engels, *Werke*, Vol. 23, Berlín, Dietz Verlag, 1968, pp. 85-98. [Edición en castellano: 'El carácter fetichista de la mercancía y su secreto', en *El Capital*, Vol. I, cap. I.4, trad. de Pedro Scaron, Madrid, Siglo XXI, 1975, pp. 87-102].

distorsionada, una consciencia falsa, acerca de la naturaleza de su condición social. De este modo el concepto de falsa conciencia se aplica tanto al mencionado fetichismo de la mercancía como, por ejemplo, a la ideología de clase, entendiendo por ello tanto el intento de las clases dominantes de imponer su ideología a las sometidas como una forma más de dominio, como la conciencia de grupo que tiene todo colectivo sometido que quiere subvertir las relaciones poder.

Dejando a un lado cuestiones tan debatidas como si Marx creía que la ideología influía sobre las condiciones de producción o a la inversa, es decir, si la estructura determina la superestructura o si ambas se codeterminan, o si cree que la ideología supone o no necesariamente una deformación o adulteración[12] en la concepción de la propia dimensión social de los individuos, podemos convenir que, en caso de suponer tal deformación, esta *no es del todo consciente*. Dicho de otro modo: el sujeto es consciente de que mantiene y defiende tal o cual ideología; es consciente también de que se identifica con unos grupos y no con otros, y de que actúa de acuerdo con tal ideología; y quizá es también cierto que esta ideología supone una conciencia equivocada, falsa o deformada acerca de las verdaderas condiciones materiales, de producción y de clase del

[12] Por ejemplo, para una exposición de la doctrina de la ideología en Marx como un proceso de ocultación de contradicciones sociales o de deformación de la imagen social del individuo puede verse Jorge Larrain o Terry Eagleton [Larrain, Jorge (1983), *Marxism and Ideology*, London, Macmillan; Eagleton, Terry (1991), *Ideology*, London, Verso]. Por el contrario, Joseph McCarney ha tratado de poner de relieve el carácter mítico de la noción de ideología atribuida a Marx como proceso defectuoso, deficiente, deformado, ilusorio, engañoso, parcial, etc. [McCarney, Jospeh (2005), 'Ideology and False Conciousness', en *Marx Myths and Legends*.

individuo. Pero este error no es voluntario; no es un error conscientemente aceptado ni autoinducido y, por tanto, no es asimilable al autoengaño. Más bien se parece al *error evaluativo* en unas ocasiones o al *engaño interpersonal* en otras. Un error evaluativo quizá debido a las especiales condiciones de producción del capitalismo, a la propiedad privada, o cualquier otro condicionante histórico; o un engaño inducido por la clase social dominante en busca de un estado de equilibro en el que perpetuar su dominio. En cualquier caso, carece de los rasgos característicos y no satisface las condiciones necesarias para dar lugar al autoengaño.

Podría estudiarse, qué duda cabe, si hay individuos que se adscriban a una ideología u otra de modo voluntario en un intento consciente por evadir los problemas, responsabilidades o conflictos asociados a su clase, religión, etnia o grupo social; por supuesto, en caso de ser el autoengaño posible, en principio uno también podría engañarse con respecto a unos valores o a la pertenencia a un grupo social u otro. Pero este tipo de autoengaño no sólo exigiría que el sujeto se adhiriese a un credo por conveniencia, sino que habría de hacerlo con sinceridad y sabiendo, además, que ese credo es falso o no se corresponde con los valores de su clase. Esto ya no es lo que habitualmente se denomina como falsa conciencia.

I.4 - Disonancia cognitiva

La teoría de la disonancia cognitiva (o "cognoscitiva", como es traducida en algunas ocasiones) fue propuesta por Leon Festinger en 1957 en su obra del mismo título *A theory of cognitive dissonance*. El trabajo de Festinger tenía algunos precedentes claros, como son los famosos experimentos realizados por Solomon Asch (1951), o Irving Janis y Bert King [Janis y King (1954); King y Janis (1956)].

Los experimentos de Asch buscaban demostrar la fuerza que la presión del grupo social ejerce sobre los individuos a la hora de expresar sus opiniones. Como es bien sabido, presentó a un grupo de sujetos una línea estándar que ellos debían comparar con otras tres, expresando en voz alta cuál de esas tres les parecía igual a la primera (Fig.1). El experimentador había acordado en privado con la mayoría de los sujetos que eligiesen una concreta (y errónea); el resto, y sólo el resto, serían los verdaderos sujetos del experimento. Evidentemente, a los individuos objeto del experimento se les colocaba en una posición retrasada, de modo que cuando tenían que responder, ya habían oído la respuesta de la gran mayoría del grupo. El resultado fue que, aunque la mayoría de ellos elegía la correcta, un número elevado (en torno al 33%) se adhería a la opinión del resto.

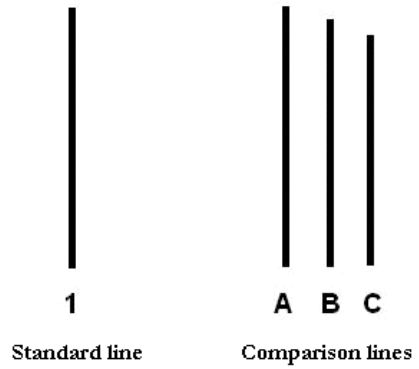

Fig. 1

Por su parte, Janis y King observaron que cuando a un individuo se le obligaba a defender un argumento contrario a su opinión privada, éste mostraba finalmente en muchas ocasiones un acercamiento a la postura a la que había sido forzado. El acercamiento era, desde luego, mucho mayor que en las personas que simplemente oían el discurso o que lo leían. Anteriormente, Herbert Kelman (1953) había intentado ir más allá. Según sus hipótesis, si a alguien se le intentaba convencer para adherirse a una opinión ofreciéndole una suma de dinero, entonces, a mayor suma de dinero, mayor correspondencia entre la opinión privada y aquella a la que había sido encaminado. Sin embargo, los datos que obtuvo en distintos experimentos no confirmaron sus hipótesis; de hecho, la correlación que halló fue la inversa. La explicación que aventuró entonces Kelman ante los inesperados resultados fue que esto era un efecto del autoconvencimiento: quienes habían recibido una menor cantidad de dinero, necesitaban un mayor autoconvencimiento para adherirse a la misma tarea, lo cual producía finalmente una mayor sintonía sincera entre la opinión privada y el curso de acción que habían sido inducidos a llevar.

Es en este contexto en el que Festinger propone su propia teoría del asunto. Según Festinger, nuestras creencias y actitudes son generalmente consistentes, esto es, nuestras actitudes se fundamentan en nuestras creencias y, de algún modo, el entramado evidencia-creencias-acciones es consistente. Sin embargo, esto no es siempre así: en algunas ocasiones hay inconsistencia; esta inconsistencia incomoda al sujeto que la detecta y, en consecuencia, trata de evitarla racionalizándola. De hecho, las inconsistencias son más normales de lo que puede parecer a simple vista:

en pocas ocasiones toda la evidencia está de un lado; pocas son las decisiones que tomamos sin que haya contras, aun cuando sean pocos, débiles o de escasa significación. No obstante, cuando éstos tienen mayor significación, el sujeto siente una incomodidad psicológica. Festinger llama a este estado "disonancia", en lugar de inconsistencia, precisamente para evitar las connotaciones lógicas; efectivamente, aun cuando no haya inconsistencia lógica, el sujeto puede creer que sí, y ésta sería una dificultad más bien psicológica y no lógica. Por idénticas razones, al estado de conformidad entre evidencias, creencias y actitudes lo denomina "consonancia".

Festinger tuvo la oportunidad de comprobar empíricamente cómo se comportaba un grupo del que se esperaba una fuerte disonancia cognitiva[13]. El caso comenzó cuando leyó una noticia en un periódico local en la que se hacían eco de un grupo que esperaba la llegada de un platillo volante que les salvase de un cataclismo inminente. Marion Keech, un ama de casa de Chicago, decía recibir mensajes de unos alienígenas del planeta Clarion a través de escritura automática, en los que se profetizaba el fin del mundo tras una gran inundación el 21 de diciembre; sólo los que creyesen en la profecía y esperasen al platillo se salvarían. Los miembros del grupo se habían deshecho de sus pertenencias, habían dejado sus trabajos, el colegio, a sus parejas, etc.

[13] Festinger, Leon, Riecken Henry W. y Schachter, Stanley (1956), *When Prophecy fails. A Social and Psychological Study of a Modern Group That Predicted the Destruction of the World*, Minneapolis, University of Minnesota Press.

Festinger supuso que, al no cumplirse la profecía, la gente sufriría una enorme disonancia cognitiva; efectivamente, se habían comprometido de un modo muy fuerte con una creencia que iba a ser demostrada como falsa, y el alto coste al que habían adquirido la creencia iba a dificultar que la cambiasen; la otra alternativa era que buscasen apoyo social para su verdad, ya que cuanta más gente apoyase esa creencia, más posible resultaba que fuese verdadera después de todo[14]. Para poder evaluar bien lo que sucedía, decidió infiltrarse en el grupo junto a sus colegas Stanley Schachter y Henry Riecken.

A medida que iba acercándose el día clave, los sujetos evitaban la publicidad del asunto; el día señalado, la gente aguardaba con esperanza la llegada del platillo a medianoche. Sin embargo tras la hora fijada, la gente se fue desilusionando a medida que pasaban las horas, hasta que la señora Keech anunció que había recibido un nuevo mensaje: al ver su fe y esperanza, Dios había salvado a la Tierra de su trágico final.

¿Cuál fue la reacción de la gente? Como Festinger y sus colegas esperaban, el grupo no abandonó su creencia. En lugar de esto, se

[14] Esta idea es, como es evidente, por completo falaz. Se trata, o bien de la conocida falacia *ad populum*, según la cual algo es verdad sólo porque mucha gente lo apoya, o bien en la falacia *ad nauseam*, según la cual el hecho de repetir constantemente un mismo mensaje lo convierte en verdadero. A veces se ha señalado que la razón psicológica que está detrás consiste en que ningún mensaje que no sea verdadero merece tanto esfuerzo en ser repetido. Por otro lado, también podría explicarse a la vista de otra falacia, la falacia *ad ignorantiam*, según la cual todo aquello que no se ha demostrado falso resulta verdadero. En nuestro ejemplo, el hecho de que más y más gente se adhiriese a la creencia en la verdad de la profecía de Clarion minimizaría las posibilidades de que alguien demuestrase su falsedad y, por tanto, demostraría *ad ignorantiam* su verdad.

hicieron más fervientes, adhiriéndose al nuevo mensaje profético y comenzando a dar publicidad al gran acontecimiento en busca de reconocimiento social de este sistema de creencias.

Según Festinger, este suceso demostró que si se adopta con fuerte convicción una creencia susceptible de ser falsada por los hechos y con implicaciones en la acción del sujeto, y su adopción supone además un fuerte compromiso, en caso de ser efectivamente negada por los hechos (y ser esto reconocido por el sujeto), tenderá a producir una disonancia proporcional a la convicción con que se había abrazado y al compromiso en forma de costes que supuso su adopción. El sujeto tenderá a mantener esta creencia manipulando su sistema de creencias hasta hacerlo consonante, y el éxito en su empresa será proporcional, en gran medida, al apoyo social que logre.

Es evidente por otro lado que las cosas no suelen ser blancas o negras; por tanto, casi todas las decisiones o elecciones en nuestra vida presentan pros y contras, es decir, suponen cierta disonancia. Sin embargo, sólo cuando esta disonancia es grande, el individuo siente un malestar psicológico que trata de reducir. El modo en que puede hacerlo es eliminando uno de los elementos que están en disonancia. Por ejemplo, un fumador que ha leído en una revista especializada que fumar es perjudicial para la salud, presentará disonancia entre la cognición de seguir fumando y la cognición de esa información acerca del efecto de esa acción en su salud. El sujeto puede reducir la disonancia de varios modos: puede dejar de fumar, por ejemplo. Pero también puede tratar de contrarrestar la

información que dice que fumar es nocivo: puede buscar información acerca de los beneficios del tabaco hasta el punto de llegar a minusvalorar sus efectos negativos o decirse que no todo el que fuma presenta síntomas de enfermedad relacionados con el tabaco, ni muere prematuramente, etc.

Los esfuerzos que haga el sujeto por reducir la disonancia serán más o menos exitosos en función de la resistencia al cambio que presenten los elementos en disonancia y de las posibilidades de encontrar nueva información consonante. Así, la máxima disonancia que puede resultar es igual a la resistencia al cambio del más débil de los elementos en disonancia. Como es evidente, si la tensión supera la resistencia de ese elemento, ese elemento cambiará, quedando eliminada la disonancia. No obstante, hay situaciones en las que no es posible eliminar la disonancia, en tanto que el sujeto encuentra razones de peso para mantener dos creencias que conllevarían dos cursos de acción disonantes. Actuar de acuerdo con una de ellas sería disonante con la otra creencia y viceversa.

La magnitud de la disonancia está en función de la importancia de la decisión tomada y del atractivo de las opciones alternativas. Esta *decrece* de modo directamente proporcional a la cantidad de elementos consonantes con la decisión tomada y su importancia ponderada, y aumenta en función de la cantidad de elementos disonantes con la decisión (o consonantes con la alternativa no elegida) y su importancia ponderada. Por otro lado, en los casos en los que se fuerza a un sujeto a llevar a cabo una acción, la disonancia disminuirá en función del valor de la recompensa o el castigo. Debido a que una mayor recompensa significa un mayor peso de los

elementos consonantes con la acción a la que uno es forzado o inducido, a mayor recompensa, menor disonancia; a mayor castigo, menor disonancia.

Según Festinger, la hipótesis de Kelman que predecía que a mayor cantidad de dinero ofrecida a un sujeto, mayor conformidad entre su idea privada y aquella a la que se le forzaba, era errónea por una razón. En realidad, lo que ese sujeto mostraría es menor disonancia. Es decir, el sujeto realiza la acción a la que se le induce con mucho más agrado cuanto mayor es la recompensa que se le ofrece, pero eso, no implicaba que su opinión privada hubiera de cambiar en modo alguno. Todavía más: Festinger y Carlsmith (1959), hallaron que la correlación era inversa.

Para demostrar esto, realizaron un experimento en el que tomaron como campo de estudio a 71 estudiantes, a los cuales se les encomendó realizar una tarea premeditadamente aburrida durante una hora y media, que consistía en meter un sacar carretes de una bandeja; a continuación tenían que girar unas clavijas en sentido horario, de cuarto en cuarto durante otra media hora. Después de esto se les contrataba y se les pedía que hablasen con otros presuntos sujetos de experimentación con el objeto de informarles antes de realizar la tarea sobre aquello en que consistía y decirles que era francamente interesante; se hicieron dos grupos: a unos se les pegaba un dólar (cantidad que no era despreciable en aquel momento) y a otros 20 dólares. Además había un tercer grupo de control, a los que únicamente se les pidió su opinión sobre la tarea.

El resultado obtenido apuntaba claramente a que aquellos que habían recibido dinero, consideraban que la tarea había sido mucho más

interesante que el grupo de control. Sin embargo, aquellos que habían recibido menos dinero eran más entusiastas con respecto al experimento.

Tanto Kelman (1953) como Janis y King (1954; 1956) explicaban diferencias conductuales similares en función del autoconvencimiento, esto es, de los mayores esfuerzos en demostrar que aquello que se les invitaba a creer era realmente el caso. Los sujetos que habían recibido menos recompensa necesitaban más de estas autoexplicaciones y racionalizaciones y, finalmente, acababan por convencerse a sí mismos. Sin embargo, Festinger y Carlsmith tenían una explicación alternativa. Para falsar la hipótesis del autoconvencimiento, grabaron a escondidas las explicaciones que los sujetos contratados ofrecían a los nuevos sujetos de experimentación y hallaron que aquellos que habían recibido 20 dólares invertían más tiempo en tratar de convencer del interés del experimento, ofrecían más razones y más convincentes. Esto parecía echar por tierra los argumentos tanto de Janis y King como de Kelman; la única explicación posible era ahora la teoría de la disonancia cognitiva ofrecida dos años antes por el propio Festinger, según la cual, cuánto mayor era la fuerza que presionaba al sujeto para adherirse a una opinión o realizar una acción, menor era la tendencia privada que mostraba en ese sentido.

La razón última es, como se ha dicho anteriormente, que el sujeto no necesita precisamente de ningún tipo de autoconvencimiento, pues el valor la recompensa es directamente proporcional a la cantidad de elementos consonantes para adherirse al curso de acción o creencia

propuesta. Dicho de modo más sencillo: no necesita convencerse de que aquello está bien o es correcto; sea o no sea correcto,

a) lo aceptará de mayor agrado cuando mayor sea la recompensa (le dedicará más esfuerzo) y, además,

b) cuanto mayor sea ésta, menos disonancia le producirá expresar una opinión o seguir un curso de acción distinto al que en principio tendía.

El sujeto que es mejor recompensado, invierte más esfuerzo en realizar la tarea. Pero este esfuerzo no lo invierte para convencerse de nada, pues la recompensa le convence por sí misma: es ella la que al ponerse del otro lado de la balanza deshace cualquier tensión o disonancia posible. Este mayor esfuerzo lo haría porque siente que hay razones para ello; la inversión de tiempo en algo que en principio es vacío o tedioso resulta consonante con la gran recompensa que recibe.

Esta es la interpretación que se ofreció desde la teoría de la disonancia cognitiva a los resultados que arrojaron los experimentos. Cabe preguntarse, no obstante, si la explicación de Festinger es absolutamente incompatible con la ensayada por Kelman o Janis y King. A nuestro juicio, no sería contrario a las tesis de Festinger el que un sujeto, debido a la poca recompensa que recibe presente mayor disonancia y, consecuentemente, encuentre más necesaria la racionalización de su acción. De este modo invertiría más tiempo y esfuerzo que los sujetos con una fuerte recompensa. Somos conscientes de que la réplica más obvia a nuestra

postura es la siguiente: los propios Festinger y Carlsmith no se cerraron a esta posibilidad y la investigaron, hallando como resultado —como hemos visto anteriormente— que era los sujetos mejor pagados los que invertían más tiempo y mejores argumentos. Sin embargo, a nuestro juicio las pruebas que se realizaron para probar esto son imprecisas; ellos evaluaron la racionalización del sujeto observando las cintas grabadas en las que trataban de convencer a otros del interés y entretenimiento que suponían las pruebas. Pero, dado que las grabaciones se efectuaron inmediatamente después de realizar la tarea, ¿no sería acaso posible que los sujetos continuasen manteniendo disonancia y la racionalizasen posteriormente? Todavía más: ¿no es posible que la racionalización sea un proceso privado que nada tiene que ver con la exposición pública de los pros y contras?

En cualquier caso, al margen de estas y otras críticas posibles, no tanto para con los experimentos en sí mismos, cuanto para con su interpretación, lo más importante para nuestro estudio es subrayar la diferencia que hay entre la disonancia cognitiva y el autoengaño. Aunque el sujeto que se autoengaña ha de presentar disonancia cognitiva, no toda disonancia cognitiva supone un caso o intento de autoengaño. Como hemos visto, la disonancia está presente casi en cada cognición diaria. Pocas cosas son blancas o negras, y casi todas nuestras cogniciones (creencias, deseos, opiniones, acciones, actitudes, etc) manifiestan cierto grado de disonancia con otras; el autoengaño, sin embargo, no sólo supone la presencia de creencias, deseos y acciones en disonancia. Más bien, quien se autoengaña tiene evidencia *fuerte* de que algo que le

resultaría profundamente dañino es el caso, y adquiere una creencia en contra del peso total de su evidencia.

Por supuesto, en algunos casos de disonancia, el autoengaño sería una vía que el sujeto buscaría para reducir la tensión psicológica, tratando de eliminar todas las cogniciones disonantes con aquello que desea que sea el caso, aun a riesgo de perder la verdad.

I.5 - Debilidad de la voluntad

> Sí, conozco los crímenes que voy a realizar, pero mi pasión es más poderosa que mis reflexiones y ella es la causante de males para los mortales.
>
> [*Medea*, 1078-1081]

Parece que el autoengaño, tal y como ha sido definido en muchas ocasiones, describe una situación en la que el sujeto está más inclinado a creer —y de hecho cree— la proposición contraria a la que estaría en principio más apoyada por la totalidad de la evidencia disponible para él. Esto acerca el problema del autoengaño al problema de la debilidad de la voluntad.

La debilidad de la voluntad describe la situación en la que un sujeto, que duda entre dos cursos de acción enfrentados y apoyados ambos por distintas justificaciones, una vez considerados todos los argumentos de los que dispone, juzga que uno de los cursos de acción es el mejor y, pese a

ello, escoge voluntariamente la otra alternativa. Nótese que no se trata de que el sujeto considere todos los argumentos que pueden darse, sino que, una vez tomados en cuenta todos aquellos de los que dispone efectivamente y colocados en la balanza, el sujeto se decante por el curso de acción menos justificado.

El término *akrasia* (en griego ακρασία; también ακρατεία o ακρατία), que podemos traducir por "incontinencia" o "falta de control", hace precisamente referencia al hecho de que un sujeto actúe en contra de aquello que, tomando en consideración todos los datos relevantes, considera su mejor juicio. Ya desde la antigüedad ha habido opiniones muy diversas al respecto. Así, por ejemplo, en las tragedias de Eurípides vemos individuos que se comportarían acráticamente:

> [Y] me parece que no obran de la peor manera por la disposición natural de su mente, pues muchos de ellos están dotados de cordura. No; hay que analizarlo de otro modo. Sabemos y comprendemos lo que está bien, pero no lo ponemos en práctica, unos por indolencia, otros por preferir cualquier clase de placer al bien. [*Hipólito*, 377-383]

Sin embargo, parece que Sócrates negó la posibilidad de que un sujeto actuase en contra de lo bueno, salvo por desconocimiento del bien o lo bueno. Según la famosa divisa, "nadie peca, salvo por ignorancia". Fue Platón quien ensayó el primer análisis teórico del asunto en el *Protágoras*, donde por boca de Sócrates defiende la tesis de que un sujeto no puede actuar deliberadamente en contra del que sería su mejor argumento, ni siquiera llevado por la inmediatez y la sensualidad de los placeres. Si actúa poniendo en práctica cosas malas, sólo puede hacerlo por ignorancia, y nunca basándose en el conocimiento.

> […] sabes entonces que muchos hombres no nos creen, ni a ti ni a mí, y que afirman que muchos que conocen lo mejor no quieren ponerlo en práctica, aunque les sería posible, sino que actúan de otro modo […] [*Protágoras* 352 d5-8].
>
> […] resulta absurda vuestra afirmación, cuando decíais que, a pesar de conocer el hombre que las cosas malas son malas, sin embargo las pone en práctica —aunque le sería posible dejar de hacerlo— arrastrado y seducido por los placeres. Y, por otra parte, también decía que el hombre, a pesar de conocer lo que es bueno, no quiere practicarlo por sufrimientos momentáneos, dominado por ellos. Cuán absurdas son esas afirmaciones […] [*Protágoras* 355 a7-b5].

Por tanto parece que Platón se posiciona frente a aquellos que describen una situación en la que "muchos", *pese a conocer lo mejor*, actúan *voluntariamente* (la pasión no les domina completamente, ya que "les sería posible" actuar de otro modo) en contra de lo que consideran mejor, de las cosas buenas. Al igual que defendió Sócrates, según Platón sólo la ignorancia justifica un comportamiento malvado o estúpido.

Aristóteles, sin embargo, dice que la tesis de Sócrates está "en contradicción manifiesta con los hechos" [*EN*, 1145b 28-29] y se propone dar cuenta de ello. Así, concede que hay situaciones en las que un sujeto, aun sabiendo que lo que va a realizar está mal, actúa de acuerdo a ello. Dice,

> [E]l incontinente sabe que obra mal movido por su pasión, y el continente, sabiendo que las pasiones son malas, no las sigue a causa de su razón [*EN*, 1145b 12-13]
>
> […] la incontinencia es contraria a la propia elección, y el vicio está de acuerdo con ella [*EN*, 1151a 6-7]

El incontinente *sabe* que obra mal —dice Aristóteles— pero, en cualquier caso, no es relevante si posee o no verdadero conocimiento, pues en muchas ocasiones aquellos que sólo poseen opiniones no están menos convencidos de la verdad de éstas que otros que poseen conocimiento [*EN*, 1145b 29]. De hecho, según Aristóteles el incontinente actúa en cierto modo cegado por las pasiones; en el momento de elegir olvida aquello que reconocía como lo mejor, conducido por las pasiones a estados similares al sueño, la borrachera y la locura. El incontinente puede aducir que tiene conocimiento en el momento de obrar, pero en realidad cuando habla se parece a un actor de teatro y su conocimiento sería más bien del tipo que tiene "un borracho que musita demostraciones y versos de Empédocles o un estudiante de ciencia principiante que ensarta frases sin saber lo que dice" [*EN*, 1147a 15-25].

En general estas son las posturas que han venido siendo defendidas, aunque de hecho, son muy pocos los autores que, con Platón, nieguen explícitamente la posibilidad de que los individuos puedan actuar acráticamente. Santo Tomás sigue a Aristóteles y coloca también la fuente de la incontinencia en las pasiones y la concupiscencia. Así, el hombre realiza a veces acciones en contra de su mejor juicio, pero no es incontinente porque su acción no es voluntaria al estar generada por una fuerte emoción, como el temor. Pero dado que la concupiscencia *inclina a la voluntad* a desear algo, quien se deja dominar por ella lo hace voluntariamente. [*Summa Theologica*, Parte II, Q.6]. Otros autores no han tratado de un modo directo sobre el tema, aunque de sus escritos pueden

sacarse ciertas conclusiones. Un caso controvertido e interesante es el de David Hume. Por un lado, afirma que a veces elegimos voluntariamente bienes menores aun a sabiendas de que hay otros mayores:

> [N]o es contrario a la razón el preferir la destrucción del mundo entero a tener un rasguño en mi dedo. No es contrario a la razón que yo prefiera mi ruina total con tal de evitar el menor sufrimiento a un *indio* o a cualquier persona totalmente desconocida. Tampoco es contrario a la razón el preferir un bien pequeño, aunque lo reconozca menor, a otro mayor, y tener una afección más ardiente por el primero que por el segundo. [Hume (1739-40), p. 563, SB 417]

Parecería por tanto que Hume defiende la posibilidad de que actuemos de modo incontinente, esto es, que ante dos posibles cursos de actuación, sigamos voluntariamente aquel que sabemos que comporta un bien menor. Sin embargo, algunos han visto a Hume como representante de un escepticismo práctico [Fleming (2006), pp. 1-7]; según esta interpretación, Hume negaría la existencia de *razones* para actuar, pues son las pasiones las que mueven nuestras acciones. Pero entonces, si la acción no se fundamenta en *razones*, ¿puede haber acciones irracionales? o ¿puede haber algo tal como "el mejor juicio" o "el más racional"? Decíamos que son las pasiones las encargadas de generar acciones; pero las acciones no pueden ser consideradas contrarias a la razón, *irracionales,* por seguir a las pasiones. En primer lugar, las pasiones no pueden ser irracionales porque, según el célebre pasaje, la razón ha de ser esclava de las pasiones, y en ningún caso le es posible oponerse a ellas. A una pasión sólo puede oponérsele otra pasión contraria. En segundo lugar, la irracionalidad sólo puede predicarse de las ideas que, en tanto que copias de las cosas existentes, pueden ser imperfectas. Las pasiones no son copias, sino "existencias", y por ello no

son ni verdaderas o falsas, ni racionales o irracionales. [Hume (1739-40), pp, 561-562, SB 415].

> […] una pasión deberá estar acompañada de algún falso juicio para ser irrazonable; e incluso, para hablar con propiedad, no es la pasión lo irrazonable, sino el juicio. [Hume (1739-40), p. 563, SB 417]

Una pasión puede resultar irracional sólo por dos motivos. A veces una pasión se ha fundamentado en la suposición de existencia de algo realmente no existente y, por tanto, al ser el juicio que acompaña a la pasión erróneo, la convierte en irracional; otras veces la pasión escoge medios insuficientes para lograr sus fines. En ambos casos la pasión es contraria a la razón, y es entonces —y sólo entonces— cuando la pasión cede su lugar a la razón.

Estaríamos por tanto ante la siguiente situación: Hume no admite *razones* para actuar; actuamos según nuestras pasiones y la razón no puede ser contraria a ellas; es de hecho esclava de las pasiones. Esto parece cerrar la posibilidad de hablar de acciones incontinentes en tanto que acciones en contra de la razón, esto es, elimina la irracionalidad del fenómeno. Sin embargo Hume afirma que a veces actuamos sabiendo que nos guiamos por un bien menor. Pero menor ¿en qué sentido?

Creo que hay al menos tres vías para responder a esto. Patrick Fleming (2008) ha ensayado una de ellas. Para Fleming, hay tres principios en Hume que, en conjunción, serían capaces de salvar la aparente contradicción de las tesis humeanas: el *principio de la pasión predominante*, el

principio de la costumbre en las pasiones, y el *principio de la idea más particular y determinada.*

Según el *principio de la pasión predominante*, cualquier emoción que acompañe a la pasión predominante se convierte en ella, reforzándola y dándole más vivacidad, aun cuando sus naturalezas sean contrarias. [Hume (1739-40), p. 567, SB 419]. El principio de la *costumbre en las pasiones* muestra que la costumbre de realizar una acción genera tanto facilidad al realizar tal acción como una tendencia a ella. [Hume (1739-40), p. 571, SB 422], Por último, el *principio de la idea más particular y determinada* nos dice que donde sea que una idea de bien o mal adquiera mayor vivacidad, generará una pasión más violenta y estimulará a la imaginación [Hume (1739-40), p. 573, SB 424].

Patrick Fleming toma un ejemplo de Christine Korsgaard para ilustrar su defensa de la supuesta respuesta de Hume a la pregunta por la debilidad de la voluntad. Según el ejemplo, Harry tiene una enfermedad y necesita ponerse una inyección. No tiene la *costumbre* de darse inyecciones contra enfermedades mortales, pero tiene un enorme deseo de vivir una larga vida. Sin embargo le horrorizan las agujas. Su deseo es ponerse la inyección, pero llegado el momento, el miedo le atenaza y no acepta la inyección. No obstante, poco después reconoce que debió haberse inyectado. ¿Cómo podría dar cuenta Hume de un caso tal? Según Fleming, Harry siente el deseo de ponerse la inyección cuando observa desde la distancia los beneficios. Ahora bien, cuando le acercan la aguja, el inminente dolor se le aparece de un modo tan claro que, por el *principio de*

la idea más particular y determinada, éste adquiere mayor fuerza motivacional. Esta pasión se hace predominante y absorbe la pasión de vivir una larga vida por el *principio de la pasión predominante.* Finalmente, cuando la idea de dolor no está presente, la idea de vivir una larga vida —a la que está más acostumbrado— vuelve a él, por el *principio de la costumbre en las pasiones.* [Fleming (2006), p. 11].

Sin negar el valor de esta reconstrucción que realiza Fleming, nos parece que aun queda algo por explicar. Si nos atenemos a la reexposición de Fleming, ¿Hume admite o no la posibilidad de la incontinencia? De hecho, Hume podría haber admitido el análisis explicativo anterior sin conceder que por ello Harry fuese incontinente. Porque, ¿en algún momento actúa Harry en contra de su mejor juicio, como exige la incontinencia?

Una posible explicación que mostraría un Harry incontinente sería la siguiente: Hume admite que un bien que racionalmente consideraríamos menor puede ser perseguido sin convertir a la acción que lo persigue en irracional, ya que la superioridad o inferioridad racional de un estado de cosas no tiene incidencia causal en la persecución de la acción, pues sólo las pasiones —y no los juicios racionales— tienen tal fuerza causal. Harry actuaría según sus pasiones, unas veces en consonancia con su juicio racional, otras en su contra. Aunque su juicio racional no prescriba cómo ha de actuar, siempre que Harry actuase en una dirección distinta, sería incontinente, pero este tipo de incontinencia en modo alguno sería irracional o indeseable. De hecho, sería perfectamente racional, y esta

racionalidad vendría dada por el hecho de que los medios que utiliza Harry sirven perfectamente para alcanzar el fin que busca. La pasión generada por el miedo a la aguja es efectivamente un medio eficaz para conseguir su fin: apartar el brazo y evitar así el dolor del pinchazo. En cada momento, Harry actuaría del modo que estima oportuno en función de sus pasiones, que se revelan como útiles y racionales precisamente por ser adecuadas para lograr sus fines, ya que sus fines en presencia de la aguja "amenazadora" son bien distintos al fin que tiene cuando no está presente dicho elemento hostil. Por tanto, la acción de Harry no sería incontinente, y teniendo a la vista además que según Hume no hay ninguna razón que justifique una acción, y que ésta debe fundamentarse en pasiones, la acción de Harry no sería, ni podría ser tampoco, irracional.

Finalmente puede ensayarse una interpretación distinta de la acción de Harry que siendo compatible con los textos humeanos y sin traicionar sus tesis, muestra un Harry incontinente y a la vez irracional: la acción que lleva a cabo podría ser irracional por estar basada en una pasión irracional, ya que ésta *no* utilizaría los medios adecuados para el fin que busca, a saber, vivir largo tiempo. Y la razón de que no use los medios adecuados sería, ahora sí, el temor por el dolor, temor —pasión— que en presencia de la aguja se le aparece con una enorme viveza. Esta pasión se revela por tanto como irracional y errónea, y genera que la acción de Harry sea incontinente, pues de no haber participado de esa pasión, habría actuado, consideradas todas las alternativas de las que disponía, siguiendo un curso de acción muy distinto y conforme a su deseo de tener una larga vida.

Por lo que respecta a las teorías sobre la *akrasia* de los últimos 40 años el referente inexcusable es Donald Davidson. Todos los autores hacen referencia a sus ensayos, y no son pocos los que definen sus posiciones con respecto a la suya.

Hemos de comenzar diciendo que Davidson considera que la debilidad de la voluntad o *akrasia* es interesante porque supone un problema de racionalidad, como lo hacen el autoengaño, la ceguera intelectual, el pensamiento desiderativo o la debilidad de la justificación. Es importante señalar que la irracionalidad no supone una total ausencia de racionalidad (arracionalidad), sino un fallo dentro de la propia racionalidad. [Davidson (1982), p. 169]. Al tratar de comprender a los demás, suponemos que no tienen creencias inconsistentes o extravagantes, y se comportan en cierto modo como nosotros, esto es, que son racionales. Esto es lo que Davidson llama, siguiendo a Quine, el *Principio de Caridad*. No nos parecen extraños distintos comportamientos siempre que permanezcan dentro de unos márgenes, pero cuando se salen de éstos, algo chirría y se nos aparece como irracional. La irracionalidad es, por tanto, una caída o ruptura de los patrones internos de racionalidad.

Al enfrentarnos a estos problemas de racionalidad, somos conscientes de que surge una doble paradoja: rechazar estos fenómenos como totalmente incomprensibles e inexplicables supone obviar su problematicidad; además, parece que son demasiado comunes y cotidianos como para adscribir irracionalidad o patologías a quienes presentan tales estados. Sin embargo, si tratamos de racionalizar el problema para

comprenderlo, disolvemos su complejidad e irracionalidad. [Davidson (1985), p. 199].

Con respecto al problema de la incontinencia o debilidad de la voluntad, Davidson cree que además de la teoría de Platón, que negaba su existencia ("doctrina de la pura racionalidad" [Davidson (1982), p. 175]), hay otras dos hipótesis principales: por un lado está la teoría de Aristóteles, según la cual, cuando un sujeto actúa acráticamente debe hacérsele de algún modo inconsciente en el momento de elección, debido a una fuerte pasión, el que consideraría *en otras condiciones* su mejor juicio. Por otro lado, otros autores como Richard M. Hare, han visto este problema como una especie de fuerza extraña que sobrepasa nuestro mejor argumento y nuestra propia voluntad trastornándonos e imponiéndosenos [Hare (1963), p. 71]. Hare cita como ejemplo las palabras de Medea "una compulsión desconocida, muy a mi pesar, me acosa" de la tragedia de Eurípides, por lo que Davidson denomina esta caracterización de la debilidad de la voluntad como el *Principio de Medea* [Davidson (1982), p. 175]. En el momento de actuar, una pulsión *extraña y externa* se hace con el control del sujeto y le empuja —involuntariamente, por tanto— en contra de lo que desea.

En opinión de Davidson ambos describen situaciones posibles, pero no son casos de incontinencia, pues ésta supone que un agente, una vez consideradas todas las opciones de las que dispone, juzgue que de entre varios cursos de acción uno es mejor y, sin embargo, siendo *consciente* de ello en el momento de actuar, *voluntariamente* actúe en contra de ese curso

de acción (o simplemente escoja uno menos apoyado). La teoría de Aristóteles presenta a un individuo que *no es consciente* del mejor juicio, mientras el *Principio de Medea elimina la voluntad* del proceso. [Davidson (1982), p. 176].

Según Davidson, un caso genuino de comportamiento acrático exige la satisfacción de las siguientes cláusulas:

- P1: Si un agente quiere hacer *x* más de lo que quiere hacer *y*, y se cree libre para hacer o bien *x* o bien *y*, entonces intencionalmente hará *x* si hace intencionalmente o bien *x* o bien *y*.

- P2: Si un agente juzga que sería mejor hacer *x* que hacer *y*, entonces quiere hacer *x* más de lo que quiere hacer *y*.

- P3: Hay acciones incontinentes.

P1 y P2 implican juntos que si un agente juzga que sería mejor para él hacer *x* que hacer *y,* y se cree libre para hacer o bien *x* o bien *y*, entonces intencionalmente hará *x* si hace *x* o *y*. Esto parece hacer imposible P3, esto es, que haya acciones incontinentes [Davidson (1970), p. 39]. Sin embargo, Davidson no quiere renunciar a ninguno de ellos.

> Mi estrategia básica no es la de intentar presentar un alegato inobjetable a favor de P1-P3 […] Más bien lo que *espero demostrar es que no se contradicen entre sí* y, por tanto, que no debemos abandonar ninguno de ellos. [Davidson (1970), p. 40. El subrayado es mío]

No resisto la tentación de decir lo siguiente. ¿Es acaso la no contradicción lógica entre los tres principios condición suficiente para no abandonar ninguno de ellos? Sin duda es condición necesaria, pero sólo habremos recorrido la mitad del camino. Supongamos que yo quiero hacer una descripción acerca de la composición de los planetas del sistema solar, y establezco: Mercurio está constituido de carbón, Venus de queso verde, Marte de grafito… Seguramente no hay ningún tipo de *contradicción lógica* entre estas situaciones, pero esto no es una razón para que no las abandone; no es suficiente para que abrace las hipótesis sin reparo. La no-contradicción lógica es condición necesaria, pero no suficiente. Hay otras condiciones igualmente necesarias, como ser empíricamente adecuadas. Davidson supone, por ejemplo, que P3 es cierta, dando por sentado que la incontinencia es un hecho. Esto es algo con respecto a lo cual muchos somos escépticos, por lo que sentar aquello que está por demostrar parece, cuando menos, cercano a una petición de principio. El propio Davidson advierte:

> No hay prueba de que existan tales acciones; pero me parece absolutamente cierto que existen. [Davidson (1970), p. 46]

Concedamos, provisionalmente y por mor del argumento, que Davidson está en lo cierto y que la debilidad de la voluntad resulta ser un hecho indubitable que hemos de explicar. El mayor problema se encuentra en P2, que ha sido objeto de numerosas malinterpretaciones e intentos de reformulación. Ninguno de éstos, sin embargo, ha arrojado luz sobre el problema. Según Davidson esto ocurre porque P1 y P2 se derivan de una tesis muy persuasiva acerca de la acción intencional: cuando una

persona actúa con una intención, en primer lugar le da un valor positivo a algún estado de cosas y, creyendo (sabiendo o percibiendo) que una acción producirá o realizará ese estado de cosas que valora, actúa en función de ese deseo o creencia. [Davidson (1970), p. 49]. Por tanto, las creencias y deseos *causan* una acción. Añade Davidson que siempre que un conjunto de razones *r* constituya *la razón* para que hagamos *x*, *r* causará que hagamos *x*, pero que *r* cause *x* no implica que *r* sea una razón para *x*. [Davidson (1970), p. 60]. Lo que nos está diciendo es que hay acciones que están causadas por conjuntos de razones que no constituyen una verdadera razón para realizar dichas acciones. Los casos que postula P3, esto es, los casos de debilidad de la voluntad serían aquellos en los que el conjunto de razones *r* que supone la causa de la acción *x*, no es una verdadera razón para *x*. Sin embargo, anteriormente Davidson ha dicho que si algo constituye una razón para hacer *x*, entonces causará *x*. Pero entonces, dado que el mejor argumento constituye siempre una *razón,* esto es, tomados todos los argumentos en consideración el conjunto razones *r* supondrá una *razón* para hacer *x*, entonces *r causará* que yo haga *x,* lo cual entra en flagrante contradicción con la observación anterior —y de paso con la posibilidad de existencia de ejemplos de incontinencia— pues, si yo tengo un conjunto razones *r* que supone una *razón* para hacer *x*, *r causará* que yo haga *x*… y esto, aún cuando haya deseos u otras razones que me empujen a hacer *y*. Dicho brevemente: ¿Cómo defenderá Davidson que un sujeto pueda comportarse incontinentemente y hacer *y* en lugar de *x*, si ha establecido que el conjunto *r* resultante de tomar en consideración todas las opciones ha de ser *razón*, y por tanto ha de causar *x*?

La única salida viable sería establecer que *r* no tomaba en cuenta todas las consideraciones. Este es el camino que propone Davidson. Sin embargo esto no supondría más que un parche que retrasaría sólo momentáneamente la anterior dificultad (pues habría, en último término, un conjunto más comprehensivo que tomase todas las razones en consideración, como realmente exige la definición de la *akrasia*) y finalmente tendría que hacer frente al problema de cómo un sujeto cuyo conjunto de razones —siendo ya omnicomprehensivo— *es razón* para hacer *x*, y en cambio no tiene la fuerza causal suficiente para que haga *x* —como veíamos que afirmaba anteriormente—, permitiendo que el sujeto haga más bien *y* o *z*. Éste es el camino explicativo que toma Davidson mediante lo que denomina *Principio de continencia,* que establece que un sujeto racional ha de tomar en cuenta todas las razones disponibles para evaluar qué curso de acción es conveniente [Davidson (1970), p. 61]. Así realiza la distinción entre los conjuntos de razones que examinan *prima facie* los cursos de acción y los que toman de forma integradora el conjunto de razones siguiendo el principio de continencia. Como hemos dicho, esto no resuelve el problema. En todo caso, Davidson tendría que explicar por qué a veces se producen exámenes parciales del conjunto de razones (¿quizá por fuerza de los impulsos y deseos?) y su propuesta se acercaría de modo irreversible a la aristotélica, en la que el sujeto no tomaba en consideración todas las razones (es decir, no había *consciencia* del mejor argumento; o dicho de otro modo: el mejor argumento sólo lo sería *prima facie*).

Quizá lo más interesante de la teoría davidsoniana consiste en el divorcio de la incontinencia y la moral, en tanto que muchas veces se ha presentado la incontinencia como asociada a los placeres sensuales, la concupiscencia y la caída en la tentación. Sin embargo, a veces, pese a que los placeres sensoriales nos ofrecen la mejor argumentación, los patrones morales o sociales establecidos hacen que nos apartemos de nuestro mejor juicio [Davidson (1970), p. 48]. Davidson sitúa así el problema de un modo claro en el ámbito de la filosofía de la acción, en lugar de la filosofía moral.

Por su parte, Kathrin Glüer ha dado una versión peculiar del asunto. Pese a que sigue a Davidson en gran parte de su argumentación, sus conclusiones son bastante diferentes. De hecho se presenta como escéptica con respecto a la noción de debilidad de la voluntad o *akrasia*. En realidad, para Glüer hay un problema de fondo en los casos que etiquetamos como comportamientos incontinentes o acráticos: un concepto de evaluación restringido, estrecho. Esto se debe, según Glüer, principalmente a dos razones. Unas veces al realizar una acción se toman en cuenta sólo las razones para un beneficio o placer a corto plazo, mientras en otras se tienen en cuenta las consecuencias a largo plazo, o cualquier otro motivo. En cualquiera de estos u otros posibles casos, al no encontrarnos con un sujeto que esté considerando todas las cosas y alternativas de las que dispone, su comportamiento no puede ser catalogado como acrático en absoluto. En otras ocasiones, el individuo niega que su mejor juicio no apunte a la acción que realiza, cometiendo un error al evaluar la evidencia "y, por tanto, autoengañándose." El error

consistiría en una mala interpretación de los deseos y su fuerza. [Glüer (2003), pp. 81-82].

Con respecto a esto tenemos que apuntar dos cosas: en primer lugar, hay otras situaciones en las que el sujeto no juzga correctamente la fuerza de los deseos, y sin embargo no está autoengañado. Los casos de *ceguera intelectual* o *pensamiento desiderativo* encajan en este modelo. De hecho, la propuesta aristotélica puede ser vista como un caso de *ceguera intelectual*, en la que la fuerza de un deseo turba y nubla nuestro recto juicio, haciéndonos olvidar algunas de nuestras razones. En segundo lugar, el autoengaño, de ser posible, no consistiría en un error al evaluar la evidencia, sino en que una vez evaluada —sea acertada o equivocadamente, poco importa aquí— el sujeto se forma una creencia en contra de lo que la evidencia apoya, cometiendo lo que Davidson llama, por analogía, *debilidad de la justificación*. Davidson ha señalado que el *autoengaño* y la *akrasia* son dos fenómenos que, aunque cercanos, son perfectamente distinguibles. A veces, aparecen juntos y reforzándose mutuamente, pero no son lo mismo. [Davidson (1985), p. 201]

Por último examinaremos la postura de Richard Holton. A él le debemos la separación de lo que, principalmente por parte de la corriente anglosajona, había sido identificado como sinónimos: la debilidad de la voluntad y la *akrasia*.

> Lo que en los círculos filosóficos anglosajones se denomina el problema de la debilidad de la voluntad se refiere a aquello que preocupó a Sócrates: el problema de cómo un agente puede elegir tomar el que cree que es el peor curso de acción, sobrepasado por la

> pasión. La expresión inglesa no traería, o al menos no en principio, este tipo de caso a la mente, sino ejemplos tales como la demora, la indecisión, falta de valentía moral y un fallo al poner en marcha planes. La palabra griega '*akrasia*', por otro lado, significa '*ausencia de control*', y esto realmente sugiere ejemplos del tipo socrático. [Gosling (1990), p. 97][15]

Holton, en cambio, reserva el término *akrasia* para caracterizar el problema clásico sobre la posibilidad de que un sujeto, una vez consideradas todas las cosas, juzgue un curso de acción como el mejor y, sin embargo, siga otro. La debilidad de la voluntad consistiría, en cambio, en una *revisión injustificada de las propias intenciones*.

Siguiendo a Michael Bratman [Bratman (1987)], Holton sostiene que las intenciones son algo distinto de las creencias y deseos —aunque el sujeto tome éstos en cuenta para su formación— y consisten en una decisión de realizar una acción. Éstas no siempre están guiadas por deseos o creencias previos, pues incluso en casos en los que no hay razones de peso a favor de ninguna alternativa, como el célebre ejemplo de Buridan, si el sujeto desea superar la indecisión, ha de tomar una decisión acerca de cómo actuará, esto es, ha de formarse una intención.

En cualquier caso, las intenciones se forman o bien para ordenar la vida intrapersonal de un sujeto y vencer las posibles inclinaciones que en el

[15] «What in Anglo-Saxon philosophical circles is called the problem of weakness of will concerns what worried Socrates: the problem of how an agent can choose to take what they believe to be the worse course, overcome by passion. The English expression would not, or at least not primarily, bring this sort of case to mind, but rather such examples as dilatoriness, procrastination, lack of moral courage and failure to push plans through. The Greek word '*akrasia*', on the other hand, means 'lack of control', and that certainly suggests the Socratic sort of example.»

momento de realizar la acción le puedan tentar o bien como método de coordinación interpersonal. Supongamos que alguien desea pintar la puerta de su casa, y no tiene preferencia por un color, esto es, no tiene razones para pintar una puerta de un color u otro; no obstante, espera la visita de unos amigos que sabe que sólo reconocerán la casa por el color de la puerta. Necesita, por tanto, tomar una decisión, ya que ha de ir a comprar la pintura ese mismo día. [Holton (1999), p. 244].

La pregunta crucial es la siguiente: ¿cuándo es entonces razonable revisar una intención? Parece que los límites están difusos, y que por tanto la respuesta será vaga. Sin embargo, podemos afirmar que un sujeto que cambia continua e injustificadamente sus decisiones puede ser considerado como inconstante o caprichoso. Como señala Holton, la debilidad de la voluntad y el capricho suponen una revisión injustificada de una resolución, así que pertenecen a un mismo género: la *irresolución;* pero no son iguales, sino dos especies distintas. La debilidad de la voluntad supone la revisión injustificada de una intención cuyo objetivo era precisamente el vencimiento de una posible futura inclinación contraria, mientras el capricho, no. Veamos esto con un ejemplo: supongamos que quiero seguir una dieta. Tengo la intención de empezar el lunes que viene, esto es, tomo la decisión de realizar esa acción el lunes. En principio no hay razones para escoger el lunes o el martes, pero mi intención de empezar un día concreto tiene un objetivo: vencer las inclinaciones contrarias que sospecho que pueda tener el día que haya de comenzar mi dieta. Si llegado el lunes reviso mi intención sin una razón justificada, mostraré una voluntad débil. Supongamos ahora que deseo ir

al cine con unos amigos. Hay tres salas de cine en la cuidad y en principio no hay más razones para acudir una que a otra; ni la distancia, ni la calidad, ni el precio muestran diferencias. Sin embargo, he de decidirme por alguno de ellos si quiero coordinarme con mis amigos e ir a ver la película finalmente. Esta intención de ir al cine *Y*, no la tengo para evitar el deseo o inclinación futura de ir a otro cine; por ello, si antes de ir al cine la reviso injustificadamente varias veces, cambiando de decisión, estaré siendo inconstante y caprichoso. Por supuesto, los límites entre la debilidad de la voluntad y el capricho son difusos, por lo que el concepto de debilidad de la voluntad adquiere mayor vaguedad. Pero esto no es un problema para Holton, ya que considera que refleja la vaguedad que se corresponde con el uso ordinario del término [Holton (1999), p. 250-251].

Frente a la debilidad de la voluntad estaría la fuerza de voluntad. La fuerza de la voluntad no consiste en el rechazo a revisar cualquier intención o resolución sino, más bien, en la capacidad de no revisar una intención a menos que sea razonable hacerlo. Aquellos sujetos que no revisan sus intenciones aun cuando sería razonable que lo hiciesen, no muestran una voluntad fuerte, sino testarudez [Holton (1999), p. 252].

Holton ofrece varias pruebas de la ventajas que supone distinguir entre debilidad de la voluntad y *akrasia*, pero no podemos entrar aquí en ellas. Mencionaremos a modo ilustrativo sólo una: a veces se nos presentan dos situaciones entre las cuales debemos elegir. Ambas son inconmensurables, esto es, no podemos compararlas. Supongamos, dice Holton, que un joven ha de decidir entre ir a la guerra a luchar contra el fascismo o cuidar

de su madre enferma. Ambos cursos de acción son inconmensurables, por lo que *ex hypothesi*, no puede haber un mejor juicio que nos dicte qué curso de acción hemos de seguir. Pero entonces seguir uno u otro no puede dar lugar a un comportamiento acrático, pues no hay ningún mejor juicio contra el que actuemos. Sin embargo, es posible que, consciente de que puedo estar indeciso eternamente y la situación es apremiante, tome la decisión de ir a luchar, pero que llegado el momento revise mi intención y decida quedarme en casa cuidando de mi madre. Al revisar injustificadamente mi intención estaría mostrando una voluntad débil, aunque como hemos dicho, no mostraría incontinencia o *akrasia*. [Holton (1999), p. 251]

Por otro lado un sujeto puede, por el contrario, actuar de un modo acrático, sin tener una voluntad débil. Supongamos que tenemos un amigo que cree que todos sus argumentos apuntan en la misma dirección: debería dejar de comer carne. No obstante, nos confiesa que no lo hace. Ciertamente actúa en contra de lo que considera su mejor juicio, esto es, acráticamente, pero dado que no hay ninguna intención de no comer carne que revise o viole, no es débil cuando se come un filete. Un caso distinto sería si el sujeto toma la decisión de dejar de comer carne y, posteriormente, rechaza esa intención sin justificación cuando se le presenta un apetecible estofado de carne. Diremos entonces que, además de comportarse de un modo acrático, tiene la voluntad débil.

Pero una vez que sabemos que la debilidad de la voluntad supone un fracaso al tratar de mantener las intenciones que nos habíamos propuesto

llevar a cabo, y que la fuerza de voluntad supone resistir las inclinaciones contrarias, la pregunta relevante no será ya acerca de cómo es posible la debilidad de la voluntad, sino la fuerza de voluntad [Holton (2003), p. 39]. Para Holton ésta es posible porque el sujeto bloquea la reconsideración futura de las resoluciones que toma. Las resoluciones (intenciones) que no son reducibles a creencias y deseos, como quieren otras teorías que Holton denomina *humeanas aumentadas,* son apartadas de la reconsideración por una capacidad denominada *poder de la voluntad* (*will-power*). En defensa de la existencia de esta facultad Holton aduce algunos experimentos recientes de la psicología experimental social. Parece que los niños desarrollan, entre los 3 y 4 años, la capacidad de bloquear sus deseos tomando en cuenta una recompensa futura mayor y, para conseguir este bloqueo, diseñan estrategias, unas más elaboradas que otras y, también, unas más exitosas que otras. [Holton (2003), p. 55].

Como vemos la debilidad de la voluntad y la *akrasia* plantean interrogantes que aun hoy mantienen su frescura. ¿Es posible que un sujeto actúe en contra de lo que considera su mejor argumento? ¿Cómo? ¿Es racional que alguien bloquee la revisión de sus decisiones? ¿Hasta qué punto?

Desde luego, me parece que la distinción conceptual ente *akrasia* y debilidad de la voluntad introducida por Holton es totalmente pertinente, por ser más fina, casar mejor con el uso común que le damos a las nociones, y salvar mejor los fenómenos.

Sin embargo, la adopción de una definición no comporta un compromiso con la existencia de aquello descrito. Antes bien: podemos describir seres fantásticos o situaciones imposibles. A veces es necesaria la caracterización de un fenómeno aun cuando presumimos incluso que no existe, y la razón sería que el concepto "folk" aparentemente casa con los fenómenos. Sin embargo descubrimos que, al caracterizarlo de un modo más fino y sacar sus consecuencias últimas, vemos que describe una situación, bien imposible, bien posible pero no real, y hemos de buscar una explicación alternativa.

La *akrasia* me parece un caso de este tipo. En mi opinión, un sujeto actúa siempre en función de su mejor juicio. Los principios que Davidson clasifica como P1 y P2 me parecen irrenunciables. De ahí se sigue que, cuando un sujeto que cree que algo es mejor, lo desea, y si es libre para hacerlo, lo hará (siempre y cuando no entre en conflicto con algo que quiere más). Por supuesto deben eliminarse de la discusión los casos de coacción, coerción, compulsión y demás estados en los que el sujeto no es libre para realizar lo que desea. Lo que sucede aquí, y creo que es la principal dificultad, es que muchas razones por las que actuamos, no se nos aparecen como tales razones. La razón de esto es el enquistamiento que sufren algunas ideas, el cual las hace muy difíciles de erradicar tras una crítica. En este caso me refiero a la división entre razones y pasiones, entre razón y sentimiento. A veces se ha señalado que la razón es una pasión más; o que al menos, no hay una oposición entre ambas. David Hume vio esto muy bien cuando afirma que la razón es esclava de las pasiones. Aquello que nos impulsa a la acción son pasiones. Las pasiones no son, en

principio, irracionales, y sólo cuando lo son, son eliminadas de nuestras consideraciones. Otras veces se dice que la pasión es una razón más: de este modo, las ocasiones en que se dice que el individuo toma una decisión "cegado" por una pasión o deseo, no supondrían ya que lo haga en contra de su mejor argumento: en ese momento, las pasiones supondrían una razón que le motivaría a realizar esa acción. En otro momento en el que el deseo no estuviese presente, sus mejores razones podrían cambiar. Esta era la razón por la que Davidson rechazaba la teoría aristotélica; le parece una situación posible, pero no un caso de debilidad de la voluntad. El punto crucial aquí es que habría que rechazar la oposición entre pasión/razón y, con ello, el carácter peyorativo que se le otorga a la pasión como enturbiadora de la razón o cegadora del recto juicio.

De hecho, a mi juicio la postura de Aristóteles, aunque él pretenda lo contrario, es bastante cercana a la de Sócrates y Platón. Estos negaban la posibilidad de que un sujeto actúe contra lo que sabe que es bueno; si hace algo malo lo hace por ignorancia. La postura de Aristóteles es similar: la incontinencia no puede ser continua, sino que es intermitente e incluye lo que llamamos ceguera intelectual: una pasión nos ciega y hace que no sopesemos correctamente las razones de las que disponemos. Pero entonces, ¿diremos que ese sujeto toma la decisión sabiendo que es equivocada, o diremos más bien que la pasión le ciega y le hace *momentáneamente ignorante* —como quería Sócrates— con respecto a algunas de las razones que debería tomar en cuenta al evaluar qué debe hacer?

Como veíamos, el propio Aristóteles compara al incontinente con un borracho, que *dice* saber, pero que en realidad *no* sabe.

Más problemática es la postura de Holton. En realidad Holton establece una diferencia teórica entre debilidad de la voluntad y *akrasia*, y parece comprometerse con la existencia de ambos fenómenos, pero no ofrece razones para defender la existencia de la *akrasia*. Con respecto a su defensa de la debilidad/fortaleza de la voluntad y de la facultad del *poder de la voluntad,* reconoce que hay espacio conceptual para otras teorías alternativas, y aunque ofrece apoyos basados en pruebas empíricas, parece más bien una hipótesis conceptual heurística que algo descubierto por introspección o investigación empírica. En cualquier caso, el bloqueo de la reconsideración de las resoluciones e intenciones parece algo razonable. El problema es que ese bloqueo es razonable hasta que hay un buen motivo para rechazarlo, y lo que un sujeto acepte como motivo depende de su testarudez, de su ceguera intelectual, de si presenta pensamiento desiderativo… y, finalmente, de si las pasiones pueden ser o no razones; es decir, que un sujeto manifieste una disposición *irrazonable*, bien a mantener su intención previa incluso ante razones convincentes (testarudez), bien a cambiar sus intenciones ante un deseo o emoción (debilidad de la voluntad), puede ser visto como algo perfectamente razonable y racional desde otras teorías alternativas, que juzgarán cada caso particular con otras aristas y detalles concretos. No me parece, en resumen, que Richard Holton aporte más luz al asunto que la de distinguir conceptualmente dos fenómenos que habían sido confundidos y entremezclados.

Así, parece que no hay ninguna teoría que haya avanzado más de lo que lo hizo Platón. El propio Davidson reconoce que no hay ninguna prueba de que existan casos de incontinencia aunque le parezca absolutamente cierto que existen. Lo que no se comprende muy bien es por qué, amparado en una intuición, rechaza la posibilidad de reinterpretar cada caso a la luz de otros fenómenos similares y se aferra a la debilidad de la voluntad, aun sin pruebas de su existencia, como única ilustración posible de tales situaciones, aun cuando ésta supone además una irracionalidad en el curso de acción del sujeto; curso de acción del que otros modos de explicación de la acción pueden dar cuenta racionalmente y sin contradicción. A este respecto, resulta especialmente iluminador este párrafo:

> Hay una considerable gama de acciones similares a las acciones incontinentes […] autoengaño, insinceridad, mala fe, hipocresía, deseos, motivos e intenciones inconscientes, etc. De hecho hay una tentación muy grande, al trabajar en este tema, de jugar al psicólogo aficionado. Nos estamos muriendo por decir: recuerda la enorme variedad de formas en las que un hombre puede creer o sostener algo, saberlo, querer algo, tenerle miedo o hacer algo. […] Sin duda explican, o al menos señalan un camino para describir sin contradicción muchos casos en los que nos encontramos hablado de debilidad de la voluntad o incontinencia [Davidson (1970), pp. 45-46]

Pero pese a afirmar esto, a continuación afirma que nosotros mismos mostramos cierta debilidad como filósofos si no nos preguntamos si hay o no casos que muestran un sujeto consciente de que no sigue su mejor argumento, y sin embargo actúa libre y voluntariamente contra él, o que escoge otro curso de acción alternativo y, a su juicio, inferior.

Según mi parecer, en cambio, si otros caminos explican estos comportamientos humanos sin contradicción lógica y en consonancia con los fenómenos, sin duda han de preferirse. Además no parece que nadie haya ofrecido hasta ahora una prueba empírica o una argumentación fuerte y sostenible de la existencia de tales fenómenos irracionales. ¿Por qué habríamos de aceptarlos sin más? La *akrasia* y la debilidad de la voluntad me parece que describen una situación imposible, o dicho de otra manera, son una mala caracterización o errónea descripción de los fenómenos que trata de investigar.

En estos casos, el sujeto siempre actúa según su mejor razón (incluyendo aquí no sólo razonamientos abstractos, sino pasiones, deseos, etc.). Cuando nos parece que no es posible que la conducta de un sujeto sea racional, su ignorancia es una explicación mucho más sencilla y convincente, aunque habría que especificar, en cada caso, a qué se debería. Unas veces será debida a falta de información, otras a ceguera intelectual, etc.

Así pues, tanto el autoengaño como la debilidad de la voluntad (problemática o no) exigen un conflicto entre dos alternativas respaldadas ambas por evidencia o argumentos, y en ambos casos el sujeto actúa en contra de la alternativa mejor apoyada. La diferencia entre ambos es que el producto del autoengaño es una creencia y el producto de la debilidad de la voluntad es una acción (o una intención dirigida a cierta acción).

I.6 - Debilidad de la justificación

Como vimos en la sección anterior, el hecho de que el sujeto que trata de autoengañarse estaría más inclinado a creer —y de hecho creería— la proposición contraria a la que estaría en principio más apoyada por la totalidad de la evidencia disponible para él, acerca el autoengaño al problema de la debilidad de la voluntad. De hecho, Davidson afirma que el autoengaño contiene un fenómeno similar al de la debilidad de la voluntad, que bautiza como debilidad de la justificación (*weakness of warrant*) y que supone un paso ilegítimo en nuestra formación de creencias.

Se trata de un error cognitivo semejante a la debilidad de la voluntad. La diferencia entre ambos es que el producto de la debilidad de la voluntad es una acción (o una intención dirigida a cierta acción) y el producto de la debilidad de la justificación es una creencia. En muchas ocasiones ambos fenómenos aparecen juntos reforzándose el uno al otro.

El sujeto que padece debilidad de la justificación, ante evidencia a favor de *p* y de *no-p*, se forma una creencia contra el peso de su evidencia total. En efecto, la debilidad de la justificación supone que el sujeto se forme una creencia al margen de su evidencia. Pero, ¿hay alguna premisa adicional oculta que convierta a la debilidad de la justificación en irracional como sucedía con la debilidad de la voluntad? Nuestra premisa adicional debería rezar más o menos como sigue:

1) La formación de tus creencias estará guiada por tu evidencia.

O bien,

2) La formación de tus creencias estará guiada por tus deseos.

Pero sabemos que 2) resulta imposible, pues los deseos no son una base legitima para la formación de creencias, pues no apuntan a la verdad. Por ello, sólo 1) es una verdadera candidata, lo cual supone afirmar que la formación legítima de creencias ha de basarse en la evidencia. Hasta aquí todo correcto; casi todos los teóricos estarán de acuerdo en exigir la evidencia como sustento racional de las creencias.

Es de suponer, aunque Davidson no lo dice explícitamente, que dado el paralelismo entre la debilidad de la voluntad y la debilidad de la justificación, la explicación de esta última discurra por similares senderos. De este modo, el sujeto podría formarse una creencia en contra de su evidencia si tomase un conjunto e de evidencia que sólo contemplara *prima facie* el conjunto total de evidencia. El *Principio de Evidencia* (paralelo también al Principio de Continencia) exigiría que el sujeto tomase en cuenta toda la evidencia disponible y relevante, y no sólo la evidencia *prima facie* total.

Como no puede ser de otro modo, creo que las críticas a la debilidad de la voluntad son por completo extensibles a la doctrina de la debilidad de la justificación, *mutatis mutandis*. En concreto, y para no repetir lo dicho, un conjunto que tome en cuenta la evidencia total sólo *prima facie*, es equivalente a un conjunto que no toma en cuenta toda la evidencia, es decir, es un subconjunto incluido propiamente en el verdadero conjunto total. La formación de una creencia en contra del mayor apoyo que represente la evidencia total de este conjunto es un fenómeno imposible.

Sin embargo, en esta ocasión podemos ir más allá incluso. Adelanto ya que desde mi punto de vista no sólo es deseable o racional que las creencias se fundamenten en evidencia: es *necesario conceptualmente* por lo que, como veremos más adelante, no puede ocurrir de otro modo. Por esto creo que la debilidad de la justificación describe una situación imposible. De hecho parece que, en realidad, lo que confunde a Davidson es la oscura suposición de que quien muestra debilidad de la justificación estaría violando algo similar a lo que Carnap y Hempel[16], entre otros, llamaron el *requerimiento de evidencia total para el razonamiento inductivo,* según el cual, cuando estamos tratando de decidir entre dos hipótesis mutuamente excluyentes, racionalmente debemos optar por aquella más respaldada por la totalidad de la evidencia disponible.

> (RET): *E apoya evidencialmente a una hipótesis H para un sujeto S en un contexto* C sii *E* confirma *H*, en relación a *K*, donde *K* es el conjunto de evidencia total de *S*'s *en* C.

Sin embargo, pese a la aparente similitud entre la fortaleza de la justificación y este *requerimiento,* hay una significativa diferencia entre ambos, ya que el segundo no se gestó en el contexto de una teoría de la creencia, sino más bien en una teoría de la ciencia, junto al problema de la lógica inductiva y la evaluación de hipótesis científicas. En este contexto, la elección de una hipótesis u otra, podría —aun cuando no fuera

[16] *vid.* Carnap (1950) y Hempel (1962, 1965). Una caracterización clásica es la siguiente: "[…] the credence which it is rational to give to a statement at a given time must be determined by the degree of confirmation, or the logical probability, which the statement possesses on the total evidence available at the time." [Hempel (1965), p. 64]

deseable— hacerse al margen de la evidencia empírica. De hecho, sabemos que la marcha de la ciencia poco tiene que ver con un progreso constante debido a una búsqueda "limpia" de la verdad sustentada únicamente en la evidencia disponible; más bien está guiada además, en muchas ocasiones, por intereses particulares, políticos, económicos o sociales. Dicho de otro modo: en la ciencia la evidencia *no nos obliga necesariamente* a elegir una u otra hipótesis (aunque sea más adecuado y racional hacerlo), sino que podemos controlar qué hipótesis es la que vamos a abrazar o tratar de corroborar. Es decir: podríamos optar deliberadamente por una hipótesis que creyésemos errónea.

Además, a diferencia de las hipótesis científicas, nuestras creencias no surgen nunca de la creatividad del sujeto; antes bien, requieren evidencia de peso para que éste pueda albergarlas. Con esto no niego la existencia de intuiciones geniales y creativas, sino que hayan de ser etiquetadas como creencias. Es cierto también, que este tipo de intuiciones puede intervenir en la formación de otras creencias, pero no son creencias ellas mismas. Por último, las hipótesis científicas pueden ser —¿qué duda cabe?—, objeto de creencia por parte de un sujeto.

El punto crucial es que las creencias no las controlamos nosotros. No serían (¿seríamos?) pocos los que tratarían de cambiar sus creencias a su antojo en busca de la felicidad si esto fuese posible. Pero no lo es. Más bien, la formación de creencias es una respuesta en cierto modo "automática" a la acumulación de evidencia; en ausencia de evidencia quizá tengamos deseos, ilusiones o esperanzas, y cuando hay poca

evidencia, sospechas. Pero las creencias necesitan de evidencia concluyente o, al menos, con un peso fuerte. Alguien podría objetar que las esperanzas o sospechas son distintos tipos de creencias, quizás con grado bajo de confirmación, o creencias degeneradas… Sin embargo, me parece que cuando tenemos poca o muy poca evidencia disponible de algún suceso cualquiera, si nos preguntasen seriamente: "¿lo crees?", nuestra respuesta sincera sería que no. Esto *no* impide que las sospechas o esperanzas sirvan, al igual que lo hacen las creencias, como guías para la acción. Quizá sea porque ambas características son comunes a creencias y sospechas o esperanzas, a saber, estar apoyadas por evidencia y ser guías para la acción que nos empujan a buscar más evidencia en una dirección determinada (dirección marcada por esas sospechas o esperanzas), lo que genere esta confusión. Ha de reconocerse, además, que los límites entre una sospecha y una creencia quizá son difusos, pero en todo caso, creo que hay razones suficientes para discriminar la creencia de las esperanzas y sospechas, pues no se trata únicamente de un mero cambio cuantitativo, sino cualitativo.

Las creencias serían, en este sentido, algo distinto de las esperanzas o sospechas y constituyen una respuesta semiautomática, fuera de nuestro control, a la acumulación de evidencia en uno u otro sentido (lo que sí podemos controlar es la acumulación de esa evidencia, adulterarla…). O dicho de otro modo: tener una creencia es sinónimo de tener un buen apoyo evidencial de algo. A mayor apoyo, mayor fortaleza de la creencia.

Pero entonces, si la formación de creencias no depende de nuestra voluntad y está siempre guiada por la evidencia más fuerte, no podemos padecer debilidad de la justificación. Por ello, si deseamos de ofrecer una explicación satisfactoria acerca del autoengaño, debemos examinar caminos alternativos y buscar la explicación de su aparente irracionalidad o ilegitimidad en procesos previos a la formación de creencia como, por ejemplo, en la acumulación de evidencia adulterada. Esto choca frontalmente con el proyecto no sólo de Donald Davidson sino de casi la totalidad de autores que han tratado de explicar el fenómeno del autoengaño y se antoja como una tarea en la que la filosofía de la mente y de la acción tienen mucho camino por delante.

I.7 - Mala Fe

> En la mala fe, no hay mentira cínica ni sabia preparación de conceptos engañosos. El acto primero de mala fe es para rehuir lo que no se puede rehuir, para rehuir lo que se es.
>
> [Sartre (1943), p. 124]

Aunque durante toda la historia del pensamiento occidental pueden rastrearse referencias al autoengaño, descripciones de hombres que se mienten a sí mismos o de lo terribles que son las consecuencias de tal ejercicio, no ha sido hasta el siglo XX que este fenómeno ha comenzado a ser estudiado en profundidad. Más allá de meras citas o incluso de comentarios en los que el autoengaño se presupone pero no se explica

(por ejemplo, en los *Sermones* de Joseph Butler [Butler (1726)] o en la *Teoría de los sentimientos morales* de Adam Smith [Smith (1759)]), fue Sigmund Freud el primer autor que encara el problema y la paradoja que suscitaría un conjunto de intenciones, deseos, creencias contradictorias dentro de un yo racional y unitario dándole una explicación conceptual (amén de unas aplicaciones prácticas).

Una de las críticas más consistentes, a menudo minusvalorada debido a la dificultad comprensiva que imponen la oscuridad introducida por el vocabulario y la complejidad del discurso en el que se entretejen los argumentos, es la efectuada por Jean-Paul Sartre en 1943 en su obra *El Ser y la Nada*, concretamente en el capítulo que reserva para exponer la "mala fe".

Aunque voy a dejar la crítica a Freud para el capítulo dedicado al análisis de las propuestas psicoanalíticas más relevantes con respecto al autoengaño o engaño a uno mismo (§II.1), es preciso ahora explicar en qué consiste la noción de "mala fe" (*mauvaise foi*) a fin de distinguirla del autoengaño. Nuestra postura es que pese a que el propio Sartre afirma que la mala fe supone una mentira a uno mismo, no debe confundirse —como a menudo se ha hecho— con el autoengaño. La razón más obvia y directa es que "engaño" y "mentira" no son equivalentes. Sin embargo, podría aducirse que en realidad es Sartre quien al no distinguir ambos fenómenos, ofrece sin percatarse de ello una explicación del autoengaño en términos de "mentira a uno mismo" o "mala fe". Trataré de explicar por qué esto no es así.

Pese a que algunas afirmaciones acerca de los objetivos de la mala fe según las cuales "en el caso de la mala fe se trata de enmascararse una verdad desagradable o presentar como verdad un error agradable" [Sartre (1943), p. 97] podrían inclinarnos a pensar que la búsqueda de tal enmascaramiento comporta un engaño a uno mismo, sin embargo Sartre afirma también que

> […] la esencia de la mentira implica, en efecto, que el mentiroso esté completamente al corriente de la verdad que oculta. No se miente sobre lo que se ignora; no se miente cuando se difunde un error de que uno mismo es víctima; no miente el que se equivoca. [Sartre (1943), p. 96]

Por tanto, el enmascaramiento total no parece posible. La mala fe tiene pues, en apariencia, la estructura de la mentira. Sólo que —y esto lo cambia todo— en la mala fe yo mismo trato de enmascararme la verdad. Así, la dualidad del engañador y del engañado no existe en este caso. Uno trata de disimular*se* algo, y por tanto, la mala fe implica por esencia la unidad de *una* conciencia. [Sartre (1943), p. 97]. Los problemas y paradojas que suscita el intento de engaño a uno mismo, los describe Sartre de forma muy precisa en el siguiente párrafo:

> [...] aquel a quien se miente y aquel que miente son una sola y la misma persona, lo que significa que yo, en tanto que engañador, debo saber la verdad que me es enmascarada en tanto que engañado. Mejor aún: debo saber muy precisamente esta verdad *para* ocultármela más cuidadosamente; y esto no en dos momentos diferentes de la temporalidad —lo que permitiría, en rigor, restablecer una apariencia de dualidad—, sino en la estructura unitaria de un mismo proyecto. ¿Cómo, pues, puede subsistir la mentira si está suprimida la dualidad que la condiciona? A esta dificultad se agrega otra que deriva de la total

> translucidez de la conciencia. Aquel que se afecta de mala fe debe tener conciencia (de) su mala fe, ya que el ser de la conciencia es conciencia de ser. [Sartre (1943), p. 98]

También indica Sartre otro aspecto muy importante común a la mentira a uno mismo y el autoengaño. La mentira directa con respecto a uno mismo no puede ser, en modo alguno, exitosa.

> Se admitirá, en efecto, que, si trato deliberada y cínicamente de mentirme, fracaso completamente en tal empresa: la mentira retrocede y se desmorona ante la mirada; queda arruinada, por detrás, por la conciencia misma de mentirme [...] hay por tanto una evanescencia de la mala fe. [Sartre (1943), p. 98]

La mentira [intento de engaño] directa hacia uno mismo produce a lo sumo un estado evanescente, poco duradero. Así, la mala fe es una estructura "metaestable" ya que el sujeto se halla en un punto de continua oscilación entre la buena fe y el cinismo. Unas veces uno tiene despertares de buena fe, otras se encierra en el más absoluto cinismo, pero no por ello la mala fe deja de ser una forma de vida psíquica. Incluso, para muchas personas es tan duradera que se convierte en el modo de vida normal. Según Sartre, ocurre también a menudo que el sujeto se persuade a medias de su mentira, pero esto son intermedios "bastardeados" entre la mentira y la mala fe. [Sartre (1943), p. 97].

Por tanto, la mala fe se nos presenta de modo totalmente paradójico, ya que pese a que no podemos rechazarla por ser el modo de vida normal de mucha gente, tampoco podemos comprenderla. [Sartre (1943), p. 98]

Sartre indica que el psicoanálisis ha hecho un esfuerzo por ofrecer una explicación coherente de esta situación, pero ésta no resulta en absoluto satisfactoria, porque, como veremos de modo más detenido más adelante (§II.1.3), reubica el problema sin hacernos ganar nada en absoluto. El problema es que la explicación que pivota en torno al inconsciente no puede resultar convincente para tales fenómenos, ya que éstos presuponen la unidad de la psique: "la esencia misma de la idea reflexiva de disimular*se* alguna cosa implica la unidad de un mismo psiquismo" [Sartre (1943), p. 102] y, además, "existe una infinidad de conductas de mala fe que rechazan explícitamente ese tipo de explicación, porque por esencia implican que no pueden aparecer sino en la translucidez de la conciencia." [Sartre (1943), p. 104] En efecto, Sartre indica que incluso los estudios de algunos psicoanalistas no ortodoxos, como Wilhelm Stekel, afirman que en todos los casos en los que se ha avanzado lo suficiente, se ha descubierto que el núcleo de las psicosis es consciente (en contra de la tesis freudiana). Es la consciencia de un placer experimentado como obsceno lo que produce un rechazo asimismo consciente y lo que les empuja a negarlo, llevando a cabo diferentes estrategias para ello.

Es entonces, tras constatar que la mala fe parece tan común como paradójica, cuando Sartre admite que la única salida a este atolladero pasa por explorar las condiciones de posibilidad de la mala fe, esto es, por dar respuesta a la pregunta: ¿qué ha de ser el hombre en su ser, si ha de poder ser de mala fe?

En primer lugar, la mala fe no es un *estado*; no es algo que alguien sufre o padece, porque esto supondría que es algo externo al sujeto. Sartre dice: "*somos* de mala fe"; la mala fe se identifica con mi yo, es mi ser. La propia estructura del *cogito* nos conduce a ella. La realidad humana "se constituye como un ser que es lo que no es y que no es lo que es" [Sartre (1943), p. 109]. ¿Qué quiere decir con esto? Según Sartre, la realidad del hombre puede ser considerada desde distintos enfoques, tales como su factididad/transcendencia o el "ser para mí"/"ser para otros". No se trata de que uno de ellos sea el correcto y el otro sea una deformación; sin duda, ambos tienen su valor, y son dignos de ser conjugados o superados en una síntesis. Pero la mala fe aprovecha esta dualidad para afirmar su identidad a la vez que mantiene sus diferencias, "mantiene la disociación para hacer un perpetuo deslizamiento del presente naturalista a la trascendencia, y viceversa" [Sartre (1943), p. 107]. ¿Qué significa esto? ¿Cómo ocurre? ¿Por qué?

Veamos un ejemplo del propio Sartre en el que expone cómo la mala fe aprovecha la dualidad facticidad/transcendencia. La *facticidad* hace referencia al ser tal y como es un momento concreto, como cosa, como algo fijo, inmóvil en su esencia; la *transcendencia* se refiere al ser como pura potencialidad. Supongamos que estamos ante una pareja que tiene su primera cita. La mujer, ante la actitud respetuosa y galante de su acompañante, elimina todo trasfondo sexual de las conductas de su pretendiente, reduciendo su significado al momento presente; éstas quedan fijadas en una permanencia cosista como una "proyección del presente en el flujo temporal", esto es, para ella, su acompañante *es*

respetuoso y atento como una mesa *es* una mesa. Ha obviado la potencialidad, las posibilidades de desarrollo que indican esas conductas, reduciéndolas a su pura inmanencia. Pero por otro lado, un deseo crudo y desnudo, inmanente, carnal y cosista le resultaría violento, y por ello trasciende éste hacia una forma elevada de respeto y admiración. Sin embargo, todo amenaza desmoronarse cuando de repente él la coge la mano; ahora ella ha de tomar una decisión: por un lado, retirar su mano supondría romper la armonía del momento; por otro, consentir supone entrar en el juego del flirteo y con ello en la facticidad y carnalidad del deseo que le asustan. ¿Qué hará? Conocido es, dice Sartre, lo que hace nuestra mujer: ni consiente ni sofoca; abandona su mano, pero *no se da cuenta* de ello; habla de su vida, de sí misma como persona, como conciencia, produciéndose "el divorcio entre alma y cuerpo". Su mano es ahora únicamente una cosa inerte entre las manos de su pareja. Según Sartre, diremos que esta mujer es de mala fe.

Por tanto, esta mujer

> […] ha desarmado las conductas de su pareja reduciéndolas a no ser sino lo que son, es decir, a existir en el modo del en-sí. Pero se permite disfrutar del deseo de él, en la medida en que lo capte como no siendo lo que es, es decir, en que le reconocerá su trascendencia. Por último, sin dejar de sentir profundamente la presencia de su propio cuerpo —quizá hasta el punto de turbarse—, se realiza como *no siendo* su propio cuerpo, y lo contempla desde arriba, como un objeto pasivo al cual pueden *acaecer* sucesos, pero que es incapaz de provocarlos ni evitarlos. [Sartre (1943), p. 106].

Como hemos señalado, hay otros modos de mala fe, además de éste resultante de la relación facticidad/transcendencia. La mala fe también aprovecha la dualidad "ser para mí"/"ser para otro".

Así pues, en todos los casos la mala fe supone una huida. La mala fe enmascara el verdadero ser del sujeto, enmascara la angustia. El hombre es angustia; la angustia que produce el vacío de una elección auténtica. La mala fe supone una forma de huir de la libertad que somos y de la responsabilidad que engendra toda elección, haciendo recaer la responsabilidad sobre alguien o algo ajeno a mí. Ésta puede hacerse recaer sobre Dios, sobre una ideología, o sobre las circunstancias. Pero en cualquier caso, la mala fe supone esa huida de la angustia que le produce al sujeto tener que elegir constantemente, mintiéndose a sí mismo diciéndose que es lo que no es o que no es lo que es. No podemos suprimir la angustia porque somos angustia. Siempre estamos optando libremente aunque lo ocultemos, y somos responsables de nuestras decisiones.

A menudo, el hombre que comete faltas no reconoce aquello que es. Esto se debe a dos cosas: en primer lugar, se niega a que su ser quede reducido a un conjunto de conductas. Pero también "lucha con todas sus fuerzas contra la aplastante perspectiva de que sus errores le constituyan un *destino*". [Sartre (1943), p. 116] Efectivamente, el hombre que dice no ser un ladrón[17] pretende que no se le considere como un ser que está ya

[17] En realidad, Sartre utiliza la homosexualidad como ejemplo de conducta reprochable e incluso enfermiza (habla de una "posible cura psíquica"). Este ejemplo poco afortunado de Sartre, que llega a exigir al homosexual que "reconozca que es un pederasta" [¿?] [Sartre (1943), p. 116] creemos que empaña pero no anula

de antemano determinado a robar. Se niega a conceder que sus conductas delictivas previas le marquen el futuro, apoyado en que ningún conjunto de conductas puede hacerse equivalente a su ser. Es decir, niega ser ladrón desde el punto de vista de la trascendencia; niega con toda la razón que las posibilidades estén cerradas para él. Sin embargo, desde este plano legítimo, se desliza hacia el plano de la facticidad, afirmando no ser ladrón del mismo modo que una mesa no es un tintero, y entonces es de mala fe [Sartre (1943), pp. 116-117]. Según Sartre, la única posibilidad de escapar radicalmente a la mala fe consiste en la asunción del ser podrido del hombre, reasumiendo que la mala fe está en el origen de todo proyecto. [Sartre (1943), nota al pie, p. 124]

Finalmente, ¿es la mala fe un modo de autoengaño? Hemos expuesto en qué consiste la mala fe. La angustia producida por la elección constante ante el abismo que produce la libertad del futuro sin realizar, nos hace tratar de huir de esta elección; pero ésta es forzada y no podemos evadirla, pues nuestro ser no está determinado de antemano. El carácter forzado de la elección produce una angustia ineliminable: somos angustia. La mala fe es la mentira que el hombre se dice para intentar hurtarse a la elección. Y en este punto es en el que Sartre da un salto infundado: niega la dualidad de la conciencia, y con ello cualquier explicación que aluda a divisiones de la mente, inconsciente o cualquier otro asunto. La consciencia es para Sartre "consciencia de ser consciencia", saber es saber que se sabe [Sartre (1943), p. 102]. Al mismo tiempo, afirma que el ser del hombre es de mala

su argumento de fondo: el hombre no reconoce ser aquello que ha venido siendo, precisamente porque se niega a estar determinado por sus conductas precedentes.

fe; incluso, la buena fe y la sinceridad han de ser en su raíz de mala fe. Pero, ¿cómo ha de ser esto posible? Sartre no da ninguna explicación del fenómeno; en su lugar, ofrece ejemplos en los que establece de antemano, por ejemplo, que la mujer abandona su mano *sin darse cuenta de ello*; establece además que toma esta conducta por la angustia que le supone la urgencia de la decisión a la que se ve forzada. Pero, ¿cómo hace esto? ¿Cómo consigue ser víctima de su propio engaño? Sartre da una explicación de la utilidad que tendría el autoengaño: evadirnos de la angustia que supone la elección forzada y constante. Pero no explica cómo sería esto posible. En el momento en que tiene que dar una explicación conceptual del mecanismo que lo haría posible, sólo hace una crítica (profunda y muy consistente) del psicoanálisis y la división de la conciencia.

Por tanto, nos parece que tomando como punto de partida los dos puntos sartrianos, a saber, unidad de conciencia y angustia ante las elecciones vitales, la explicación más coherente y consistente sería la siguiente.

Si el engaño presupone una dualidad que negamos de antemano en razón de la unidad de la conciencia, el hombre no puede engañarse a sí mismo, no puede autoengañarse. Por tanto, ante la urgencia y el carácter forzoso de la elección, sólo puede *fingir* ser lo que no es y hacer *como si* no tuviera que elegir constantemente, tratando de evadir la angustia que le produce esa elección constante; se dice una mentira y trata de actuar como si esta fuera el caso. Pero decirse una mentira a uno mismo y actuar

"como si" fuese el caso no es engañarse; a lo sumo, es *tratar de* engañarse. No cabe duda de que en muchos momentos todos podemos sentir la angustia de una elección forzada y difícil; sin duda la vida consiste en buena parte en esto y, trataríamos de engañarnos, pero eso no es lo que está en cuestión. Que nos mintamos puede ser inútil, baladí, pero no resulta conceptualmente imposible. Ahora bien, no constituiría en modo alguno engaño, sino tentativa de engaño (condenada de antemano al fracaso a causa de la unidad psíquica).

SEGUNDA PARTE

II - APROXIMACIONES AL PROBLEMA CONCEPTUAL

II.1 - Aportaciones desde el psicoanálisis

II.1.1 - Represión e inconsciente

> Es un suceso cotidiano, aun en personas sanas, que se engañen acerca de los motivos de su obrar y sólo devengan conscientes de ellos con posterioridad, toda vez que un conflicto entre varias corrientes de sentimiento les cree las condiciones para ese estado de confusión.
>
> [Sigmund Freud (1906), vol. IX, p. 55]

El estudio de la obra de Sigmund Freud resulta central en el análisis del autoengaño. En nuestra opinión, si hasta el siglo XX sólo se ha hablado del autoengaño en relación con lo pernicioso e inmoral de su ejercicio y no se ha tratado en profundidad la problemática asociada a la posibilidad teórica y práctica del fenómeno, ha sido debido a que se carecía del aparataje conceptual necesario. Esto no es un caso aislado; del mismo modo, hasta

la irrupción de las teorías psicoanalíticas[18] no reciben una explicación sistemática tanto fenómenos asociados a sujetos insanos, —e.g., algunos trastornos mentales como las neurosis, psicosis e histeria, que desde ese momento son tratados como enfermedades—, como sanos —los sueños reciben ahora una "explicación" coherente dentro de la vida anímica del sujeto, así como otros fenómenos tales como los *lapsus linguae* o los actos fallidos.

La principal razón por la que las precedentes teorías encontraban un escollo insalvable en todos estos fenómenos resultándoles intratables, era el presupuesto que equiparaba el psiquismo y la consciencia. Efectivamente, ya algunos filósofos románticos habían acentuado el papel del inconsciente en la vida psíquica del sujeto, pero es Freud —con permiso de Schopenhauer— quien haciendo uso de la división de la masa psíquica, emprende un largo camino junto a Charcot y Breuer, entre otros, en un intento de dar cuenta de modo satisfactorio de los casos patológicos a los que se enfrentaba en la praxis clínica.

[18] Realmente Arthur Schopenhauer había propuesto en *El Mundo como Voluntad y Representación* (1819/1844) muchas teorías sorprendentemente similares a las defendidas por Freud. Su concepto de voluntad tiene muchas conexiones con el inconsciente freudiano; además, ofrece una etiología de la locura que anticipa la teoría de la represión, y los puntos de vista de ambos sobre la sexualidad son extremadamente semejantes. Además ofrecen explicaciones similares en la explicación de los sueños. Parece que las conexiones entre ambos fue más allá de la mera influencia cultural, y que Freud leyó a Schopenhuer más, y mucho antes, de lo que siempre reconoció; en todo caso, si no fue de modo directo, sabemos que los maestros de Freud sí estuvieron en contacto con la obra de Schopenhuer. Para una visión más detallada de este asunto puede verse Young, Christopher y Brook, Andrew (1994), 'Schopenhauer and Freud', *International Journal of Psychoanalysis,* vol. 75, pp. 101-118.

Aunque ni podemos ni vamos a hacer un desarrollo exhaustivo del sistema teórico freudiano, sí conviene repasar algunos de los hitos más importantes de su teoría, así como una serie de conceptos elementales para el posterior análisis del autoengaño.

I. BREVES NOTAS SOBRE METAPSICOLOGÍA

Como es bien sabido, hay dos grandes tópicas en Freud. En la primera, Freud considera que el aparato psíquico está dividido en dos sistemas, el consciente y el inconsciente (que se subdivide a su vez en preconsciente e inconsciente). Sin embargo, más adelante, sin modificar absolutamente todo el planteamiento, cambiará este esquema como herramienta de trabajo por el también célebre Yo-Ello-Superyó. Veamos esto un poco más detenidamente:

Si lo consciente queda definido como las representaciones que se hallan presentes en nuestra conciencia y son objeto de nuestra percepción en un momento dado, denominaremos "inconsciente" al conjunto de representaciones que no percibimos, pero de cuya existencia estamos, sin embargo, ciertos basándonos en indicios y pruebas de otro orden. Según Freud, la hipótesis tan ampliamente criticada del inconsciente resulta evidente e innegable a la luz de ciertos argumentos.

En primer lugar, el conjunto de elementos correspondientes a la vida psíquica consciente resulta altamente incompleto y falto de coherencia si

no suponemos la existencia adicional de otros elementos inconscientes que les confieran apoyo explicativo.

> […] una ganancia de sentido y de coherencia es un motivo que nos autoriza plenamente a ir más allá de la experiencia inmediata. [Freud (1915b), p. 163].

En segundo lugar, tomando como herramienta de trabajo esta hipótesis, podemos diseñar métodos de actuación terapéutica que producen efectos en la vida consciente del sujeto; por último, la cantidad de representaciones que mantenemos de modo consciente en un momento concreto es muy limitada. Por tanto, la mayor parte de lo que denominamos como conocimiento consciente ha de permanecer por largos periodos en modo de latencia, es decir, de inconsciencia momentánea.

Por supuesto, como el propio Freud admite, puede aducirse que este tipo de fenómenos latentes no son parte del contenido psíquico del sujeto, sino "restos somáticos" de anteriores procesos psíquicos. Sin embargo, con este razonamiento estaríamos, o bien cayendo en una petición de principio, o bien mostrando una discrepancia meramente nominal, de nomenclatura. En este último caso, como en todo caso de diferencias de nomenclatura, nada puede contraargumentarse. Sin embargo, sí puede decirse algo: ¿resulta teóricamente provechosa la alternativa nominal? Según Freud, no. En primer lugar, rompe la continuidad en la vida psíquica del sujeto; en segundo lugar no es capaz de dar cuenta de algunos fenómenos para los que el psicoanálisis tiene propuestas; por último,

paraliza de antemano la investigación en un campo que quizá produzca en el futuro mayores avances.

Por tanto, según Freud, hemos de abrazar la hipótesis de que la vida psíquica del sujeto se compone de elementos tanto conscientes como inconscientes.

> Así como Kant nos alertó para que no juzgásemos a la percepción como idéntica a lo percibido incognoscible, descuidando el condicionamiento subjetivo de ella, así el psicoanálisis nos advierte que no hemos de sustituir el proceso psíquico inconsciente, que es el objeto de la conciencia, por la percepción que ésta hace de él. [Freud (1915b), p. 167].

Sin embargo, es preciso hacer aquí la primera puntualización. En diversos textos, Freud usa indiferentemente el término inconsciente en un sentido lato y en un sentido estricto. En su sentido lato, inconsciente designa todo lo no-consciente. Sin embargo, Freud especifica en su artículo de 1912 "Algunas observaciones sobre el concepto de lo inconsciente en el psicoanálisis" que aquellas representaciones que no son conscientes pero que pueden ser rescatadas y activadas en la consciencia serían parte del preconsciente, reservando el término 'inconsciente' para aquellas representaciones que no sólo no son conscientes, sino que no pueden serlo por estar reprimidas. Hasta ese momento, se consideraba que las ideas que se presentaban a la conciencia tenían una mayor vivacidad e intensidad, frente a aquellas de las que sólo se tenía un recuerdo o de las que simplemente no éramos conscientes. Ahora Freud nos indica que algunas de ellas, pese a su enorme intensidad y eficacia (capacidad de producir efectos en la actividad del sujeto), se mantienen

lejos de la conciencia. En sentido estricto, estás son las que forman parte del inconsciente, y coinciden plenamente con las representaciones reprimidas.

He aquí el otro concepto fundamental en Freud: la represión. En su artículo de 1915 titulado precisamente "La represión", Freud afirma que "la esencia de la represión consiste exclusivamente en rechazar y mantener alejados de lo consciente a determinados elementos". Esta afirmación ha provocado desconcierto, en tanto que en esta misma época afirma en otros pasajes que lo reprimido no agota todo lo inconsciente. Así, en otro artículo titulado "Lo inconsciente", también de 1915, Freud advierte:

> Todo lo reprimido tiene que permanecer inconsciente, pero queremos dejar sentado desde el comienzo que lo reprimido no recubre todo lo inconsciente. Lo inconsciente abarca el radio más vasto; lo reprimido es una parte de lo inconsciente. [Freud (1915b), p. 161]

Sin embargo, a nuestro juicio, una lectura atenta del texto muestra enseguida que el concepto de inconsciente que del que habla aquí Freud es el amplio, incorporando también, por tanto, las representaciones latentes del preconsciente. En el resumen que Freud hace de su teoría en 1923 en *El Yo y el Ello*, afirma:

> Lo reprimido es para nosotros el modelo de lo inconsciente. Vemos, pues, que tenemos dos clases de inconsciente: lo latente, aunque susceptible de conciencia, y lo reprimido, que en sí y sin más es insusceptible de conciencia […] en el sentido descriptivo hay dos clases de inconsciente, pero en el dinámico sólo una. Para muchos fines expositivos este distingo puede desdeñarse, aunque, desde luego, es indispensable para otros. Comoquiera que fuese, nos hemos habituado bastante a esta ambigüedad de lo inconsciente, y hemos salido airosos

> con ella. Hasta donde yo puedo ver, es imposible evitarla. [Freud (1923), p. 17]

En resumen desde el punto de vista *descriptivo* hay dos clases de inconsciente, es decir, al constatar aquello que no es objeto de conciencia, tanto lo latente capaz de hacerse consciente como lo reprimido son parte del inconsciente; sin embargo, desde el punto de vista *dinámico* sólo hay una clase de inconsciente, esto es, lo preconsciente no es esencialmente inconsciente en tanto en cuanto puede, en el movimiento del devenir de la consciencia, llegar a hacerse consciente.

De hecho, Freud afirma que todo acto psíquico comienza por ser inconsciente (léase preconsciente); posteriormente puede hacerse consciente o no. Si la razón de que no pase a ser consciente es que al tratar de hacerse consciente, encuentra una resistencia infranqueable que lo mantiene en estado no consciente, pasará a formar parte del inconsciente (en sentido estricto).

Para explicar esto podemos poner un ejemplo: Supongamos que voy caminando por la calle. Sin duda, mi cerebro capta muchos datos sensoriales a los cuales no atiende. Todos ellos son inconscientes, pero pueden llegar a hacerse conscientes, es decir, son preconscientes. De hecho, algunos de ellos (aquellos sobre los que centro mi atención) se hacen transitoriamente conscientes. Determinados estímulos pueden hacer cambiar mi foco de atención constantemente o focalizarla sobre un elemento concreto. A cada momento, algunas de aquellas representaciones que sin duda estaban siendo captadas de modo preconsciente pueden

hacerse conscientes. Sin embargo, si a una de ellas le es vetado el acceso a la consciencia, esto es, es reprimida, pasará a formar parte del inconsciente. ¿Pero cómo se produce la represión?

Según Freud, cuando un estímulo externo nos resulta doloroso, el método de defensa más adecuado sería la fuga. Sin embargo, cuando el estímulo es interno, "de nada vale la huida, pues el yo no puede escapar de sí mismo" [Freud (1915a), p. 141]. Trataremos de explicar, del modo más sencillo posible, cómo se produce este proceso según Freud.

En principio, el individuo tiene una serie de instintos, de impulsos libidinales orientados a la obtención de placer y la conservación (alimentarse, sexo, etc.) que generan grandes descargas de energía. Estas descargas de energía son asociadas a cierta representación (imagen, señal acústica, etc.) y esto produce una "carga" (también llamada "investidura", o "catexis") de tal representación. El individuo procede entonces a descargar esa energía, satisfaciendo el impulso libidinal o deseo. Sin embargo, en algunas situaciones,

> […] la satisfacción de la pulsión sometida a la represión sería sin duda posible y siempre placentera en sí misma, pero sería inconciliable con otras exigencias y designios. Por tanto, produciría placer en un lugar y displacer en otro. [Freud (1915a), p. 142].

Así, pese a que toda satisfacción de un impulso genera en principio placer, la satisfacción de ese impulso generaría un displacer mayor —en todo acto de represión esto es condición esencial— y por tanto, es refrenado y la representación asociada a la satisfacción de ese impulso es reprimida.

Es importante hacer notar que según la teoría freudiana, los procesos psíquicos implican la circulación y equilibrio de una energía pulsional cuantificable (esta perspectiva de la actividad psíquica es conocida como *economía*). En este sentido, cuando una representación es reprimida, por el acto de represión le es retirada la carga de energía con la que había sido cargada o catectizada (produciéndose así una *ausencia de catexis*). Dado que la energía ni se crea ni se destruye, la consecuencia inmediata de esto es que la energía sobrante queda libre y en disposición para asociarse a otra representación.

Detengámonos en este punto. Según Freud, hay dos momentos en la represión: una represión primitiva, y la represión propiamente dicha.

La primera fase de la represión consiste en que no se le permite el acceso a la consciencia a la representación de un impulso. En ese primer momento, el impulso queda fijado, ligado a esa representación.

> La segunda etapa de la represión, la represión propiamente dicha, recae sobre retoños psíquicos de la agencia representante reprimida o sobre unos itinerarios de pensamiento que, procedentes de alguna otra parte, han entrado en un vínculo asociativo con ella. A causa de ese vínculo, tales representaciones experimentan el mismo destino que lo reprimido primordial. La represión propiamente dicha es entonces una fuerza opresiva ('nachdrängen') posterior.[19] [Freud (1915a), p. 143].

[19] Hemos modificado ligeramente el texto. En concreto, en la traducción que manejamos —la *Standard Edition* de James Strachey— dice literalmente: "La represión propiamente dicha es entonces un «esfuerzo de dar caza»". Nos inclinamos aquí, para la traducción del término "nachdrängen", por la versión de Biblioteca Nueva realizada por Luis López Ballesteros que en todo momento hemos tenido a la vista. [Freud, Sigmund (1886/1939), *Obras Completas*, Madrid, Biblioteca Nueva, 9 tomos, 1974].

Aquí vemos una de las características principales de la represión: el retorno de lo reprimido. Efectivamente, la pulsión lucha por salir a la conciencia, pero al serle impedida la salida, busca otros medios; presenta ramificaciones, se disfraza o desplaza o entra en conexión asociativa con otras representaciones.

No podemos entrar en detalles aquí de cómo se producen estos mecanismos; baste con decir que no sólo aparecen en la vida psíquica de los individuos que presentan patologías, sino en casos de individuos sanos. Los olvidos de nombres o los actos fallidos son ejemplos de ello. Especialmente conocido es el caso de los sueños, en los que el inconsciente, aprovechando la baja vigilancia y relajamiento de la censura, trata de disfrazarse para pasar a la consciencia (esta teoría es célebremente recogida en *La interpretación de los sueños*, 1900).

Un aspecto importante de la represión es que ésta no es del todo inconsciente, pero tampoco en modo alguno consciente. Así, Freud señala lo siguiente:

> […] se comete un error cuando se destaca con exclusividad la repulsión que se ejerce desde lo consciente sobre lo que ha de reprimirse. En igual medida debe tenerse en cuenta la atracción que lo reprimido primordial ejerce sobre todo aquello con lo cual puede ponerse en conexión. Probablemente, la tendencia a la represión no alcanzaría su propósito si estas fuerzas [atracción y repulsión] no cooperasen, si no existiese algo reprimido desde antes, presto a recoger lo repelido por lo consciente. [Freud (1915a), p. 143].

Resulta obvio que lo primitivamente reprimido ha de ejercer esta función de atracción desde el inconsciente al que ha sido relegado.

Con esto quedan expuestos, de modo somero, los puntos más importantes respecto de la primera tópica (Consciente/ Preconsciente / Inconsciente). Freud, en torno a 1923 (aunque ya había signos desde 1914) instaura lo que es conocido desde entonces como la segunda tópica (Yo/Ello/Superyó).

La razón para este giro se halla en lo siguiente: desde sus comienzos, Freud había tratado las alteraciones patológicas como resultantes del conflicto entre lo consciente y lo inconsciente. Tomando prestado de Georg Groddeck el término '*Es*' ('*Ello*'), que éste había acuñado para denominar a las fuerzas desconocidas e indomables que nos habitan, Freud denomina Ello a lo inconsciente, los impulsos y lo reprimido. El 'Yo' ('*Ich*') correspondería en principio a lo consciente y preconsciente. Sin embargo, los resultados clínicos apuntaban a un dato sorprendente. A menudo los pacientes tropezaban con dificultades cuando se les pedían ciertas labores o sus asociaciones cesaban a medida que se iban acercando al núcleo del suceso traumático, esto es, manifestaban resistencias. Aunque el sujeto fuese consciente del displacer experimentado y pudiese inferir que estaba siendo víctima de resistencias, no lograba darles nombre ni describirlas. Incluso refería no saber nada si se le indicaba la presencia de tales resistencias. Pero entonces,

> […] esa resistencia seguramente parte de su yo y es resorte de este […] Hemos hallado en el yo mismo algo que es también inconsciente, que se comporta exactamente como lo reprimido [Freud (1923), p. 19].

Este es el gran descubrimiento que queda reflejado en la segunda tópica: también en el Yo hay algo inconsciente.

> También una parte del yo, Dios sabe cuán importante, puede ser inconsciente, y es seguramente inconsciente. Y esto Inconsciente del yo no es latente en el sentido de lo Preconsciente, pues sí así fuera no podría ser activado sin devenir consciente, y el hacerlo consciente no depararía dificultades tan grandes. Puesto que nos vemos así constreñidos a estatuir un tercer Inconsciente, no reprimido, debemos admitir que el carácter de la inconsciencia pierde significatividad para nosotros. Pasa a ser una cualidad multívoca que no permite las amplias y excluyentes conclusiones a que habríamos querido aplicarla. Empero, guardémonos de desdeñarla, pues la propiedad de ser o no consciente es en definitiva la única antorcha en la oscuridad de la psicología de las profundidades. [Freud (1923), p. 19-20].

Esta parte inconsciente del Yo es lo que Freud denomina el 'Superyó' ('*Über-Ich*'). Trataremos de hacer de nuevo un resumen sencillo del proceso genético de tales instancias.

Según la nueva concepción freudiana, el niño nace siendo puro Ello: pura pulsión libidinal. En un primer momento, es guiado por el "principio del placer", que le instiga a la satisfacción de estos impulsos, para lo cual incluso hace uso de la fantasía. Sin embargo, la realidad se le impone: por medio de la fantasía el impulso no ha sido satisfecho. Este choque con la realidad es que el introduce el "principio de la realidad", según el cual el individuo ahora ha de tener en cuenta el entorno para la satisfacción de sus deseos. A partir de ese momento se forma el Yo, que no es sino el producto del choque del Ello con la realidad, y por tanto una especificación del Ello, para el cual las percepciones son lo que para el Ello eran los impulsos.

En este movimiento de impulsos y búsqueda de satisfacción, el niño es posteriormente coartado por una fuerza externa: sus padres. Ahora ni siquiera puede hacer todo lo que la realidad le permite para satisfacer sus impulsos. Este fenómeno, conocido en la terminología psicoanalítica como 'castración' hace referencia, en particular, al hecho de la amenaza al niño con cortarle su miembro si ejercita la masturbación. Sin embargo, en un sentido más amplio, representa toda restricción paterna, según la cual el niño interioriza las normas sociales. Este es el punto de formación del Superyó, que Freud colocaba en la supresión del complejo de Edipo: el niño, que en principio tenía un sentimiento de amor hacía su madre y de aversión hacia su padre (porque el padre suponía un obstáculo para llegar a su madre), recibe la prohibición paterna de dormir con la madre. Esta imposición es interiorizada posteriormente por el niño, que acepta la prohibición y se identifica con el padre. A partir de entonces, se forma el Superyó, que no es sino la suplantación del complejo de Edipo.

> […] el superyó del niño no se edifica en verdad según el modelo de sus progenitores, sino según el superyó de ellos; se llena con el mismo contenido, deviene portador de la tradición, de todas las valoraciones perdurables que se han reproducido por este camino a lo largo de las generaciones. [Freud (1932), p. 62].

Algunos psicoanalistas, como Melanie Klein colocan el punto de formación del Superyó mucho antes, en la "fase oral": aquella en la que el bebé obtiene placer en llevarse cosas a la boca) desde el momento en que se distingue entre la introyección de objetos "buenos" y "malos"; o Sándor Ferenczi que la sitúa en la "fase anal" en la que al niño se le enseña a controlar sus esfínteres por medio también de una prohibición. En

cualquier caso, el Superyó es el Yo ideal (*ideal Ich*), la voz de la conciencia, el imperativo categórico, que funciona como juez del Yo, dictándole qué le está permitido hacer y qué le está vedado, e introduciendo angustia si no sigue sus dictámenes. Sin embargo, también ejerce de representante del Ello, en tanto que no es sino una especificación del Yo (y por tanto, subespecificación del Ello) y el Ello busca la satisfacción de todos sus impulsos.

Ahora sabemos de dónde provienen las exigencias a las que anteriormente hacíamos referencia y que hacían de la satisfacción de algunos impulsos algo displacentero: el Superyó. En palabras de Freud, «el superyó es, genéticamente, heredero de la instancia parental; a menudo mantiene al yo en severo vasallaje» [Freud (1927), p. 160][20]. El sujeto se haya entre los impulsos ciegos en busca de placer (Ello) y las exigencias de las normas parentales y sociales interiorizadas por el niño (Superyó); en esta tensión, el Yo es un vasallo entre dos señores que ejercen presión para que siga sus dictámenes. Por esta razón, el Yo no sólo debe acomodarse a las limitaciones físicas que le impone la realidad, sino que lo que en principio sería placentero, es displacentero al chocar frontalmente con el comportamiento que le exige el yo ideal.

[20] Freud, Sigmund (1927), 'Der Humor', *Almanach*, pp. 9-16. Reeditado en *Gesammelte Werke*, vol. 14, pp. 383-389. [Edición en castellano: 'El humor', en *Obras completas*, vol. 21, pp. 157-162].

II. AUTOENGAÑO

Con estas herramientas de trabajo, Freud está capacitado para tratar la cuestión del autoengaño. Sin embargo, nunca estudio esta cuestión de modo explícito. Hay varias referencias al engaño a uno mismo a lo largo de su obra y, sin duda alguna, Freud creía que a la base de muchos trastornos había un engaño a sí mismo por parte del paciente; sabemos, además, que Freud afirmó que

> […] no hace falta estar gobernado por un delirio para actuar de ese modo; más bien es un suceso cotidiano, aun en personas sanas, que se engañen acerca de los motivos de su obrar y sólo devengan conscientes de ellos con posterioridad, toda vez que un conflicto entre varias corrientes de sentimiento les cree las condiciones para ese estado de confusión. [Freud (1907), p. 55][21]

Sin embargo, no queda claro que Freud esté pensando en un autoengaño en sentido fuerte. Ya hemos comentado que "estar engañado" es equivalente, en una de sus acepciones, a "estar equivocado", "estar en el error"; para hacer de este uso un uso en sentido fuerte, hemos de exigir dos cosas: intención por parte del sujeto y consciencia.

Como sucede en la cita anterior, sus palabras pueden ser interpretadas en otros casos en el sentido de un error provocado por uno mismo, pero no queda claro en modo alguno que sea voluntario y en todo caso, afirma específicamente que es un proceso inconsciente, aunque pueda devenir en consciente posteriormente. Presumiblemente, si el sujeto se hace consciente

[21] *Der Wahn und die Traume in W. Jensens «Gradiva»*, Leipzig/Viena, H. Heller. Reeditado en *Gesammelte Werke*, vol. 7, pp. 31-125. [Edición en castellano: 'El delirio y los sueños en la «Gradiva» de W. Jensen', en *Obras completas*, vol. 9, pp. 7-77].

abandonará entonces su estado de engaño, por lo que parece que las palabras de Freud pueden ser analizadas como impidiendo, al menos, un autoengaño consciente. Si el autoengaño se produce, ha de ser por medio de un mecanismo parcialmente inconsciente.

La terapia psicoanalítica se fundamenta en sacar a la luz consciente lo reprimido (que constantemente empuja y quiere salir, produciendo displacer) evitando que se produzca ese displacer; en disociar la representación de ese afecto (retirando la catexis), de modo que la representación pueda emerger en la vida consciente sin generar angustia u otro sentimiento que ponga en peligro la salud psíquica del paciente. En lo que se refiere a los procesos de defensa de los sujetos sanos (Freud se esforzó siempre por mostrar una continuidad entre los procesos de la vida cotidiana y los procesos patológicos) el sujeto que se autoengaña buscaría, en primer lugar, eliminar la angustia (como más tarde harán Mark Johnston o Annette Barnes) o cualquier otra poderosa fuente de displacer. En último término, el éxito sería total si lograse disociar la representación (por ejemplo, las representaciones que indiquen que su hijo se droga, que su mujer le pone los cuernos, que no logrará el amor de su pretendida, o que tiene una enfermedad terminal) del sentimiento. Esto, sin embargo, no parece a primera vista posible, pues no se trata de una asunción de un hecho traumático, sino de la confrontación de nuestras más altas aspiraciones con la realidad; para decirlo con Freud, de nuestro Superyó con la realidad. No obstante el propio Freud indicó que hay varios tipos de represión. Así pues, trataremos de ensayar una respuesta tentativa a nuestro problema del modo más fiel posible a los textos freudianos.

Hemos dicho ya que Freud nunca expuso una teoría al respecto; sin embargo, creemos que no resulta difícil suponer cual habría sido aproximadamente su respuesta. Sin duda, hay un postulado de partida en Freud que muestra cómo el Yo trata siempre de preservar un equilibrio psicológico y tiende a reducir tanto la angustia como cualquier otro sentimiento que ponga en peligro tal equilibrio.

Sabemos también que Freud introdujo, como hemos señalado anteriormente, un "principio de realidad", según el cual el sujeto no puede tratar de satisfacer sus necesidades, impulsos e instintos al margen de los límites que le impone la propia realidad y, por tanto, ha de acomodarse a ella. Así pues, hacerse una imagen distorsionada de la realidad sería, en principio, inadecuado para poder alcanzar la satisfacción de las demandas del Ello. Hemos indicado también que, en determinadas circunstancias, la satisfacción de un impulso puede acarrear displacer por contravenir los preceptos interiorizados por el individuo que conforman su yo ideal. Pero queremos indicar otro aspecto de displacer importante: a veces, el principio de realidad nos presenta una imagen que generaría displacer por sí misma. El Yo se hallaría entonces en una situación en la que aceptar la representación que le impone el principio de realidad supondría una angustia o dolor tal que, en un intento de salvaguardar su equilibrio, se produce todo un proceso de defensa contra esa representación dolorosa o angustiosa de la realidad (lo que se reprime no es la realidad, sino su representación). Como ya quedó dicho, la defensa es en un primer momento inconsciente; posteriormente, será sólo parcialmente consciente

(pues lo primitivamente reprimido sigue ejerciendo de polo de atracción de las ramificaciones y representaciones asociadas a ello).

Así pues, ante una realidad que resulta insoportablemente angustiosa (presumiblemente por contravenir los preceptos del Superyó o del Yo ideal) se produce una defensa inconsciente por medio de la cual los vínculos establecidos entre tal representación y cierta energía pulsional (es decir, la energía en forma de la atención que le prestamos, los visos de realidad y verdad que le otorgamos, etc.) le son retirados (ausencia de catexis) y, como consecuencia, parte de esa energía que queda libre (la energía ni se crea ni se destruye) es reconducida hacia otra representación y vinculada a ella (contracatexis). Este tipo de representación con la que es contracatectizada puede ser derivada de la representación consciente, es decir, un substitutivo (casos de fobias y otros trastornos mentales) o una formación reactiva (es decir, una representación contraria, como por ejemplo, un marido fiel o un hijo modélico). Dicho breve y sencillamente: se relegan ciertos elementos al inconsciente, prestando entonces la atención y crédito a una representación opuesta placentera.

Restan sólo por matizar ciertas cuestiones: ¿Este tipo de engaño y represión, de ser posibles del modo descrito, pueden ser efectivos y duraderos? Sabemos que lo reprimido tiende a intentar volver a la conciencia, bien cuando consiste en una pulsión proveniente del Ello, bien cuando es una representación impuesta por la propia realidad; esto es lo que se conoce como "el retorno de lo reprimido". Lo reprimido trata de vencer las resistencias, se transforma, desplaza su energía a representaciones

similares, etc. Y de hecho, algunas ramificaciones y elementos asociados a la represión primordial u originaria se presentan de modo consciente. En palabras de Freud,

> […] ni siquiera es cierto que la represión mantenga apartados de lo consciente a todos los retoños de lo reprimido primordial. Si estos se han distanciado lo suficiente del representante reprimido, sea por las desfiguraciones que adoptaron o por el número de eslabones intermedios que se intercalaron, tienen, sin más, expedito el acceso a lo consciente. Es como si la resistencia que lo consciente les opone fuese una función de su distanciamiento respecto de lo originariamente reprimido. [Freud (1915a), p. 144].

La pregunta que ciertamente se nos impone, ya fue planteada por el propio Freud, cuando a continuación apostilla:

> ¿Hasta dónde tiene que llegar la desfiguración, el distanciamiento respecto de lo reprimido? Es algo que no podemos indicar en general. Ahí opera un fino sopesamiento cuyo juego se nos oculta; empero, las modalidades de su acción eficaz nos hacen colegir que se trata de detenerse antes que se llegue a determinada intensidad en la investidura de lo inconsciente, rebasada la cual lo inconsciente irrumpiría hacia la satisfacción. La represión trabaja, entonces, de manera *en alto grado individual*. [Freud (1915a), p. 145].

Sin embargo esta explicación no consigue iluminar algunas oscuridades cruciales en la explicación. En concreto, ¿cuál es la instancia que realiza tal sopesamiento o valoración? ¿En dónde se halla situada? Es decir, ¿pertenece al terreno de lo consciente, de lo preconsciente o de lo inconsciente?

Por otro lado, dado que el objetivo de la represión no era otro que hacer desaparecer el displacer que ponía en peligro el equilibrio psíquico,

no será el impedimento de acceso a la vida consciente de la representación lo más importante, ni lo que determine el éxito del proceso defensivo. Por ello,

> Si una represión no consigue impedir que nazcan sensaciones de displacer o de angustia, ello nos autoriza a decir que ha fracasado, aunque haya alcanzado su meta en el otro componente, la representación. [Freud (1915a), p. 148].

Así pues vemos que, desde una perspectiva estrictamente freudiana, parece que en la explicación del autoengaño resulta difícil explicar:

a) Cómo se produce el mecanismo de represión.

b) Cómo, qué instancia y en qué lugar del aparato psíquico (consciente, preconsciente o inconsciente) se efectúa el fino cálculo para estimar cuando las ramificaciones se alejan lo suficiente como para no ser ya objeto de represión.

c) El éxito del autoengaño. Pues aun cuando se sortee el problema de la constante amenaza de la realidad (principio de realidad) y el sujeto evite que la representación (en forma de imagen mental, sospecha, creencia o conocimiento) acceda a la consciencia, lo reprimido trata de acceder constantemente, introduciendo angustia en la vida consciente del sujeto si no logra su objetivo. Por tanto, el objetivo principal no se habría logrado.

Desde esta perspectiva, si sólo se consigue suprimir la representación, creencia o pensamiento dolorosos, pero la angustia reaparece, el sujeto puede presentar ya sustitutivos y síntomas de un estado patológico. Esto,

aunque es en principio posible, nos aleja del asunto que pretendíamos abordar: el autoengaño en tanto que fenómeno paradójico, pero bastante usual, cotidiano y, en todo caso, propio de sujetos sanos.

II.1.2 - Los mecanismos de defensa

Como es bien sabido Anna Freud, hija de Sigmund Freud, continuó la labor iniciada por su padre en la investigación de la psique y sus trastornos, tomando un camino que ha sido bastante criticado por ciertos sectores dentro del psicoanálisis. Si bien, como es lógico, aquí no vamos a entrar en la valoración de estas cuestiones, sí nos parece interesante examinar algunas de las ideas de Anna Freud, en tanto que sus estudios contribuyeron al conocimiento de los procesos defensivos, y éstos parecen estar a primera vista a la base de los mecanismos de autoengaño.

La noción de defensa había sido introducida en el terreno psicoanalítico por su padre en el artículo de 1894 "Las neuropsicosis de defensa", aunque el propio Freud en un primer momento no hace un uso muy técnico del concepto, llegando en ocasiones a mezclarlo con el de represión. No obstante, Freud reconocerá en *Inhibición, síntoma y angustia* (1926) que la represión no es sino uno de los muchos mecanismos de defensa que son puestos en marcha ante elementos angustiosos o displacenteros. En este momento define la defensa o proceso defensivo (*Abwehrvorgang*) como

> […] la designación general de todas las técnicas de que el yo se vale en sus conflictos que eventualmente llevan a la neurosis, mientras que «represión» sigue siendo el nombre de uno de estos métodos de

> defensa en particular, con el cual nos familiarizamos más al comienzo, a consecuencia de la orientación de nuestras indagaciones. [Freud (1926), p. 153]

En este contexto, Anna examinó en *El Yo y los mecanismos de defensa* (1936) los distintos mecanismos que el Yo emplea para defenderse de la irrupción en la vida consciente de ciertos contenidos que generarían displacer. Este estudio había estado centrado, en un primer momento del desarrollo del psicoanálisis, en el inconsciente. A través de la hipnosis, se trataba de anular la parte consciente del paciente (y sus resistencias) para poder acceder libremente a su inconsciente, encontrar la representación reprimida y reinsertarla en la vida anulando con ello el síntoma. Sin embargo, según Anna Freud, esto generaba problemas posteriores: el Yo había quedado fuera del proceso terapéutico y, una vez finalizada la hipnosis, cualquier sobrevenida de la representación patógena generaba nuevas defensas del Yo que desbarataban el éxito conseguido [Anna Freud (1936), p. 33].

Estos resultados fueron los que rápidamente inclinaron a Sigmund Freud a abandonar el método de la hipnosis, sustituyéndolo por el de la libre asociación en el que el paciente verbaliza el flujo de pensamientos que le pasan por la mente. Con este método, el Yo del sujeto no se ponía al margen; simplemente se lo invitaba a no ejercer ningún tipo de defensa, a abandonar la crítica de las ideas que se le ocurriesen y la necesidad de establecer conexiones lógicas.

> Por así expresarnos, se le pedirá al yo que calle, y bajo la promesa de que en su acceso a la conciencia sus derivados no encontrarán los

> obstáculos acostumbrados, se invitará a hablar al ello. [Anna Freud (1936), p. 34]

El paciente comienza a mostrar resistencias a medida que el analista se acerca al núcleo de lo reprimido. Es en este momento, según Anna Freud, cuando se producen los mayores avances; a través de las resistencias, de las transformaciones, transferencias, etc. el analista puede estudiar los mecanismos de defensa empleados por el Yo. Por esta razón, Anna Freud inaugura una línea en la que el Yo, que había sido dejado a un lado frente al estudio del Ello, el inconsciente y sus formaciones, captará toda la atención, ya que es sólo a través del estudio de sus mecanismos de defensa y de la relación dialéctica con Ello, como podemos acceder a lo reprimido y a la fuente de angustia. Esta línea se conoce como *Ego-Psychology,* y se ha desarrollado principalmente en el ámbito anglosajón.

I. MECANISMOS DE DEFENSA

Como ya indicara Freud, ante determinadas situaciones que pueden poner en peligro el estado emocional del sujeto, su yo (que es en cualquier caso la instancia psíquica que cumple las funciones de *agente* y está encargada del control del sistema motor) se defiende de la representación o afecto displacentero poniendo en marcha algún mecanismo. Hay cuatro motivos de defensa:

1. *Angustia objetiva*: ante amenazas externas

2. *Angustia instintiva*: amenazas internas, de los instintos

3. *Angustia de conciencia*: ante las severas exigencias del superyó

4. *Necesidad del yo de lograr o mantener su síntesis* ante impulsos contradictorios o conflictos intra o intersistémicos. [Anna Freud (1936), pp. 64-65, 71]

Las amenazas son, a veces, la simple previsión de no-cumplimiento. Cuando el yo se da cuenta de que la realidad externa no va a procurarle los medios para satisfacer las demandas de sus impulsos, puede iniciar un proceso de negación, por ejemplo, de la realidad y posteriormente una creación fantasiosa de una realidad placentera. En otras ocasiones, la amenaza es literal: algún otro agente, humano o animal, amenaza la satisfacción. Estos motivos de angustia externos son denominados *objetivos*. Frente a ellos están los *subjetivos*, que provienen del propio sujeto: a veces el sujeto percibe sus impulsos como demasiado fuertes, continuos o recurrentes, ante lo cual también inicia un proceso defensivo: en esto consiste la angustia por los instintos, instintiva. En otras ocasiones, es el Superyó el que con los remordimientos o problemas de *conciencia*, genera angustia ante la contemplación de un impulso, y trata de impedirle su satisfacción; ésta es la angustia de conciencia. Por último, en otras ocasiones dos o más impulsos son incoherentes o inconsistentes, pues su satisfacción resulta contraria o contradictoria (no se pueden satisfacer ambos). A estos conflictos podríamos denominarlos intrasistémicos, pues se producen dentro de uno de los sistemas, bien del Ello bien del Superyó. Otras veces los conflictos se producen entre ambos sistemas: ninguno tiene la fuerza suficiente para imponerse con claridad; así, el impulso no se

satisface porque el Superyó se opone, pero éste tampoco es capaz de evitar su acceso a la conciencia. Esta tirantez o conflicto intersistémico puede amenazar la integridad del yo, y genera una angustia ante la que se produce el proceso defensivo.

Los mecanismos de defensa que había manejado Sigmund Freud son diez: represión, regresión, formación reactiva, aislamiento, anulación, proyección, introyección, vuelta contra sí mismo, transformación en lo contrario y sublimación o desplazamiento del objeto. [Anna Freud (1936), p. 54]. A éstos Anna Freud añade la identificación con el agresor, el altruismo, la racionalización y la intelectualización. A continuación se ofrece una breve descripción de cada uno de ellos:

- *Represión*: impedimento de acceso a la conciencia desde el inconsciente de la representación de un impulso cuya satisfacción podría introducir displacer o angustia.

- *Regresión*: es el retorno a una etapa de desarrollo mental anterior.

- *Formación reactiva*: La representación dolorosa es substituida por una agradable.

- *Aislamiento*: es la separación de una representación y los afectos o sentimientos dolorosos asociados a ella.

- *Anulación o Negación*: el individuo trata factores obvios de la realidad como si no existieran.

- *Proyección*: es el mecanismo por el cual afectos, ideas, sucesos o acciones dolorosas son proyectadas hacia otras personas o cosas cercanas pero que el individuo siente ajenas.

- *Introyección*: es el mecanismo por el cual el sujeto incorpora afectos, ideas, sucesos o acciones placenteros que en realidad pertenecen a otras personas o cosas cercanas.

- *Transformación en lo contrario*: La representación, suceso o afecto doloroso es sustituido inmediatamente por su contrario.

- *Sublimación*: el impulso es canalizado a una nueva y más aceptable salida, hacia objetos social o moralmente mejor considerados.

- *Desplazamiento*: la representación o afecto conectado a una persona o hecho en particular se vincula a otra persona o hecho.

- *Racionalización*: es la sustitución de las verdaderas razones de un suceso o acción por otras falsas pero aceptables, cuando las primeras resultan dolorosas.

- *Restricción del yo*: El individuo, ante la posibilidad de fracaso, cede de antemano en todo empeño por lograr alcanzar sus objetivos.

- *Altruismo*: Es un tipo de proyección. El individuo renuncia a la satisfacción de sus propios impulsos, proyectándolos en otros, a quienes ayudar en su satisfacción.

- *Identificación con el agresor*: Es un tipo de introyección. El sujeto agredido introyecta los comportamientos de los que ha sido víctima.

- *Intelectualización*: Es un tipo de sublimación típico de la pubertad en el que la pulsión se reconduce principalmente hacia la actividad artística y la investigación intelectual.

Anna Freud señala que "los motivos que determinan al yo a la elección de un señalado mecanismo son poco conocidos" y apunta a que posiblemente la represión sea un mecanismo exclusivo ante los impulsos sexuales; o quizá el resto de mecanismos sólo se ponen en acción una vez que la represión no ha logrado su cometido o que simplemente retorna lo reprimido [Anna Freud (1936), p. 58]. Su padre creía que la represión, más bien, era parte o ayuda de los otros procesos defensivos. En cualquier caso,

> El punto crucial es que la angustia del yo —sea como temor ante el mundo externo o como temor ante del superyó— activa el proceso defensivo. [Anna Freud (1936), p. 63]

El yo ejerce una función de negociación entre las exigencias del Ello y las demandas del Superyó. Tiene que arreglárselas para lograr el equilibrio psicológico y evitar la angustia, satisfaciendo en la medida de lo posible los impulsos del ello sin contravenir los principios del Superyó.

> El yo triunfa cuando sus funciones defensivas cumplen su propósito; cuando con su ayuda logra limitar el desenvolvimiento de la angustia y del displacer y asegurar al individuo —inclusive en circunstancias difíciles— alguna satisfacción por medio de las transformaciones

> instintivas necesarias; por tanto, cuando en la medida de lo posible, logra establecer una armonía entre el ello, el superyó y las fuerzas del mundo. [Anna Freud (1936), p. 142]

II. AUTOENGAÑO

Anna Freud tampoco estudió directamente el fenómeno del autoengaño. Si hemos considerado oportuna la inclusión de sus estudios en este trabajo sobre el autoengaño se debe a que la puesta en marcha de estos mecanismos de defensa del yo ante un suceso displacentero —sean instintos, amenazas, códigos sociales, legales o morales, etc.— supondría un engaño del sujeto para consigo mismo con respecto a su realidad, que es transformada en otra más llevadera. La cantidad de mecanismos estudiados por Anna pondría de relieve la variedad de tipos de autoengaño y su riqueza, así como la complejidad del problema: no se trataría ya simplemente de un sujeto engañándose a sí mismo; los métodos de engaño pueden ser muchos. Ahora quien se autoengaña quizá simplemente *niega* en palabras o actos la realidad; o quizás *reprime* el suceso traumático (impidiéndole constantemente el acceso a la consciencia), o lo *aísla* (elimina los afectos asociados a él, resultando entonces tal suceso accesible a la consciencia pero inocuo); o lo *proyecta* (atribuyéndole la falta o problema a otro), lo *transforma en su contrario*, o lo *sublima* (es decir, transforma un impulso en otro moralmente más elevado).

Por supuesto, en la medida en que la represión es el mecanismo más importante y posibilitador de otros, el *inconsciente* tiene un papel fundamental para la ejecución de estas defensas. Por tanto, la explicación

del autoengaño basada en la apelación a estos procesos defensivos heredará en su mayoría los problemas teóricos que planteaba la instancia del inconsciente y la actividad de la represión.

Algunos modelos defensivos, como la sublimación, exigen además la formación de una conciencia moral o Superyó, lo cual supone que son defensas más tardías cronológicamente; otras como la negación y creación de fantasías se dan desde una edad muy temprana. Así, no nos resulta en absoluto extraño observar cómo un niño relata fantasías en las que parece dar rienda suelta a sus deseos e impulsos; sin embargo, como Anna Freud hace notar,

> La capacidad del yo de negar la realidad, hállase en radical contradicción con otra función muy apreciada por él: la capacidad de reconocer la realidad y valorarla críticamente. En la primera infancia esta contradicción no opera aún como trastorno. [Anna Freud (1936), p. 78]

No obstante, Anna subraya que en la edad adulta, la sustitución de la realidad por eventos fantasiosos se considera, además de peligrosa por no ser adaptativa, una patología grave[22]. Asimismo, conseguir mediante este tipo de defensas un equilibrio psíquico no supone una garantía de estabilidad. Por un lado, lo reprimido busca constantemente emerger a la

[22] «Este mecanismo de defensa pertenece a una fase normal del desenvolvimiento del yo infantil. Pero si la encontramos en la vida ulterior, será indicio de un grado avanzado de enfermedad psíquica. En ciertos episodios agudos de confusión psicótica, el yo del individuo no se comporta de otra manera frente a la realidad. Bajo los efectos del un shock como, verbigracia, la impresión de una súbita pérdida de un objeto de amor niega el hecho real y sustituye algún aspecto de la realidad insoportable por una ilusión de algo deseado» [Anna Freud (1936), p. 78].

conciencia; por otro, la realidad constantemente amenaza con desmontar el proyecto defensivo. Además, cualquier evento traumático puede hacer aflorar otros sucesos previamente reprimidos, devolviendo al sujeto al estado de angustia del que se había estado defendiendo.

> Cuando a causa de alguna catástrofe —pérdida del objeto de amor, enfermedad, miseria, guerra—, el individuo sufre un cambio inesquivable en su forma de vida, el yo ha de enfrentarse nuevamente con la primitiva situación de angustia. [Anna Freud (1936), p. 95]

Sin embargo, aun cuando la censura tenga éxito en mantener alejadas de la conciencia las representaciones, afectos o situaciones dolorosas, la represión puede llegar a ser peligrosa, pues

> La disociación del yo, producida por la sustracción a la consciencia de porciones totales de la vida afectiva e instintiva, es susceptible de destruir en forma definitiva la integridad personal. [Anna Freud (1936), pp. 57-58]

En resumen: siempre desde una perspectiva psicoanalítica, las fuentes generadoras de tensión, displacer o angustia son varias: internas (fuerza de los instintos, patrones morales interiorizados) o externas (negativas, amenazas, límites reales, sociales o legales, etc.). El sujeto dispone de diversos medios de protegerse de la realidad dolorosa. Unos operarían desde la instancia consciente para alejar representaciones de la consciencia (como la evitación de un estímulo doloroso externo); otros (como el aislamiento, la proyección o la conversión en el contrario) exigirían mecanismos parcialmente ocultos o inconscientes, que relegarían asimismo la representación o evento doloroso al inconsciente. Por otro lado, algunos de ellos, como la negación y transformación en lo contrario

por medio de la fantasía, son comunes en la infancia, pero revelarían patologías graves en la edad adulta. El resto, aunque pueden ser empleados por sujetos sanos, podrían desembocar fácilmente en trastornos neuróticos y psicóticos; sólo el desplazamiento se produciría en la vida sana de los sujetos (por ejemplo en los sueños, para evadir la censura). Es importante señalar esto, pues en casos de doble personalidad u otros trastornos psicológicos el autoengaño pasaría a ser objeto de estudio médico y no ya una paradoja filosófica.

En todo caso, la represión parece, además del mecanismo más importante de defensa, el más eficaz, pues retiraría de la franja consciente los datos dolorosos. Sin embargo, no sólo resultaría desventajosa adaptativamente al no transmitirnos una idea fiel de la realidad, sino que daría lugar a un estado continuamente amenazado por ésta y, lo que es más grave, pondría en peligro la unidad o integridad del yo, pudiendo ocasionar alguna patología severa (como la esquizofrenia o la doble personalidad).

II.1.3 - Una crítica demoledora al psicoanálisis

> La explicación por lo inconsciente, por el hecho de que rompe la unidad psíquica, no puede dar razón de los fenómenos que, a primera vista, parecen pertenecerle.
>
> [Sartre (1943), p. 104]

En el parágrafo §I.7 exponíamos la concepción sartriana de la mentira a uno mismo, que él denomina "mala fe" (*mauvaise foi*) y que, como explicamos anteriormente, consideramos significativamente distinta del

autoengaño. Indicábamos que el propio Sartre señala que aceptaría de buen grado definir la mala fe como *mentirse a uno mismo*, a condición de que inmediatamente a continuación se distinguiese el mentirse a uno mismo de la mentira a secas. Como vimos, la mentira a secas está vinculada a la pretensión de engaño a otro, y por tanto a la duplicidad de conciencias que posibilita tal engaño; por ello, la posibilidad del mentirse a uno mismo —dentro de una unidad psíquica— supone para Sartre tales dificultades lógicas o psicológicas, que la explicación ha de transcurrir necesariamente por unos derroteros totalmente alejados de nuestra concepción usual o intuitiva del autoengaño. La mala fe consistía en que el hombre jugaba a ser lo que no era (dado que por definición uno sólo puede intentar ser lo que no es), y a no ser lo que era (pues rehusaba reducirse a sus conductas pasadas), pero en ningún caso Sartre habla de un proyecto (exitoso o no) de engaño a uno mismo.

Por otro lado, cuando Sartre escribe *El Ser y la Nada* tampoco le resultan prometedoras —aunque sí un esfuerzo meritorio— las distintas explicaciones que había ofrecido la doctrina psicoanalítica para tratar de solventar los problemas conceptuales suscitados por la mentira o engaño a uno mismo mediante la apelación a la escisión de la masa psíquica, el inconsciente, los mecanismos de defensa y la represión. Efectivamente, el mayor problema que plantea el autoengaño según Sartre es el de la unidad psíquica y la intencionalidad. Un [auto]engaño sin intención parece ya menos engaño, y el autoengaño en un sujeto cuya masa psíquica está de algún modo escindida, reintroduce la dualidad que permite el autoengaño a costa de alejar el problema de los casos típicos de autoengaño. Por

supuesto el psicoanálisis parte del supuesto de que esta escisión no es en modo alguno patológica (aunque posibilite desvíos defensivos patológicos) sino constitutiva, pero este supuesto es duramente atacado por Sartre por medio de una finísima argumentación.

El psicoanálisis ha sustituido, según Sartre, la noción de mala fe por "la idea de una mentira sin mentiroso": ahora no se trata de explicar cómo el sujeto se engaña a sí mismo, sino cómo *es mentido*, ya que al escindir la masa psíquica, puede colocar una instancia frente a otra, recuperando así la dualidad de engañador y engañado que se imponía como condición esencial del engaño, a través del "Ello" y el "Yo".

Sin embargo, Sartre advierte de que esta operación no es tan sencilla como parece a simple vista, y que esconde dificultades importantes. En primer lugar, Freud señala en su experiencia clínica resistencias al acercarse al núcleo reprimido de las neurosis… Pero entonces cabe preguntarse, ¿qué parte del paciente es la que se resiste así?

Desde luego —Sartre reflexiona— no puede ser el "Yo", pues esta instancia consciente se encuentra en la misma posición de *prójimo* que el psicoanalista frente a sus reacciones. A lo sumo, el Yo del paciente puede observar el grado de probabilidad de las hipótesis del analista en función de la extensión de hechos que pueden explicar. Por supuesto hemos de suponer que el paciente no miente o trata de engañar al analista, pues en caso contrario inmediatamente se resolvería el presunto problema: no habría autoengaño, sino engaño a otro (al analista en este caso); supongamos, por tanto, que acude por propia voluntad para descubrir

algo que desconoce. Pero tampoco puede tratar de engañarse conscientemente a sí mismo:

> ¿Se dirá que el enfermo se inquieta por las revelaciones cotidianas que le hace el analista y que trata de hurtarse a ellas a la vez que finge a sus propios ojos querer proseguir la cura? En tal caso, ya no es posible recurrir al inconsciente para explicar la mala fe: ésta está ahí, en plena conciencia, con todas sus contradicciones. Pero no es así, por otra parte, como el psicoanalista entiende explicar esas resistencias: para él, son sordas y profundas, vienen de lejos, tienen sus raíces en la cosa misma que se quiere elucidar. [Sartre (1943), p. 101]

Por tanto, las resistencias no provienen del Yo; pero tampoco pueden venir del impulso reprimido, ya que éste sería más bien un colaborador del analista que lucha por salir y expresarse en la conciencia clara, y sabemos que hace uso de argucias y trata de engañar a la censura. Así pues, el único plano que resta es el de la censura. Sólo la censura puede interpretar las preguntas o hipótesis del analista como más o menos próximas a los verdaderos impulsos que ella se afana en reprimir, porque ella sola *sabe* lo que reprime.

> […] la censura, para aplicar su actividad con discernimiento, debe conocer lo que ella reprime. Si, en efecto, renunciamos a todas las metáforas que representan la represión como un choque de fuerzas ciegas, forzoso es admitir que la censura ha de *elegir* y, para elegir, ha de *representarse*. ¿De dónde provendría, si no, el que deje pasar los impulsos sexuales lícitos, que tolere a las necesidades (hambre, sed, sueño) expresarse en la conciencia clara? ¿Y cómo explicar que pueda *relajar* su vigilancia, que hasta puede ser *engañada* por los disfraces del instinto? Pero no basta que discierna las tendencias malditas; es menester, además, que las capte como algo *que debe reprimirse*, lo que implica en ella, por lo menos, una representación de su propia actividad. En una palabra, ¿cómo podría discernir la censura los impulsos reprimibles sin

> tener conciencia de discernirlos? […] Todo saber es conciencia de saber. [Sartre (1943), p. 102]

Pero a continuación, Sartre advierte: la conciencia de la censura ha de ser conciencia de ser conciencia de reprimir *para no ser conciencia de eso mismo*. Es decir, la propia censura tiene que esconderse sus propósitos: dado que no puede ser consciente de que se engaña, ha de ser una conciencia autónoma que se engañe escondiéndose el propio proyecto, esto es, que se autoengañe. Por tanto todo el rodeo no nos ha hecho ganar nada, pues para dar cuenta del autoengaño el psicoanálisis ha introducido entre lo consciente y lo inconsciente una instancia [la censura], que no es sino *una conciencia de mala fe*. Esta crítica de Sartre es lo que se conoce como la *falacia homuncular*: al tratar de solucionar un problema, éste se reproduce recursivamente en el interior del sujeto. Esto no soluciona nada, pues requiere a su vez una explicación que nuevamente retrotraerá el problema a un nivel inferior, y así *ad infinitum*. El intento de restablecer la dualidad que posibilita el engaño —estableciendo incluso una trinidad (*Yo, Ello, Superyó*) bromea Sartre— ha quedado reducido a mera apariencia verbal.

Además, otra dificultad añadida es que localizando la doble actividad de atracción y repulsión del instinto en el nivel de la censura el psicoanálisis no es capaz de dar cuenta del fenómeno total. Por un lado, el impulso es ciego, inconsciente… pero trata de acceder a la conciencia. Para ello, según la teoría psicoanalítica, se disfrazaría (desplazamiento, condensación, etc.) pero, ¿por qué y cómo disfrazarse si no hay conciencia de ser reprimido, conciencia de ser rechazado por ser lo que es, y un proyecto de disfraz? Un proyecto de disfraz implica recurrir a la finalidad,

y ésta ha de ser consciente. Por otro lado, ¿Cómo dar cuenta del placer y angustia producidos por la satisfacción de un impulso si no hay una comprensión del mismo como deseado y prohibido a la vez? Esto es, ¿cómo vincular los procesos inconscientes del Ello (deseos ciegos) y del Superyó (imperativos morales o sociales inconscientes) en un mismo proyecto consciente?

> […] por haber rechazado la unidad consciente de lo psíquico, Freud se ve obligado a sobrentender por doquiera una unidad mágica que vincula los fenómenos a distancia y por encima de los obstáculos […] La explicación por la magia no suprime la coexistencia —al nivel inconsciente, al nivel de la censura y al de la conciencia— de dos estructuras contradictorias y complementarias, que se implican y se destruyen recíprocamente. Se ha hipostasiado y "cosificado", pero no evitado, la mala fe. [Sartre (1943), p. 103]

Finalmente, tras estas severas críticas a las dificultades asociadas a la censura, represión y el inconsciente, Sartre añade las impresiones —a las que ya hicimos referencia—, de un psicoanalista heterodoxo que afirma lo siguiente: "Cada vez que he podido llevar suficientemente lejos mis investigaciones, he comprobado que el núcleo de la psicosis era consciente"[23].

Sin duda el psicoanálisis plantea interrogantes y presenta sombras en sus postulados; esto es obvio desde el momento en que defiende la existencia de una instancia inconsciente a la que nos está vedado el acceso

[23] Stekel, Wilhelm (1920), *Die Geschlechtskälte der Frau* (Eine Psychopathologie des weiblichen Liebeslebens), Wien und Berlin, Urban und Schwarzenberg. [en castellano, *La mujer frígida*]. Citado en Sartre (1943), p. 104. [Erróneamente, Sartre se refiere a 'Steckel', en lugar de 'Stekel'].

y cuyos contenidos nos son, por tanto, constitutivamente incognoscibles. Sin embargo, toda la arquitectónica de la crítica sartriana pivota en torno a un supuesto que tampoco es inmune a un ataque teórico: la transparencia de la conciencia y la autoridad de primera persona, y con ello, la inexistencia de intenciones y proyectos inconscientes. Sartre introduce, de modo no fundamentado, que toda "conciencia es conciencia de ser" y que todo conocimiento implica "saber que se sabe". Esto podría conducir también a un regreso *ad infinitum*. Revisaremos estas objeciones más adelante.

II.2 - Primeros acercamientos

II.2.1 - Los primeros análisis del problema

Los primeros análisis pormenorizados del autoengaño surgen en la década de los 60, concretamente en el ámbito anglosajón. Que sea precisamente en este momento y lugar se debe, muy probablemente, a que la recepción de las ideas de Sartre abre un intenso debate.

Pese a que *El Ser y la Nada* fue publicado originalmente en 1943, la primera edición en inglés fue ofrecida por Metheuen & Co. en 1956. Esta versión llegó, como es evidente, a un sector más amplio del público en el ámbito anglosajón, y en seguida provocó las más diversas reacciones. En este contexto podemos situar el florecimiento de diversos intentos de atacar ya directamente el paradójico fenómeno del autoengaño. Como es natural, las primeras tentativas publicadas en diversas revistas filosóficas

especializadas produjeron a su vez críticas, réplicas y contrarréplicas, dando lugar a la ingente bibliografía de la que disponemos hoy día. Raphael Demos, Canfield y Gustavson, John King-Farlow, Frederick Siegler y Stanley Paluch son los pioneros en exponer sus enfoques. Veremos a continuación cuáles fueron sus propuestas.

Raphael Demos constata que en el lenguaje ordinario, engañar y mentir no son estrictamente equivalentes; según él, mientras en el engaño sólo cuenta el resultado, esto es, que el engañador produzca, *independientemente de su intención*, una creencia falsa en el engañado, en la mentira la intención juega un papel esencial. [Demos (1960), p. 588]. Pero con este cuadro, es posible que una mentira exitosa no dé lugar a un engaño, si quien miente está equivocado en su creencia: efectivamente, si cree algo falso, al mentir dice algo verdadero, y por tanto no induce a error a la víctima de su mentira. Para evitar estas dificultades, Demos decide tomar engaño y mentira como sinónimos en el análisis del autoengaño,

> Diremos que 'B miente a (engaña a) C' significa: B trata de inducir una creencia errónea en C, B tiene éxito al llevar a cabo esa intención, y finalmente B sabe (y cree) que lo que le dice a C es falso. Incluye las tres cosas: intención, resultados y conocimiento. Procederé a la discusión de 'mentirse a sí mismo' o 'engañarse a uno mimo' a partir de este significado de mentir y engañar.[24]

[24] «I will say that 'B lies to (deceives) C' means: B intends to induce a mistaken belief in C, B succeeds in carrying out this intention, and finally B knows (and believes) that what he tells C is false. All three: intention, results, and knowledge, are included. From this meaning of lying and deceiving I will proceed to a discussion of 'lying to oneself' or 'deceiving oneself'.» [Demos (1960), p. 588]

En su acercamiento al autoengaño, Demos aclara qué es aquello que no va a considerar como casos típicos de este fenómeno. Según él, quedan excluidas las situaciones en las que alguien cree algo falso porque tiene unos niveles de aceptación demasiado bajos debido a que ignora algo, como cuando se dice que un científico se engaña al creer que es un excelente científico porque su creencia acerca de los niveles de excelencia en ciencia son muy bajos —quizá debido a la falta de competencia—. En el autoengaño se trata, más bien, de impulsos o pasiones que influyen en nuestras creencias. Raphael Demos rechaza hablar de que estos arrollen (*overwhelm*) la evidencia; más bien es uno mismo quien no hace lo suficiente por resistir ese impulso o pasión. Además, la *exigencia de responsabilidad* es un rasgo crucial. De este modo, si bien la mera auto-causación no es condición suficiente para el autoengaño (un cleptómano no es considerado responsable, aunque la causa de su manía sea interna) la autocausación es condición necesaria. Así, cuando la causa es externa, como en los casos de la hipnosis y las drogas, tampoco se trataría de autoengaño, sino de ilusiones o delirios quizá auto-provocados. También distingue el autoengaño de los casos es los que el sujeto meramente finge[25] y simplemente trata de *hacer creer* (make-believe) que tiene tal o cual creencia. [Demos (1960), p. 590]

[25] Al fingir, en algunas ocasiones especiales uno puede meterse tanto en el papel que acabe creyéndose en parte el asunto, pero esto es algo que de ocurrir, es pasajero; de no ser pasajero, daría lugar a un corte tal con la realidad que podría ser catalogado como patología. Dice Demos: "Conceivably, in playing a game or putting on an act, we might be unwittingly 'taken in' by our performance; be, so to say, possessed by it like a poet 'in frenzy speaking'. Now, indeed there is believing.

Según Demos, el autoengaño resulta problemático porque exige que el sujeto crea a la vez dos proposiciones contradictorias (o crea y no crea la misma) siendo consciente de ello. Esto parece violar la ley de contradicción:

> Creer y descreer constituyen una actitud favorable y desfavorable; son contrarias y, por tanto, es lógicamente imposible que existan al mismo tiempo en la misma persona respecto del mismo asunto. Cuando B se miente a sí mismo, llega a creer algo que sabe que es falso; aceptar esto como la descripción de un hecho es admitir la violación de la ley de contradicción.[26]

Aunque estudia varias soluciones a este problema, Demos considera insatisfactorias algunas de ellas, como la que propone que el sujeto puede mantener creencias contradictorias si ocurren en momentos distintos o sucesivos o la que indica que la creencia agradable está en la mente consciente y la desagradable reprimida en la inconsciente. Según Demos, ambas no dan cuenta de los casos en los que parece que alguien mantiene dos creencias contradictorias conscientemente y al mismo tiempo. La solución que ofrece Demos es similar al argumento ofrecido por

---

Such states are usually temporary, but there are people who live out their whole life as if it were a play, who 'dramatize' everything. Yet I doubt that such an attitude entails total belief. Such total belief would be what I have called a delusion, as with people who are insane and "cut off from reality." On the contrary, those with a dramatic temperament, the enthusiasts and the like, preserve some sense of reality; in some 'corner of the mind' they know that it is all an act." [Demos (1960), pp. 590-91]

[26] «Believing and disbelieving are pro and con attitudes; they are contraries and therefore it is logically impossible for them to exist at the same time in the same person in the same respect. When B lies to himself he comes to believe what he knows to be false; to accept this as the description of a fact is to admit a violation of the law of contradiction.» [Demos (1960), p. 591]

Aristóteles con respecto a la *akrasia*. Según él, hay dos tipos de consciencia: la consciencia simple y la consciencia acompañada de atención [Demos (1960), p. 593]. Por tanto, el concepto clave es el de "poner atención", "fijarse" (*notice*). Uno, pese a que tiene una creencia de modo consciente, evita focalizar su atención sobre la evidencia y la creencia evidencial, lo que favorece que el deseo cumpla su papel en la adulteración y posterior adquisición de la creencia deseada. De este modo, el sujeto mantiene dos creencias contradictorias: relegando una de ellas a alguna "esquina de su mente" [Demos (1960), p. 595]. Estas ideas serán retomadas, como veremos más adelante, por Herbert Fingarette y servirán más tarde de trasfondo conceptual en los ensayos de la psicología social llevados a cabo por Ruben Gur y Harold Sackeim en la búsqueda de confirmación empírica del fenómeno.

Canfield y Gustavson, en cambio, advierten que el aspecto paradójico del autoengaño es debido únicamente a las interpretaciones de este fenómeno como un caso particular de "engaño a otros". La estructura del engaño interpresonal es la siguiente [Canfield y Gustavson (1962), p. 32]:

(i) Jones sabe que P es falso.

(ii) Jones intenta hacer creer a Smith que P es verdad.

(iii) Jones consigue hacer creer a Smith que P es verdad.

El problema surge cuando substituimos a Smith por Jones, y convertimos el proceso y estado de engaño en reflexivo:

(A) Jones sabe que P es falso, intenta hacerse creer que P es verdadero y lo consigue.

Según Canfield y Gustavson esto suena muy extraño y produce paradojas en la medida en que uno acepte la verdad de

(α) la noción de 'engaño' es la misma en el engaño interpersonal y en el autoengaño.

Por ello proponen que examinemos un caso similar, el caso de las "órdenes a otros" frente a las "órdenes a uno mismo" (el dominio a otros frente al autodominio). Según estos autores, ambas situaciones poseen una lógica diferente. Mientras en el primer caso al intentar que otro realice algo, 1) se lo pedimos u ordenamos, 2) él comprende nuestra petición y 3) la obedece o acata, en el caso del autodominio, no nos damos órdenes que posteriormente reconocemos como tales y acatamos. Más bien, cuando hablamos de autodominio, nos referimos al hecho de realizar acciones *bajo circunstancias desfavorables* (p. ej., pese a que estamos cansados, enfadados, enfermos, tenemos otras apetencias, etc.).

Canfield y Gustavson proponen interpretar el autoengaño como un caso especial de autodominio [Canfield y Gustavson (1962), p. 34]: se trataría de formarnos (u olvidar) una creencia en condiciones desfavorables, esto es, cuando la evidencia nos empuja a lo contrario. Sin embargo, como ellos mismos reconocen, no dicen nada acerca de cuáles habrían de ser los mecanismos psicológicos que permitiesen estas maniobras; se limitan a indicar que con esta interpretación no sería

necesaria la postulación de la existencia de un yo inconsciente al no haber ninguna extrañeza en la cláusula A [Canfield y Gustavson (1962), p. 35, nota 1] y que el fenómeno no sería en absoluto paradójico si prescindimos de la cláusula (α). [Canfield y Gustavson (1962), p. 36].

Estas ideas de Canfield y Gustavson fueron sometidas a crítica casi inmediata cuando John King-Farlow (1963) indicó dos problemas principalmente: el primer lugar, estos autores trataban de caracterizar el engaño (interpersonal) por medio de una serie de condiciones necesarias y suficientes; en segundo lugar, criticaban los esfuerzos de otros autores —entre ellos Jean-Paul Sartre—, por analizar a su vez el autoengaño bajo el modelo de engaño interpersonal. Por su parte, King-Farlow pone de relieve las muchas maneras en las que puede darse el engaño. A veces —dice— uno miente, otras desencamina al otro, en unas ocasiones no da toda la información relevante, o lo hace pero haciendo un énfasis equivocado; en otras ocasiones le da la información adecuada, pero enterrada en medio de un montón de información o meros datos irrelevantes, de modo que el oyente pierde de vista el asunto… Todas estos modos de engaño tienen algo así como un 'aire de familia', pero la supresión, el ocultamiento, la distorsión, la falsa evaluación de los hechos…etc. son modos distintos de engaño. Por tanto, hablar de "*el* sentido", o de "un caso *típico*" del verbo "engañar" en el engaño a otro es bastante desorientador y, finalmente, lo que equivocadamente empuja a Canflied y Gustavson a realizar un intento de establecer una especie de condiciones necesarias. [King-Farlow (1963), pp. 131-132]

King-Farlow subraya que ni el hecho de que el sujeto sepa que *p* es falso ni el éxito en crear una creencia falsa pueden ser condiciones necesarias. Por ejemplo, un dogmático apasionado puede engañarme completamente sobre la probabilidad de que *p* sea verdad *porque* argumenta con mucho ardor y sinceridad que un asunto de salvaje especulación o una pura estupidez es con seguridad verdad.

Según King-Farlow, Canfield y Gustavson hacen caso omiso de los ejemplos que propone Sartre, y simplemente afirman dogmáticamente que no se han encontrado nunca casos de autoengaño y que en todos los casos quien se autoengaña simplemente no disponía de la información necesaria [King-Farlow (1963), p. 134]. La explicación de esta reticencia a admitir los ejemplos sartrianos habría que buscarla en la aversión que Canfield y Gustavson sienten hacia la teoría del Ello/Yo/Superyó, cuya causa es a su vez un *prejuicio unitario* en la concepción de la conciencia [King-Farlow (1963), p. 135][27]. Sin embargo, según King-Farlow, nadie debería decir ni una sola palabra sobre el asunto del autoengaño sin tomar en cuenta seriamente los argumentos y ejemplos de Sartre [King-Farlow (1963), p. 134].

En esta disputa entra de lleno Frederick Siegler quien, aunque no cita a Canfield y Gustavson, parece que es de ellos de quienes toma el conjunto de condiciones necesarias y suficientes al definir la noción de "engaño".

[27] Nótese que ya vimos que Sartre también presenta este 'prejuicio unitario', y que también niega de modo taxativo el autoengaño en sentido fuerte. Parece, por tanto, que King-Farlow interpretaría la mala fe Sartriana de un modo distinto al nuestro, más como un verdadero engaño a uno mismo que como una simple mentira a uno mismo.

Siegler hace notar que no siempre que se adscribe un engaño reflexivo podemos hablar de autoengaño; por un lado, hay usos de "me engañaba cuando creía que tal y cual" que responden más a autoinculpaciones o reprimendas a uno mismo por cosas equivocadas que hice o pensaba, que a reproches por un autoengaño. En otras ocasiones, se trata de caracterizar una creencia que mantenía y que ahora considero estúpida, equivocada o irrazonable. Por otro lado, los usos en plural del tipo: "os engañáis al pensar que tal y cual" tampoco son acusaciones o atribuciones de autoengaño a otros, sino que lo que atribuimos a ese conjunto de individuos es meramente una creencia injustificada, acompañada quizá de un reproche o atribución de responsabilidad por mantener tales creencias. [Siegler (1963), pp. 33-34]

Siegler nos plantea la cuestión acerca de si "saber" y "engaño" tienen el mismo significado en las atribuciones interpersonales y en las reflexivas. Para ellos nos pide que supongamos que un sujeto A, aunque desea que *no-p* sea el caso, cree en t^1 que *p* lo es. Sin embargo más tarde, en t^2, comienza a tomar en cuenta cierto conjunto de datos D como evidencia E de que *p* es falso. Finalmente en t^3, A cree que *no-p* es el caso. Es posible, como Siegler señala, que en ausencia de ese deseo, A nunca hubiese considerado ese conjunto de datos como E. Sin embargo, esto no prueba ni que A tratase de creer intencionalmente que *no-p*, ni que tratase de ver como verdadero lo que creía que era falso [Siegler (1963), p. 34]. Por tanto, este tipo de ejemplos sólo mostraría que, de ser posibles este tipo de procesos, "hay algunos usos de expresiones de engaño reflexivo que son un tanto diferentes de los usos en expresiones de engaño

interpersonal. Pero, ¿no es posible que haya casos de autoengaño que sean lógicamente idénticos a los casos de engaño interpersonal?" [Siegler (1963), p. 35]. En la búsqueda de este tipo de fenómeno, distingue dos posibles sentidos de autoengaño: uno fuerte, que exige que el sujeto mantenga la creencia inicial verdadera aun después de adquirir la falsa, y otro débil, que sólo exigiría que éste cambie injustificadamente de creencia. [Siegler (1963), p. 36]

Al estudiar el caso fuerte, observamos que a veces nos parece que un sujeto se comporta de modo contradictorio y nos da la impresión de que se debe a que dispone de evidencia contraria o incluso contradictoria. Sin embargo, Siegler nos alerta de la posibilidad de *confundir el hecho de que un sujeto se comporte de modo aparentemente contradictorio con el hecho de que posea creencias contradictorias*; efectivamente, el sujeto puede estar confuso, pero que posea evidencia en conflicto no prueba que mantenga creencias contradictorias; más bien, en casos semejantes la evidencia no parecería tan clara como para que haya creencia alguna y, desde el momento en que no hay creencias, no podemos hablar de autoengaño. [Siegler (1963), p. 38] Parece, por tanto, imposible establecer o descubrir cómo un sujeto podría mantener dos creencias contradictorias.

Sin embargo, aunque no podamos probar la existencia de casos fuertes o estrictos, aun podemos examinar si puede darse un autoengaño débil, esto es, sin la exigencia de creencias simultáneas contradictorias; bastaría con que el sujeto sea el único responsable de un voluntario cambio de creencia que sabe injustificado. Para ello hemos de asegurarnos de que el

sujeto cambia de creencia por voluntad propia y que sabe que aquello que cree es algo falso; tanto si el injustificado cambio de creencia fuese a causa de una tortura como si no creyese que aquello que ahora abraza es falso, no podríamos hablar de autoengaño. En lo casos que clasificamos como autoengaño nos preguntamos cómo alguien puede creer algo tan obviamente falso y creemos que ha de ver tan claramente lo contrario que debe creerlo [Siegler (1963), p. 41]. Pero en realidad,

> […] en el mejor de los casos, encontramos evidencia para "Jones creía p y ahora cree no p" donde no *nos* parece que haya nada que justifique el cambio de creencia. No encontramos evidencia para "Jones cree p y no p", ni para "Jones cree no p como resultado de su propio intento procedimiento para inducirse una creencia que creía falsa" [en cursivas en el original][28]

Por tanto, el autoengaño parece diluirse en la mera *adscripción externa* de un estado mental irracional que figura la siguiente pregunta retórica: "¿Cómo podría Jones creer tal sinsentido? Realmente, no puede" [Siegler (1963), p. 43]. De este modo, Siegler se muestra escéptico con respecto a la verdadera existencia del autoengaño.

Una postura también escéptica es la de Stanley Paluch, quien distingue entre engañarse a uno mismo y estar autoengañado. Con respecto al primer fenómeno, Paluch dice que no es extraño que alguien reemplace un conocimiento con una creencia incompatible con él; sin embargo, si

[28] «…at best, we find evidence for "Jones believed p and now he believes not p" where there seems to *us* to be nothing to justify the change in belief. We could not find evidence for "Jones believes p and not p", or for "Jones believes not p as a result of his own procedure intended to induce a belief which he believed to be false.» [Siegler (1963), p. 42]

esta empresa resulta exitosa, el sujeto no sabrá ya en t^2 lo que sabía en t^1, y por tanto, no cree una cosa y su contraria al mismo tiempo. En este tipo de situaciones, me habría engañado si me he conducido al error. Obviamente, Paluch está considerando aquí lo que, a nuestros ojos, representa un caso de error en nuestras creencias. Lo que resulta singular es que se produzca tras haber abrazado previamente otra que era adecuada [y que hemos sustituido por ésta falsa] pero esto no supone ningún problema de racionalidad si no lo hemos hecho contra el peso de nuestra evidencia.

Sin embargo, Paluch está más interesado en el caso en que el sujeto se halla 'en estado de autoengaño', esto es, el caso que define como aquel en el que *el sujeto cree lo opuesto a lo que sabe*.

Paluch indica que hay modelos explicativos del autoengaño en los que el uso de "engaño" o "engañar" no es el estándar. Mientras Sigmund Freud había propuesto en la *Psicopatología de la vida cotidiana* una explicación según la cual el sujeto sabe inconscientemente que *p*, pero mantiene conscientemente la creencia de que *no-p*, Raphael Demos había tratado de dar cuenta del fenómeno mediante una distinción similar a la freudiana —pero inspirada en Aristóteles— entre un conocimiento *latente* de que *p*, y una creencia *actual* de que *no-p*. Ambos modelos supondrían para Paluch *modelos debilitados* de autoengaño. La razón de esta debilidad se encuentra en que en ambos el sujeto no puede ser *ex hypothesi* consciente del autoengaño. Por un lado, en el modelo freudiano el psicoanalista puede encontrar un conocimiento inconsciente en el paciente incoherente con

sus creencias, pero el sujeto no es consciente de tal incoherencia; el psicoanalista podría hacer que el sujeto fuese consciente, pero entonces se desharía ya la incoherencia, abandonando por tanto el estado de autoengaño. Por otro lado, bajo el modelo de Raphael Demos, si el conocimiento *latente* del sujeto se *actualizase*, éste abandonaría el estado de autoengaño. Por esta razón estos modelos son débiles; el engaño no tiene las mismas características que en el engaño interpersonal, en los que el engañador necesariamente ha de saber y ser consciente de que aquello que quiere hacer creer al engañado es falso. La pregunta entonces es en qué sentido el "engaño" del engaño interpersonal y del autoengaño son análogos: la respuesta para Paluch puede encontrarse en la capacidad de salir del engaño, de desengañarse. En el autoengaño sucede que,

1) "X tiene la capacidad de desengañarse" y,

2) "X no actualiza esta capacidad"

Paluch prevé una posible crítica: aunque tanto en el engaño interpersonal como en el autoengaño el sujeto tiene la capacidad de desengañare, en el autoengaño el sujeto, además, es consciente de que tiene esta capacidad. Sin embargo no es posible que alguien sea consciente (*aware*) de una capacidad que se supone inconsciente (*unconscious*) o latente [Paluch (1967), p. 274].

> Cuando le digo a alguien que se está engañando a sí mismo, le estoy sugiriendo que debería revisar sus creencias. ¿Cómo puedo distinguir el caso en que descubre conocimiento latente del caso en que cae en la cuenta y adquiere un conocimiento nuevo? Una objeción continua al psicoanálisis freudiano es que Freud no consiguió establecer una diferencia clara entre la conseguir que el paciente admita algo que,

> inconscientemente, sabía y conseguir que el paciente acepte una hipótesis sobre su inconsciente ofrecida por el psicoanalista. Se puede señalar un punto similar contra el modelo de Demos.[29]

La opacidad de tales conceptos como "conocimiento inconsciente" o "latente" nos fuerza a movernos hacia una caracterización del autoengaño más débil aún. Ahora el autoengaño nos presenta a un hombre que tiene una creencia, pero podría creer (o podría esperarse que creyese) otra cosa. Ha de tener la capacidad de hacerlo. Por tanto, un sujeto X se autoengañaría si:

(1) X cree que *p,* y *p* es falsa.

(2) X conoce la evidencia que cuenta en contra de la verdad de que *p*.

(3) X tiene algún motivo para descartar esa evidencia.

(4) En ausencia del motivo, X vería que *p* es falsa y su negación verdadera.

(5) Si el motivo fuese claro para X, vería que no tiene base para albergar su creencia.

(6) X es libre para advertir el carácter de su motivo.

---

[29] «When I tell someone that he is deceiving himself I suggest that he ought to review his beliefs. How am I to distinguish between the case where he realizes latent knowledge and that where "the penny drops" and he acquires new knowledge? It is a standing objection against Freudian psychoanalysis that Freud failed to make clear the difference between getting the patient to admit something that, unconsciously, the patient knew and getting the patient to accept an hypothesis about his unconscious provided by the analyst. A similar point can be made against Demos's model.» [Paluch (1967), p. 275]

El sujeto ha de tener evidencia contraria pero, cegado por algún motivo, no atiende a esa evidencia. Sin embargo, le sería posible hacerlo y, si lo hiciera, aceptaría la creencia contraria.

Paluch realiza dos aclaraciones a esto:

1) Adelantando una idea que explotará Davidson, afirma que el autoengaño *no* puede tener que ver con un mero error al considerar la fuerza de la evidencia.

2) Cuando alguien se autoengaña, no es que no sea capaz de abandonar su creencia.

Lo primero sería más bien torpeza (dullness) evaluativa que autoengaño; lo segundo, fijación, idea fija o delirio [Paluch (1967), p. 276].

El mayor problema que observamos nosotros es que si uno no sospecha del motivo por el que alcanzó una creencia, no efectuará un autoanálisis; e incluso no se entiende muy bien en qué consistiría afirmar que tal sujeto tiene esa capacidad. Si por el contrario sospecha del motivo, parece que no hay lugar para hablar de autoengaño, en tanto que las "creencias bajo sospecha" son, en el mejor de los casos, una estirpe peculiar de creencias. [Paluch (1967), pp. 276-277].

Paluch, consciente de todas estas dificultades, admite finalmente:

> Los modelos que he discutido en este artículo son, hasta donde sé, los líderes entre los modelos competidores de autoengaño no-paradójicos (o aparentemente no-paradójicos). Es cualquier cosa menos obvio que pudiera decirse de cualquiera de ellos que refleje casos reales de

> autoengaño. De hecho, no es ni mucho menos claro que haya casos estrictos de autoengaño que reflejar. No hay duda acerca del hecho de que hablamos de gente que se engaña a sí misma, pero lo que no está claro es que haya instancia alguna en la que este modo de hablar sea el mejor o no pueda ser reemplazado por otras descripciones que no incluyan la noción de engaño en absoluto.[30]

II.2.2 - El papel de la atención

> 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.
>
> [Fingarette (1969), p. 1]

Con estas palabras comienza Herbert Fingarette su libro sobre el autoengaño, el primero —que nosotros sepamos— de semejantes características, pues esta obra sería la primera que, como indica su título *Self-Deception*, versa íntegramente del tema que nos ocupa. Según Fingarette, todos los esfuerzos por aclarar el concepto de autoengaño han acabado en descripciones paradójicas, lo cual ha supuesto una oscuridad

[30] «The models I have discussed in this paper are, so far as I know, the leading contenders among non-paradoxical (or seemingly non-paradoxical) models of self-deception. It is anything but obvious that any of them could be said to mirror actual cases of self-deception. Indeed, it is anything but obvious that there are strict cases of self-deception to be mirrored. There is no doubt about the fact that we talk about people deceiving themselves, but what is not clear is that there are any instances where this way of speaking is best or could not be replaced by other descriptions which no not involve the notion of deception at all.» [Paluch (1967), p. 277]

que *infecta* nuestra concepción de lo que es una persona, lo que es conocerse a sí mismo y lo que es actuar con responsabilidad [Fingarette (1969), p. 1]. El objetivo de Fingarette no consiste —ni podría hacerlo— en ofrecer una demostración irrefutable acerca de la validez de sus propuestas; más bien, busca que los provechosos e interesantes resultados que se siguen tanto de aceptar y aplicar su enfoque, como de ponerlo en relación con otras propuestas para problemas similares, le confieran apoyo. [Fingarette (1969), p. 6].

Fingarette comienza haciendo un análisis de los intentos previos de dar una explicación satisfactoria del asunto. Según Fingarette, un problema que ha lastrado la investigación de este concepto es el intento de examinarlo bajo el modelo del engaño, concretamente, del engaño interpersonal. En este sentido, analiza las propuestas de algunos de los autores que hemos visto previamente, como Demos, Siegler, Canfield y Gustavson y Paluch. En opinión de Fingarette ninguno de ellos consigue ofrecer una versión sólida y convincente a causa de dos problemas: o bien se elimina la paradoja del autoengaño, pero lo que queda ya no es autoengaño, o bien el autoengaño está aún frente a nosotros, pero no se ha conseguido eliminar su carácter paradójico. [Fingarette (1969), p. 13]

Según Fingarette, la principal dificultad en la propuesta de Raphael Demos reside en que no logra distinguir entre la inconsistencia de creencias y el autoengaño, pues la mera inconsistencia entre creencias es bastante común y no problemática. Como hemos visto anteriormente, Demos cree que el sujeto que se autoengaña no advierte una de las

creencias que mantiene. Sin embargo, la pregunta crucial es la siguiente: ¿hay intención de "no percatarse" por parte del sujeto? Demos habla en unas ocasiones de que el sujeto "no se percata" [Demos (1960), p. 594] o "se distrae" [Demos (1960), p. 593] con respecto a una de sus creencias, pero afirma en otras que la "ignora deliberadamente" [Demos (1960), p. 593]. Sin embargo, según Fingarette, si el sujeto lo hace por descuido o confusión, no será ya autoengaño; y si lo hace deliberadamente, no se evita la paradoja [Fingarette (1969), p. 16].

Frederick Siegler señalaba que el autoengaño no tiene por qué basarse en el modelo del engaño interpersonal, y que las "expresiones de engaño reflexivo" pueden tener otras funciones. Cuando uno dice: "aunque me negué a creerlo, *supe* todo el tiempo que tal y tal", muy bien puede ser que realmente no lo *supiese*. Esta puntualización de Siegler, en principio inocua (pues sólo mostraría la posibilidad de error al autoatribuirse autoengaño) se hace relevante cuando Siegler parece reducir todo caso posible de expresión reflexiva de engaño a otras funciones distintas al autoengaño. En este sentido, al decir: "*Supe* todo el tiempo que…", el sujeto podría estar fingiendo ante los demás; o ver en ello un modo de disculparse o minimizar el error que ha cometido en cierta tarea; o expresar miedos, esperanzas o aprensiones pasadas, más que conocimiento alguno; o expresar un nuevo modo de interpretar el pasado, basándose en su memoria. Por otro lado, cuando un sujeto dice: "me he estado engañando", puede estar expresando con ello simplemente una resolución a cambiar algún aspecto de su vida en el futuro. También puede ser la expresión de un juicio acerca de lo que debería haber conocido mejor,

pero de hecho no lo hizo. Además, cuando por ejemplo una mujer, ante una evidencia contraria abrumadora, dice que su hijo no puede ser un asesino, quizá no expresa realmente que no cree algo, sino una dificultad para comprenderlo y asumirlo. O bien sólo está expresando una creencia debida a que desea fuertemente que así sean las cosas (pensamiento desiderativo).

Fingarette cree que todos estos casos son plausibles, pero que ninguno de ellos constituye autoengaño, sino un intento de engañar a otro, mala memoria, un error de juicio, espolearse de cara a una dura empresa en el futuro, pensamiento desiderativo, una promesa de cara al futuro, o la expresión de una perplejidad, por ejemplo. Echa de menos una verdadera explicación del autoengaño en Siegler, y le hace varios comentarios en tono de reproche, sin percatarse —en mi opinión— de que no es que Siegler no sea capaz de definir con precisión el autoengaño, sino que éste encuentra irresoluble el fenómeno bajo ciertas descripciones y trata deliberadamente de reducir el supuesto autoengaño a otros fenómenos no paradójicos.

De hecho, nos parece que Fingarette malinterpreta la afirmación de Siegler según la cual cuando atribuimos autoengaño a otro, le atribuimos una creencia errónea que es irrazonable mantener, y le exigimos responsabilidad por ello. Fingarette se centra en lo "irrazonable" e interpreta esta afirmación de Siegler de modo similar a la tesis de Canfield y Gustavson según la cual, todo lo que ocurre en el autoengaño es que la persona cree u olvida algo en circunstancias adversas (es decir, en contra

de gran cantidad evidencia significativa). Sin embargo, como ya vimos en §II.2.1, nos parece que Siegler es escéptico con respecto a la verdadera existencia de casos en los que el sujeto crea algo en contra de la evidencia, y que más bien opina que somos nosotros quienes, incapaces de hacer casar la evidencia que creemos que tiene el sujeto con la creencia que expresa, le atribuimos autoengaño en ausencia de una explicación mejor.

En cualquier caso, Fingarette cree que quien mejor ha visto el asunto ha sido Terence Penelhum, ya que no sólo rechaza el modelo de engaño interpersonal, sino que le añade a la tesis de Canfield y Gustavson un rasgo crucial. Penelhum acepta de Canfield y Gustavson que el concepto de autoengaño exige como condición necesaria la posesión de una creencia en circunstancias adversas, esto es, en contra de la evidencia, pero cree que ésta no es condición suficiente: el sujeto ha de ser, además, consciente de que la evidencia *es adversa*, esto es, ha de *apreciar el peso de la evidencia* [Fingarette (1969), p. 23]. El problema es que Penelhum, en un intento de modificar las tesis de Demos, afirmaba que el autoengaño es un "estado de conflicto":

> Quien se encuentra en este estado satisface parcialmente los criterios para creer y también para descreer —en particular, tenderá a afirmar su descreencia en aquello a lo cual ve que apunta la evidencia[31]

Esto introduce problemas adicionales en la investigación del fenómeno, ya que no está nada claro qué habría de significar "satisfacer parcialmente

[31] «Someone in this state does partially satisfy the criteria for belief and also those for disbelief—in particular he will tend to declare his disbelief in that to which he sees the evidence points» [Penelhum (1964), p. 258-259].

los criterios para creer". No está claro si Penelhum se refiere a "medias creencias" (*half-beliefs*), y además no realiza un análisis ulterior del asunto. Lo que hace en su lugar, es establecer el conjunto de condiciones *conjuntamente suficientes* para el autoengaño:

(1) Una creencia en contra de fuerte evidencia.

(2) El conocimiento por parte del sujeto de la evidencia.

(3) El reconocimiento del sujeto de la importancia de la evidencia.

Como hace notar Fingarette, en (1) ya no habla de "satisfacción parcial de creencias", sino de creencia (completa). Pero el mayor escollo que se presenta ahora es que anteriormente había establecido que (2) y (3) eran condiciones suficientes para que el sujeto aceptase una creencia evidencial. Pero entonces, dado que (1) establece que el sujeto posee una creencia contraevidencial y (2)-(3) son condiciones suficientes para que el sujeto posea una creencia evidencial, Penelhum ha acabado por reintroducir, en otros términos ligeramente modificados, aquello que pretendía negar: que el autoengaño consista en la posesión de creencias contradictorias. [Fingarette (1969), p. 25]

El elemento que pierde de vista Penelhum y que Fingarette considera crucial es el de *propósito*. El sujeto realiza toda esta maniobra de engaño *a propósito* (*purposefulness*). Es la intención del sujeto lo que diferencia el autoengaño de la testarudez en mantener una creencia, de cualquier disfunción neurológica, o de un presunto estado de hipnosis, por ejemplo.

> El sujeto se *persuade* para creer en contra de la evidencia *para*, de algún modo, evadir la displacentera verdad a la que ha visto que apunta la evidencia [Fingarette (1969), p. 28][32]

Fingarette cree que la razón de que Penelhum no se percate de la importancia del "propósito de engaño" que encierra el autoengaño, es la importancia que éste otorga a los *motivos*. Reconoce que, sin duda, generalmente hay un motivo en la base del autoengaño, pero cree que Penelhum no ha llegado a advertir que, incluso en ausencia de motivo, lo que es esencial al autoengaño es el *propósito de engañarse* [Fingarette (1969), p. 28]. Es más, Fingarette cree ha sido debido al ansia por resolver la presunta paradoja de las creencias contradictorias, por lo que Demos, Siegler, Canfiel y Gustavson y Penelhum han dejado de lado la *ignorancia intencionada* inherente al autoengaño.

Según Fingarette, es muy común hablar de quien se autoengaña como alguien que no *percibe* su propio engaño o que no puede *ver* a través de la cortina de humo que él mismo ha creado. También se dice que este sujeto *cree* la historia que cuenta, mientras *in foro interno sabe* que no es cierta. O que hace que le *parezca* así. Quien se autoengaña no está al tanto (*unaware*) de su propio engaño, y puede haber deseos y fantasías *inconscientes*. Este tipo de lenguaje en términos de percepción pertenece a una familia "cognición-percepción" en contraste con una familia "volición-acción". [Fingarette (1969), p. 33]

[32] «[…] our subject *persuades* himself to believe contrary to the evidence *in order to* evade, somehow, the unpleasant truth to which he has already seen that the evidence points» [Fingarette (1969), p. 28]

Frente a este proyecto de estudio el concepto de autoengaño, Fingarette propone dos cosas:

a) Definir ciertos términos bajo el enfoque volición-acción.

b) Un nuevo enfoque "volición-acción".

No se trataría de eliminar del análisis del autoengaño todos los conceptos que estaban ligados al enfoque cognición-percepción, sino de cambiar su acento. Concretamente, cree que sería provechoso interpretar el término "conciencia" y sus variantes, bajo el enfoque volición-acción. Generalmente, términos como "saber", "estar al tanto de" o "ser consciente de", están ligados al verbo "ver" [Fingarette (1969), p. 34]. Según Fingarette, el hecho de que seamos seres con mucha movilidad y de que la visión sea el sentido que nos proporciona la mejor y mayor cantidad de información para movernos, es posiblemente la razón por la cual la consciencia se vincula a la visión. Sin embargo, el verbo ver es en gran medida un verbo pasivo; aún cuando hay un elemento activo en la visión, a saber: podemos dirigir nuestra vista a un lado u otro —lo cual queda reflejado en el verbo activo "mirar"—, la visión es esencialmente pasiva. Nociones como la de consciencia, hacerse consciente, perder la consciencia, etc., son interpretadas bajo este modelo, lo cual ha llevado a ver la consciencia como algo que nos sucede, que no podemos dejar de tener, que no puede ejecutarse mejor o peor, sistemáticamente o al azar [vid. White (1964), pp. 39-40].

Fingarette propone precisamente abandonar esta interpretación de la consciencia como un registro pasivo, como un reflejo del mundo. Para él

la consciencia es algo *activo*, algo que *hacemos*. En este sentido, la consciencia es una *destreza* y, como tal, algo que aprendemos y que "rutinizamos" [Fingarette (1969), p. 37].

Todos tenemos diversas relaciones y compromisos con el mundo; tenemos diversos modos de experimentar el mundo y a nosotros mismos dentro de ese mundo (conductas, aspiraciones, deseos, esperanzas, miedos, percepciones, recuerdos, etc.); sin embargo, no siempre somos conscientes de todas estas relaciones que mantenemos y, del mismo modo que no decimos explícitamente todo lo que hacemos —ni podríamos hacerlo—, sino que seleccionamos aquello que decimos, y no lo hacemos de modo arbitrario, lo mismo vale, *mutatis mutandis*, para la explicitación (*spell-out*) de nuestras relaciones con el mundo. En esto consiste la consciencia en un "sentido fuerte"[33]: en una explicitación de nuestros compromisos con el mundo, en atender explícitamente a algo [Fingarette (1969), pp. 39-40].

El punto crucial para Fingarette es que no debe darse por supuesto la consciencia, sino su ausencia; la consciencia surge a partir del ejercicio de una destreza a causa de una razón especial que requiere nuestra atención explícita [Fingarette (1969), p. 41]. Sin embargo, a veces el sujeto tiene

[33] A diferencia de la consciencia en "sentido débil", tal y como se usa en casos como los siguientes: "aunque se pegó un buen golpe, permaneció plenamente consciente" o "perdió la consciencia"; o incluso, "¿Eses consciente de que estás barajando las cartas? – Sí, por supuesto soy consciente de que lo hago; pero eso no me ha distraído de lo que me dices, ya que es lo único a lo que presto atención y nada más llega a mi mente". [Fingarette (1969), p. 46]

razones precisamente para no explicitar alguno de los compromisos o relaciones que mantiene con el mundo:

> [El autoengaño] consiste en la situación en la cual hay una razón primordial para *no* explicitar algún compromiso, donde con destreza tomamos esto en cuenta y sistemáticamente evitamos explicitar el compromiso, y donde, a su vez, nos abstenemos de explicitar el ejercicio de esta destreza para explicitar. En otras palabras, evitamos hacernos explícitamente conscientes de nuestro compromiso, y evitamos hacernos explícitamente conscientes de que lo estamos evitando. Éste es el caso que nos llevará a dar nuestro primer paso principal dentro de la región del autoengaño, del inconsciente freudiano, de la mala fe sartriana o de otras variantes de nuestro tema principal [Fingarette (1969), p. 42].[34]

Pero no es algo que el sujeto haga sin percatarse de ello y nos parece que, en cierto modo, podría hacerlo explícito:

> Quien se autoengaña es "incapaz" de admitir para sí la verdad (incluso aunque sabe en su corazón que es así). Hay un tipo de autenticidad en su "ignorar"; no es simple hipocresía, mentira o engaño a otros. Incluso sentimos que en algún sentido, podría admitir la verdad simplemente si así lo deseara. Creo que podemos arrojar algo de luz ahora sobre esta paradójica "ignorancia deseada" [Fingarette (1969), p. 46].[35]

---

[34] «[Self-deception] is the situation in which there is overriding reason *not* to spell-out some engagement, where we skillfully take account of this and systematically avoid spelling-out the engagement, and where, in turn, we refrain from spelling-out this exercise of our skill in spelling-out. In other words, we avoid becoming explicitly conscious of our engagement, and we avoid becoming explicitly conscious that we are avoiding it. It is this case which will take us our first major step into the region of self-deception, of the Freudian unconscious, of Sartrian Bad Faith, and of other variants on our main theme».

[35] «The self-deceiver is "unable" to admit the truth to himself (even thought he knows in his heart it's so). There is a kind of genuineness to his "ignoring"; it is not

Por tanto, cuando un sujeto se autoengaña, encuentra que hay una razón fuerte y preponderante para no explicitar —ni siquiera a sí mismo— alguno de sus compromisos con el mundo. Por esa razón, adopta la política de no explicitar tal compromiso, y lo que es más: esa misma política le obliga a no explicitar que ha realizado esa evaluación de la situación, y que ha adoptado esta táctica para realizar tal maniobra [Fingarette (1969), p. 47]. Para llevar a cabo esta maniobra, el sujeto ha de hacer la historia que cuenta natural, cercana a los hechos y coherente con el resto de sus compromisos. Para ello rellena los huecos con su imaginación o inventiva. Evidentemente, esto supone en una destreza que puede ser mayor o menor y, además, puede entrenarse. [Fingarette (1969), pp. 48-49].

Este análisis del autoengaño, que muestra una discrepancia entre lo que el sujeto hace (los compromisos que mantiene con el mundo) y lo que dice hacer (aquellos que explicita), conduce a otro asunto: el de la sinceridad del sujeto, o su "insincera sinceridad". Como Fingarette hace notar, ninguno de los análisis filosóficos previos de Demos, Siegler, Canfield y Gustavson, Penelhum, etc., examina el asunto desde la perspectiva de la autenticidad, la alienación del yo de sí mismo, de la responsabilidad o de la "yoidad", tal y como lo habían hecho Sartre, Freud o el propio Kierkegaard [Fingarette (1969), p. 74].

simple hypocrisy, or lying, or duping of others. Yet we also feel that in some sense, he could admit the truth if only he would. I believe we can now throw some light on this paradoxical "wishful ignorance"»

Fingarette nos avisa de que resulta problemático afirmar que el sujeto sea insincero consigo mismo. La razón es que resulta paradójico

> proponer que uno excluye a propósito cualquier cosa que dejase a la vista sus racionalizaciones, ya que esto presupone que uno está enterado de estas racionalizaciones *como* tales, que uno es capaz de identificar consideraciones en contra, y luego —precisamente debido a que uno ha identificado las consideraciones en contra y su fuerza como tales— que uno rechaza considerarlas! [Fingarette (1969), p. 78][36]

Estas palabras nos recuerdan, aunque Fingarette no lo cite, a una de las críticas que Sartre le hacía al psicoanálisis:

> Pero no basta que discierna las tendencias malditas; es menester, además, que las capte como algo *que debe reprimirse*, lo que implica en ella, por lo menos, una representación de su propia actividad. En una palabra, ¿cómo podría discernir la censura los impulsos reprimibles sin tener conciencia de discernirlos? [Sartre (1943), p. 102]

Según Fingarette, en todo autoengaño el sujeto es "incitado" a un tipo de compromiso con el mundo que, en parte o en todo, no puede confesar (*avow*) como propio, ya que hacerlo traería consecuencias terriblemente angustiosas y presumiblemente destructivas para su persona.

Por otro lado, el proceso de maduración de un niño consiste precisamente en la incorporación y articulación de los distintos compromisos que adquiere con el mundo. El niño, que inicialmente persigue compromisos independientemente como si se tratara de

[36] «to propose that one *wilfully* excludes whatever would expose one's rationalizations, for this presupposes that one is cognizant of those rationalizations *as* such, that one is able to identify counter-considerations, an then —precisely to the extent that one has identified the counter-considerations and acknowledge their force as such— that one refuses to consider them!»

proyectos autónomos, va desarrollando la capacidad de integrarlos en una compleja unidad. La formación de la persona, del "yo personal" va unida al desarrollo de esta destreza. De este modo, un niño de 2 años sería un "yo agente", pero no una persona. [Fingarette (1969), p. 86].

En este sentido, el autoengaño supone para Fingarette una "regresión" en la que el sujeto realiza un compromiso con el mundo de modo aislado, incluso después del surgimiento del "yo personal". Y este aislamiento del resto de proyectos, valores, gustos, sensibilidades, etc., es lo que provoca la ausencia de confesión (*avowal*) de tal proyecto por parte del "yo personal", que no es capaz de reconocer este compromiso "como suyo". La consecuencia ulterior es la ausencia de responsabilidad por parte del sujeto. Dado que no es un proyecto que identifique como suyo, no acepta responsabilidad ninguna sobre él. [Fingarette (1969), p. 87]. Quien se autoengaña *no es moralmente responsable*, ya que se ha producido en él una subversión de la agencia personal y de la capacidad moral, y por ello ha perdido capacidad de control sobre lo que hace [Fingarette (1969), p. 140].

Su teoría, pues, supone una novedosa propuesta en el planteamiento conceptual del problema. Fingarette cree que si su teoría pudiera casar con la ontología de la consciencia sartriana y con las teorías psicoanalíticas (que no dejaban de estar basadas en casos empíricos, clínicos), ya tendría mucho camino recorrido. En este sentido, acepta de Sartre la importancia del planteamiento de la cuestión en términos de elección y responsabilidad, más que de conocimiento e ignorancia. Sin embargo, rechaza la jerga sartriana por considerarla en ocasiones *esotérica* [Fingarette

(1969), p. 94] e indica que el postulado de Sartre de la transparencia de la consciencia le empujaba a considerar las opacidades de la consciencia como un rechazo del sujeto a responsabilizarse de las elecciones tomadas. Fingarette cree que estas opacidades se deben más bien al rechazo a reflexionar sobre determinados eventos de la consciencia, y cree además que este punto está contenido implícitamente en Sartre, aun cuando él mismo no hubiese reparado en ello: una consciencia totalmente transparente nos abocaría a una jerarquía infinita de reflexiones, que Sartre consideraba como absurdo e imposible. Por esta razón ha de haber, según Fingarette, una consciencia capaz de no reflexionar sobre sí misma, esto es, una consciencia que no sea explícita [Fingarette (1969), pp. 97-98]. De Freud toma prestada Fingarette mucha terminología, así como buena parte de su razonamiento; nos habla constantemente de catexis, contracatexis, eliminación de catexis e hipercatexis, y su idea de una consciencia no explícita es deudora de la noción del preconsciente freudiano. Sin embargo, está de acuerdo con Sartre en que la teoría freudiana tiene algunas limitaciones y cree que pese a que Freud acertó al postular la división de la psique, se equivocó al describir su naturaleza [Fingarette (1969), pp. 127-128]. Concretamente, apunta que uno de los últimos escritos de Freud, el inacabado *La escisión del yo en el proceso defensivo* (1938), podría haber iluminado algunos asuntos oscuros. Freud comienza este texto con estas sugerentes palabras:

> Por un momento estoy en la interesante situación de no saber si lo que voy a comunicar ha de apreciarse como algo hace tiempo consabido y evidente, o como nuevo por completo y sorprendente. Me inclino, empero, a creer lo segundo [Freud (1938), p. 275]

Aunque por desgracia Freud no vivió lo suficiente como para desarrollar estas ideas, Fingarette cree que Freud llegó en los últimos días de su vida a la misma conclusión a que ha llegado él: el Yo, en el proceso defensivo, se escinde dando lugar a un sistema que es, en gran medida, igual que el Yo. Esta escisión no es algo que el Yo padezca, sino algo que *hace*. Lo que diferencia a este sistema del Yo, es que contiene el motivo de la defensa, propósitos, sentimientos, percepciones e impulsos, así como el sentimiento de culpabilidad generado. Por lo demás, tiene todas las características del Yo, y por tanto puede ser considerado como un *Contra-Yo*. De hecho, Fingarette afirma que este sistema es la otra cara de la moneda del proceso defensivo; es la cara *estructural* del proceso *económico* conocido como contracatexis (que examinamos ya en §II.1.1). El Yo no reconocería al Contra-Yo ("*Ese no soy yo*"), y este rechazo sería la causa del rechazo a reconocer el propio compromiso [Fingarette (1969), pp. 128-129].

Como hemos dicho, Fingarette trata de acomodar su propia visión del asunto a las teorías de Freud y Sartre. Según él, Sartre y Freud difieren principalmente en el enfoque moral. Ambos evitan —como hace el propio Fingarette— plantear el asunto en términos de conocimiento e ignorancia, de "conocer", "ignorar", "ver", etc., ya que en esos términos la paradoja epistémica resulta insalvable, y esta conduciría a una paradoja moral. Pero mientras Freud habla con un lenguaje no-teleológico de mecanismos de defensa, sistemas, trasferencia de energías, etc. que pretende ser clínico, y por ello neutral desde el punto de vista moral, Sartre enfoca el problema desde una perspectiva teleológica, en términos de elección, responsabilidad, integridad y libertad [Fingarette (1969), pp. 135-138].

Sin embargo, Fingarette no acepta la conclusión sartriana acerca del autoengaño: Sartre cree que es la falta de integridad lo que hace que el sujeto trate de evitar la responsabilidad que acarrean las constantes elecciones que hemos de realizar. Al contrario, Fingarette cree que es la integridad lo que empuja al autoengaño. A menor integridad, menor necesidad de engañarse ("cuanto más se acerca uno a la santidad, más sufre"). Cuanto menos íntegro es un sujeto, menos necesita unidad y coherencia en sus compromisos y, por tanto, menos necesidad tiene de evitar reconocer un compromiso como suyo. De hecho, Fingarette opina que es la integridad que le suponemos al sujeto que se autoengaña lo que hace que atemperemos nuestra condena para con él [Fingarette (1969), p. 139; cf. Rorty (1988), p. 25].

Finalmente, Fingarette subraya que la declaración de los propios compromisos e incluso la "agencia personal" o integración de los compromisos en una unidad, son condición necesaria —aunque no suficiente— de la responsabilidad. De este modo, a diferencia de un niño de 1 año, un niño de 5 años es capaz de reconocer sus compromisos con el mundo, pero no suele considerársele responsable [Fingarette (1969), p. 146]. Más aún: los sociópatas son consideradas personas que reconocen sus acciones y compromisos y actúan incluso de modo inteligente, pero no reconocen responsabilidad moral alguna en sus actos [Fingarette (1969), p. 147]. Por ello, el reconocimiento de los propios compromisos es condición necesaria para el reconocimiento de responsabilidad, pero no suficiente. Fingarette reconoce que esto conduce a problemas que no va a tratar [Fingarette (1969), p. 149].

Treinta años después, Fingarette abandona el léxico psicoanalítico y algunas de sus tesis más cercanas a este enfoque, pero en el fondo su postura no ha variado mucho. Así, en su artículo "Self-Deception Needs No Explaining" defiende que la razón por la que el autoengaño se nos aparece como un fenómeno desconcertante, inconsistente e incluso autocontradictorio reside en que cuando tratamos de explicarlo caemos en un malentendido con respecto a su propia naturaleza. En realidad el autoengaño representa un modo de actuar totalmente cotidiano y las técnicas que empleamos al autoengañarnos son comunes a las que empleamos en muchas otras actividades no problemáticas de nuestra vida mental ordinaria, aunque los resultados sean inusuales. Lo que debemos hacer es tratar de explicar cómo funciona en realidad la mente, y veremos como el carácter problemático del autoengaño desaparece [Fingarette (1998), pp. 289-90].

La clave para Fingarette está en distinguir entre "tomar algo en cuenta" ("*take into account*") y "focalizar nuestra atención" ("*focus our attention*"). En nuestra vida diaria realizamos muchas de nuestras acciones sin focalizar nuestra atención en ellas, como quien escribe algo y está cogiendo el lápiz de un determinado modo, está escribiendo dentro de los márgenes, siguiendo una línea horizontal imaginaria o siguiendo unas reglas sintácticas, gramaticales y ortográficas. Todas estas cosas las tenemos en cuenta cuando escribimos, pero no necesitamos focalizar nuestra atención sobre ellas. De hecho, buena parte de lo que implica aprender a escribir consiste en tomar en cuenta todas estas cosas sin necesidad de focalizar nuestra atención sobre ellas [Fingarette (1998), p. 291].

Todo esto es algo que hacemos de un modo inconsciente, sin focalizar nuestra atención en ello. Pero además hay muchas otras cosas sobre las que no focalizamos nuestra atención, como el ruido de los coches que pasan o el del frigorífico en la cocina; estos ruidos están siendo registrados por nuestros oídos, pero no les hacemos caso, y si no lo hacemos es precisamente porque resultarían perjudiciales para la tarea que estamos realizando. Sin embargo Fingarette afirma que no podemos decir que nuestro cerebro no esté tomándolos en cuenta, pues a veces nos llega de repente un sonido que reconocemos, como el coche de nuestra mujer, y focalizamos nuestra atención en ello saliendo a recibirla. Ciertamente estábamos concentrados en aquello que deseábamos escribir y no focalizábamos nuestra atención en los ruidos que nos resultaban irrelevantes o incluso molestos para nuestra tarea, pero sí estábamos registrándolos todos ellos y teniéndolos en cuenta, como demuestra el que podamos volver nuestra mirada y focalizar nuestra atención sobre el ruido que sí nos resulta relevante: el coche de nuestra mujer llegando a casa [Fingarette (1998), p. 292].

Por tanto, el proceso mental por el que focalizamos nuestra atención sobre aquello que nos interesa y evitamos concentrarnos sobre aquellos datos que nos resultan innecesarios, irrelevantes o molestos para nuestros propósitos (pese a que los tenemos en cuenta) es un proceso totalmente normal y cotidiano que empleamos de un modo inteligente y adaptativo ya que sólo somos capaces de focalizar nuestra atención sobre unos pocos datos y siempre despreciamos aquellos que, pese a ser registrados por nosotros, resultan superfluos e incluso molestos para nuestra actividad.

Lo que sucede con el autoengaño es que, como ocurre con muchos otros de nuestros procesos cotidianos, hace uso de esta técnica, pero el resultado que se produce es distintivo y característico. En los casos de autoengaño, ante determinadas situaciones que sabemos resultarían dañinas o perjudiciales para nosotros, decidimos no focalizar nuestra atención sobre ellas y sesgamos nuestra evidencia o nuestros recuerdos pensando en otra cosa, con lo que nuestras creencias se forman sobre una evidencia (nuevos datos, recuerdos...) adulterada. Por tanto creemos algo que en principio, si focalizásemos nuestra atención sobre la evidencia relevante disponible, no estaríamos dispuestos a creer.

Una pregunta importante para Fingarette es si el sujeto que se autoengaña puede ser sincero con respecto a sus creencias y sentimientos si sabe que está evitando ciertas situaciones o recuerdos. Su respuesta es que sí. Para Fingarette, aquí sucede algo similar a lo que nos ocurre cuando nos encontramos viendo una obra de teatro. Ciertamente sabemos que quienes estamos viendo son en realidad actores, que sus nombres no son ni Otelo ni Desdémona y que aquello que dicen no es espontáneo y ha sido escrito por otro. Sin embargo, suspendemos estas creencias y no somos conscientes de ellas. De hecho, ante ciertas situaciones no nos vemos en nuestra butaca y experimentamos sinceros sentimientos de dolor, ternura, alegría o compasión[37]. Del mismo modo, para Fingarette

[37] Es relevante señalar aquí que aquellas personas que no son capaces de suspender esas creencias no son capaces tampoco de generar esos sentimientos de un modo espontáneo y sincero, y por tanto, no disfrutan del cine o teatro.

quien se autoengaña tiene una creencia y actúa de acuerdo a ella experimentando todo ello de modo sincero [Fingarette (1998), p. 299].

Lo que sí es cierto es que cuando vemos una obra de teatro y vemos en peligro a Desdémona ni le gritamos avisándola ni salimos corriendo hacia el escenario en su ayuda, esto es, tenemos ciertas reservas en nuestras emociones; éstas no son totales[38]. Esto se debe a que, aunque estamos evitando focalizar nuestra atención sobre el hecho de que aquello no es más que una representación, sí que lo tenemos en cuenta de un modo más o menos inconsciente. Del mismo modo las creencias que tenemos como producto del autoengaño no son totales, en el sentido de que somos conscientes de algún modo de que estamos evitando cierta evidencia o recuerdos, y por tanto estamos segando la evidencia que tenemos para ellas. De hecho es esta creencia parcial y consciencia de que estamos sesgando la evidencia lo que nos permite continuar alimentando el motor de nuestro engaño y seguir sesgando la evidencia.

[38] Podría parecer que lo niños no actúan de este modo, ya que ante los guiñoles o marionetas, sí gritan, avisan... sin embargo hay que puntualizar esto. En primer lugar, parece que los niños saben que los guiñoles y marionetas no son entes animados autónomos, y sin embargo gritan y avisan. Pero creo que esto se debe al carácter participativo intrínseco a este tipo de representaciones; el niño sabe que tiene que gritar y avisar, y entra en el juego (los mayores quizá se sentirían avergonzados de realizar muchas cosas que a un niño le están permitidas). Pero hay que subrayar que el niño que grita y avisa no irrumpe en el escenario para romper la marioneta que representa al ser malvado. En otras ocasiones creo que el niño simplemente puede no ser consciente de que los que están frente a él son actores, y por eso su reacción es sincera y, a diferencia de lo que ocurre con los adultos —que son conscientes—, incontrolada. Esto ocurre sobretodo en el teatro, donde la cercanía y realidad física de los actores puede hacer difícilmente distinguible para un niño al personaje y la persona.

TERCERA PARTE

III - PRINCIPALES TEORÍAS: EL ESTADO DE LA CUESTIÓN

III.1 - Enfoques intencionalistas

A partir de los años 80 se produjo una revitalización de los estudios sobre el fenómeno del autoengaño. El enorme prestigio e influencia de algunos autores como Jon Elster (*Ulises y las sirenas, Uvas Amargas, El yo multiple, Ulises desatado*), David Pears (*Motivated Irrationality*), Bernard Williams (*Problems of the Self*), Donald Davidson (*Problems of Rationality*) o Alfred Mele (*Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*), entre otros, y el hecho de que todos ellos publicasen obras dedicadas por completo al estudio de la racionalidad e irracionalidad, así como la incorporación en el debate filosófico de los nuevos aportes de la neurociencia y la informática o el estudio de la conciencia —a través de autores como Paul Churchland, John Searle o Roger Penrose, entre muchos otros también— produjo un interés renovado en las cuestiones relacionadas con la naturaleza de la conciencia, la unidad del yo, la racionalidad y los distintos problemas asociados a estas nociones. Entre ellos, el autoengaño, por constituir el caso de irracionalidad máxima (hasta el punto de que Jon Elster ha llegado a decir que "la resolución de este

problema representa la prueba de fuego para cualquier teoría acerca de la naturaleza humana" [Elster (1979), p. 286]) ha atraído la atención y los esfuerzos explicativos de una cantidad ingente de teóricos.

Por otro lado, las propuestas de David Pears y sobre todo de Donald Davidson han ejercido una enorme influencia como polo atrayente, a la vez que ha representado una referencia inexcusable, dando lugar a que la inmensa mayoría de las teorías de los últimos 30 años puedan considerarse enfoques intencionalistas o no intencionalistas. Si bien podemos rastrear rasgos de intencionalidad en caracterizaciones anteriores, es a partir de los 80 cuando son los propios autores los que conscientemente se posicionan o bien aceptando, o bien rechazando, la idea de que la intención es el rasgo más fundamental en la explicación del fenómeno. Éste es ahora el verdadero criterio de demarcación entre teorías, aunque dentro de los intencionalistas haya autores de muy distinto pelaje.

Así, unos aceptan la idea de que el autoengaño implica la posesión de creencias contradictorias (Pears, Davidson o Rorty) mientras otros lo niegan rotundamente (Audi, Bach, Talbott o Bermúdez); otros, como es el caso de Brian McLaughlin, por ejemplo, afirman que el autoengaño comporta dos tipos de creencias: el sujeto cree *p* y cree *no-p*, pero no del mismo modo [McLaughlin (1988), p. 50]. Algunos sostienen que necesariamente las paradojas del autoengaño nos indican que hemos de suponer algún tipo de división o escisión mental (Pears, Davidson, Rorty), mientras otros no ven necesario suponer tal cosa (Audi, Bach, Bermúdez) o directamente defienden con vehemencia la unidad del yo (Talbott).

Tampoco todos aquellos que respaldan la observación de creencias contradictorias y el divisionismo, creen que sea necesario suponer que alguna de las creencias haya de ser inconsciente (Davidson), mientras otros consideran que la inconsciencia o el olvido son esenciales para el éxito del autoengaño (Audi, Pears, Bach).

En cualquier caso, desde el intencionalismo moderado de Audi, reflejado en su enfoque cognitivo-volitivo en el que el sujeto tiene un control parcial de sus creencias inconscientes, hasta el intencionalismo fuerte de Davidson, para quien el concepto de autoengaño exige la consideración de la intención como elemento constitutivo esencial, hay una compleja trama de posiciones que hemos tratado de reflejar escogiendo a los autores más emblemáticos y dedicándoles un análisis detallado.

III.1.1 - Un control parcial de las creencias inconscientes

> La consideración del autoengaño, lejos de disminuir la aparente racionalidad y responsabilidad moral del agente humano, nos anima a ampliar el dominio de ambas.
>
> [Audi (1982), p. 156]

Según Audi, aunque el término "autoengaño" es vago y se ha usado de diversas maneras, al menos en la literatura filosófica, quizá el aspecto más importante en el que difieren los distintos enfoques consiste en tratar el

fenómeno como un proceso cognitivo o volicional. El enfoque que propone Audi es más cercano al cognitivo —aunque también utiliza nociones volitivas—, y le da un papel central a la *creencia inconsciente*. [Audi (1982)].

Dado su carácter equívoco o multívoco, la primera tarea consiste entonces para Audi en fijar el fenómeno que va a estudiar, analizando para ello el amplio espectro de situaciones similares y distinguiendo el autoengaño entre ellas. En este sentido, cuando la conducta de un sujeto no nos parece ajustada a la evidencia de la que dispone, podemos pensar que está autoengañado. Sin embargo, podría estar sin duda alguna simplemente guardando las apariencias en unos casos o podría no conseguir percibir el significado de la evidencia en otros.

En otras ocasiones, el individuo muestra oscilaciones en sus creencias; cabe que éstas se produzcan de modo no problemático debido a que el apoyo evidencial del que dispone un sujeto puede cambiar con rapidez. Finalmente otras veces un individuo pasa a creer *conscientemente, y en contra de su evidencia,* que *p,* y ya no creerá que *no-p*. Cuando el sujeto muestra este tipo de actitud desconectada por completo de la realidad, entra en delirio (*delusion*) y no está autoengañado. [Audi (1982), p. 140].

Audi señala que los casos problemáticos e interesantes son aquellos que tienen que ver con el autoengaño (*self-deception*) y no con el engaño auto-causado (*deception that is self-caused*) o el engaño a uno mismo (*deceiving oneself*). Por un lado, uno está engañado simplemente cuando cree algo que

es falso; pero en la medida en que esto puede causárselo uno mismo[39], habría que distinguir el engaño que es auto-causado (*deception that is self-caused*) del autoengaño (*self-deception*) [Audi (1982), p. 143; (1997), p. 104; cf. Elster (1979), p. 292-293; (1983). p. 216; Johnston (1988), pp. 76-78]. Por otro lado, es al menos dudoso que exista alguien que pueda controlar totalmente a voluntad el autoengaño, ya que autoengañarse no es como alzar un brazo [Audi (1982), p. 145; cf. Williams (1973)]. Así, el engaño a uno mismo (*deceiving oneself*) como una decisión de engañarse directamente de un plumazo resulta aparentemente imposible.

En los casos paradigmáticos de autoengaño, el sujeto declara sinceramente, o está dispuesto a declarar sinceramente, que *p,* pero *sabe inconscientemente que no-p*. En algunos otros casos, quizá el sujeto no tenga siquiera un conocimiento inconsciente de que *no-p*, sino una mera *creencia inconsciente* de que *no-p*. [Audi (1982), p. 135]. En ambas situaciones, lo que explica por qué el conocimiento o la creencia de que *no-p* es inconsciente, es el deseo de evitar ciertos pensamientos dolorosos. Este deseo explica también por qué el sujeto trata de adquirir la creencia de que *p*. [Audi (1982), p. 136].

Este proceso puede ser en algún caso el resultado de una "tendencia instintiva" —como la tendencia del individuo a no pensar en su muerte—, pero el autoengaño suele tener una estrecha relación con el pensamiento

[39] Por ejemplo, contando con mi malísima memoria, trato de engañarme y escribo una entrada falsa en mi diario acerca de una futura cita que, meses después, olvido que era falsa, resultando víctima de mi propia estrategia. Esto sería para Audi un engaño auto-causado, pero no autoengaño.

desiderativo en tanto que, por lo general, surge en parte porque el sujeto *quiere que algo sea el caso* [Audi (1982), p. 136]. También se halla relacionado de algún modo con la debilidad de la voluntad por inclinar al sujeto a creer aquello que en principio tiene menor apoyo evidencial.

El autoengaño consiste en ciertas conductas, como alejar de la mente determinados tipos de evidencia, o negar con sinceridad algo que inconscientemente se cree o sabe [Audi (1982), p. 142]. En este sentido, el sujeto puede (a) alejar de su mente la evidencia en contra de *p*, (b) manipular la evidencia contra *p* hasta reducir su aparente plausibilidad, (c) manipular la evidencia de modo que favorezca *p* para aumentar su plausibilidad, (d) buscar nueva evidencia a favor de *p* sabiendo que está inflando la evidencia, y (e) hacer cosas que sólo resultan adecuadas si *p* es el caso. Pero precisamente por ser manifestaciones de debilidad de la voluntad, muestran que pueden estar bajo el control del sujeto. [Audi (1982), pp. 146-148].

Por tanto, S se autoengaña con respecto a *p* si y sólo si,

(1) *S* sabe *inconscientemente* que *no-p* (o tiene razones para creer, e inconsciente y verdaderamente cree que *no-p*).

(2) *S* declara con sinceridad, o está dispuesto a declarar sinceramente, que *p*; y

(3) *S* tiene al menos un deseo que explica en parte tanto por qué es inconsciente la creencia de que *no-p* como por qué *S* está dispuesto a negar una creencia de que *no-p* y a declarar que *p*,

> incluso cuando se le presenta lo que él ve como evidencia contra ello. [Audi (1982), p. 137].

No se trata de que haya un subagente —la voluntad— que determine cuándo el sujeto entra o permanece en el autoengaño [Audi (1982), p. 144]. Como vimos, Audi considera que es dudoso que exista alguien que pueda controlar totalmente a voluntad el autoengaño y, en cualquier caso, no pretende determinar si uno puede manipular o no a voluntad las creencias inconscientes, pues esto es algo que ha de decidir la psicología empírica. Más bien, lo que quiere subrayar es que si esto fuera posible, según su teoría se seguiría que el sujeto tendría un control meramente parcial e indirecto del autoengaño por medio de la modificación de sus creencias inconscientes [Audi (1982), p. 144].

Por creencia inconsciente Audi se refiere a aquellas que, aun siendo en muchos aspectos como las conscientes, tienen dos características que las distinguen: en primer lugar, el sujeto no sabe que las tiene y, si llegan a manifestarse en su vida consciente, es muy improbable que las reconozca si no hace un escrutinio especial de sus creencias u otro individuo le ayuda; en segundo lugar, es muy improbable que el sujeto *explique* las acciones que en verdad realiza en virtud de ellas, como si estuvieran realmente causadas por ellas. Sin embargo, tiende a manifestarlas en la vida consciente, por ejemplo mediante *lapsus linguae*. [Audi (1982), p. 137].

Audi se pregunta cómo podemos entonces hacer justicia a la analogía entre engaño interpersonal y autoengaño, y si quien se autoengaña ha de creer tanto *p* como *no-p*. Aunque sólo hay una persona en lugar de dos,

Audi indica que es *como si* fueran dos, ya que opera a dos niveles: a nivel inconsciente, el individuo cree que *no-p*, pero a nivel consciente declara que *p*. [Audi (1982), p. 141]. Efectivamente, Audi trata de recoger la tensión esencial en el autoengaño a la vez que escapa a la paradoja estática o doxástica: según él, la declaración sincera de que *p* no entraña que el sujeto crea que *p*[40]. Esto se debe a que, aunque es posible que alguien crea proposiciones que son contradictorias, no es posible que alguien crea dos creencias que él mismo cree contradictorias.

> No me parece que el que S esté en estado de autoengaño con respecto a p y crea inconscientemente que no-p, entrañe que también crea —aunque conscientemente— que p. Todo lo que mi enfoque requiere con respecto a la actitud positiva de S hacia p, es que S esté dispuesto a declararlo sinceramente. Ya que la declaración sincera que hace S de que p es en general una razón excelente para pensar que lo cree, podemos querer decir, en virtud de su declaración sincera, que "conscientemente lo cree". ¿Deberíamos decir, sin embargo, que realmente cree que p cuando inconscientemente sabe o cree que no-p? Creo que no. En este respecto, mi enfoque difiere de aquellos que exigen la creencia de que p y de que no-p.[41]

---

[40] La tensión que Audi quiere reflejar —y que según él no captan los enfoques no-intencionalistas como el de Alfred Mele— es aquélla que muestra cómo el sujeto sabe que la evidencia apunta en un sentido, y sin embargo afirma sinceramente lo contrario [Audi (1997), p. 104]. Mele se ha defendido criticando tanto la idea de que un sujeto pueda declarar sinceramente que *p* sin creer realmente que *p* como, por ende, el que esta idea sirva para escapar a la paradoja estática [Mele (1982); (1987b); (1997), p. 131].

[41] «It does not appear to me that *S*'s being in self-deception with respect to *p* and unconsciously believing not-*p* entails his believing—albeit consciously—that *p*. All my view requires regarding *S*'s positive attitude toward *p* is that *S* be disposed sincerely to avow it. Since *S's* sincerely avowing *p* is generally excellent reason to think he believes it, we may want to say, in virtue of his sincerely avowing it, that he "consciously believes" it. Should we say, however, that he actually does believe *p*

Esto se debe a que la contradicción entre proposiciones del tipo *p* y *no-p* es tan obvia para cualquiera, que será percibida y rechazada. En todo caso, Audi afirma que incluso aunque el sujeto pudiera mantener tal contradicción, siempre que sea posible es preferible explicar los datos de un modo alternativo que no implique algo irracional y, aparentemente, inexplicable. [Audi (1982), pp. 139, 147].

Este enfoque, por tanto, trata de solventar las paradojas estática y dinámica rechazando la idea de que el sujeto tenga creencias contradictorias, al tiempo que contempla la tensión entre las declaraciones del sujeto y su evidencia que los no-intencionalistas pasarían por alto. La razón por la que Audi puede ser considerado como intencionalista, es que otorga a la voluntad un papel central en todo el proceso, si no directo (debido a la aparente imposibilidad de creer a voluntad, idea que Audi comparte con Bernard Williams), sí de un modo indirecto.

III.1.2 - *Pensar* en p *y pensar* que p

Para Bach el autoengaño es un fenómeno complejo y por ello el enfoque que ofrece pretende recoger toda esa complejidad a la vez que trata de escapar de las paradojas.

En primer lugar, Bach se propone dejar sentado qué fenómenos caen fuera del centro de nuestra discusión. El autoengaño no es ni (1)

when he unconsciously knows or believes not-*p*? I think not. In this respect, my account differs from those requiring beliefs that p and that not-*p*.» [Audi (1982), p. 138].

pensamiento desiderativo, pues a diferencia de aquel, éste no implica ningún tipo de razonamiento o racionalización, ni (2), ceguera intelectual, pues quien se autoengaña ve todo demasiado bien, ni (3) un caso de pensamiento sesgado, ya que cuando acusamos a alguien de estar prejuiciado o sesgado, simplemente queremos decir que está influenciado por sus sentimientos, ni (4) un tipo peculiar de irracionalidad similar a cuando estamos cansados, confusos, en estado de shock o bajo los efectos del alcohol, pues en este tipo de casos no es necesario que haya una motivación que nos autosatisfaga, como sucede en el autoengaño. [Bach (1981), pp. 351-352]. Tampoco está de acuerdo con la descripción que hace David Pears según la cual quien se autoengaña tiene, en contra de su evidencia, "el deseo de creer que *p*"; Bach cree que quien se autoengaña no tiene el deseo de tener una creencia, sino el deseo de que las cosas sean de tal o cual manera [Bach (1981), p. 353]. Por esta razón, Bach rechaza la idea de que el autoengaño sea una cuestión de creencias; en concreto, rechaza caracterización de quien se autoengaña como alguien que posee creencias contradictorias. La propuesta de Bach consiste en presentar a quien se autoengaña como un individuo que "desea que *no-p* mientras cree que *p*, y lo que hace es evitar el pensamiento recurrente o sostenido de que *p*."[42]

La clave para Bach está en distinguir entre "creer" y "pensar". Él reserva el término "creer" para los usos disposicionales de creencias y "pensar" para las ocurrencias de creencias. Así, pensar que *p* no sería ni

[42] «[the] self-deceiver desires that no-p while believing that p, and what he does is to avoid the sustained or recurrent thought that p.» [Bach (1981), p. 354]

necesario ni suficiente para creer que *p*. No es condición necesaria porque de hecho todos tenemos muchas creencias en las que no pensamos, así como otras que jamás se nos han pasado por la mente, como que los canguros son más grandes que las cacatúas. Y es suficiente porque no es lo mismo "pensar *en p*" (*think* of *p*) que "pensar *que p*" (*think* that *p*). Efectivamente uno puede pensar en *p* y, a la vez, (a) pensar *que* no es posible que *p* sea el caso, (b) *que* es seguro que *p* es el caso, o (c) *que* no tiene ni idea acerca de si *p* es o no el caso. Todas ellas son compatibles con el pensamiento *en p*. Por ejemplo, un sujeto puede *pensar en* que mañana llueva (pensando quizás en las consecuencias que tendría para sus planes) y además *pensar que* no lloverá; también puede *pensar en* que mañana llueva y *pensar que* lloverá o *pensar en* que mañana llueva y *pensar que* no podría pronunciarse dado el estado cambiante del tiempo ese día [Bach (1981), p. 354]. Esta distinción es en la que se apoya Bach para describir el autoengaño de modo intencional, sin caer en las paradojas.

En su opinión, normalmente la creencia de que *p* es el caso hace que en cualquier momento en el que *p* venga a la mente del sujeto, es decir, cada vez que piense *en p*, el sujeto *piense que p*. Sin embargo, lo que ocurre en el caso del autoengaño es que esa tendencia es superada: en estos casos, una creencia dolorosa hace saltar un mecanismo por el que el sujeto reinterpreta la evidencia. Hay tres tipos de mecanismos: racionalización, evasión e inserción. [Bach (1981), pp. 357 y ss.]

La racionalización (*rationalization*) supone que el sujeto trata de dar una explicación a la nueva creencia o pensamiento *de que p* que adquiere.

Aunque tenga evidencia en contra, trata de racionalizar esta evidencia para hacerla casar con esta nueva creencia. Esto es, además, el proceso normal en las ciencias —dice Bach: cuando las teorías comienzan a fallar, no se abandonan sin lucha ante la primera evidencia contraria. Ante datos recalcitrantes, los científicos no reajustan sus teorías, sino que examinan qué puede ir mal en esos datos. Antes de realizar cambios importantes tratamos de explicar esos datos y, sólo si es imposible dar cuenta de ellos sin construir epiciclos teóricos, abandonamos la teoría. Esto pasaría también en la vida diaria y, de este modo, tratamos de racionalizar nuestras creencias antes de abandonarlas ante el primer signo de evidencia contraria [Bach (1981), p. 358; cf. Quine y Ullian (1978); Rorty (1988), p. 19]. Lo que sucede en el autoengaño es que la creencia es contraevidencial, el proceso de racionalización se hace más complejo y no siempre funciona. Es entonces cuando el sujeto pasa a una de las otras estrategias.

La evasión (*evasion*) es la estrategia a la que todos los autores hacen referencia y consiste en sesgar la evidencia. Es muy utilizada por el sujeto que se autoengaña, que simplemente evita la adquisición de evidencia contraria y resalta la evidencia favorable para el apoyo de la creencia contraevidencial que desea abrazar.

Por último, si las dos estrategias precedentes resultan infructuosas, queda la inserción (*jamming*). Esta hace referencia a que en algunas ocasiones, ante el pensamiento *en p*, al sujeto le es imposible evitar el pensamiento *de que p*, y lo único que puede hacer es intentar traer a su mente el pensamiento de que *no-p*, actuar *como si no-p* fuese el caso y tratar

de imaginar las consecuencias favorables que se desprenderían de que *no-p* fuese el caso, con la esperanza de llegar a convencerse de que *no-p*. Lo que hace aquí el sujeto es inducirse a la fuerza el pensamiento *de que no-p*.

Mediante estos mecanismos el sujeto evita *intencionalmente* el recurrente pensamiento *de que p* es el caso, movido por el deseo de que *no* sea el caso que *p* y por la creencia de que *p*. Por tanto, el autoengaño no exige necesariamente que haya un cambio de creencia; no es necesario que el sujeto que cree que *p* es el caso pase a creer que *no-p* lo es (aunque puede hacerlo), sino que es suficiente con que ante el pensamiento *en p*, el sujeto piense *que no-p* [Bach (1981), p. 357].

Es por tanto la creencia (dolorosa) de que *p* junto con el deseo de que *no-p* lo que lleva al sujeto a evitar pensar *que p*. Este proceso es intencional, pero esta intención no incluye ni la intención de autoengañarse ni la de violar los propios patrones de racionalidad internos. El sujeto sólo evita el pensamiento *de que p* cuando *p* viene a su mente, pero no es consciente de que lo que le facilita evitar el pensamiento de que *p* sea caso es el deseo de que *no-p* combinado con la creencia (originaria) de que *p*. Esto se le mantiene oculto (en esta exigencia de inconsciencia de la incoherencia e inconsistencia del proyecto es en lo que se acercará a David Pears), pues en caso contrario el sujeto sería consciente de la irracionalidad del proceso que le permite evitar pensar que *p,* y por tanto no podría llevarlo a cabo. Sin embargo no hay nada que impida que el sujeto pueda hacerse consciente en algún momento de que este proceso le hace romper sus

patrones de racionalidad y, por ello, a veces el sujeto se hace consciente de todo el asunto y el autoengaño fracasa.

III.1.3 - Teoría de Subsistemas

III.1.3.1 - División, inconsciente y consistencia

La postura defendida por David Pears tiene algunas similitudes con la de Kent Bach pero, como veremos, se distingue de ella en importantes cuestiones. En concreto, Pears no acude a la distinción ente creer y pensar y sí cree que el autoengaño puede incluir en algunos casos la observación de creencias contradictorias.

David Pears es uno de los introductores, junto a Davidson, de la teoría de subsistemas. La importancia de su teoría queda constatada, entre otras cosas, por el hecho de que Davidson admite una importante deuda intelectual con sus ideas.

Según Pears, el autoengaño se presenta como paradójico por dos razones principalmente. La primera paradoja puede formularse del siguiente modo: ¿cómo puede uno formarse una creencia contra el peso total de la evidencia contraria? Para que esto resulte realmente paradójico tiene que tratarse además de un proceso consciente y libre, pues casos como la hipnosis o el error al apreciar el impacto de los razonamientos en la elección de una creencia u otra están libres de irracionalidad [Pears (1982), p. 266]. La segunda paradoja apunta a la dificultad que supone el

hecho de que el (auto)engaño parece exigir que se mantenga la creencia opuesta a la nueva que se quiere generar.

A Pears esta segunda paradoja le parece superficial, ya que parece deberse a un hecho puramente léxico, que dependería de la elección del nombre que se le asigne al fenómeno: si el nombre "autoengaño" tiene una connotación autocontradictoria, sería mejor no aplicar este nombre a la formación desiderativa de creencias. Sin embargo, Pears señala que los mecanismos y estrategias subyacentes en los casos que consideramos como autoengaño son mucho más complejos que en los ejemplos de pensamiento desiderativo, y por tanto es necesario buscar una explicación distinta que dé cuenta de este fenómeno [Pears (1982), p. 266].

Evidentemente David Pears acepta la posibilidad del autoengaño, y rechaza la tentación de interpretar éste en los términos de engaño interpersonal por las terribles consecuencias conceptuales que hemos visto, pero añade además otra razón: mientras la búsqueda de la verdad es una meta que uno debe perseguir (es la meta que la racionalidad apoya) el deseo de dar a otros la oportunidad de alcanzar la verdad no es ni mucho menos universal [Pears (1986), p. 63]. El autoengaño, entonces, no puede ser como el engaño a otro, pues aunque a otros les neguemos la oportunidad de alcanzar la verdad, no podemos negarnos a nosotros mismos tal meta.

Las estrategias que el sujeto sigue para autengañarse son principalmente tres, dirigidas hacia el *input*, el *output* y lo que hay dentro de la propia mente. En el primer caso, algún deseo, dolor o miedo hace que el sujeto

sesgue la evidencia que le llega (esta es la más común y la más fácil de realizar, en tanto que no poseemos una creencia anterior a la que desbancar o con la que podamos entrar en conflicto). En el segundo caso, lo que hacemos es, pese a que la evidencia apunta a la verdad de *no-p,* actuar como si *p* fuera el caso, lo cual puede ayudarnos a creer que *p* ciertamente es el caso. Por último tenemos el caso en el que nuestros deseos atacan la creencia instalada en el sistema principal asaltando directamente la evidencia que apoya esta creencia.

Pears distingue además entre dos tipos de autoengaño: los que presentan holgura o margen (*latitude cases*) y los que no (*non-latitude cases*) [Pears (1986), p. 63]. El tipo de autoengaño que presenta margen es aquel en el que la evidencia que apoya a *p,* pese a ser mucho mayor que la que apoya a *no-p*, no llega a ser concluyente y, por ello, el sujeto tiene cierto "margen de maniobra"; hay, por así decirlo, cierta holgura, porque aun sería posible que nueva evidencia favoreciese a *no-p*. En estos casos es posible autoengañarse "abiertamente" (*openly*), pero esto requiere que el sujeto sea consciente, no sólo del origen de su creencia, sino también de los mecanismos utilizados. En tanto que el origen suele ser un deseo, y los deseos no apuntan a la verdad (meta que según Pears todos perseguimos como seres racionales), el sujeto podrá hacerse consciente de la irracionalidad de su creencia y, por esta razón, este tipo de autoengaño "abierto" suele estar condenado al fracaso. Este es el motivo de que aún en los casos que presentan holgura, el sujeto no se engañe siempre abiertamente, y a veces se oculte la creencia originaria, los motivos por los

que sesga la evidencia y los métodos a través de los cuales lo hace. En este caso, sí puede tener éxito.

En los casos en los que no hay margen, bien porque la evidencia es definitiva, bien porque hay un argumento lógicamente válido que conduce directamente a una creencia que no deseo, parece que ni el conjunto de la evidencia concluyente ni la inferencia lógica *pueden* aparecer de modo consciente en el sujeto. Esta es la razón por la que, sobre todo en estos casos en los que no hay margen, el sujeto suele tender a ocultarse la creencia originaria, los motivos por los que sesga la evidencia y los métodos a través de los cuales lo hace. La pregunta es: cuándo no hay margen, ¿cómo puede alguien seguir manteniendo el autoengaño? Según Pears parece que la única solución es aceptar, al menos de un modo general, la hipótesis de Freud de la partición de la mente.

> Es difícil evitar la conclusión de que [el individuo] debe ser inconsciente de su irracionalidad y, de modo similar, de que en muchos casos en los que hay holgura, debe ser inconsciente del hecho de que su creencia está causada por un deseo. Esta fue la idea principal de Freud y parece que debe ser gran parte de la verdad.[43]

Pears admite entonces, como harán Donald Davidson o Amélie Rorty, que ha de haber unos límites que separen dichos elementos, e infiere así la existencia de un sistema principal y unos subsistemas, porque la aparente inconsistencia e irracionalidad nos obliga a ello. Sin embargo, Pears

[43] «It is difficult to avoid the conclusion that he must be unconscious of its irrationality and similarly that in many cases where there is latitude he must be unconscious of the fact that his belief is caused by a wish. This was Freud's main idea and it seems that it must be a large part of the truth» [Pears (1986), p. 69]

advierte que uno ha de evitar pensar que la idea de que hay un subsistema segregado del sistema principal reintroduce la dualidad entre engañador y engañado:

> La relación entre el subsistema y el sistema principal no es como la relación entre dos personas, una de las cuales quiere engañar a la otra.[44]

Como veremos Mark Johnston, siguiendo a Talbott, criticará duramente la idea de un subsistema que tenga capacidad intencional y discriminatoria para discernir qué genera y qué no genera inconsistencia. Concretamente cree que no se entiende cómo es posible que el sistema principal no se percate de la manipulación a la que es sometido; si se percata, o bien todo se va al traste, o bien hay una connivencia entre ambos sistemas. Pero esto reintroduce el problema: el sistema principal parece ahora conocedor *en parte* del plan del subsistema, y Johnston se pregunta si hemos de admitir un subsistema dentro del subsistema para explicar ese conocimiento parcial; así, o bien hay una cadena infinita de subsistemas que no explicaría nada, o se pone fin a la cadena con un último subsistema "lo suficientemente estúpido" como para no percibir el engaño [Johnston (1988), p. 65; cf. Talbott (1995), nota 44, p. 64].

Pears responde —de un modo muy poco convincente a mi juicio— que el subsistema simplemente inserta la creencia deseada en el sistema principal antes de que haya una creencia que desplazar:

[44] The sub-system and main system is not quite so like the relation between two people one of whom wants to deceive the other [Pears (1991), p. 402].

> El subsistema no tiene que persuadir al sistema principal para que adopte una creencia contraevidencial por el placer que recibirá al hacerlo, porque simplemente la genera directamente en el sistema principal. Se sigue que el sistema principal no es simplón, sino vulnerable a la manipulación interna […][45]

Para Pears, lo que se produce en la mente del sujeto es un aislamiento (*insulation*) de la nueva creencia contraevidencial que genera la inconsistencia con la antigua creencia, y es este aislamiento lo que le concede la inmunidad frente a la evidencia contraria. Pero pese a que Pears cree que el sujeto ha de ocultarse necesariamente los motivos por los que sesga la evidencia y los métodos a través de los cuales lo hace, precisa sin embargo que la línea que separa sistema y subsistema no es, como en Freud, la que hay entre consciencia e inconsciencia [Pears (1986), p. 69]. El subsistema no sólo contiene la creencia contraevidencial y el deseo, temor, o sospecha que la genera, sino que además debe incluir *todo que sea necesario* para que el propio subsistema sea *autoconsistente*.

Pears insiste en que algunos autores exigen consistencia al sistema principal, y se olvidan de que el subsistema ha de ser consistente también. El dibujo que presenta Pears no son dos conjuntos uno incluido en otro (pues la creencia nueva no puede coexistir con la otra sin desencadenar una fatal inconsistencia); pero tampoco disjuntos, pues muchos elementos del sistema principal son necesarios para la consistencia del subsistema,

[45] The sub-system does not have to persuade the main system to adopt the counter-evidential belief for the sake of the pleasure that it will get from it, because it simply generates it directly in the main system. It follows that the main system is not gullible but vulnerable to inner manipulation [Pears (1991), p. 402].

del mismo modo que los elementos del subsistema que no generen inconsistencia con el sistema principal pueden (y deben) pertenecer al sistema principal; por tanto, "los dos subsistemas habrían de ser concebidos como círculos solapados"[46]. De este modo hay elementos del subsistema que son conscientes, al menos porque pertenecen al sistema principal también, mientras otros elementos (sólo los que generan directamente inconsistencia) no lo son. Por tanto, al subsistema pertenecen todos los elementos que no pueden aparecer en el sistema principal porque generarían inconsistencia, más aquellos elementos del sistema principal que el subsistema necesita para presentar él también consistencia interna [Pears (1986), pp. 74-75]. Queda claro por tanto, por qué la línea divisoria entre sistemas no puede ser la que separa consciencia e inconsciencia.

El divisionismo de Pears tiene continuidad en Davidson, aunque con ligeros matices; en primer lugar, Pears no excluye la posibilidad de que los subsistemas sean centros separados de agencia [Pears (1991), pp. 404-405], cosa que Davidson niega explícitamente. En segundo lugar, Pears cree que el sujeto ha de esconderse a sí mismo los métodos y motivos por los que alcanzó la creencia placentera, mientras para Davidson, como veremos a continuación, no habrá ninguna razón para suponer que algún elemento del subsistema haya de ser inaccesible o inconsciente para el sujeto. Por

[46] «[…] the two systems should to be pictured as overlapping circles» [Pears (1986), p. 73; cf. Davidson (1982), nota 6, p. 181].

estas razones, el divisionismo davidsoniano es, como veremos a continuación, mucho más modesto.

III.1.3.2 - Caridad vs interpretación: divisionismo moderado

El enfoque de Donald Davidson ha sido tomado sin duda como el principal punto de referencia por la mayoría de las propuestas posteriores a su artículo "Deception and Division" de 1985, a pesar de que el propio Davidson reconoce abiertamente las fuertes deudas que tiene su postura con los postulados freudianos y, sobre todo, con David Pears:

> Las diferencias entre mi concepción y la de Pears son pequeñas comparadas con las similitudes. Esto no es accidental, ya que mi exposición es deudora de ambos artículos[47]. [Davidson (1985), nota al pie, p. 114]

Sin embargo, pese a estas palabras de Davidson, su obra presenta —a nuestro juicio— un enfoque complejo y fino que le hace merecedor del título de campeón de la interpretación intencionalista del autoengaño. En la exposición de su teoría vamos a observar el tratamiento que ofrece de varias cuestiones cruciales en el abordaje del autoengaño: el problema de la irracionalidad, el holismo, la división de la mente y el propio autoengaño.

Para Davidson, aunque el autoengaño no supone normalmente un problema para quienes se autoengañan, sí lo es para la psicología filosófica, pues representa un ejemplo de irracionalidad en nuestra vida

[47] Davidson se refiere a 'Motivated irrationality' y 'The Goals and Strategies of Self-deception.' [Pears (1982) y (1986)]

diaria. La irracionalidad es algo que un sujeto atribuye a otro al tratar de explicar infructuosamente la relación entre sus creencias, intenciones y acciones.

El único modo en que podemos llevar a cabo esta tarea es mediante el *Principio de Caridad*, que nos obliga a ser lo más benevolentes que nos sea posible en la interpretación de las actitudes de los demás, evitando la adscripción de incoherencias o creencias extravagantes, y establece que los demás son en buena medida como nosotros, esto es, hemos de suponer que lo que ocurre en la mente de otros individuos es similar a lo que ocurre en la nuestra y que actúan de acuerdo a patrones de racionalidad semejantes a los nuestros. Es porque al sujeto le presuponemos racionalidad, por lo que la irracionalidad no es una falta de racionalidad, sino un fallo dentro de la propia racionalidad; concretamente, una ruptura de los patrones cognitivos del propio sujeto. Esta ruptura se produce, en el caso del autoengaño, debido a que el sujeto no sólo mantiene dos creencias contradictorias, sino que además parece que una de ellas sostiene (de algún modo) a la otra. Pero veamos esto detenidamente.

I. LA IRRACIONALIDAD

Como decimos, para Davidson lo irracional no es meramente lo no-racional, lo que permanece fuera del ámbito de lo racional; la irracionalidad es "un fallo dentro de la casa de la razón" [Davidson (1982), p. 169]. En este sentido, buena parte de lo que se considera irracional no constituye una paradoja. Muchos podrían mantener que es irracional,

dados los peligros que entraña, intentar escalar el Everest sin oxígeno (o incluso con él). Pero no hay problema en explicar esto si se toman en cuenta todos los deseos, ambiciones y actitudes en conjunto y que la persona actúa a la luz de su conocimiento y valores. Davidson admite que quizá sea irracional creer en la astrología, en platillos volantes o brujas, pero indica que tales creencias pueden tener explicaciones estándar si están basadas en lo que sus creyentes creen que es evidencia. Por el contrario, el tipo de irracionalidad que constituye un problema es aquel que supone un fallo de coherencia o consistencia en el patrón de creencias, actitudes, emociones, intenciones y acciones de una persona. Ejemplos de esto son el pensamiento desiderativo, actuar contra el mejor juicio propio o *akrasia*, creer algo que uno mantiene que está desacreditado por el peso de la evidencia y el autoengaño. [Davidson (1982), p. 170]

Davidson afirma que cualquier interpretación satisfactoria del problema ha de abrazar algunas de las más importantes tesis de Freud, pese a que muchos filósofos denuncien que entraña errores fundamentales. Al intentar explicar estos fenómenos, los freudianos han sostenido tres cosas:

1) La mente contiene un número de estructuras semi-independientes caracterizadas por atributos mentales como pensamientos, deseos y recuerdos.

2) Las partes de la mente son agentes en aspectos importantes, no sólo en tener (o consistir en) creencias, necesidades y otros rasgos psicológicos, sino en que éstos pueden combinarse para causar otros eventos en la mente o fuera de ella.

3) Algunas de las disposiciones, actitudes y eventos mentales habrían de contemplarse bajo el modelo de disposiciones y fuerzas físicas cuando afectan a, o son afectadas por, otras subestructuras de la mente. [Davidson (1982), pp. 170-171]

Esta doctrina no es compartida por algunos filósofos —por ejemplo, Sartre— debido a que, por un lado, la idea de que la mente puede estar dividida se ha considerado ininteligible, ya que parece requerir que los pensamientos, apetencias e incluso acciones se atribuyan a algo menor y distinto de la persona completa: cada parte se convertiría en un pequeño agente y la mente única se convertiría en "un campo de batalla donde fuerzas contrapuestas luchan, se engañan unas a otras, se esconden información y planean estrategias". Por otro lado, hay muchas dudas con respecto a su metodología explicativa, pues ensancha en exceso el campo teleológico aumentando el número de fenómenos que pueden explicarse de modo racional (sueños, olvidos, *lapsus linguae*, angustias, psicosis, etc.). Las críticas se centran en que el psicoanálisis simplemente trata de hacer algo imposible al intentar dar cuenta de mecanismos psicológicos en términos de leyes causales.

Frente a esto, Davidson cree que algunos elementos importantes en el tratamiento freudiano de la irracionalidad son correctos [Davidson (1982), pp. 171-172]. Los tres elementos que Davidson rescata de la teoría psicoanalítica son: la partición de la mente, la existencia de una estructura considerable en cada parte cuasi-autónoma, y las relaciones causales no-lógicas entre las partes. Además, Davidson señala que muchos fenómenos mentales que son normalmente accesibles a la consciencia, no son en

otros momentos ni conscientes ni fácilmente accesibles a la consciencia. Subraya que esto es, en cierto modo, independiente de la teoría de la división de la mente, pero la fortalece al dar cuenta de un mayor rango de fenómenos. [Davidson (1982), pp. 185-186]

Con respecto a la tesis freudiana de la relación causal entre eventos mentales, Davidson afirma que no hay conflicto inherente entre explicaciones de razón y explicaciones causales. Desde el momento en que creencias y deseos causan las acciones para las que son razones, las explicaciones de razón incluyen un elemento causal esencial. La mayoría de nuestros deseos, esperanzas, apetencias, emociones, creencias y miedos depende "de una inferencia simple (por lo general desapercibida) desde otras creencias o actitudes". [Davidson (1982), p. 174]

Así, al explicar una acción son necesarios dos elementos: un valor, meta o actitud del agente, y la creencia de que actuando de un modo concreto se puede alcanzar ese valor o meta. Entre ellos ha de haber una *relación lógica:* las creencias y deseos tienen un contenido, y estos contenidos han de ser tales que impliquen que hay algo de valor o deseable en la acción. Asimismo, las razones que un agente tiene para actuar deben ser, si van a explicar la acción, las razones *por las que* actuó; las razones han de haber jugado un papel *causal* en la ocurrencia de la acción. Por tanto, toda actitud racional (creencia, deseo, acción, etc.) se explica aludiendo a la conexión *causal* y *racional* entre distintos eventos mentales.

La contrapartida de esto es que Davidson describe la irracionalidad como la mera *conexión causal* entre dos eventos mentales o entre un evento

mental y acción; como una relación causal que no constituye una razón. Sin embargo, dejando a un lado ciertos fenómenos especiales[48], Davidson señala que el hecho de que haya una relación meramente causal, y *no* de razón —lógica—, entre eventos mentales no siempre constituye un caso de irracionalidad:

> Hay, sin embargo, un modo en el que un evento mental puede causar otro evento mental sin ser una razón para ello, y donde no hay problema ni irracionalidad alguna necesariamente. Esto puede ocurrir cuando causa y efecto tienen lugar en mentes diferentes. Por ejemplo, deseando que entres en mi jardín, planto una bonita flor allí. Tú te encaprichas por ver mi flor y entras en el jardín. Mi deseo causó tu encaprichamiento y la acción, pero mi deseo no fue una razón para tu encaprichamiento, ni una razón por la que actuaste. (Quizá incluso ni siquiera tenías conocimiento de mi deseo). Hay fenómenos mentales que pueden causar otros fenómenos mentales sin ser razones para ellos. […] Sugiero que esta idea puede aplicarse a una mente simple y una persona." [Davidson (1982), p. 181]

Ahora bien, si Davidson no quiere renunciar a clarificar la irracionalidad de quien se autoengaña y con este fin aplica a un solo individuo esta explicación de la conexión meramente causal entre eventos mentales de dos individuos, se verá obligado a admitir otro de los postulados freudianos, a saber: la partición de la mente.

> De hecho, si es que vamos a explicar la irracionalidad, parece que debemos asumir que la mente puede estar particionada en estructuras cuasi-independientes que interactúan. [Davidson (1982), p. 181]

[48] En algunos casos particulares estas condiciones para la racionalidad no son necesarias, como podemos observar, por ejemplo, en algunos fenómenos como el de *asociación* (e.g., cuando recuerdo un nombre al canturrear cierto tono). Este tipo de casos suponen que un evento mental cause otro para el que no constituye una razón, sin generar por ello irracionalidad. [Davidson (1982), p. 186]

Efectivamente, la división de la mente es el único modo en que Davidson puede reproducir en un único individuo la dualidad inherente en el caso de causalidad mental interpersonal. Para constituir una estructura de este tipo, cada parte de la mente ha de mostrar un grado de consistencia o racionalidad mayor que el que se le atribuye al todo. Y añade en una nota al pie:

> Aquí, como en otros lugares, mi enfoque altamente abstracto de partición de la mente se separa del de Freud. En concreto, no tengo nada que decir sobre el número o naturaleza de las divisiones de la mente, su permanencia o etiología. Sólo me interesa defender la idea de una compartimentación mental, y argumentar que es necesaria si vamos a explicar una forma común de irracionalidad. Debería subrayar quizá que frases como "partición de la mente", "parte de la mente", "segmento", etc., son engañosas si sugieren que lo que pertenece a una división de la mente no puede pertenecer a otra. Mi dibujo consiste en territorios solapados. [Davidson (1982), nota 6, p. 181; cf. Pears (1986), p. 73]

Y aquí, el holismo davidsoniano genera una tensión entre este divisionismo mental y el Principio de Caridad interpretativa: al tratar de explicar la conducta de un sujeto en términos de deseos, creencias, proyectos, motivos, etc., lo que un sujeto declara acerca de sus estados mentales no supone un acceso más directo a sus estados mentales que otros comportamientos suyos, ya que aquello que declara también ha de interpretarse. Debido a que la cantidad de elementos que influyen en su comportamiento es tan grande que nos resulta imposible tomarlos todos en cuenta (son "demasiadas incógnitas para las ecuaciones"), la estrategia entonces pasa por asumir que la persona en cuestión es, en gran parte, como nosotros; podemos admitir diferencias en cuestiones que no sean centrales, pero a causa del *carácter holísta* de lo mental (es decir, que una

proposición no tiene sentido sola, y que su atribución supone la atribución de muchas más) necesitamos compartir la mayor parte de creencias para comprender su acción. Sin embargo, el Principio de Caridad interpretativa está en oposición a la partición de la mente, ya que mientras la partición de la mente es una hipótesis que permite la coexistencia en la misma mente de creencias contradictorias, el Principio de Caridad establece que la inconsistencia o incoherencia conduce a la ininteligibilidad; esta dificultad es conocida como la paradoja de la irracionalidad:

> [S]i explicamos algo irracional demasiado bien, alcanzamos una forma encubierta de racionalidad, y si simplemente asignamos incoherencia de un modo sospechosamente demasiado sencillo, simplemente estaremos poniendo en entredicho nuestra capacidad para diagnosticar irracionalidad, al alejarnos del trasfondo de racionalidad necesario para justificar cualquier diagnóstico.[49]

La salida a este atolladero está, según Davidson, en ver esto como una cuestión de grado: mientras las diferencias entre nuestras creencias y las del sujeto y las incoherencias sean suficientemente pequeñas, no supondrán un gran problema a la hora de predecir y explicar su comportamiento, pero cuando se hacen demasiado grandes, nos resulta del todo incomprensible [Davidson (1982), pp. 183-184].

Debemos poner especial cuidado en la tarea de explicar la acción de los sujetos. Concretamente, hemos de tomar en cuenta que

[49] «[I]f we explain it too well, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, we merely compromise our ability to diagnose irrationality by withdrawing the background of rationality needed to justify any diagnosis at all.» [Davidson (1982), p. 184]

> puesto que una creencia no puede mantener su identidad al perder sus relaciones con otras creencias, no es posible que la misma proposición sirva para interpretar actitudes particulares de dos personas distintas y guarde al mismo tiempo, con las demás actitudes de una de ellas, relaciones muy diferentes de las que guarda con las de la otra. [Davidson (1985), pp. 105-106]

Por tanto las diferencias se explican, siempre que sean suficientemente pequeñas, sobre un fondo de creencias compartidas. Cuando las desviaciones respecto de los patrones de racionalidad son demasiado grandes "es más verosímil que se encuentren en el ojo del intérprete que en la mente del sujeto interpretado". [Davidson (1985), p. 105] Como decíamos anteriormente, si la actitud de un sujeto nos resultase por completo incomprensible, le atribuiríamos *arracionalidad*, mientras que si lo comprendemos en gran medida y en algún punto nos resulta ininteligible, le atribuimos *irracionalidad* en ese aspecto. Uno debe poder pensar en gran parte de modo racional para poder actuar o ser irracional. [Davidson (1993), p. 218]

Davidson, previendo las críticas de *falacia homuncular* a su divisionismo (dificultad que ya vimos que había señalado Sartre al propio Freud y que Mark Johnston, principalmente, les indicará los intencionalistas), advierte de un modo bastante oscuro y sin más explicación que

> La idea de divisiones cuasi-autónomas no exige un pequeño agente en la división; de nuevo, los conceptos operativos son el de causa y razón. [Davidson (1982), p. 185]

Si la partición de la mente en dos o más estructuras semi-autónomas es el único modo de explicar cómo un pensamiento o impulso puede causar

otro sin ser una razón, el fallo o caída (*breakdown*) de la razón es lo que muestra dónde trazar las fronteras de esa subdivisión. [Davidson (1982), p. 185]

II. AUTOENGAÑO

Según Davidson, el autoengaño no consiste en "mentirse a uno mismo". Aunque pudiera parecer lo contrario, mentirse a uno mismo resulta más difícil de explicar que el autoengaño, pues la mentira es un tipo de actitud especial que supone una insinceridad con respecto a la representación de las propias creencias; por tanto, el autoengaño no consiste en mentirse a uno mismo porque, aunque la mentira es también un *acto intencional* que requiere un *motivo*, quien miente sólo logra engañar si su audiencia no discierne su intención de dar una imagen falsa de lo que cree. Así, tomar la noción de mentirse a uno mismo de modo demasiado literal generaría un problema: el sujeto habría de realizar un acto con la intención tanto de que la intención fuese reconocida por él mismo, como no reconocida, y esto daría al traste con su propósito al implicar una intención de autoanulación. [Davidson (1985), p. 111; (1993), p. 215; cf. Sartre (1943), p. 98] Por esa razón,

> […] sería mejor hablar de "mentirse a sí mismo" como un tipo de metáfora —una metáfora muerta, ya que se usa muy a menudo; sería a lo sumo, entonces, un modismo. [Davidson (1993), p. 215]

Sabemos que la mentira es sólo un modo de engaño. Así, Davidson buscará explicar el autoengaño como un tipo de engaño distinto a la mentira, un engaño menos directo.

Por otra parte, el autoengaño tampoco consiste meramente en la posesión de creencias contradictorias. Davidson cree que en raras ocasiones un sujeto está seguro (*certain*) de que una proposición es verdadera a la vez que está seguro también de que la contradictoria de esa proposición es verdadera. Esto puede ocurrir cuando el sujeto no es lo suficientemente competente para captar la incoherencia entre ambas creencias, bien por falta de conocimientos, bien porque la dificultad o cantidad de datos que requiere lo hace imposible; pero esto no es en principio irracional ni constituye autoengaño.

Sin embargo, en otras ocasiones un sujeto tiene una gran evidencia a favor de una cierta proposición que le inclina a creer que es el caso y, a causa del enorme dolor o desasosiego que esto le genera, tiende a favorecer y resaltar la evidencia a favor de la falsedad de esa proposición, lo cual hace que más tarde esté más inclinado a apoyar la negación de tal proposición. Esta situación en la que el sujeto acaba por estar más inclinado a creer una proposición contraria a la que estaría en principio más apoyada por la totalidad de la evidencia disponible, acerca el problema del autoengaño al problema de la debilidad de la voluntad, ya que el sujeto contravendría también lo que le indica el *requisito de evidencia total*; la diferencia es que en esta ocasión no se trata de un problema evaluativo del par creencia-acción, sino de un problema cognitivo del par evidencia-creencia. Por esa razón no habría una voluntad débil, sino una justificación débil. No es necesario reexponer aquí la caracterización de la debilidad de la voluntad que vimos en §I.5.

Lo interesante y paradójico, es que la creencia placentera a la que llega el sujeto está sustentada por la original displacentera, y ambas conviven.

Tómense las siguientes proposiciones:

(1) Davidson cree que es calvo.

(2) Davidson cree que no es calvo.

(3) Davidson cree que (es calvo y no es calvo).

(4) Davidson no cree que es calvo.

El sujeto que se autoengaña tiene una creencia (1), que le genera malestar; esta creencia *causa* de algún modo que más tarde crea (2). Así cree y descree (*disbelieve*) la misma proposición; Davidson subraya que es tentador mantener que (2), esto es, descreer una proposición, entraña (4), o sea, no creerla; pero en ese caso entonces el problema es nuestro: podríamos estar diciendo que el sujeto cree y no cree la misma proposición y este tipo de descripción haría la situación contradictoria[50]. Podemos no obstante escapar de esta dificultad rechazando esa implicación.

Pero por otro lado, dado que el sujeto cree (1) y (2), podríamos inferir que también cree (3). Nuevamente ha de resistirse esta tentación, ya que supondría atribuir al sujeto la creencia de una contradicción limpia y

[50] C*a*p ∧¬C*a*p, donde C= creer, a= Davidson y p = "soy calvo".

obvia[51]. Hemos de distinguir entre creer proposiciones contradictorias y creer una contradicción[52] [Davidson (1985), pp. 99-100; (1993), pp. 216-217]

La razón por la que constituye un verdadero problema tanto el que el sujeto crea proposiciones contradictorias como cualquier otra forma de irracionalidad, radica en que uno ha de aceptar cierto grado de holismo si pretende explicar la racionalidad. Efectivamente, si las creencias fuesen elementos atómicos que pudieran ser añadidos individualmente, cambiados o eliminados sin tomar en cuenta su entorno proposicional, como mantienen —a juicio de Davidson— Jerry Fodor y Ernie Lepore[53], entonces cualquier grado de inconsistencia sería posible. Davidson cree por el contrario que la mera posesión de actitudes proposicionales supone un alto grado de consistencia, y que la identificación de creencias depende en parte de sus relaciones lógicas con otras creencias; en este sentido, las inconsistencias imponen una tensión en la atribución o explicación de creencias [Davidson (1993), p. 217].

Los casos centrales de autoengaño se caracterizan por que (a), el sujeto sigue siendo consciente de que el conjunto de su evidencia apoya la creencia contraria a la que mantiene y (b), es la conciencia de este hecho lo que motiva sus esfuerzos por librarse del temor que le produce. [Davidson (1985), p. 113] Por tanto, está claro que el autoengaño incluye debilidad de

[51] $Ca(p \wedge \neg p)$, donde C= creer, a= Davidson y p = "soy calvo".

[52] Esto es, Davidson nos insta a distinguir entre $Cap \wedge Ca\neg p$ y $Ca(p \wedge \neg p)$.

[53] Davidson se refiere a Fodor, J. y Lepore E. (1992), *Holism: A Shopper's Guide*, Blackwell, Oxford.

la justificación, pues el sujeto acepta una proposición sobre *sólo una parte* de lo que considera como la evidencia relevante [Davidson (1985), pp. 106-107] Además, hay un motivo para el autoengaño y, a diferencia de la debilidad de la voluntad o de la debilidad de la justificación en las que no forma parte de su análisis el hecho de que la desviación tenga un motivo (aunque pueda tenerlo), el estudio del motivo es central en el caso del autoengaño.

El autoengaño no sólo tiene una causa (todo la tiene), sino que ha de tener una razón: alguien tiene una creencia que encuentra desagradable, dolorosa o desalentadora, y por tanto tiene una razón para cambiar las cosas. Así, actúa o piensa de un modo que causa el rechazo de la creencia desagradable. Sin embargo, uno ha de percatarse de los dos sentidos en los que uno puede decir que tiene una razón para abrazar una creencia: una cosa es tener una razón para creer algo (un motivo) y otra tener una evidencia a la luz de la cual es razonable creer algo [Davidson (1993), p. 216; cf. (1985), p. 107].

Davidson nos advierte que tampoco es irracional en sí misma una acción intencional que tienda a hacernos felices o a aliviar nuestras penas, ni se convierte en tal si los medios empleados incluyen el intento de disponer las cosas con vistas a tener cierta creencia. Según Davidson a veces puede ser inmoral hacer esto con otra persona, y esto mismo se aplica a las creencias autoinducidas; sin embargo, lo que no es necesariamente irracional cuando se le hace a otra persona tampoco lo es cuando el objeto es el propio yo futuro. Es más, las creencias de este tipo

pueden ser buenas a veces; las relaciones sociales entre familiares y amigos se mantienen a menudo gracias a una mejor opinión de la que estaría justificada, y algunos esposos y esposas mantienen la estabilidad familiar gracias a que pasan por alto indicios de infidelidad.

Una creencia autoinducida es irracional si y sólo si el sujeto *sigue pensando* que la evidencia en contra de la creencia en mejor que la evidencia en favor de la misma, pues entonces estamos ante un caso de *debilidad de la justificación*. Pero si el sujeto ha olvidado la evidencia que originariamente le llevó a rechazar la creencia que ahora abriga, o si la nueva evidencia parece ahora lo bastante buena como para compensar la antigua, el nuevo estado mental no es irracional. [Davidson (1985), p. 108]

Además, una creencia autoinducida es irracional porque la creencia original sustenta la creencia contraria, donde "sustentar" no significa ofrecer fundamento racional, sino meramente causal. El conocimiento o creencia original ha de jugar algún papel causal en el autoengaño. [Davidson (1985), p. 112; (1993), p. 216][54] Hay por tanto una causa para el autoengaño que no constituye una verdadera razón [Davidson (1985), pp.

[54] Davidson explica en otra parte que lo que evita que el ejemplo de la entrada falsa en la agenda sea un caso de autoengaño es que cuando la persona ve la entrada y cree que la cita será tal día, no juzga que la totalidad de su evidencia apoya que la cita no será ese día. La intención que inicia la acción que conduce a la creencia no es a misma que la que sostiene la creencia. En el enfoque davidsoniano, por tanto, aunque la persona está engañada, no está autoengañada. En este sentido, las palabras de Davidson recuerdan a las de Bela Szabados: "[T]o say of someone that he is self-deceived is *not* merely to say that he is in a certain state of mind; ascriptions of self-deceit also involve an appraisal as to how the person in question got into that state of mind and how he sustains himself in it" [Szabados (1974a), p. 54]

112, 117; cf. (1982), pp. 173, 181, 186; (1993), p. 220] Esta relación causal entre las acciones que llevan al autoengaño y el propio estado de engaño *no puede ser accidental*, pues de otro modo una persona estaría autoengañada si leyese una noticia falsa en un periódico. [Davidson (1985), p. 110]

El autoengaño puede implicar, además de debilidad de la justificación, pensamiento desiderativo [Davidson (1985), p. 109, 112]. La diferencia es que en el pensamiento desiderativo el autoengaño toma siempre la forma del deseo, mientras que en el autoengaño no siempre sucede así. De este modo, alguien se puede autoengañar en creer algo que le resulte especialmente doloroso, como ocurre en los casos de esposos inseguros y demasiado celosos, hipocondriacos, etc.[55] [Davidson (1985), p. 109; (1993), p. 216; cf. Mele (2001), pp. 4-5; Barnes (1997), p. 54]

En resumen, los casos centrales de autoengaño podrían ser definidos, así pues, del siguiente modo: el sujeto A tiene una evidencia sobre la base de la cual la verdad de *p* es más probable que su falsedad, y la creencia de que *p* es el caso o el pensamiento de que debería creer que *p* es el caso, motiva que A *intencionalmente* favorezca la evidencia en favor de *no-p* o minimice la evidencia en favor de *p,* con lo que acaba por creer que *no-p* es el caso. Además —y esto es lo que según Davidson hace del autoengaño un problema—, el estado mental que motiva el autoengaño y el estado que produce coexisten. Por consiguiente, el autoengaño es una especie de

[55] Este tipo de autoengaño al que Alfred Mele se refiere como autoengaño retorcido (*twisted self-deception*) [Mele (2001), pp. 4-5; cap. 5], es para Davidson y Mele un caso peculiar de autoengaño.

debilidad de la justificación autoinducida donde la razón (motivo) para inducir una creencia determinada es el dolor o malestar que produce la creencia de su contradictoria. Por ello el autoengaño es un fenómeno intencional que exige un estado en el que ambas creencias se presenten de modo consciente al sujeto. Además en algunos casos, aunque no en todos, el autoengaño envuelve pensamiento desiderativo, ya que la razón que mueve el proceso es que el sujeto desea que aquella creencia que se induce sea el caso (o teme que no lo sea).

La irracionalidad proviene para Davidson, como hemos dicho, del hecho de que el sujeto considera simultáneamente creencias contradictorias[56], y parece además que una está causada y, en el caso más grave, sostenida por la otra. Lo que ha de explicarse es cómo el sujeto puede mantener ambas creencias ($Cap \wedge Ca\neg p$), sin ponerlas en conjunción ($Ca(p \wedge \neg p)$), ya que aunque es posible creer un conjunto de proposiciones contradictorias entre sí, no es posible creer la conjunción de aquéllas cuando la contradicción resulta obvia [Davidson (1985), p. 115; (1993), p. 217; cf. Audi (1982), p. 139]. Las creencias pueden coexistir si de algún modo permanecen separadas y *sin que se permita contemplarlas de un*

[56] Davidson dice que, en este respecto, su opinión coincide con la de Jon Elster [Davidson (1985), nota 9, p. 115]. Sin embargo, parece ser un error de lectura de Davidson, pues en el lugar indicado por Davidson, Elster dice textualmente: «self-deception should be clearly distinguished from the simultaenous entertainment of incompatible beliefs» [Elster (1979), p. 174] ("el autoengaño debe ser claramente distinguido del mantenimiento simultáneo de creencias contradictorias" [Elster (1979), p. 289]). También en otro lugar dice Ester "No se trata de que el autoengaño implique alimentar simultáneamente creencias contradictorias, lo cual haría del autoengaño algo imposible" [Elster (1983), p. 214]).

único vistazo. [Davidson (1993), p. 220] Por ello, debemos aceptar la idea de que existen unos límites entre distintas partes de la mente, y debemos postular que esos límites separan esas creencias contradictorias. La caída del patrón de racionalidad es la que nos indica la necesidad de marcar una línea divisoria dentro del sistema principal, dando lugar a un subsistema.

El paso irracional consiste precisamente en trazar ese límite que mantiene separadas las creencias contradictorias. La *causa* de esta delimitación es el deseo de evitar lo que el *requerimiento de evidencia total* recomienda, pero esto nunca puede ser una *razón*, pues no hay nada que pueda resultar una razón para obviar las mejores normas de racionalidad de un individuo. [Davidson (1985), p. 116-117]

Estos límites no se descubren por introspección, sino que son *ayudas conceptuales* que hacen coherente la descripción de ciertos fenómenos irracionales, entre ellos el autoengaño. Además, a Davidson no se le ocurre ninguna razón para suponer que alguno de los territorios delimitados sea inaccesible a la consciencia.

Esta idea de la división de la mente es deudora de la teoría psicoanalítica —aunque como señalamos no tiene todos sus rasgos— y se hace eco de una larga tradición: Platón, Aristóteles, Agustín o Butler han propuesto partes semiautónomas del alma en sus filosofías de la mente. Pero Davidson subraya que no asume

> […] que las divisiones sean fijas, o que merezcan tales nombres como consciencia, coraje, intelecto o *ello*.

Además, no cree que los límites, permanentes o temporales, sean

> territorios autónomos separados; los territorios se solapan […] La imagen a la que querría invitar no es, por tanto, la de dos mentes separadas capaces de actuar como agentes independientes; la imagen es más bien una única mente no completamente integrada; un cerebro que sufre, quizá, de una lobotomía temporal autoinfligida. [Davidson (1993), pp. 220-221]

En realidad, Davidson sólo necesita suponer que hay un subsistema que mantiene la nueva creencia contraevidencial separada de toda la evidencia contraria y de la antigua creencia que se mantiene en el sistema principal.

La teoría de sistemas y subsistemas ha sido duramente criticada y acusada de introducir elementos explicativos que han sido interpretados como maniobras *ad hoc* por unos [Bermúdez (1997)], como falacias homunculares por otros [Johsnton (1988)] o como exotismos mentales [Mele (2001)]. Sin embargo, tras las críticas Davidson se defiende subrayando que su postura no trata de ser una explicación psicológica del autoengaño, sino conceptual:

> Este enfoque altamente abstracto de la estructura lógica del autoengaño no pretende, ni lo hizo nunca, ser una explicación psicológicamente reveladora de la naturaleza y etiología del autoengaño. Su modesto propósito fue eliminar, o al menos mitigar, las características que en principio hacen el autoengaño inconcebible. [Davidson (1993), p. 221]

Por otro lado, Davidson es consciente de las dificultades, pero también está seguro de que ningún enfoque dará cuenta de todos los casos [Davidson (1993), p. 221; cf. Mele (2001), p. 65], ya que el autoengaño

> […] se presenta en muchos grados, desde sueños ordinarios, pasando por ensoñaciones semidirigidas, hasta completas alucinaciones; desde la

> imaginación normal de las consecuencias de acciones ponderadas hasta los delirios psicóticos, desde el inocuo pensamiento desiderativo hasta el error autoinducido elaboradamente. Sería un error tratar de dibujar trazos claros entre estos continuos. [Davidson (1993), p. 230]

Los análisis filosóficos no toman en cuenta todos estos detalles y omiten el color que confiere interés a los casos particulares, y es normal preguntarse si habrá realmente un fenómeno que se ajuste a una definición tan "acartonada y formal" [Davidson (1993), p. 221], pero Davidson no cree que estos análisis tengan que ser falsos por ser pálidos y racionales. [Davidson (1993), p. 230]

III.1.3.3 - Integración y desintegración del yo

> Las atribuciones de autoengaño son atribuciones de incoherencia, no atribuciones incoherentes.
>
> [Rorty (1972), p. 395]

Amélie O. Rorty, ha sido señalada a menudo como representante de una postura divisionista fuerte o, al menos, más fuerte que las de Donald Davidson o David Pears.

Aunque Rorty cree que el autoengaño es un fenómeno variado y rico, dice estar interesada sobre todo en los casos en los que el sujeto niega tener las creencias que le son claramente atribuibles, y considera —al igual que la mayoría de autores—, que el autoengaño le debe su carácter paradójico a que es un fenómeno que "se multiplica": el sujeto no sólo ha de engañarse con respecto a algo, sino que ha de engañarse con respecto a

los movimientos envueltos en ese autoengaño, sus inusuales focalizaciones y desvíos de la atención. Además, requiere también actitudes de segundo orden, como el reconocimiento del conflicto entre creencias, y algunas estrategias *ad hoc* encaminadas a reconciliar tales conflictos [Rorty (1988), p. 12]. Si el yo es un todo racionalmente integrado que escanea y corrige automáticamente sus creencias, entonces el autoengaño es paradójico.

Según Rorty, las condiciones que hacen posible el autoengaño son las siguientes:

1) La persona cree que *p,* y
2) O bien (a): la persona cree que *no-p*. Generalmente esto supone que la persona cree que *q*, lo que (dadas sus creencias y sus arraigados hábitos de inferencia) debería reconocer como equivalente a *no-p*.
   O bien (b): la persona niega que crea que *p*.
3) Si el autoengaño no se reduce a error, la persona ha de reconocer a algún nivel que posee creencias en conflicto. Generalmente, la atribución de tal reconocimiento es una inferencia a la mejor explicación del comportamiento e inferencias de esa persona.
4) Si el autoengaño no se reduce a conflicto, la persona ha de negar a algún nivel que sus creencias estén en conflicto. A veces esto se logra por medio de una estrategia *ad hoc* con el fin de reconciliar el conflicto aparente. Quien se autoengaña normalmente no trata de suspender el juicio, ni de determinar cuáles de sus creencias son defectuosas.
5) La atribución de autoengaño presupone una teoría de lo que una persona normalmente cree, percibe, reconoce, infiere; presupone

> que acepta cánones de racionalidad y que está alerta con respecto al tipo de evidencia que pesa en contra de su creencia. [Rorty (1988), p. 25, nota 1]

En principio, podría pensarse que la división del yo en múltiples subsistemas acabaría con el problema: si el yo es un sistema vagamente organizado y compuesto de subsistemas autónomos, parece posibilitar el autoengaño y el fenómeno quedaría así desmitificado y naturalizado, e incluso, explicado en cierto modo. Sin embargo, como veremos, a pesar de las apariencias esta segunda imagen del autoengaño socava totalmente la posibilidad del autoengaño. Amélie Rorty planteará una teoría divisionista, pero desea salvar el fenómeno.

Como decíamos, si el yo está esencialmente unificado o al menos fuertemente integrado (y, por tanto, es una entidad capaz de efectuar reflexiones críticas orientadas a la verdad, con todas sus funciones accesibles en principio y corregibles entre ellas), no puede engañarse a sí mismo. Para dar cuenta del fenómeno, uno puede tratar de debilitar esta imagen del yo, cuyas características ideales serían la unidad, transparencia, aspiración de veracidad, y reflexividad; ahora serían más bien *integración* en lugar de *unidad*, la *conectividad sistemática* en lugar de la *transparencia*, *principios de racionalización* en lugar de *veracidad* y, finalmente, la condición de *reflexividad* se convierte en un ideal regulativo [Rorty (1988), p. 15]. Ha de subrayarse que Rorty asume que, sobre todo bajo ciertas condiciones de opacidad —como cuando una persona está cansada o enferma—, no hay nada de misterioso o particular en que un sujeto cometa errores sin intención alguna. En todo caso, presentar al yo como orientado de modo

predominante hacia la verdad no nos fuerza a admitir algún tipo de incapacidad para mostrar ignorancia o cometer errores, efectuar juicios poco cuidadosos o no regulados, evidenciar fallos de atención, realizar inferencias ilógicas y mantener creencias volubles o conflictos no reconocidos. Un yo complejo y unificado puede sufrir todas estas debilidades: pero es en principio capaz de percatarse de sus desórdenes y, bajo circunstancias normales, su corrección no necesita de una motivación especial [Rorty (1988), p. 13]. Y, sin embargo, ella cree que hay procesos de irracionalidad que no podemos explicar como meros errores o fallos.

El problema es que aquí hay cierta tensión, ya que sólo si están en funcionamiento las capacidades de racionalidad crítica y reflexiva, se puede atribuir autoengaño. Si estas capacidades no se hallan en acto, los presuntos casos de autoengaño se reducen a casos de ignorancia, conflicto o error, pues donde no hay presunción o capacidad de racionalidad, no puede haber *a fortiori* fallos de racionalidad. [Rorty (1988), p. 16]

Es entonces cuando Rorty inicia el giro hacia la visión del yo como un conjunto de subsistemas. Según ella, no hubiéramos sobrevivido como las criaturas que somos si nuestras únicas capacidades como investigadores fueran las de unidad, transparencia y crítica. Ni siquiera hubiésemos sobrevivido si la racionalidad crítica fuese nuestro único ideal regulativo, dominante sobre otros [Rorty (1988), p. 16]. Dejando a un lado que esta tesis acerca de nuestra supervivencia es cuando menos controvertida, la pretensión de Rorty es defender, como dice Daniel Dennet siguiendo a Gazzaniga, que "la mente no es un sistema unificado, sino un paquete de

subsistemas parcialmente autónomos unidos bajo un yugo" [Dennett (1992), p. 111][57] y que, aunque para algunos propósitos un *monitor central panóptico* pudiera resultar adaptativo, también poseemos subsistemas que hacen uso de mecanismos que se disparan automáticamente. Estos modos de compartimentación, focalización automanipulada, insensibilidad selectiva, persistencia ciega o astuta insensibilidad, tienen enormes beneficios.

Por tanto, bajo esta segunda imagen, el yo

> está dividido en subsistemas homunculares que están compuestos a su vez de subsistemas cada vez más simples e independientes, alcanzando finalmente un nivel de funciones proto-intencionales relativamente mecánicas, subpersonales, y especializadas.[58]

Del mismo modo que existen patrones de dominancia en la atención visual (por ejemplo, el rojo domina sobre el gris, o el movimiento irregular sobre los objetos estáticos), ciertas actitudes resultan "magnéticas" con respecto a la atención, como el miedo, la agresión, los lazos amorosos, el vínculo con los hijos y las acciones y reacciones al poder, estableciendo *patrones generalizados de predominancia en nuestra atención*. El hecho de que estos elementos capten de modo fuerte nuestra atención, resulta generalmente

[57] Daniel Dennett (1992), "The Self as a Center of Narrative Gravity", en Frank S. Kessel, Pamela M. Cole y Dale L. Johnson (eds.), *Self and Consciousness: Multiple Perspectives*, Hillsdale (NJ), Lawrence Erlbaum Associates, pp. 103-115.

[58] «[The] self is divided into homuncular subsystems that are themselves composed of increasingly simple, independent subsystems, eventually reaching a level of relatively mechanical, subpersonic, proto-intentional functions» [Rorty (1988), p. 19]

beneficioso; sin embargo, en algunas ocasiones pueden anteponerse a otras consideraciones que serían más apropiadas en ese contexto o situación concretos. Así, cuando una persona tiene miedo, está absorbida por el amor, o envuelta en dolor, quizá no sea capaz de tomar en cuenta e integrar cierto material que sería relevante pero queda en la periferia de su atención y, de este modo, lo que no resulta destacado, puede parecerle *subjetivamente poco importante* [Rorty (1988), pp. 17-18; cf. Fingarette (1998)]. De modo similar, hay otra estrategia en principio beneficiosa que puede dar lugar tangencialmente al autoengaño: la inercia de las creencias en contra de la evidencia. En principio, uno tiene la inercia de mantener sus creencias ante la aparición de evidencia contraria con el fin de evitar una constante revisión y cambio de creencias. [Rorty (1988), p. 19; cf. Quine y Ullian (1978); Bach (1981), p. 358].

Sin embargo, irónicamente, tratar de controlar ciertas emociones y su capacidad magnética no siempre trae consecuencias deseables. Consciente de su tendencia a la hipocondría, y sabiendo que los médicos de diagnóstico son susceptibles de miedos y enfermedades, una médica que posee evidencia de padecer ella misma un cáncer, podría adoptar la política general de intentar ignorar o, al menos evitar, observar su condición física. Su negativa al cáncer no tiene por qué ser una maniobra cínica y *ad hoc;* podría racionalizarse como una *política justificada* que consistiría en evitar caer en casos de hipocondría. Por traer generalmente grandes beneficios, una política que como consecuencia no pretendida —aunque en parte predecible— permita autoengaño, puede ser a veces la política más razonable, una vez consideradas todas las cosas. El

autoengaño es irracional; la política que "tangencialmente le da hospitalidad", no. [Rorty (1988), p. 14]

Resumamos dónde estamos: bajo la primera imagen del yo (*modelo del yo unificado*), una buena parte del pensamiento es simplemente clasificado como erróneo, irracional, en conflicto e ignorante. Sin embargo, cuando la irracionalidad cumple un patrón, y seres supuestamente racionales continúan mostrando una resistencia inesperada a la corrección, necesitamos una verdadera explicación. Por su parte, la segunda imagen del yo (*modelo del conjunto de subsistemas*) ha logrado dar cuenta de los fallos de integración como errores en la integración de las funciones de los distintos subsistemas, y con ello ha desmitificado y naturalizado el problema… pero a costa de hacerlo desaparecer. El problema "se ha evaporado" [Rorty (1988), p. 21] porque con la división se ha abandonado la identidad de engañador y engañado [Rorty (1988), p. 22]. Podemos preguntarnos entonces: ¿ha llegado Rorty a un callejón sin salida? Según ella, nosotros atribuimos con cierta frecuencia autoengaño a los demás, e incluso a nuestros yoes pasados. Pero, si todo esto es ilusorio, ¿por qué esta ilusión es tan persistente? ¿Por qué puede rastrearse a lo largo de toda la historia, en biografías, novelas, casos de estudio, etc.?

La razón se encontraría, según Rorty, en que no podemos renunciar a ninguna de estas imágenes ya que, por un lado, el yo como sistema unitario se revela como ideal regulativo necesario en la revisión y corrección de errores y creencias falsas (pues si los subsistemas funcionasen de modo totalmente independiente, nada garantizaría que el

sujeto se formase creencias con contenido contradictorio a cada momento), y esto es condición esencial para que nos veamos como agentes racionales y creyentes responsables; pero por otro lado, cualquier imagen seria del yo debería incluir otras características además de la racionalidad crítica, porque "una criatura cuyas únicas creencias y motivaciones se derivasen de la racionalidad crítica sería una criatura muy aburrida y de vida corta" [Rorty (1988), p. 23]. Sigo pensando que Rorty podría tratar de argumentar esto con algún tipo de dato etológico o antropológico.

En todo caso, cada una de las imágenes intenta representar importantes rasgos de la otra, y salvar los fenómenos que la otra niega. El defensor de una de ellas tratará quizá de superponer la una a la otra, o reducir una a la otra; por ejemplo, Rorty cree que en principio parecería que un modo de salvar el autoengaño sería presentar al yo como un conjunto de subsistemas que le otorgarse un rango especial a las capacidades ejercidas en la crítica reflexiva. Sin embargo, Rorty es consciente de que esto no es más que superponer la imagen del yo unificado, con su visión panóptica y capacidades reflexivas y legislativas, sobre la imagen del yo como conjunto de subsistemas [Rorty (1988), p. 23]. Y tampoco el yo con capacidad crítica puede ser *un sistema más* entre el conjunto de subsistemas. Así mismo, tampoco es cierto que el yo unificado sea una experiencia subjetiva o de primera persona, mientras el yo como conjunto de subsistemas sea una experiencia objetiva o de tercera persona [Rorty (1988), p. 24]. De hecho, es esta superposición de imágenes la que genera, en opinión de Rorty, la aparente intratabilidad del autoengaño.

Según Rorty toda teoría del yo ha de incluir ambas imágenes de modo complementario, porque el funcionamiento normal del yo ha de incluir ambos modos; es precisamente cuando uno de ellos falla, cuando se produce el autoengaño:

> Sólo aquellos que, a pesar de su compromiso real y efectivo con la primera imagen, están realmente compuestos de subsistemas relativamente autónomos, pueden no conseguir integrar sus creencias. Así que sólo una persona cuyo yo esté presuntamente integrado e interprete su sistema-de-subsistemas-relativamente-independientes a través de la *primera imagen del yo*, sólo una persona que trate la independencia de sus subsistemas constitutivos como fallos de integración, es capaz de autoengaño. Es una enfermedad que sólo pueden padecer los que son presuntamente de mentalidad fuerte.[59]

Generalmente Amélie Rorty es considerada como la abanderada del caso más extremo de divisionismo o escisión de la mente. Es cierto que según su postura, sólo una división de lo mental puede dar cuenta de las maniobras que estarían implícitas en el autoengaño (y las maniobras subsiguientes con el propósito de ocular las primeras, etc.). Pero no es menos cierto que ella afirma explícitamente que si el sujeto presentase una escisión absoluta, no podríamos atribuirle autoengaño, pues éste exige algún tipo de negligencia epistémica bajo la forma de un continuo obviar

[59] «Only those who, despite their effective, actual commitment to the first picture, are actually composed of relatively autonomous subsystems can fail to integrate what they believe. So only a presumptively integrated person who interprets her system-of-relatively-independent-subsystems through the first *picture of the self*, only a person who treats the independence of her constituent subsystems as failures of integration, is capable of self-deception. Not everyone has the special talents and capacities for self-deception. It is a disease only the presumptively strong minded can suffer.» [Rorty (1988), p. 25, cf. Fingarette (1969), p. 139].

la evidencia y no aceptar la responsabilidad. En pocas palabras: el sujeto ha de ser capaz de tener todo esto en cuenta de algún modo "panópticamente" para que pueda atribuírsele autoengaño y no mero error. Esto presupone, sino ya unidad, sí al menos algún tipo de integración de los subsistemas por lo que queda claro que el divisionismo que propone Rorty no es tan extremo como en ocasiones ha querido presentarse.

En cualquier caso es cierto que la salida a este atolladero que ensaya exige que el sujeto, que realmente está constituido de subsistemas, crea erróneamente que su yo está unificado, y por tanto no vea la necesidad de integrar la información que realmente maneja cada subsistema. Éste tipo de desorden que ha de presentar quien se autoengaña bajo la forma de una compartimentación de la actividad de los subsistemas, es lo que ha llevado a otros autores a interpretar a Rorty como defensora de un fuerte divisionismo.

III.1.4 - Un proyecto intencional en una mente coherente y unificada

En su extenso artículo de 1995, 'Intentional Self-deception in a Single Coherent Self', menos reconocido de lo que sin duda merece, Talbott presenta una postura diferente e interesante con respecto al problema del autoengaño, en tanto que defiende la necesidad de incorporar la intención en el proceso, pero rechaza tanto el divisionismo como la posesión de creencias contradictorias por parte de quien se autoengaña.

Para Talbott, habría poca motivación para ser intencionalista si hubiese una buena teoría no-intencionalista, pero no cree que así sea. El mejor argumento para defender una teoría intencional es que las teorías no-intencionales no son capaces de explicar lo siguiente: si el autoengaño es ciertamente un proceso no intencional, ciego… ¿cómo explicar que a veces se ponga en funcionamiento y otras no? ¿Por qué ante la angustia que me produce el pensamiento de que mis frenos pueden fallar, no me formo simplemente la creencia de que están en perfectas condiciones? Parece que si este mecanismo fuese ciego, no habría sido favorecido por la selección natural.

> Creo que esta selectividad del autoengaño sólo se puede explicar adecuadamente bajo la suposición de que implica adulteración intencional.[60]

Por tanto, Talbott cree que lo que separa el autoengaño de muchos otros procesos psicológicos similares es la intención. Por ejemplo, los procesos de represión y defensa de tipo freudiano se corresponderían más bien con un modelo no-intencional y, sin negar la existencia de este tipo de mecanismos para la reducción de angustia, Talbott cree que no constituirían autoengaño [Talbott (1995), p. 65].

Según él, la resistencia por parte de algunos teóricos a aceptar un modelo intencional se debe a que la mayoría de los modelos intencionales

[60] «I believe that this selectivity of self-deception can only be adequately explained on the supposition that it involves intentional biasing» [Talbott (1995), p. 62; cf. Bermúdez (1997), p. 108; (2000), p. 317].

son divisionistas, es decir, a que generan más problemas de los que solucionan:

> Estoy de acuerdo con Mele [(1987[a]), pp. 139-43] y Johnston [(1988), pp. 82-86] en que en las explicaciones del autoengaño, tales divisiones del yo parecen dar lugar al menos a tantos problemas explicativos como los que solucionan.[61]

Los problemas que indica Talbott apuntan a que este tipo de divisiones parecen tener una peligrosa tendencia a multiplicarse fuera de control *ad infinitum*. Así, ¿por qué el sistema víctima controla lo que controla? ¿Por qué la parte que engaña controla lo que controla? ¿Cómo se producen los cambios de control, dado que no son al azar? ¿Cómo se explica que el sub-yo estratega esté activo cuando es preciso? Parece que los divisionistas habrían de introducir sucesivas divisiones (introducir un super-estratega que indique cuándo ha de estar activo el sub-yo estratega). [Talbott (1995), nota 44, p. 64; cf. Johnston (1988), pp. 64-65].

No obstante, Talbott no defiende la tesis de la transparencia de la consciencia o que todo saber sea consciencia de saber; antes bien, considera que es evidente que cualquier explicación adecuada de algunos fenómenos, como la diferencia ente memoria a corto y largo plazo, requiere algún tipo de división. Sin embargo, a este tipo de divisiones las llama Talbott "divisiones inocentes", ya que no son postulados *ad hoc* para resolver un problema particular, sino que son comunes a muchos

[61] «I agree with Mele [(1987), pp. 139-43] and Johnston [(1988), pp. 82-86] that in explanations of self-deception, such divisions of the self seem to raise at least as many explanatory problems as they solve» [Talbott (1995), p. 64].

fenómenos cotidianos. Así, aunque cree que son necesarias algunas divisiones inocentes en la vida mental, considera su teoría acerca del autoengaño como anti-divisionista. [Talbott (1995), p. 29]

Habría cuatro tipos de divisiones necesarias, todas ellas "inocentes":

1) Entre dos tipos de recuerdos. Aquellos que son accesibles en un momento o contexto particular y los que no.

2) Entre el yo intencional (identificado con la *persona*) y los mecanismos o procesos *sub-intencionales o subpersonales*. El sujeto quizá inicia estos procesos, pero no juega otro papel más allá. Talbott pone con el fin de aclarar el asunto el siguiente ejemplo: imagina que metes un comando para buscar un archivo en el ordenador. Inicias este proceso, pero no eres tú quien busca, ni sabes qué archivos son escaneados; únicamente recibes el resultado al final.[62]

3) Entre estados mentales transparentes y no transparentes. La gente a veces no es capaz de catalogar bien una acción o emoción propia; su conocimiento de sus propios estados mentales y sus causas no es infalible.

[62] No hará falta señalar que este ejemplo es, de entrada, problemático, ya que el ordenador no es parte constitutiva nuestra; pero entonces, para este viaje no necesitamos estas alforjas: igualmente hay casos en los que yo doy una orden a otro sujeto (es decir, intencionalmente ejecuto una orden) y posteriormente no tengo capacidad para controlar las acciones de ese sujeto. Esto, evidentemente, sólo funciona de un modo claro bajo el supuesto de la dualidad de agentes (sea persona-ordenador como en el ejemplo de Talbott, o persona-persona, etc.). En resumidas cuentas: si es necesario introducir dos agentes (como parece que sucede en el ejemplo) la división no resulta ya inocente, y si es verdaderamente inocente y supone un sujeto unificado, o no se entiende. Lo que uno desearía de Talbott es que probase, como pretende, que esto sucede dentro de un yo unificado y coherente.

4) Entre estados mentales conscientes e inconscientes, es decir, aquellos de los que se tiene noticia frente a aquellos de los que se es ignorante.

Las divisiones que postulan los intencionalistas no son de este tipo; la razón que ha empujado a los intencionalistas a aceptar tales divisiones es el hecho de tratar el autoengaño bajo el modelo del engaño interpersonal o de *mentirse* a uno mismo, lo cual es un error, porque conduce inmediatamente a lo que se ha llamado "paradoja doxástica" [Mele (1987b)]; en este sentido, si alguien permite que el autoengaño pueda involucrar creencias contradictorias *p* y *no-p*, es casi inevitable que uno tenga que postular algún tipo de partición o división en el yo o los yoes de la vida mental con el objetivo de aislar las creencias contradictorias la una de la otra. [Talbott (1995), pp. 28-29]

> En mi enfoque, la creencia autoengañosa de que *p* no implica mentirse a uno mismo con respecto a *p*. En cambio, implica la adulteración de los propios procesos cognitivos para favorecer la creencia en *p*, debido a un deseo de creer que *p* independientemente de si *p* es cierta.[63]

La intencionalidad que Talbott considera necesaria en el autoengaño reside, por tanto, en la *adulteración intencional* de la evidencia. Con esto no quiere decir que la creencia de que *p* sea intencional, ya que Talbott está de acuerdo con Bernard Williams y William Alston en que la creencia no está bajo control voluntario directo. Lo que es intencional es la adulteración

[63] «On my account, self-deceptive belief that p does not involve lying to oneself about p. It instead involves intentionally biasing one's cognitive processes to favor belief in p, due to a desire to believe that p regardless of whether p is true.» [Talbott (1995), p. 30]

que ejerce en sus procesos cognitivos para favorecer la creencia de que *p*, *al margen de que p sea o no el caso.*

Sin embargo, el hecho de que sea un proceso intencional no quiere decir que nos demos cuenta de ello, pues según Talbott, muchos de nuestros estados mentales no son *transparentes* para nosotros, esto es, podemos estar equivocados sobre ellos. Talbott indica que la suposición equivocada de la transparencia de la conciencia es una de las razones que forzarían a un divisionismo fuerte (por ejemplo a Amélie Rorty, tal y como vimos en la sección anterior §III.1.3.3).

Así pues, pese a que podemos confiar en nuestros procesos cognitivos como orientadores a la verdad porque son el resultado de la selección evolutiva, el sujeto puede adulterarlos intencionalmente. Evidentemente, el éxito de esta tarea exige que el sujeto no sea consciente de la maniobra; y el hecho de que no todos nuestros procesos mentales sean accesibles y transparentes facilita este objetivo. [Talbott (1995), pp. 33-34]

El autoengaño es para Talbott una especie de inmunización doxástica. Del mismo modo que resulta racional que uno se ponga la vacuna correspondiente para inmunizarse frente al polio al margen de que crea que vaya a contraerlo o no, es racional para quien se autoengaña buscar una inmunización doxástica frente a *no-p* al margen de que *p* sea o no el caso. La diferencia radica en que uno no debe saber que su creencia es producto de la inmunización, lo cual implica que ha de inmunizarse frente al conocimiento de esa inmunización. Esto da lugar a una jerarquía de autoengaños por medio de creencias y deseos. [Talbott (1995), pp. 37, 43, 68]

ETC.

Nivel-3: **Deseo de no creer que *r*, independientemente de si *r* es verdad** (donde *r* = que uno no crea que *q* se debe al deseo de no creer que *q* independientemente de si *q* es verdad). Es producto del deseo de nivel-2 de no creer que *q,* independientemente de si *q* es verdad. Esto da lugar a:

Intención de adulterar los propios procesos cognitivos en favor de *no-r,* lo cual, si tiene éxito, produce o sostiene:

Creencia de que *no-r*.

Nivel-2: **Deseo de no creer que *q*, independientemente de si *q* es verdad** (donde *q* = que uno no crea que *p* se debe al deseo de no creer que *p* independientemente de si *p* es verdad). Es producto del deseo de nivel-1 de no creer que *p,* independientemente de si *p* es verdad. Esto da lugar a:

Intención de adulterar los propios procesos cognitivos en favor de *no-q,* lo cual, si tiene éxito, produce o sostiene:

Creencia de que *no-q*.

Nivel-1: **Deseo de no creer que *p*, independientemente de si *p* es verdad.** Esto da lugar a:

Intención de adulterar los propios procesos cognitivos en favor de *no-p,* lo cual, si tiene éxito, produce o sostiene:

Creencia de que *no-p*.

[Talbott (1995), p. 44]

Para Talbott, una de las características más sobresalientes del autoengaño es la resistencia emocional que muestra el sujeto a la evidencia que podría plantear alguna duda sobre la verdad de la creencia protegida

[Talbott (1995), p. 43; cf. Freud (1923), p. 19], y es esta resistencia emocional por parte del sujeto lo que nos inclina a atribuirle algún tipo de reconocimiento de que *p* es el caso, ya que, si realmente no reconociese que *p* es el caso, ¿por qué iba a presentar tal resistencia emocional? [cf. Freud (1923), p. 19][64]

El sujeto a veces obvia ciertas explicaciones, haciendo casar todos los datos con la creencia que protege. El hecho de que obvie otras explicaciones es lo que Talbott llama "selectividad de la explicación".

> Los seres humanos son y deben ser selectivos en sus explicaciones. Nadie, al tratar de explicar cierta evidencia, considera nunca toda posible explicación de la evidencia. El resultado es que, en cada caso particular, alguien que trata de explicar la evidencia dada tomará algún trasfondo de creencias como algo fijo, y formulará hipótesis explicativas que explicarían la evidencia, dado que esas creencias del trasfondo fijado son verdaderas.[65]

[64] Sobra decir que puede haber muchas razones por las que alguien muestre resistencia emocional frente a algo que niega. El mismo Talbott lo reconoce en una nota al pie en la que indica cómo Alfred Mele le ha brindado un ejemplo según el cual alguien le espeta a gritos a un individuo: "¡tu madre lleva botas de combate!". Que el sujeto muestre una resistencia emocional no implica que muestre que sabe que su madre es tal y tal, y que se autoengaña al negar el insulto. [Talbott (1995), nota 23, p. 45]. De hecho, Talbott cree que no hace falta suponer ningún tipo de reconocimiento a ningún nivel: basta con que el sujeto piense que podría ser el caso (que su pareja le es infiel, que no aprobará el curso, que perderá su trabajo, etc.), para que las desastrosas consecuencias imaginadas le hagan resistirse emocionalmente a tal conclusión.

[65] «[H]uman beings are and must be selective in their explanations. No one, in trying to explain some given evidence, ever considers every possible explanation of the evidence. The result is that, in any particular case, someone attempting to explain the given evidence will take some background beliefs as fixed, and formulate explanatory hypotheses that would explain the evidence, given that the fixed background beliefs are true» [Talbott (1995), p. 41].

Es la evidencia en conjunto la que conduce a la conclusión dolorosa o desagradable, pero el sujeto nunca la considera en conjunto; para cada pieza de evidencia disonante ofrece una explicación coherente por separado, por lo cual se queda en anécdota y lo olvida. Sin embargo, cuando todo fracasa, los recuerdos llegan de golpe a la mente y se dice: "¿cómo pude haber estado tan ciego?" [Talbott (1995), pp. 39-40]. Talbott tiene la honestidad de reconocer que para dar cuenta de esta situación hay muchas respuestas posibles [Talbott (1995), p. 40].

El autoengaño sería, por tanto, la adulteración intencional de los procesos cognitivos con vistas a la obtención de una creencia independientemente de que sea verdadera o no, y su posterior inmunización y aislamiento. El sujeto también aísla la razón y el modo por los que procede a la inmunización de la creencia.

Para explicar cómo podría desarrollarse este proceso sin incluir incoherencias que lo hagan conceptualmente imposible, Talbott pretende demostrar en primer lugar cómo podría darse el autoengaño en un *agente intencional idealmente coherente,* en el agente racional de la teoría de la decisión bayesiana, dado que éste cumpliría una serie de restricciones mucho mayor que la de un ser humano. Para ello, aúna las exigencias del agente bayesiano con las restricciones para la coherencia epistémica.

(1) Un agente bayesiano racional asigna grados de creencia a determinadas proposiciones en términos de la probabilidad subjetiva de que sean el caso, y grados de utilidad al describir estados de cosas. El agente racional trata de *maximizar la utilidad esperada* (*maximize the Expected Utility*) de creer tal o cual

proposición, tratando de ajustar la probabilidad y utilidad de que las alternativas que toma en cuenta sean el caso.

Suponiendo que damos valores entre 0 y 1 para expresar los grados de probabilidad de que una proposición sea verdadera (0 sería falsa, 1 verdadera) y que:

(a) Prob (*p*) = 1 - prob (*no-p*), y

(b) Si $C_A(p)$, entonces C_A [Prob(*p*) > 0,5],

se sigue que un agente bayesiano no podría creer a la vez que *p* y *no-p*.

(2) Coherencia epistémica: Si un sujeto cree que *p*, cree que cree que *p*, y cree además que el proceso o método por el que cree que *p* no es fiable, entonces deja de creer que *p*.

Talbott no está seguro de que los procesos cognitivos humanos estén gobernados por las condiciones de coherencia epistémica, ya que si éstos fueran siempre epistémicamente racionales, no sería posible el autoengaño. En todo caso cree que, aunque los humanos puedan violar este principio, generalmente se adaptan a él. [Talbott (1995), p. 49].

Además, Talbott asume que:

(3) Los procesos cognitivos son relativamente independientes. Aunque puede influir en estos procesos, uno no adquiere creencias a voluntad.

(4) Hay una relativa fiabilidad de los procesos cognitivos, como muestra el hecho de que en la mayoría de ocasiones estemos perfectamente adaptados al medio.

(5) El yo ha de poder interferir en los procesos cognitivos que procesan la evidencia para formarse sus creencias. También ha de poder interferir en las creencias acerca de los propios estados mentales (creencias, deseos, etc). De estas interferencias, Talbott cree que el sesgo en el acopio de evidencia y en la atención son los que más claramente están sujetos a control intencional.[66]

(6) No transparencia: el yo sólo podría tener éxito en la estrategia de engañarse si pudiera hacerlo sin darse cuenta. No se trata de que no haya un acceso especial a la información sobre las propias creencias, deseos, etc., sino de que *no es un acceso infalible*.

Por tanto, un *yo idealmente coherente* tendrá una *motivación* para interferir en los procesos de formación de creencias siempre que la *utilidad esperada* resultante de interferir sea mayor que la utilidad de no hacerlo: UE [I(p)] > UE [¬I(p)]. En tal caso, *intencionalmente* interfiere en los procesos, pero no es consciente de ello, no se percata y viola circunstancialmente el principio de coherencia epistémica; esto sólo puede suceder si se acepta la restricción de que el yo no sea transparente y pueda estar equivocado en la valoración de la fiabilidad de sus procesos cognitivos.

Con respecto a los seres humanos, Talbott es consciente de que no satisfacemos las fuertes condiciones de coherencia bayesiana. Sin embargo, si puede ser concebido y explicado el autoengaño para un sujeto con unas exigencias de coherencia tan altas, en el caso de los seres

[66] Más adelante veremos los problemas que comporta este tipo de tesis. Como adelanto piénsese, por ejemplo, en el famoso problema señalado por Elster [(1983), p. 214] ilustrado de este modo: "trata de no pensar en un oso blanco". Inmediata e inevitablemente la imagen de un oso blanco viene a nuestra mente. La desviación selectiva en la focalización de la atención no es una tarea sencilla cuando estamos tratando de evitar algo.

humanos será mucho más sencillo. Además, Talbott cree que la actividad intencional humana puede ser modelada por esta teoría; si bien no tiene mucho sentido pensar que un sujeto que se autoengaña esté realizando finos cálculos de utilidad esperada, sí podemos tratar de explicar sus acciones como un agente bayesiano que trata de maximizar la utilidad.

III.1.5 - El problema selectivo

La exposición de las ideas propuestas por José Luis Bermúdez acerca del autoengaño se hace necesaria, y se justifica, más por la cantidad de referencias que ha suscitado que por la originalidad de sus ideas. Ciertamente, Bermúdez es señalado en la literatura pertinente como un autor destacado del grupo intencionalista debido a que reta a las teorías no-intencionalistas en general, y a Alfred Mele en particular, a responder a un gran desafío conocido como "problema selectivo", aunque a nuestro juicio —y como veremos al final de la exposición— sus ideas tienen algo más que un aire de familia con las de Talbott.

Bermúdez representa al igual que Talbott una postura intencionalista anti-divisionista, y defiende el enfoque intencionalista porque cree que es capaz de salvar mejor los fenómenos que el enfoque no-intencionalista. De hecho, en la medida en que Bermúdez cree que el engaño interpersonal puede producirse sin intención, deja abierta la posibilidad de que los autores no-intencionalistas presenten teorías del autoengaño bajo el modelo de engaño *interpersonal no-intencional*. [Bermúdez (2000), n. 3, p. 315]

Sin embargo, Bermúdez opina que mientras las explicaciones intencionalistas son defendidas mediante inferencias a la mejor explicación, los enfoques anti-intencionalistas sólo argumentan de modo indirecto alegando incoherencias en las explicaciones intencionalistas [Bermúdez (2000), p. 309]. Estas acusaciones le parecen infundadas, ya que las dos principales paradojas comúnmente asociadas al autoengaño le parecen inocuas.

Con respecto a la paradoja estática (que surge porque el sujeto parecería abrazar creencias contradictorias), Bermúdez cree que no es convincente, puesto que el enfoque intencional simplemente exige que el sujeto tenga éxito en su intento de causarse determinada creencia. No se trata de que el sujeto intente creer algo falso; basta con que el sujeto intente creer algo, independientemente de que sea o no falso [Bermúdez (1997), p. 107; cf. Talbott (1995), pp. 37, 43, 68]. De hecho, ni siquiera es problemático que el sujeto tenga creencias contradictorias, ya que del mismo modo que un sujeto puede mantener *inferencialmente aisladas* dos creencias cualesquiera, esto es, un sujeto puede creer que *p* y creer que *q* en un mismo tiempo *t*, sin creer en *t* que *p* y *q*, Bermúdez sostiene que un sujeto podría creer que *p*, creer que *no-p*, y no creer que *p* y *no-p*. Además, considera plausible pensar que un sujeto que se autoengaña comience creyendo que *p* y acabe creyendo que *no-p* sin que haya un momento en que crea que *p* y *no-p*.[67]

[67] Como veremos más adelante, esto es duramente atacado por autores como David Kipp (1980).

Por otro lado, en respuesta a la paradoja dinámica (el sujeto ha de ser víctima y verdugo a la vez), Bermúdez argumenta —apoyándose en Freud— que hacer algo intencionalmente no implica hacerlo conscientemente [Bermúdez (1997), p. 108; (2000), p. 314], ya que en algunos casos hay intenciones inconscientes. En este sentido, aunque se apresura a subrayar que ni de lejos el autoengaño intencional involucra *siempre* intenciones inconscientes, sí afirma que en muchas ocasiones quien se autoengaña puede perder el contacto con la intención que dio origen al proceso de adulteración de los mecanismos para la adquisición de la creencia, sobre todo teniendo en cuenta que estos procesos suelen llevar largo tiempo. La acción se precipita por una intención consciente, pero esto no implica que mientras el sujeto lleva a cabo la acción tenga constantemente presente y consciente la intención que le dio lugar [Bermúdez (2000), p. 314].

Según Bermúdez, bajo un enfoque intencionalista hay tres maneras en las que un sujeto puede autoengañarse al adquirir una creencia. De esta forma, según el modo en que caractericemos los casos paradigmáticos, el autoengaño implicaría:

1) la intención de provocarse la adquisición de cierta creencia. *S* intenta provocarse la adquisición de la creencia de que *p*.

O bien,

2) la intención de provocarse la adquisición de una creencia falsa. *S* cree que *no-p* pero intenta provocarse la adquisición de la creencia de que *p*.

O bien,

> 3) mantener que una creencia es falsa y aún así la intención de provocarse la adquisición de esa creencia. *S* cree que *no-p* pero intenta provocarse la adquisición de la creencia falsa de que *p*. [Bermúdez (2000), p. 310]

Lo común a estos tres casos y lo que los convierte en autoengaño, es que el sujeto se forma una creencia en virtud de una intención, sabiendo además que no habría adquirido esa creencia en ausencia de esa intención: ha manipulado además los propios mecanismos de formación de creencias.[68] [Bermúdez (2000), p. 312]. Supongamos que me apetece dar un paseo esta tarde, pero sé que si voy, acabaré matando mosquitos, ya que no soporto que se posen sobre mí y me picoteen, y el lugar está lleno de ellos a esas horas. Sin embargo, me formo la intención de ir, voy y acabo matando mosquitos. ¿Los he matado intencionalmente? Según Bermúdez, es plausible que no sea así, dado que matar esos mosquitos no fue en ningún momento ni un objetivo, ni un medio para conseguir mi objetivo: dar un paseo.

Este asunto es importante, pues Bermúdez cree que lo mismo que sucede entre intenciones y acciones, puede aplicarse al autoengaño, al par intenciones-creencias: quizá yo tengo la intención de adquirir la creencia de que *p*, y quizá sepa que *p* es falsa, pero de ahí no se sigue

[68] Tanto en el autoengaño como en el pensamiento desiderativo el sujeto cree que *p* porque desea que *p* sea el caso. La diferencia entre ambos fenómenos consiste, según Bermúdez, en que en el autoengaño el sujeto desea además creer que *p*, y tiene la intención de formarse la creencia de que *p*. [Bermúdez (2000), p. 312]

necesariamente que yo tenga la intención de creer una falsedad. [Bermúdez (2000), p. 311]

Como vemos, en opinión de Bermúdez el autoengaño puede explicarse sin paradojas desde un enfoque intencionalista. Pero ¿por qué rechaza en principio un enfoque no intencionalista? Bermúdez cree que hay dos fuertes razones: el problema de la revisión y el problema selectivo.

Con respecto al primero, Bermúdez se pregunta por qué es concebible que entre dos individuos que tienen un deseo de que *p* con la misma fuerza, uno de ellos se negase a revisar su creencia frente a evidencia desfavorable mientras el otro no. Por supuesto, no puede ser una cuestión meramente motivacional en función del deseo, ya que *ex hypothesi*, ambos tienen el mismo deseo con la misma fuerza. Sólo la intención por parte de uno de ellos de abrazar la creencia placentera parece ser capaz explicar la diferencia.

Más grave es aún el problema selectivo: ¿por qué unas veces nuestros deseos dan lugar a autoengaño y otras no?

> El autoengaño es paradigmáticamente selectivo […] Hay situaciones de todo tipo en las que, pese a que deseamos fuertemente que sea el caso que *p*, no estamos de ningún modo sesgados en favor de la creencia de que *p*. ¿Cómo vamos a distinguir estos casos de las situaciones en las que nuestros deseos dan lugar a un sesgo motivacional? Llamo a esto el *problema selectivo*.[69] [70]

[69] «Self-deception is paradigmatically selective […] There are all sorts of situations I nwhich, however strongly we desire it to be the case that pm we are not in any way biased in favour of the belief that p. How are we to distinguish these

La respuesta de Mele a la objeción de que no siempre que lo intentamos conseguimos engañarnos —que veremos en §III.2.2.3 pero adelantamos ahora—, se basa en que los costes psicológicos que se seguirían de creer falsamente que algo es el caso, son mayores en unos sujetos que en otros. Por ejemplo, un sujeto A que cree falsamente que su esposa le es infiel, tendría unos costes emocionales mucho mayores que los que tendría su vecino B, con respecto a la creencia de que la esposa de A le es infiel.

Sin embargo, Bermúdez indica que, a diferencia del problema de la revisión, el problema selectivo no versa sobre el distinto modo en que afectan los deseos a los sesgos motivacionales de distintas personas

from situations in which our desires result in motivational bias? I will call this the selectivity problem.» [Bermúdez (2000), p. 317]

[70] Bermúdez es normalmente señalado como el autor del "problema selectivo" y parece además que, como él mismo dice, haya sido él quien ha acuñado esta expresión pero, en realidad, ya estaba en Talbott: «I believe that this selectivity of self-deception can only be adequately explained on the supposition that it involves intentional biasing» [Talbott (1995), p. 62]. De hecho, Bermúdez no presenta a nuestro juicio ninguna idea original relevante; es más, los cuatro pilares de su exposición están explícita y claramente expresados ya en Talbott: (a) el problema selectivo hace poco plausibles las teorías no-intencionales [Bermúdez (1997), p. 108; (2000), p. 317; cf. Talbott (1995), p. 62], (b) una explicación no-intencionalista no sería imposible; simplemente no hay ninguna adecuada por el momento [Bermúdez (2000), p. 315; cf. Talbott (1995), p. 38], (c) los intencionalistas que adoptan una postura divisionista se encuentran con problemas más difíciles de explicar que los que tratan de solventar [Bermúdez (1997), p. 107; cf. Talbott (1995), p. 64], y (d) la intencionalidad del autoengaño no implica adquisición a voluntad de una creencia que se sabe falsa, sino de una creencia al margen de su valor de verdad [Bermúdez (1997), p. 108; cf. Talbott (1995), p. 30]. Esta similitud nos parece algo más que un aire de familia; Bermúdez únicamente se separa de Talbott en que cree que el autoengaño sí puede incluir creencias contradictorias, siempre que sean inferencialmente aisladas.

[Bermúdez (2000), pp. 317-318)]; lo que indica esta objeción es que el mero deseo de que *p* no da lugar por sí sólo al sesgo cognitivo, siendo otras cosas iguales. Aunque mis motivaciones hagan que el umbral de aceptación sea más bajo y la evidencia disponible para mí sea marginal, sólo me autoengaño con respecto a una pequeña proporción de ellas. Bermúdez se pregunta por qué esas y no otras.

Evidentemente, los intencionalistas tienen una buena respuesta: es la intención la que produce la selección, ya que el sujeto no trata de engañarse con respecto a todo. Aun así, Bermúdez no cree que sea imposible que las teorías no-intencionalistas consigan desarrollar una respuesta a esta objeción. [Bermúdez (2000), p. 318] De hecho, Mele criticará duramente esta objeción devolviendo la pelota al tejado intencionalista. Veremos esto en la sección §III.2.2.3. Por otro lado, Annette Barnes ha indicado que parece tanto posible como plausible pensar que existen otras disposiciones universales que, en muchas circunstancias, anulan la disposición universal para estar predispuesto a favor de las creencias reductoras de angustia [Barnes (1997), p. 81]. También le pretaremos atención más cuidadosa a esto en §III.2.2.2.

III.2 - Enfoques no-intencionalistas

Las teorías que acabamos de analizar suponen una muestra de la exigencia intencionalista compartida por la mayoría de las propuestas explicativas del fenómeno, aunque con unas u otras modulaciones; frente a ellas se sitúa otro grupo de teorías más o menos homogéneo que podríamos subsumir bajo la denominación de *enfoques no-intencionalistas.* Evidentemente, el principal elemento común y distintivo por el que las hemos clasificado bajo este rótulo, es su rechazo a exigir *intención* por parte del sujeto como elemento esencial, bien en los procesos de adulteración y sesgo de evidencia, bien en la adopción de creencias placenteras ante evidencia desfavorable. A menudo se ha señalado que estas explicaciones parecen ser, antes que una propuesta original atendiendo a la mejor explicación, más bien el fruto de la insatisfacción ante los pobres resultados obtenidos por las distintas versiones intencionalistas en la resolución de las dificultades y paradojas.

Aunque pueda parecer algo obvio, la inconsciencia por parte del sujeto de que están produciéndose en él ciertos procesos al evaluar la evidencia o al adquirir o retener una creencia no debe confundirse con la ausencia de intención; ésta es la razón por la que las teorías de Audi, Pears o Bach no se consideran *no-intencionales,* pese a que exijan que el sujeto haya de ser *inconsciente* de ciertas cosas al tratar de autoengañarse.

Las teorías a las que hacemos referencia en el presente apartado no son de este tipo. Más bien, la estrategia que siguen es la de suavizar las fuertes

exigencias cognitivas de los intencionalistas ya que, según estos autores, ninguno de los enfoques intencionales consigue sortear las paradojas que planteaba la concepción del autoengaño bajo el modelo del engaño interpersonal. Incluso aquellos intencionalistas que negaban la necesidad de postular que el sujeto adopte creencias contradictorias (del modo que sea) con el objetivo de salvar la paradoja estática o doxástica, no logran explicar cómo es posible que la intención de engañarme no acabe por autodestruir su propio proyecto, es decir, no salvan la paradoja dinámica o de la estrategia.

Entre las teorías no-intencionales vamos a ver las de Thomas S. Champlin, Mark Johnston, Annette Barnes y Alfred Mele. Mientras Thomas Champlin propone una explicación de sabor existencialista que plantea el autoengaño como una conducta *deshonesta no deliberada* que enmascara la realidad y nos hace caer en el error, Mark Johnston y Annette Barnes sostienen, en una línea muy freudiana, que el autoengaño no es intencional, sino sub-intencional, en tanto que comporta un *propósito* de reducir la angustia producida por una creencia o la imaginación de un determinado estado de cosas desagradable. Presentan sin embargo diferencias notables que veremos a continuación. Finalmente, Alfred Mele es el referente de la postura no intencionalista o, como él gusta de decir, "deflacionaria". Según Mele, el autoengaño sería un fenómeno *motivado*, pero no necesariamente —aunque puede serlo— intencional. El sujeto, debido al dolor que le produce una cierta creencia, está motivado para adulterar la evidencia de determinados modos, y los niveles de escrutinio de evidencia y umbrales de aceptación de creencias varían también de una

forma no consciente y, por supuesto, no intencional, haciendo mucho más fácil para el individuo abrazar la creencia falsa pero placentera.

Veamos pormenorizadamente qué propone cada uno y cómo responden a las dificultades más importantes señaladas por los intencionalistas.

III.2.1 - Conductas deshonestas para con uno mismo

Uno de los primeros autores que rechaza el modelo intencional de autoengaño, es Thomas Stephen Champlin, quien opina que ha sido un gran error entender el autoengaño en términos de engaño intencional interpersonal, ya que este modelo requiere el diseño de una estrategia mediante la que inducirse una creencia que uno sabe que es falsa, lo cual supone una imposibilidad cuasi-lógica [Champlin (1994), p. 55]. En realidad lo que no es posible es que el sujeto sea tan hábil como para llevar a cabo este tipo de tretas engañadoras dirigidas contra sí mismo; esta incapacidad no se debe a falta de pericia, sino a esta imposibilidad cuasi-lógica consistente en que el sujeto crea a la vez que *p* es cierto y es falso.

Ya en un artículo de ya 1976, Champlin subrayó que a diferencia de otras actitudes, engañar tiene un sentido doble: decimos que un hombre engaña a su mujer, decimos que engaña al portero al lanzar un penalti y decimos que engaña a otro al hacerle creer algo falso. Quizá esa mujer engañada está al corriente de todo; es irrelevante. Lo que queremos decir es que su marido no le es fiel. Cuando engañamos a alguien para que crea algo, como trato de engañar al portero *para* que crea que lanzaré el penalti

hacia la derecha, puede haber engaño incluso aunque finalmente mi disparo sea defectuoso y vaya efectivamente a la derecha. Sin embargo, si le engaño *en* creer que va a la derecha, entonces su creencia ha de ser falsa. Hay, no obstante, otro modo en el que puedo engañar a alguien; a veces, sin querer, le doy una a alguien información que considero verdadera pero resulta ser falsa; como consecuencia de ello, le hago caer en el error. Champlin cree que en esto consistiría precisamente el engaño interpersonal no intencional y opina que construir la explicación del autoengaño sobre este modelo de engaño interpersonal no-intencional evitaría las paradojas asociadas a la idea de que el sujeto intente creer algo que sabe falso. [Champlin (1994), p. 55]

Champlin considera que el autoengaño implica un proceso no-intencional por el cual el sujeto se induce al error, a tomar por verdadero aquello que no lo es. La creencia se la induce a sí mismo y de un modo inconsciente. Sin embargo, hemos de tener cuidado —nos advierte Champlin— de no equivocarnos, pues hay veces en las que el sujeto cae en el error de modo inconsciente y no estamos ante un caso de autoengaño. Esto sucede cuando tenemos un error de cálculo o un error producido por nuestra percepción (como cuando creo ver a alguien que resulta ser otro, creo oír algo, ver un oasis...). Otras veces, es el propio sujeto el que se induce a error sin querer: un militar experto en camuflaje disfraza intencionalmente un arma para que parezca un árbol y engañar así a los reclutas que están a su mando; sin embargo, posteriormente él mismo no distingue el arma y cae también en la trampa. Su acción al camuflar el arma es intencional bajo la descripción "engañar a los

reclutas", pero no es intencional bajo la descripción "engañarme a mí mismo". Por tanto, aunque efectivamente podemos decir que el sujeto se engaña, pues su acción ha sido la causante de su propio engaño, no se ha autoengañado: para que haya autoengaño es necesario algo más que una equivocación autoinducida [Champlin (1988), p. 12; (1994), p. 57; cf. Barnes (1997), p. 127][71]

Lo que distingue el autoengaño de estos errores perceptivos (errores que nos llevan a formarnos creencias equivocadas en virtud de esa evidencia defectuosa) es que en el autoengaño hay una conducta *deshonesta* por parte del sujeto. Esta conducta deshonesta no es deliberada —dice Champlin—, pero sí enmascara la realidad. Por tanto, el autoengaño es para Champlin una conducta no deliberada pero deshonesta que nos lleva inducirnos una creencia errónea y tomar por realidad aquello que es mera apariencia, y por tanto engañarnos [Champlin (1994), p. 57].

III.2.2 - Fenómeno motivacional

III.2.2.1 - Pensamiento desiderativo y represión

Uno de los autores más críticos con la concepción intencional del autoengaño ha sido Mark Johnston. Según él, toda caracterización del autoengaño que describa al sujeto como engañador y víctima del engaño al mismo tiempo, como un sujeto que sabe y no sabe, etc., constituye

[71] Barnes también considera que este ejemplo no constituye autoengaño, ya que la creencia del sujeto no es una respuesta a ninguna angustia, ni cumple *a fortiori*, la función de reducirla. [Barnes (1997), p. 127].

simplemente una contradicción en *nuestra* descripción del estado mental de ese sujeto. [Johnston (1988), p. 63]

No obstante, Johnston acepta que es concebible que un sujeto planee el olvido de algo. De este modo, se sabe que uno puede tomar ciertas pastillas que causan amnesia retroactiva [Johnston (1988), p. 76]. Sin embargo, aunque este tipo de proyectos son ciertamente intencionales, no merecerían tanto el nombre de autoengaño (*self-deception*) cuanto el de engañarse a uno mismo (*deceiving oneself*). [Johnston (1988), p. 78; cf. Audi (1982), p. 143]

En su opinión, las distintas paradojas asociadas al autoengaño surgen porque algunos teóricos como Jean-Paul Sartre, Bernard Williams o Donald Davidson, entre otros, tienden a sobre-racionalizar procesos mentales que, aunque tienen un propósito, no son intencionales. Así, critica duramente tanto el enfoque David Pears, según el cual un sistema protector actúa como un mentiroso paternalista que induce creencias intencionalmente para proteger al sistema principal [Johnston (1988), p. 79], como el de Davidson, que cataloga como irracionales aquellos eventos mentales que, según él, presentarían una conexión causal, pero no racional. Para Johnston el problema radica en la constricción holista de la conducta: dado que su teoría interpretativa presupone que la racionalidad es una característica tanto constitutiva como exhaustiva de los fenómenos mentales, si la conexión no es racional, entonces no pueden ser eventos mentales en absoluto. De hecho, Johnston subraya que Davidson admite que el autoengaño y el pensamiento desiderativo suponen *prima facie* serios

contraejemplos a su teoría interpretativa [Johnston (1988), p. 67]. Ya hemos visto cómo a causa de esta dificultad aparentemente insalvable, Davidson se veía obligado a abrazar también la teoría de subsistemas.

La respuesta de la teoría de subsistemas a esta paradoja es, como sabemos, la siguiente: distintos subsistemas desempeñan los diferentes papeles de engañador y engañado, así que no es necesario suponer un único sujeto que crea y no crea, o sepa y no sepa a la vez. [Johnston (1988), p. 63] Estas tesis divisionistas, que Johnston clasifica como homuncularistas —en alusión a la falacia homuncular que cometerían— no sólo evitaban una contradicción en su descripción de cómo alguien puede autoengañarse, sino que también parecerían capaces de explicar, en términos de compartimentación mental, cómo es que quien se autoengaña podría creer proposiciones que son contradictorias [Johnston (1988), p. 64].

Sin embargo, Johnston sostiene que la explicación homuncularista simplemente reemplaza una descripción contradictoria por un sujeto lleno de puzles. ¿Cómo podría el subsistema engañador ejercer las capacidades para perpetrar el engaño? Supóngase que un sujeto está borracho, pero realmente llega a convencerse de que está bien y puede conducir el coche hasta su casa. Johnston se pregunta si el subsistema que engaña al sistema principal tiene más tolerancia al alcohol y el sistema engañado está demasiado borracho en el momento del engaño; o es que el subsistema engañador está especialmente activo o simplemente se activa cuando está borracho. Además, ¿por qué habría de estar el subsistema interesado en el

engaño? ¿Es por su propio bien o es que sabe también lo que le conviene al sistema engañado? [Johnston (1988), p. 64]

¿Cómo ocurre todo sin que el otro subsistema se percate de nada de algún modo? Porque si se percata, entonces sólo podría ser exitoso si hubiese una *connivencia* entre los dos subsistemas; esto, sin embargo, reintroduce el problema, ya que el subsistema engañado sería parcialmente agente y paciente del engaño. Johnston se pregunta entonces si para dar cuenta de este conocimiento parcial hemos de reconocer a su vez *dentro* del subsistema engañado otros dos subsistemas, uno engañador y otro engañado; en tal caso sólo habría dos opciones: o bien una cadena infinita de subsistemas, o bien un último elemento en el regreso, consistente en un "subsistema suficientemente estúpido" como para no percatarse de las estrategias de engaño y que no haya la necesidad de connivencia. [Johnston (1988), p. 65]

La consecuencia es que no se entiende en absoluto cómo la creencia protectora simplemente aparece de repente (*pops- up*), ni por qué el sistema principal habría de aceptar una creencia que no se corresponde ni con su input perceptivo ni con los datos de su memoria; quizá se argumente que el subsistema protector distrae, sugestiona o camela al sistema principal, pero esto deja abierta la cuestión de por qué el sistema principal es tan fácil de distraer, sugestionar y tan sumiso [Johnston (1988), p. 84].

La única respuesta que parece adecuada es rechazar tanto el homuncularismo como el enfoque intencional, y ensayar una teoría anti-intencionalista. Para Johnston, nuestra mente trabaja de tal modo que el

deseo angustioso de que *p* predispone a creer que *p* es el caso. [Johnston (1988), p. 85]. Este proceso sirve a algún interés de quien se autoengaña y por tanto es en cierto modo *sub-intencional*, pero no intencional. [Johnston (1988), p. 65] A este tipo de proceso sub-intencional lo denomina "tropismo mental".

Los tropismos mentales son acciones no intencionales que, sin embargo, tienen un propósito. Hay diferentes ejemplos en otros fenómenos cotidianos, como el paciente que presenta el síndrome de Korsakoff y, a fin de rellenar los huecos de su memoria, inventa historias que la hagan coherente. Esto no es algo intencional, sino sub-intencional, aunque tiene un propósito. Otro ejemplo es el de las "uvas amargas"[72]: un sujeto amolda sus deseos *ad hoc* a lo que consigue, de modo que reduce las posibilidades de frustración. Pero quizá aquel con el que estamos más familiarizados es el proceso automático de filtración de información del que hablamos en §II.2.2, y por el cual algunos elementos, de entre todo lo que captan nuestros sentidos, nos resultan más llamativos en función de nuestros intereses. Si bien es cierto que la eficacia del pensamiento o creencia desiderativos sólo es inteligible en términos de la conexión racional entre las actitudes implicadas, la propia génesis del fenómeno no está mediada por conexión racional alguna. Este tipo de actividades son

[72] Este fenómeno se denomina así en referencia a la famosa fábula conocida como "La zorra y las uvas". Como es de sobra conocido, tras varios intentos fracasados por alcanzar unas uvas que le parecen apetitosas, la zorra se rinde y afirma que no merecían la pena, pues están demasiado verdes, eso es, ajusta *ad hoc* sus deseos a lo que consigue. Para un estudio más profundo de este fenómeno, véase Elster (1983).

en cierto modo ciegas, aunque tengan un propósito, y es la presunta consecución de ese propósito lo que les conferiría su carácter racional. Oponiéndose a la hipótesis interpretativa davidsoniana, Johnston afirma que la racionalidad no es una característica ni constitutiva ni exhaustiva de los fenómenos mentales [Johnston (1988), p. 88-89]. Por el contrario, el proceso mental sub-intencional involucrado en el pensamiento desiderativo y el autoengaño son un ejemplo de regularidad mental no-accidental: un deseo angustioso de que *p* o, de modo más general, la angustia respecto de *p*, genera la creencia de que *p* [Johnston (1988), p. 66].

El resultado tanto del pensamiento desiderativo como del autoengaño consiste en una *creencia motivada*, esto es, adoptada conforme a lo que uno desea que sea el caso, en lugar de a la evidencia. La diferencia entre ambos fenómenos es que mientras el sujeto que presenta pensamiento desiderativo acepta una proposición sin poseer evidencia suficiente, quien se autoengaña la acepta *en contra de la evidencia*. El autoengaño es, por tanto, una *especie* de pensamiento desiderativo: es una creencia motivada en contra de evidencia desfavorable [Johnston (1988), p. 67].

> Debemos entender el pensamiento desiderativo, no como resultado efectivo o potencial del razonamiento práctico que ocurre en el inconsciente o en alguna parte homuncular de la mente, sino como el mecanismo mental o tropismo por el cual un deseo de que *p* y la angustia acompañante de que *no-p* fijan las condiciones para la respuesta gratificante (a causa de la reducción de angustia) de llegar a creer que *p*.[73]

[73] «[W]e should understand wishful thinking, not as the actual or potential outcome of practical reasoning occurring in the unconscious or in some

Johnston reconoce que hay algunas situaciones en las que el pensamiento desiderativo tiene un carácter intencional, a saber: aquellas en las que la convicción del sujeto de que algo es el caso hace más probable que así sea. El producto de este tipo de pensamiento desiderativo son las creencias conocidas como "creencias que se autocumplen". Por ejemplo, el éxito de un individuo que debe dar un salto importante para salvar un precipicio depende, en gran parte, de que calme su angustia convenciéndose de que lo conseguirá. Este tipo de creencia aparece ya en Blaise Pascal o William James —como veremos detenidamente más adelante—, y Johnston lo denomina "pensamiento positivo" (*positive thinking*) para distinguirlo del pensamiento desiderativo en general [Johnston (1988), p. 69]. Sin embargo, en la mayoría de los casos el pensamiento desiderativo no involucra pensamiento positivo y, en estos otros casos, ni la intención ni la voluntad juegan ningún papel. El tratamiento del pensamiento desiderativo —y por tanto del autoengaño para Johnston— como resultado *no intencional de un tropismo mental* supone el rechazo de todas las teorías que representan el fenómeno de modo intencional.

Johnston ya ha explicado por qué rechaza tanto la posesión de creencias contradictorias, como las descripciones intencionalistas y

homuncular part of the mind, but as a mental mechanism or tropism by which a desire that *p* and accompanying anxiety that *not-p* set the conditions for the rewarding (because anxiety-reducting) response of coming to believe that *p*.» [Johnston (1988), p. 73]

homunculares, pero se encuentra con otra dificultad: aunque en el pensamiento desiderativo no hay necesidad de suponer una división dentro del agente, en el autoengaño propiamente dicho, la actitud "perversa" de quien se autoengaña (según la cual adopta una creencia a pesar de que reconoce a algún nivel que la evidencia está en contra) *nos fuerza a suponer que ha de haber alguna división*, ya que no hay modo de entender cómo podría producirse una reducción de la angustia si la creencia que la reduce y la evidencia que genera la angustia son copresentes. [Johnston (1988), p. 75]

Ahora podemos entender por qué Johnston no habla de teorías divisionistas en general cuando lanza sus ataques; simplemente ataca su carácter homuncular, y por ello las denomina homuncularistas. Johnston defiende que el proceso de reducción de angustia *no* es intencional, pero cree que ha de haber alguna división que mantenga evidencia contraria y creencia separadas. Las estrategias para alejar la evidencia contraria serían la *racionalización* (reevaluación y reexplicación de la evidencia), la *evasión* (evitar pensar en el asunto) y la *sobrecompensación* (focalizar la atención en razones inventadas esperando a que brote el apoyo evidencial), pero ninguna de estas estrategias puede estar, de nuevo, en el sistema principal, ya que evitarían la reducción de angustia. Johnston encuentra ese elemento de división en la *represión* freudiana. [Johnston (1988), p. 75]

El escollo que aparece inmediatamente entonces es el siguiente: ahora, sea cual sea la entidad encargada de ejercer el papel censor de la represión, heredará los problemas que presentaban las hipótesis homunculares que

Johnston había criticado cuando se enfrentaba a las paradojas del autoengaño. Sea donde sea allí donde se ejerce la censura, ¿cómo es posible que sea consciente de aquello que debe reprimir sin ser consciente de ello? Esta dificultad es la misma que Sartre le señalaba al análisis freudiano, como vimos en §II.1.3 [Sartre (1943), p. 102]. Johnston cree, sin embargo, que ha de aplicársele el mismo remedio que a las paradojas del autoengaño: la verdadera dificultad sólo surge cuando interpretamos de modo sobre-racionalista la represión bajo el modelo intencional, pero desaparece si observamos un modelo sub-intencional, dirigido, con un propósito (*purposeful*). Ese propósito será de nuevo la reducción de angustia [Johnston (1988), p. 76]

Por tanto, Johnston contempla el autoengaño como una especie de pensamiento desiderativo en el que un sujeto está motivado, por su angustia de que no *p* sea el caso, a creer que *p* en contra de su evidencia. No hay intención de engañarse ni de abrazar una creencia falsa, aunque el propósito de las subsiguientes maniobras sea reducir esa angustia. Sin embargo, pese a que Johnston es no-intencionalista mantiene una postura divisionista con respecto a la mente, obligado por la necesidad de mantener alejadas la creencia reductora de angustia, la evidencia angustiosa y la creencia evidencial. La teoría de la represión freudiana será el apoyo teórico al que recurre para explicar esta división.

Johnston afirma que la superioridad del enfoque no-intencional no sólo se reflejaría en que es capaz de dar cuenta del fenómeno sin caer en incoherencias o acudir a nociones más farragosas aún, sino en que a la

teoría de subsistemas se le hace más difícil explicar la cobardía mental o epistémica de la que generalmente se acusa a quien se autoengaña. Efectivamente, pese a que la teoría de subsistemas contempla un elemento intencional que parecería ser la clave para la atribución de responsabilidad, el engañador es el subsistema protector, y *el sistema principal es sólo una víctima* que no tiene noticia de lo que ocurre en el subsistema y, por tanto, no puede en modo alguno evitar el engaño. Por el contrario, el enfoque no-intencional da cuenta mejor de este asunto al considerar al sujeto responsable de no ser capaz de hacer más esfuerzos por contener su angustia y encarar aquello que la produzca. [Johnston (1988), p. 85].

Por su parte, el mayor problema que se le ha señalado al enfoque no intencional de Johnston es que no consigue distinguir el autoengaño del pensamiento desiderativo. Evidentemente, esto no es un inconveniente para el propio Johsnton, sino una parte esencial de su postura; a continuación veremos otro enfoque similar e inspirado en el propio Johnston, pero que introduce modificaciones conceptuales importantes a fin de paliar tales objeciones: la teoría de la reducción de la angustia de Annette Barnes.

III.2.2.2 - Reducción de angustia y propósito funcional

Annette Barnes es autora de uno de los pocos libros dedicados por completo al autoengaño que no consiste en un compendio de artículos de distintos autores. Esto le da la posibilidad de ofrecernos un estudio amplio, vertebrado y con más matices.

Barnes propone su modelo no intencional porque cree que contemplar el autoengaño como un tipo de engaño interpersonal es descaminado, ya que no es necesario suponer que quien se autoengaña se cause una creencia contraria a la que en principio mantiene. Sin embargo, afirma que reducir el autoengaño a pensamiento desiderativo como hace Mark Johnston no le hace justicia al primero, pues el autoengaño no siempre toma la forma del deseo, como ocurre en el pensamiento desiderativo. [Barnes (1997), p. 54; cf. Davidson (1985), p. 109; (1993), p. 216; Mele (2001), pp. 4-5]

Nadie podría negar que podemos ser, y a veces somos, engañados por las apariencias; en estos casos el engaño se produce sin que haya nadie que intente generarlo, o incluso sin que nadie quiera que ocurra, pero Barnes cree que el autoengaño no puede ser concebido como un engaño por las apariencias. Tampoco consistiría en un engaño autoinducido, como ya han señalado Brian McLaughlin, Mark Johnston o Donald Davidson. Si el aislamiento de creencias que se produce en el autoengaño es intencional, ha de ser por medio de caminos indirectos, como dirigir la atención hacia la evidencia favorable a *p* o evitar la evidencia contra *no-p*. [Barnes (1997), p. 27] Tomar una píldora, la hipnosis etc., no son opciones realistas [Barnes (1997), nota 21, p. 27]

Algunos filósofos han mantenido, como hemos visto, que en los casos paradigmáticos autoengaño en creer que *p*, la persona ha de creer también que *no-p* (Davidson), o al menos reconocer que la evidencia está en contra de *p* (Johnston). Sin embargo, la explicación del autoengaño bajo el

modelo del engaño a otro resultaría menos dificultosa si en nuestra definición de engaño interpersonal, cuando *A* engaña intencionalmente a *B* para que crea que *p*, no fuese necesario que

(a) *A* sepa o crea acertadamente que *p* es falso, y

(b) *A* intencionalmente le haga creer a *B* que *p* es cierto,

pues, en ese caso, no surgiría la paradoja doxástica (que el individuo crea dos cosas cuyo contenido proposicional es obviamente contradictorio). [Barnes (1997), p. 5] De hecho, Barnes defiende que no son esenciales unas actitudes cognitivas tan fuertes para los casos centrales de autoengaño. [Barnes (1997), p. 40][74]

Como vimos, Davidson articula su teoría sobre la noción de debilidad de la justificación autoinducida; la irracionalidad se seguía del paso que permite por un lado obviar lo que recomendaba el requisito de evidencia total y, por otro, mantener creencias contradictorias sin contemplarlas la vez. Sin embargo, Scott-Kakures ya señaló que incluso si el aislamiento de creencias contradictorias es producto de ciertas acciones intencionales del sujeto, eso no demuestra que quien se autoengaña viole conscientemente sus patrones de racionalidad como quiere Davidson [Scott-Kakures (1996), pp. 41-42]. Barnes añade que si la división de la mente y el aislamiento

[74] Alfred Mele tampoco cree que sea necesario que quien se autoengaña reconozca que la totalidad de su evidencia favorece que *no-p*, ya que aunque dice que "la persona autoengañada, de nuevo a causa de su deseo, *cree* por lo general en contra su mejor evidencia", inmediatamente añade "o contra la mejor evidencia que hubiese tenido, o podría haber adquirido fácilmente, si no fuera por el deseo en cuestión" [Mele (1987a), p. 136]

consiguiente se asemejan al olvido (y Barnes no ve a qué otra cosa podrían parecerse) quien se autoengaña llegaría, en el momento de engañarse, a no ser consciente de la evidencia contraria, con lo que *no* rompería conscientemente sus patrones de racionalidad. [Barnes (1997), p. 30]

Por su parte, Talbott y Bermúdez solventaban la paradoja dinámica o de la estrategia mediante "intenciones inconscientes": no es necesario suponer que el sujeto sabe que trata de engañarse y a la vez no lo sabe, porque su intención de engañarse le es simplemente desconocida. Sin embargo, esta oscura noción es atacada por Barnes. Tal y como ella entiende "intencional", una acción intencional es aquella que se ejecuta por una razón que tiene el agente, una razón de la que el agente puede ser consciente no-inferencialmente, esto es, una razón que el agente puede reconocer directamente, sin necesidad de hacer inferencias [Barnes (1997), p. 88]. Talbott mantiene que existen intenciones o razones no reconocidas (como las reglas griceanas que todo el mundo sigue en las conversaciones sin ser consciente de ello) [Talbott (1995), p. 36]. Sin embargo, Barnes señala que quien se autoengaña nunca dice: "Sí, he intentado (deliberadamente) hacer eso", sino algo así como "Sí, esa adulteración debe haber sido lo que ha ocurrido". Las palabras "debe haber sido" reflejan la *aceptación de una conclusión* sobre la base de una *inferencia*. La razón por la que no se puede decir que el sujeto tenga intenciones no-reconocidas es simplemente porque no es capaz de explicitarlas de otro modo que no sea inferencialmente; sin embargo, toda intención ha de poder ser reconocida no-inferencialmente por el sujeto [Barnes (1997), pp. 93-94].

Talbott también afirmaba que, aunque no es imposible en principio que existan mecanismos no intencionales, la acción racional como mecanismo de maximización de la utilidad esperada (*maximizing expected utility*) es siempre intencional, y es más probable que sólo exista un tipo de mecanismo para estos fines. Sin embargo, Barnes observa que hay varios mecanismos psicológicos y fisiológicos no-intencionales que también maximizan la utilidad: el mecanismo que trae recuerdos útiles a la memoria o el de la digestión son ejemplos de ellos.

La postura de Mark Johnston le parece más interesante a Barnes. Johnston cree que la partición de la que habla Davidson es en principio innecesaria, ya que la necesidad de su postulado es únicamente el producto de una sobre-racionalización del problema; una vez que se comprende que *el hecho de que las acciones del sujeto sirvan a algún interés suyo no las convierte automáticamente en intencionales*, esto es, que los procesos implicados en el autoengaño no son intencionales aunque tengan un propósito (*purposive*), las paradojas desaparecen [Johnston (1988), p. 65]. Johnston sólo se ve obligado a suponer algún tipo de división para que la evidencia contraria no siga generando angustia: si la angustia de que *no-p* sea el caso simplemente diese lugar a la creencia de que *p*, la evidencia contraria continuaría generando angustia. El sujeto debe reprimirla, y la racionalización, focalización selectiva, evasión o sobrecompensación son modos en los que se produce la represión que el sujeto necesita.

Barnes considera que Johnston está en lo cierto tanto al establecer que uno no puede entender por qué la gente se engaña a sí misma sin asumir

que ganen algo al hacerlo, como al considerar que la ganancia siempre implica alguna reducción de la angustia que se había producido por la no satisfacción de sus deseos. Sin embargo, Johnston se equivocaría al sostener que la angustia que reduce la creencia autoengañosa de que *p* es siempre la angustia de que *no-p*. Así, Barnes complica la fórmula de una variable de Johnston introduciendo una segunda variable "*q*", donde *q* puede ser, pero no es necesario que sea, igual a *p*. La creencia que genera el autoengaño de que *p* puede reducir la angustia de que *no-q* porque el sujeto cree que si *p* entonces *q* (o, quizás, si *p* entonces probablemente *q*). Barnes opina que esta fórmula de dos variables puede dar cuenta de todos los casos de autoengaño. [Barnes (1997), pp. 35-36]

Por tanto, el motivo por el que el individuo sesga su evidencia es que tiene un *deseo angustioso* de que *q* sea el caso. Esto ocurre sólo en caso de que desee que *q* y esté angustiado porque *q* no sea el caso [Barnes (1997), p. 38]. Este proceso no es intencional: simplemente, la creencia producida por la adulteración de la evidencia tiene el propósito de reducir la angustia. Además, la creencia generada por el autoengaño no sólo causa una reducción de la angustia, sino que esa es su principal función.

Una creencia autoengañosa de que *p* cumple la función de reducir la angustia de que *no-q* cuando

(1) la creencia de que *p* es causada por el deseo angustioso de que *q* sea el caso, y

(2) el propósito de la ocurrencia de la creencia de que *p* es reducir la angustia de que *no-q*. [Barnes (1997), p. 59]

Tener una creencia autoengañosa tiene un propósito (*purpose*), pero no es la intención de quienes se autoengañan reducir su angustia de que *no-q* creyendo que *p*.

Una creencia autoengañosa es un efecto cuyo propósito consiste en alterar su causa. Este tipo de efectos no son extraños en la naturaleza: la alta temperatura corporal en los humanos causa sudoración; el propósito de la sudoración es reducir la temperatura corporal. Los niveles bajos de agua en los humanos causan sensación de sed; las sensaciones de sed cumplen la función de hacer que los organismos afectados realicen cosas que pongan remedio a esa condición. Así, aunque muchos filósofos parecen perplejos ante ellas, las relaciones causales que son funcionales son muy corrientes.

La gente tiene creencias autoengañosas porque reducen su angustia [Barnes (1997), p. 60]. Pero que una creencia reduzca la angustia no significa que satisfaga el deseo. Para que un deseo de que *q* sea satisfecho, *q* debe ser el caso y la persona que tiene tal deseo ha de ser consciente de que *q*. Sin embargo, la angustia de una persona de que *no-q* puede reducirse sea o no *q* el caso. Lo que afecta directamente a la angustia de uno no es aquello que sea verdad acerca de los objetos de su angustia, sino lo que uno *crea* cierto de ellos. Ésta es la diferencia entre que un deseo se satisfaga y la creencia de que un deseo ha sido satisfecho. Lo que reduce la angustia es la creencia de que un deseo ha sido satisfecho. [Barnes (1997), p. 62]

> Tener un deseo angustioso de que *q* es estar angustiado porque *no-q* y desear que *q*. Estar angustiado porque *no-q* es estar incierto sobre si *q* o *no-q*, y desear que *q*. No hay, además, ninguna razón para que la persona desee que *q*; desea que *q* por sí misma. Hay sin embargo por lo general alguna razón, *r*, que es la causa de la incertidumbre. La persona cree que si *r*, entonces *no-q* (o probable o posiblemente *no-q*).
> Dado que la función de la creencia autoengañosa de que *p* es reducir la angustia de que *no-q*, y uno está por lo general angustiado porque *no-q* debido a que uno cree que si *r* entonces *no-q*, una creencia autoengañosa de que *p* que le permita a uno creer que *no-r* iría dirigida a la fuente de la angustia. De otro modo, una creencia autoengañosa de que *p* que le permita a uno creer que *r* pero también creer que no es el caso que si *r* entonces *no-q*, iría dirigida también a la fuente de angustia. Vimos, además, que la reducción de angustia es posible fuera o no satisfecho el deseo de que *q*. [Barnes (1997), p. 67]

Un deseo angustioso de que *q* causa una creencia de que *p* haciendo que la persona esté predispuesta (sesgada, prejuiciada) para adquirir la creencia de que *p*. La persona, por supuesto, no reconoce que se esté produciendo un sesgo en su proceso de adquisición de creencias, sino que cree que la creencia de que *p* está justificada. [Barnes (1997), p. 76]

Hay, no cabe duda, frecuentes casos de gente que adquiere sin intención creencias falsas sin que éstos se consideren casos de autoengaño. Las emociones, por ejemplo, en ausencia de cualquier deseo angustioso, causan que la gente se haga creencias falsas; decimos sin embargo que esta gente se ha descaminado ella misma sin intención. Para hablar de autoengaño es necesario un *deseo angustioso* y una creencia que no sólo reduzca la angustia, sino cuya *función* sea esa. [Barnes (1997), pp. 125-126]

Con respecto al autoengaño, es generalmente reconocido que hay sesgo de la evidencia; la disputa versa sobre si el sesgo es intencional o no intencional, y acerca de cuál es la fuente de tal sesgo. Mientras Johnston cree que el origen del sesgo está en el deseo angustioso de que *p*, Mele cree que es el mero deseo de que *p*. Por su parte, Davidson mantiene que es el deseo por evitar un sentimiento, emoción o estado perturbador que ha generado la creencia de que *no-p*, y Talbott cree que es el deseo de creer que *p* al margen de que *p* sea o no el caso. Annette Barnes defiende, sin embargo, que la fuente de la adulteración de la evidencia es un deseo angustioso de que *q* (pudiendo ser *q*, aunque no necesariamente, igual a *p*). [Barnes (1997), p. 79]

La explicación no intencional que acude a deseos angustiosos ha sido atacada principalmente por dos motivos:

En primer lugar, Talbott ha argumentado que el principal obstáculo para un análisis no intencional de la adulteración en el autoengaño es que no puede explicar por qué los deseos angustiosos no siempre sesgan los procesos cognitivos. No obstante Barnes responde que aunque la gente que tiene deseos angustiosos está en cierto modo predispuesta para ser parcial, no siempre alcanza la creencia, pues parece tanto posible como plausible pensar que existen otras disposiciones universales que, en muchas circunstancias, anulan la disposición universal para estar predispuesto a favor de las creencias reductoras-de-angustia. La gente está, por ejemplo, universalmente dispuesta a actuar para protegerse sí misma frente a los peligros. Por ese motivo, si tengo un deseo angustioso de que

los frenos de mi coche estén en buen estado, mi deseo más general de protegerme de resultar herido en un accidente de coche anula mi inclinación por una creencia que reduzca esa ansiedad. [Barnes (1997), p. 81] Por otra parte, pese a que el sesgo funciona no-intencionalmente, el propósito de reducción de angustia asegura que el sesgo no sea ni aleatorio ni al azar. [Barnes (1997), pp. 95-97]

La segunda dificultad ha sido señalada por Martha Knight:

> […] lo que es problemático en cualquier caso de este tipo es que la especificación del deseo motivacional será casi siempre *post hoc* y sujeta a razonamiento circular. Más aún: cuando se ofrecen numerosos deseos plausibles, *post hoc*, para explicar el autoengaño, a menudo no hay un modo claro para determinar qué descripción de la motivación de la persona es más apropiada. [Knight (1988), pp. 182-183]

Barnes se defiende de esta crítica señalando que la dificultad para decidir entre diferentes deseos angustiosos plausibles, qué deseo o deseos angustiosos están operando, no es razón para creer que no hay *un* deseo angustioso operando [Barnes (1997), nota, 29, p. 47]. Barnes insiste en que no depende de su análisis del autoengaño que un deseo angustioso *particular* sea el correcto; presumiblemente otras creencias podrían reducir también la angustia de quien se autoengaña. [Barnes (1997), pp. 46-47] Así, el que una creencia particular reduzca la angustia en una persona depende su conjunto total de creencias, deseos, angustias, esperanzas, etc., así como de hechos del mundo. [Barnes (1997), pp. 47-48]

Otra característica fundamental del autoengaño para Annette Barnes, también subrayada por David Sanford, es que el autoengaño exige una

falsa conciencia por parte del sujeto, y que ésta supone esencialmente algún tipo de equivocación (*misapprehension*). Quien se autoengaña se equivoca al captar la estructura de sus actitudes y toma, como dice Sanford, "el tener una actitud como explicación de tener otra, cuando la verdadera explicación es alguna otra cosa" [Sanford (1988), p. 169]

Supongamos que Agatha es una glotona. Le encanta el helado y no deja pasar la más mínima ocasión para comerse uno. En determinada ocasión se le pregunta por qué quiere un helado y responde con sinceridad: "me apetece algo frío". Supongamos que la verdadera razón es su glotonería. En este caso hay una mala interpretación de los propios deseos. En primer lugar, hay una *sobre-estimación* del papel que juega su deseo de tomar algo frío. Sin negar que pueda apetecerle algo frío, la verdadera razón de que se tome el helado habría que buscarla en su glotonería. Pero además, el que nos dé esa explicación alternativa se debe a su *deseo angustioso* de tener una verdadera razón para tomarse el helado (no tener una explicación alternativa y aparecer como una glotona le produce angustia). Agatha *subestima* el papel que tiene ese deseo en la formación de su creencia. Hay, por tanto, sobre-estimación y subestimación de la fuerza motivacional de los propios deseos.

En otras ocasiones, simplemente hay subestimación del papel motivacional de la angustia: el sujeto cree que su creencia está generada por la evidencia. Mi perro se ha perdido y veo uno en mi jardín. Creo —equivocadamente— que es mi perro. Supongamos que subestimo cuánto contribuye mi deseo angustioso de que el perro sea el mío a la razón por la

que creo que es mi perro. Niego que mi deseo *angustioso* juegue ningún papel causal en la formación de mi creencia de que el perro de mi jardín es mi perro; creo que mi creencia está justificada, ya que creo ver clara y directamente que el perro es el mío. En este caso hay subestimación de la influencia de mis deseos en mi evidencia y creencia.

Como hemos visto, el sujeto que se autoengaña se equivoca al valorar la fuerza motivacional de sus deseos (angustiosos); en este sentido, la sobre-estimación puede aparecer en los casos de autoengaño, pero no es necesaria. La subestimación, por el contrario, es un elemento esencial [Barnes (1997), pp. 106-108]. Quien se autoengaña presenta una falsa conciencia que consiste en que no es capaz de reconocer hasta qué punto los deseos angustiosos influyen en la formación de su creencia autoengañosa [Barnes (1997), nota 48, p. 58].

En resumen, hay varios modos en los que un sujeto puede estar engañado. Si uno está engañado *en creer* que *p*, entonces:

1. La creencia que tiene de que *p,* es falsa, y
2. O bien

    (a) ha sido *intencionalmente* engañado para creer que *p* por otro o por él mismo, o

    (b) se ha autoengañado a sí mismo, donde autoengañarse tiene un propósito pero *no es intencional*, o

    (c) ha sido engañado por las apariencias, donde el engaño ni tiene un propósito ni es intencional.

Sin embargo, del mismo modo que es posible que la gente engañe a otros para que crean algo que finalmente resulta ser cierto, también es posible que la gente se engañe para creer algo que sea cierto. [Barnes (1997), p. 115]

Esto no significa que el proceso de engañarse, sea lo que sea lo que implique este proceso, no sea suficiente para el estado de autoengaño, sino que el proceso no es suficiente para *un estado particular* de autoengaño [Barnes (1997), p. 115]. Esto es, si uno se engaña para creer que *p*, y *p* es cierta, entonces uno no puede estar autoengañado *en creer que p,* aunque *en algún punto habrá de estar engañado.*

Un sujeto se autoengaña *para creer que p*, si y sólo si:

(1) Tiene un deseo angustioso de que *q* que causa que tenga un sesgo a favor de creencias que reduzcan su angustia de que *no-q*. Este sesgo o parcialidad en el pensar, juzgar, o razonar, le causa que crea que *p*.

(2) El propósito (*purpose*) de la creencia de que *p* es reducir la angustia de que *no-q*.

(3) No sesga o no es parcial intencionalmente.

(4) No consigue hacer una estimación suficientemente alta del rol que su deseo angustioso de que *q* juega en la adquisición de su creencia de que *p*. Cree que su creencia de que *p* está justificada. [Barnes (1997), p. 117]

En los casos en los que la persona no sólo se autoengaña *para creer* que *p* (*into believing that p*), sino que está además autoengañada *en creer* que *p* (*in*

believing that p), *p* ha de ser falsa, y el sujeto nunca puede estar justificado en creerlo.

> Es decir, alguien puede autoengañarse para creer que *p*, pero no puede estar autoengañado en creer que *p* si *p* es el caso. […] El estar autoengañado *para creer que p* es condición necesaria para estar autoengañado *en creer que p*, pero no es condición suficiente; estar autoengañado *para creer que p* es condición suficiente para estar engañado *en creer algo*. [Barnes (1997), pp. 118-119]

De este modo, cuando alguien se engaña y por ello llega a creer *algo falso*, está autoengañándose a sí mismo (*self-deceiving oneself*). Barnes no usa de modo pleonástico "autoengañarse a sí mismo", sino como una expresión técnica. El prefijo "auto" sirve para diferenciarlo del engaño *intencional* a uno mismo (e.g., la píldora que causa amnesia o la entrada falsa en el diario); el *autoengaño a uno mismo* es una especie de autoengaño que le deja a uno autoengañado. [Barnes (1997), p. 117]

Si me autoengaño a mí mismo *en* creer que *p*, entonces

(1) Tengo una creencia falsa de que cierta proposición (*p* o alguna otra cosa) es el caso.

(2) He llegado a mi creencia de que *p* de un modo que me convierte en un objetivo apropiado de crítica.

(3) Tengo un deseo angustioso de que *q* que causa que yo crea que *p*, y el propósito de mi creencia de que *p* es reducir mi angustia de que *no-q*.

(4) No consigo hacer una estimación suficiente de cuánto contribuye mi deseo angustioso de que *q* a la razón por la que creo que *p*. Creo equivocadamente que mi creencia de que *p* está justificada.

Barnes dedica también atención en su libro al aspecto moral del autoengaño: cree que "la mayoría de nosotros está de acuerdo en que hay algo inaceptable en el autoengaño, pero no se pone de acuerdo en qué". Esta falta de acuerdo se debe en parte a la falta de acuerdo en qué es el autoengaño. A quienes se autoengañan se les ha acusado de falsedad de corazón, insinceridad e hipocresía; sin embargo, todas estas actitudes requieren la intencionalidad que Barnes y algunos otros niegan. Barnes considera que la acusación más acertada es la de *cobardía epistémica*. En este sentido, a alguien que se autoengaña se le acusa, como señala Johnston, "de cobardía mental, de huir de la ansiedad o angustia, de fracaso al contener la angustia o falta de valentía epistémica" [Johnston (1988), p. 85].

Además se pregunta si todo (auto)engaño es *prima facie* moralmente malo. Ella cree que por lo general reconocemos un beneficio en las maniobras de quien se engaña en determinadas circunstancias. Sin embargo, cuando vemos que otra persona es capaz de enfrentarse a esa misma realidad directamente, sin evitar deshonestamente los hechos, su demostración de valentía epistémica (*epistemic bravery*) nos resulta más admirable [Barnes (1997), pp. 174-175]. El engaño y autoengaño tienen así un carácter epistémicamente inaceptable, pero no siempre son moralmente reprobables. ¿Qué podría hacerlo moralmente reprobable? Según Barnes, la mejor candidata sería la siguiente característica: traicionar la confianza que el otro pone en mí y lo que espera de mí; violar el derecho que el otro tiene a esperar mi honestidad. Barnes señala que el otro no siempre es merecedor de ese derecho a esperar mi honestidad; por ejemplo, en situaciones competitivas (el lanzamiento de un penalty o los

planes de marketing de una empresa) o en escenarios bélicos: el enemigo no evita violar el derecho del adversario a esperar honestidad informándole de los planes de ataque o defensa. Pero entonces, Barnes argumenta que dado que no siempre se viola un derecho, ni el engaño ni el autoengaño son siempre *prima facie* moralmente malos. [Barnes (1997), pp. 159-160]. No obstante, a mi juicio, para afirmar tal cosa, debería demostrar que el autoengaño supone algo similar a esas situaciones competitivas o bélicas.

Por otro lado, Barnes es consciente de que toma la noción de "sesgo motivado" como primitiva, esto es, no entra a discutir cuál es su naturaleza, sino que la da por sentada. Hay otras nociones primitivas interesantes que podrían explorarse, como la de propósito, o explicación funcional. Pero arguye que toda investigación tiene sus límites [Barnes (1997), pp. 119, 122]. Sin negar que sea necesario siempre autoimponerse un límite explicativo, nuestra impresión es que la exposición de Barnes depende de modo crucial de esas nociones, y por desgracia son bastante oscuras, sobre todo la de propósito como algo distinto a la intención.

III.2.2.3 - Un enfoque deflacionario

Si en los últimos años han ganado peso las explicaciones "no intencionalistas" que se alejan del modelo de engaño interpersonal, y describen el fenómeno como un proceso motivacional evitando así las paradojas ya mencionadas, buena parte del mérito, si no casi todo, lo tiene el que es sin duda alguna el más influyente de estos enfoques: el ofrecido por Alfred Mele.

Ya en su artículo de 1983, titulado precisamente 'Self-Deception', Alfred Mele señala que los intentos por *purgar* al autoengaño de aspectos paradójicos, han dado lugar irónicamente a explicaciones que han resultado más problemáticas de lo que de antemano cabría pensar. En este sentido, Alfred Mele tratará de explicar de modo no paradójico el fenómeno a lo largo de sus numerosas contribuciones al asunto y eludiendo nociones psicológicas de dudosa filiación, como "medias creencias", "múltiples yoes" o "saber en el propio corazón" [Mele (1983), p. 365].

Su obra *Self-Deception Unmasked* (2001), redactada a petición de Harry Frankfurt para la serie "Princeton Monographs in Philosophy" —una colección de estudios histórico-sistemáticos sobre diversos asuntos filosóficos—, supuso para él la posibilidad de recopilar y reestructurar las ideas que había ido planteando a lo largo de veinte años en su intento por clarificar el problema del autoengaño a través de diferentes publicaciones. Además, junto al de Annette Barnes[75], constituye uno de los pocos libros que, sin ser una colección de artículos de varios colaboradores, está destinado por completo al asunto del autoengaño. Esta característica junto con su actualidad (al estar redactado en 2001, recoge un amplio abanico de teorías previas) le confiere un gran interés para todo aquel que quiera

[75] No deja de ser curioso que de los cuatro libros dedicados por completo al problema conceptual del autoengaño —sin ser compendios—, dos de ellos pertenezcan precisamente a autores no intencionalistas, en clara minoría gremial. Los otros dos, pertenecen a un enfoque volitivo cercano a los no-intencionalistas, y el otro es escéptico. Los enfoques intencionalistas, en un principio mucho más numerosos, con mucha más fuerza y repercusión, han expuesto sus ideas bien en pequeños artículos, bien en capítulos de libros dedicados a problemas más generales.

acercarse a un tema tan controvertido como el del autoengaño, ofreciéndole una extensa panorámica que le ayudará a familiarizarse con la problemática, los principales argumentos y los conceptos asociados al tema. Por estas razones le dedicaremos especial atención.

Mele parte de la aclaración de que tal y como él ve el autoengaño, es una *noción explicativa*, no una verdad conceptual [Mele (2001), p. 10]; sin embargo, la interpretación del autoengaño bajo el modelo del engaño a otro o interpersonal, produce dos paradojas, una *estática* (*static puzzle*), consistente en que el sujeto ha de creer al mismo tiempo que *p* y *no-p* son el caso y otra *dinámica* (*dinamic puzzle*), según la cual para llegar a este estado ha de diseñar una estrategia mediante la que, a pesar de mantener la creencia —apoyada en la evidencia— de que *p* es el caso, se convence en contra de toda esa evidencia de que *no-p* es el caso.

En primer lugar, examina la posibilidad de demostración empírica de algún caso de autoengaño estricto (*strict self-deception*). Según Mele, el autoengaño estricto exigiría que el sujeto mantuviese al mismo tiempo la creencia de que *p* es el caso y la creencia de que *no-p* también lo es. Mele advierte que, pese a que *p* y *no-p* supone una contradicción lógica, esto no implica que si un individuo cree que *p* es el caso, sea imposible lógicamente que crea además que *no-p*. Ahora bien, a pesar de que sea lógicamente posible, Mele no encuentra casos empíricos (ni imaginarios que sean verosímiles) de autoengaño estricto y reta de nuevo —como ya hiciera anteriormente [Mele (1997)]— a que alguien le plantee un ejemplo de ello. Expone varios intentos de algunos autores, todos ellos casos

fácilmente refutables. Por supuesto, algunos casos como la hipnosis o trastornos como la doble personalidad están excluidos, así como los casos en los que un sujeto mantiene creencias contradictorias de cuya inconsistencia no es consciente, bien sea por incompetencia lógica bien por incapacidad de manejar a la vez tantos supuestos. Finalmente, Mele conviene que, aunque el autoengaño estricto no es conceptualmente imposible, los casos comunes de autoengaño sólo conllevan la adquisición (o retención) de una creencia falsa apoyada en una evidencia motivacionalmente sesgada.

Para dar cuenta de este proceso ha de introducir otros dos conceptos clave, el de *sesgo* (*bias*) y el de *agencia* (*agency*). Por un lado, para que le sujeto sea capaz de abrazar una creencia que es en principio contraevidencial, ha de sesgar de algún modo la evidencia; esta adulteración viene generada por algún motivo, esto es, es motivacional. Generalmente, la razón que lleva al sujeto a la adulteración de la evidencia disponible es el dolor o malestar insoportable que le produce el estado de cosas al que apunta la evidencia.

Nadie pone en duda que la adulteración sea motivacional, es decir, que hay un motivo para adulterar la evidencia; sin embargo, sí que hay desacuerdo con respecto a otro punto: ¿es realmente consciente el sujeto de que está sesgando la evidencia? ¿Es un proceso voluntario? Este aspecto es el que Mele denomina el problema de la *agencia* [Mele (2001), p. 13].

Peacocke y O'Shaughnessy plantean un problema adicional acerca de cómo establecer cuánto control ha de requerirse para poder hablar de "intención". Mele zanja este asunto estipulando que hablará de intención

de realizar algo siempre que haya un intento (*try*) por parte del sujeto. Sin embargo, hay otra controversia más profunda y decisiva:

Supongamos que un sujeto *S* quiere hacer *A*. Sin embargo, sabe que *B* es una consecuencia probable de *A*. Finalmente *S* hace intencionalmente *A*. Pero entonces, incluso cuando *S* ni intenta que *B* sea el caso, ni quiere *B*, ni como fin, ni como medio, ni como efecto colateral, ¿hace intencionalmente *B*? Pese a que destacados autores, como Michael Bratman [(1997), caps. 8-10] o Harman (1976) han defendido que sí, Mele afirma que no; según él, sólo si *S* intenta *B*, hace intencionalmente *B*. Por nuestra parte, estamos de acuerdo con Mele; supongamos que alguien nos chantajea y nos amenaza con que si continuamos realizando esta valiosa tesis bajo nuestra firma en lugar de cederle amablemente a él todo nuestro trabajo, nos robará durante todo el próximo semestre el *tupper* en el que traemos la comida cada día. Evidentemente, la intención de negarnos a ceder nuestro trabajo no supone una intención de que nos roben la comida.

Las implicaciones que parece tener esta postura son las siguientes: no es necesario que el sujeto tenga la intención directa de engañarse para que finalmente resulte engañado. Concretamente, hay tres tipos de acciones intencionales en la manipulación de evidencia:

1) Actividades no intencionales: focalizar NO intencionalmente sólo sobre ciertos datos.

2) Actividades intencionales: focalizar intencionalmente sólo sobre ciertos datos.

3) Actividades intencionales como parte del intento de engañarse a uno mismo.

Algunos creen que la tercera es parte esencial del autoengaño. Mele en cambio defiende que 3 no es necesaria; bastaría con 1 ó 2.

En este sentido, Mele critica un ejemplo que propone Christian Perring en defensa del autoengaño intencional. En este ejemplo un hombre, tras tener fuerte evidencia de infidelidad por parte de su mujer, trata de ocupar su mente en otras cosas.

> Sam ha estado casado y se ha divorciado en dos ocasiones anteriormente. Ambos divorcios fueron amargos [...] Tiene fuertes convicciones religiosas y no podría estar con Sally si supiese que le ha sido infiel. Se dice a sí mismo: "No voy a pensar sobre esto". Enciende la televisión, bebe unas cervezas, y no piensa en la evidencia otra vez durante la noche, o de hecho, durante varias semanas, hasta que se encuentra cara a cara con ella de nuevo [...] Sam hace esto para mantener la calma y evitar el dolor que le produciría pensar en otro divorcio [...] En el momento en que toma la decisión de evitar evaluar la evidencia de la que dispone, la búsqueda de la verdad pasa a un segundo plano ante su necesidad de mantener equilibrio psicológico.[76]

Según Perring, aquí se muestra claramente cómo una conducta totalmente intencional conduce a Sam a un engaño, es decir, cómo se

[76]«Sam has been married and divorced twice before. Both divorces were bitter [...] He has strong religious beliefs and could not stray with Sally if he knew she were unfaithful. He says to himself, "I am not going to think about this". He turns the TV on, drinks a few beers, and does not think about the evidence again that evening, or indeed, for several weeks, until he is directly faced with it again [...] Sam does this to maintain his calmness and to avoid the pain of thinking about another divorce. [...] [A]t the time he makes the decision to avoid assessing the evidence he has, the search for truth takes second place to his need to maintain psychological equilibrium» [Perring (1997), p. 123].

autoengaña intencionalmente. Sin embargo, Mele advierte que aunque es cierto que la conducta de Sam es intencional, lo que intenta no es engañarse, sino dejar de pensar en asuntos desagradables; esto tiene como consecuencia un engaño, pero ésta no era la intención de Sam, y por tanto el autoengaño no es intencional. Sam hace intencionalmente *A* (dejar de pensar en la evidencia dolorosa) y esto tiene como efecto *B* (no tomar en cuenta las pruebas de que le es infiel y formarse la creencia, por tanto, de que no le es infiel), pero Sam no hace intencionalmente *B*.

En cualquier caso, Mele no niega totalmente que la intención pueda tener un papel en el autoengaño. Más bien su postura es que *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. [Mele (2001), pp. 16, 126, nota 12] Un caso, según el cual un individuo sumamente olvidadizo que se siente avergonzado por la penosa intervención que hizo en clase un día, procede a introducir una entrada en su diario alabando su magnífica actuación con el propósito voluntario y consciente de caer en su propio engaño cuando meses más tarde, habiendo olvidado ya el asunto, crea que ese día fue un héroe en clase, es ciertamente verosímil, pero sumamente extraño y alejado de los casos paradigmáticos de autoengaño.

Si obviamos este tipo de casos, Mele indica que la intención resulta contraria al proyecto de autoengaño. La razón es que las *intenciones directas* de autoengaño acabarían con el proyecto (el sujeto sería consciente de que trata de engañarse y estaría al corriente de todo el plan de engaño), y las

intenciones ocultas no resultan necesarias para explicar algunas actividades de adulteración. Por ejemplo, Bermúdez, Barnes, Martin y Talbott defienden la existencia de intenciones inconscientes. Según Mele, es posible que existan intenciones ocultas "freudianas" con efecto en nuestra conducta y, por tanto, que sean un posible motor del autoengaño; pensar en ellas *no* es incoherente. Pero dado que no tenemos justificación empírica ni necesidad conceptual de hacer uso de ellas para dar cuenta del fenómeno, *sí* es injustificado como explicación de los casos estándar de autoengaño [Mele (2001), p. 17].

Donald Gorassini ha defendido vehementemente otra modalidad de autoengaño intencional como un fenómeno muy cotidiano. Según Gorassini, en muchas ocasiones deseamos que algo sea el caso, actuamos "como si" lo fuese, y finalmente llegamos a la conclusión de que esto es así. Por ejemplo, un hombre que carece de amabilidad pero desea ser amable, podría decidir actuar como si lo fuese, llegando a convencerse de que es amable.

Esta idea recuerda a la famosa divisa Aristotélica, "el hábito engendra la virtud" [*EN* 1103a 31-1103b 2]. Es por medio de la enseñanza (virtudes dianoéticas) y de la costumbre (virtudes éticas) por lo que nos hacemos virtuosos [*EN* 1103a 16]; pero dado que podemos conseguirlo por la costumbre, no somos virtuosos por naturaleza, en tanto que lo que es por naturaleza no cambia con la fuerza de la costumbre, como la piedra que se mueve hacia abajo no puede moverse hacia arriba, aunque se la intentara acostumbrar la lanzándola hacia arriba mil veces. [*EN* 1103a 20-22] Nos

hacemos justos y templados practicando la justicia y la templanza [*EN* 1105a 17], nos hacemos valientes despreciando los peligros y una vez que somos valientes, afrontamos con valentía nuevos peligros. [*EN* 1104b 1-4]

Pascal, de un modo similar, ante la incapacidad de aceptar la creencia en Dios sobre la base de argumentos racionales de tipo probabilístico, recomendaba hacer "todo como si creyeran, tomando agua bendita, haciendo decir misas, etc. Naturalmente esto os hará creer y os embrutecerá" [Pascal (1670), p. 127].

La respuesta de Mele sigue la misma línea, pero se resiste a aceptar la conclusión de que la intención afecte al engaño:

> [A] partir de los hechos de que estos agentes deseen que *p* sea el caso, actúen intencionalmente como si *p* debido en buena medida a su deseo de que *p* sea el caso, y lleguen a creer que *p* en gran parte como consecuencia de su conducta intencional, no se sigue que estén intentando engañarse a sí mismos para creer que *p* o intentado hacerse más sencillo el creer que *p*.[77]

En efecto, que una acción intencional tenga como efecto algo, no quiere decir que hayamos hecho ese algo intencionalmente. En este caso, autoengañarnos.

Aunque Mele no lo cita, esta misma argumentación estaba ya mucho antes en Frederick Siegler:

---

[77] «[F]rom the facts that theses agents want it to be true that *p*, intentionally act as if *p* owing significantly to their wanting *p* to be true, and come to believe that *p* largely as a consequence of that intentional behavior, it does not follow that they ware trying to deceive themselves into believe that *p* or trying to make it easier for themselves to believe that *p*.» [Mele (2001), p. 20]

> Supóngase que A cree que p en t^1, y desea que p sea falso. En t^2 comienza a considerar ciertos hechos (H) como evidencia (E) de que p es falso. En t^3 cree que p es falso, y sobre la base de los H que tomó como E. Ahora bien, podría ser que si A no hubiese deseado que p fuese falso, no hubiese tomado H como E. Pero de aquí no se sigue que A intentase hacerse creer que p, que lo que creía cierto, fuese falso. Ni se sigue que el hecho de que tome H como E sea evidencia de que A intentase tal cosa. [...] La razón puede ser que A deseara tanto esa evidencia, y estuviese tan angustiado por encontrarla, que a causa de ello tomó H como E cuando no lo era.[78]

Sin embargo, aunque estamos de acuerdo con Alfred Mele en esta crítica, creemos que hay otra más profunda. Al igual que sucede con algunos ejemplos aristotélicos, nuestra respuesta al desafío de Gorassini consiste en lo siguiente: muchas virtudes o características no dependen de otra cosa que de ser efectivamente practicadas. Es decir, un sujeto no se está engañando al creer que es amable porque actúa de modo amable, ya que ser amable no es un rasgo esencial, natural, sino una conducta. Actuar como si se fuese amable ya le convierte a uno en amable. Otro asunto es el de la honestidad o la hipocresía: uno puede ser amable de modo hipócrita, pero si se comporta de modo amable con la gente, por mucha

[78] «Suppose A believes p at t^1, and wishes that p were false. At t^2 he begins to regard certain facts (F) as evidence (E) that p is false. At t^3 he believes that p is false, and on grounds of F which he takes as E. Now it may be that if A had not wished that p were false he would not have taken F as E. But from this it does not follow that A intended to get himself to believe that p, which he believed to be true, was false. Nor does it follow that his taking F as E is evidence that A intended such a thing. [...] The reason may be that he wished so much for evidence, and was anxious to find it and consequently took F as E when it wasn't.» [Siegler (1963), pp. 34-35]

hipocresía que encierren sus actos, no deja de ser amable. Esta persona será amable e hipócrita o falsa.

Hemos visto por tanto el problema que Mele señala en las "teorías de la agencia" o intencionalistas: el individuo se hace consciente de que la evidencia adulterada es ilegítima. Pero la interpretación antiagencial tiene su propio problema: si no hay intención de engaño ulterior, no se ve de forma clara qué mecanismo empuja al sesgo o adulteración de la evidencia.

El punto de partida de Mele son los experimentos en psicología social, los cuales nos mostrarían que adulteramos motivacionalmente nuestras creencias en función de deseos [Brown y Dutton (1995), p. 1290; Kunda (1990), p. 483]. Sin embargo, en una línea similar a la de Mele, Jeffrey Foss ya había indicado que los deseos no tienen por sí mismos fuerza explicativa; necesitan de creencias que identifiquen cuáles han de ser los medios para satisfacer el deseo: esa es la lógica del deseo-creencia [Foss (1997), p. 112]. Mele admite que en algunas ocasiones es difícil encontrar creencias que cumplan tal función mediadora para algunos ejemplos, como el caso en el que se realiza una encuesta en la que el 25% de los universitarios cree estar entre el top 1% (cosa que, evidentemente, pueden pensar, pero es imposible). Sin embargo, nos parece que esta dificultad es dependiente, obviamente, de que el ejemplo de la encuesta constituya verdaderamente un caso de autoengaño. Esto no está muy claro, pues esa encuesta puede estar revelando un error perceptivo o evaluativo, más que un autoengaño colectivo. En cualquier caso, Mele se apoya en diversos

estudios sociológicos que muestran que los deseos pueden contribuir a la producción de creencias motivacionalmente adulteradas o sesgadas —incluyendo las "autoengañosas"—, de formas bien conocidas y que se ajustan al "modelo *anti*-agencial" [Mele (2001), p. 24]. Concretamente, pueden hacerlo de cuatro modos:

1) Mala interpretación negativa.
2) Mala interpretación positiva.
3) Focalización/atención selectiva.
4) Acopio selectivo de pruebas.

Mientras la *mala interpretación negativa* consiste en malinterpretar la evidencia de modo que no evaluamos como peso en contra de *p* aquello que, en ausencia de nuestro deseo de que *p* sea el caso, reconoceríamos fácilmente como contrario a la verdad de *p,* la *mala interpretación positiva* consiste en tomar como favorable a *p* aquello que en ausencia de nuestro deseo de que *p* sea el caso, evaluaríamos como contrario a la verdad de *p*. Por otro lado, la *focalización/atención selectiva* consiste en que debido a nuestro deseo de que *p* sea el caso, únicamente focalizamos nuestra atención sobre aquello que favorece la verdad de *p*, y el *acopio selectivo de pruebas* supone que, debido a nuestro deseo de que *p* sea el caso, pasamos por alto evidencia en contra de *p* fácilmente obtenible y en cambio buscamos con ahínco evidencia a favor de *p* mucho menos accesible [Mele (2001), pp. 26-27]. Mediante estos mecanismos, adulteramos la evidencia que más tarde sirve de apoyo para la creencia falsa, y por tanto, para el autoengaño.

Pero si el autoengaño no consiste en el *intento* exitoso de engaño a uno mismo, en la *autoinducción voluntaria* de creencias que sabemos (o creemos) falsas, ¿en qué consiste entonces? Según la teoría deflacionaria de Alfred Mele, las condiciones que en conjunto resultan suficientes para el autoengaño son las siguientes:

(1) La creencia que *S* adquiere de que *p* es el caso, es falsa.

(2) S trata los datos relevantes (o aparentemente relevantes) para la verdad de que *p* es el caso, de modo sesgado.

(3) La adulteración de dichos datos es una *causa no desviada* (*nondeviant cause*), de que *S* adquiera la creencia de que *p* es el caso.

(4) El *cuerpo de datos* que *S* posee en el momento en que se forma la creencia, da mayor justificación a que *no-p* sea el caso que a *p*.

Las condiciones de Mele no son una lista exhaustiva de condiciones necesarias, sino un conjunto de condiciones suficiente para hablar de autoengaño [Mele (2001), pp. 51, 120]. *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 (2001), p. 4; (1997), p. 91].

En primer lugar, la creencia ha de ser falsa, pues en caso contrario no podemos hablar de engaño. O al menos, el sujeto puede autoengañarse *para creer* que *p*, pero no está autoengañado *en creer* que *p*.

Además, los datos que el sujeto toma para formarse la creencia han de ser sesgados por el propio individuo. 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 [Mele (2001), p. 31].

La teoría de Friedrich [*Primary error detection and minimization* (PEDMIN)] afirma que al evaluar hipótesis el sujeto está mucho más ocupado en tratar de detectar y minimizar la cantidad de errores que en buscar la verdad [Friedrich 1993, p. 299]. Esta tesis tiene un claro sabor popperiano, al postular que en el proceso de evaluación de hipótesis no tratamos de verificar distintas propuestas, sino de falsarlas en busca de una hipótesis crecientemente robusta. Esta estrategia suele ser inconsciente, e incluso quizá la selección natural ejerce presión de modo que favorece el desarrollo de estrategias de evaluación automáticas. Según Friedrich, otra aplicación de PEDMIN relevante es la evaluación de hipótesis que tienen que ver con uno mismo, la autoimagen y la autoestima. [Friedrich (1993) p. 314; cf. Lacan (1966/84), p. 87].

Por su parte, 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 asociada. Por supuesto, el umbral para aceptar la creencia que se adapta a los deseos del sujeto es mucho más bajo que el que es necesario superar para forjar una creencia que le resulte dolorosa. [Trope y Liberman (1996), p. 253]

Hay un umbral de confianza que se compone de dos límites o umbrales: uno de aceptación (cantidad de evidencia mínima favorable necesaria para aceptar la verdad de la hipótesis y abandonar el proceso evaluativo) y otro de rechazo (cantidad máxima de evidencia disconfirmante para la hipótesis que resulta tolerable antes de abandonar su evaluación). Estos umbrales no son fijos ni requieren la misma cantidad de evidencia, ya que dependen de los "costes asociados" a una posible aceptación o rechazo erróneos (costes emocionales, psicológicos, etc.) [Trope y Liberman (1996), p. 253]. Mele afirma que estos costes son subjetivos ya que los objetivos perseguidos por cada individuo son a menudo insospechados y sorprendentes [Mele (2001), p. 35; cf. Helm (1994)]. Así, si deseo que *p* sea el caso (y la creencia de que *no-p* es el caso resulta dolorosa para mí) esto puede producir dos efectos:

1) El umbral para confirmar y aceptar que *p* es el caso será mucho más bajo y alcanzable o satisfacible que para aceptar que *no-p* (y rechazar que *p*) lo es.

2) Los costes que se siguen de aceptar erróneamente que *p* es el caso son mucho menores de los que se siguen de aceptar erróneamente que *no-p* lo es. Por tanto, no sólo será más fácil satisfacer los requisitos evidenciales para aceptar que *p* es el caso, sino que además estamos más dispuestos de antemano a apostar por *p*, debido a que los costes potenciales de un error al creer que *p* son mucho menores que los costes potenciales que se siguen de abrazar equivocadamente la creencia de que *no-p* es el

> caso. Evitamos el error más costoso [Mele (2001), p. 36; véase también Trope y Lieberman (1996), p. 240] y nos es más fácil satisfacer la conclusión deseada [Trope y Lieberman (1996), p. 252].

Aunque éste sea el modelo básico FTL, no es necesario abrazarlo en todas sus consecuencias y, de este modo, una versión moderada de FTL nos puede ayudar a explicar o sustentar las 4 estrategias de sesgo de la evidencia mencionadas.

La tercera condición hace referencia a que esta adulteración ha de ser la *causa* de la formación de la creencia errónea o autoengañosa (*self-deceptive*). Sin embargo, Mele hace hincapié en que ha de ser causa no-desviada (*nondeviant*). Esta demanda se debe a que hay algunas dificultades acerca de la intencionalidad y la causalidad en la acción. Un ejemplo del que podemos valernos nos muestra como Vera, una muchacha ducha en el uso de armas de fuego, trata de alcanzar cierta diana. Para ello, hace un cálculo cuidadoso, se concentra y dispara. Sorprendentemente, sus cálculos son del todo desacertados, pero aún nos sorprendemos más cuando observamos cómo la bala golpea contra una pared, rebota sobre una cañería y acaba por alcanzar la diana. Por supuesto, Vera tenía la intención de dar a la diana y, de hecho, la ha dado. Sin embargo, no estaríamos dispuestos a decir que la ha dado voluntaria o intencionalmente. Este tipo de causas *desviadas* son excluidas del autoengaño tal y como lo concibe Alfred Mele, aunque quizá habría de entender un engaño así producido como "engaño causado a uno mismo"

frente a "autoengaño" [Audi (1982), p. 143; (1997), p. 104; cf. Elster (1979), pp. 292-293; (1983), p. 216; Johnston (1988), pp. 76-78].

Precisamente Robert Audi propuso en un artículo aparecido en el volumen que la revista *Behavioral and Brain Sciences* [vol. 20 (1), 1997] dedicó al autoengaño, un caso que presuntamente satisfacía las condiciones suficientes que Mele había establecido y sin embargo no era un caso de autoengaño. En dicho ejemplo, un hombre trata de dar explicación a un accidente de avión; como desea que se haya producido por un fallo mecánico antes que por la explosión de una bomba, sólo comenta el asunto con gente que cree precisamente eso, quedando la evidencia de la que dispone totalmente adulterada. Sin embargo, ese intento de hablar con gente que opina que se trata de un accidente, hace que comente el tema con Eva, quien logra convencerle de que se trata de un ataque terrorista, pese a que la evidencia de la que él disponía apoyaba más la hipótesis del error mecánico (hipótesis que, supongamos, es el caso). Audi indica que podría imaginarse además que los datos son complejos y que la dificultad en su evaluación es genuina.

Estaríamos ante un caso en el que, tras un proceso de adulteración motivacional, el sujeto obtiene una creencia en contra del mayor peso de la evidencia. Dado que el ejemplo cumple todos los requisitos que Mele exige, habría de considerarse como un caso genuino de autoengaño. Sin embargo, Audi se pregunta si no sería concebible interpretar esto como un error, incluso por parte de un individuo muy racional. "¿Habría de ser necesariamente el resultado de una adulteración?" [Audi (1997), p. 104].

Alfred Mele no está dispuesto a aceptar esta conclusión. Según Mele, es cierto que no habría autoengaño, pero esto se debe a que no cumple la cláusula 3, es decir, a que la causa es desviada. No es una causa directa porque el proceso de adulteración intencional no tiene como resultado la formación de una creencia errónea (del mismo modo que el disparo intencional de Vera no tiene como resultado el alcance de la diana del modo en que la intención de Vera había previsto) [Mele (1997), p. 131; (2001), pp. 122-123]. Dicho sea de paso, esta réplica ya había sido prevista por el propio Audi:

> Mele podría hacer notar que mi manejo motivacionalmente sesgado de la evidencia no es una causa directa de mi creencia falsa. Esto es cierto, pero el caso podría ser revisado con el fin de llegar al mismo resultado *directamente* (en algún sentido plausible de dirección), como donde un subconjunto propio de mis datos me convenciese del mismo modo en que lo hace el argumento de Eva.[79]

Mele no contesta a esta objeción. Sin embargo, no resulta difícil suponer por qué. Resulta irónico el sinfín de ocasiones en las que un autor inventa un ejemplo que él mismo cataloga en el mismo artículo como ineficaz, añadiendo con una tranquilidad pasmosa que "podría ser modificado" a fin de hacerlo lo suficientemente convincente, robusto o adecuado. Evidentemente, si es posible hacer tal cosa, podría haberla

[79] «Mele might note that my motivationally biased handling of evidence is not a direct cause of my false belief. That is true, but the case might be revised to yield the same result *directly* (for some plausible notion if directness), as where a proper subset of my data do the same convincing that Eva's argument does» [Audi (1997), p. 104].

hecho él mismo de antemano. Si no lo ha hecho, alguna dificultad verá en tal empresa.

En este caso concreto, la dificultad estriba en que, una vez que se ha aceptado que las cláusulas de Mele pueden ser suficientes para algún tipo de autoengaño, se hace muy difícil distinguir entre el error y el autoengaño. Así, toda creencia apoyada en evidencia sesgada motivacionalmente, será producto de un estado de autoengaño. Seguramente Audi es consciente de esta dificultad, y por ello exige posteriormente como condición necesaria que el sujeto muestre cierta *tensión* entre la afirmación sincera de que *p* y el conocimiento de que *no-p*. Esta demanda atañe ya a la cláusula 4).

Efectivamente algunos autores indican que la cláusula 4) tal y como es formulada por Mele resulta demasiado débil. Como ya hemos visto, algunos teóricos del autoengaño (por ejemplo el propio Audi, Davidson, Pears o Bach) indican que esa tensión es necesaria en el autoengaño. De este modo, exigen que no sólo el cuerpo de datos dé más apoyo a *no-p*, sino que el sujeto *reconozca este hecho*. Los autores que defienden esta tesis suelen defender una concepción intencional del autoengaño.

Baumaister y Cairns (1992), presentaron un caso de estudio en el que a algunas personas se les daban informes sobre su persona. Éstas no les prestaban atención cuando eran negativos, atendiendo mucho más tiempo a la información positiva. Posteriormente Baumaister y Newman (1994) observaron que si se les decía que el informe sería público, esta vez prestaban más atención a la información negativa. [Baumeister y Newman

(1994), p. 12] ¿Tenían la *intención* de engañarse? ¿Supone este caso un problema explicativo para FTL? Según Alfred Mele, no; más bien, habría que entender estos resultados en el sentido en que lo hacen Baumeister y Cairns: algunos sujetos consideran, en primer lugar, muy desagradables o dolorosos ciertos informes negativos y reaccionan no atendiendo a ellos, debido a que sienten una aversión a las críticas. Esto no quiere decir —aclara Mele— que esta aversión les lleve a intentar engañarse o hacerse más sencillo creer que lo que dicen los informes es falso [Mele (2001), p. 48]. Más bien lo que sucede es que si se les dice que los informes serán públicos, tratarán de rebatirlos, ya que el deseo de rebatir y refutar los posibles ataques públicos hacia su persona supera dicha aversión. [Mele (2001), p. 49]

Por otro lado Mele indica que, como hemos visto también anteriormente (§III.2.2.2), Annette Barnes cree que para poder adscribir autoengaño a un sujeto, es necesario que dicho sujeto tenga el propósito de reducir su angustia (*reduction of anxiety*). Quien se autoengaña muestra un deseo angustioso de que *p* sea el caso y está igualmente angustiado ante la idea de que *no-p* lo sea.

Barnes indica que su teoría, al igual que la de Mark Johnston, difiere de la de Mele precisamente en que mientras la fuerza motivacional recae en la teoría de Mele en un mero deseo, ellos creen que se trata de un *deseo angustioso*. Mele acepta que en algunos casos el deseo que siente el sujeto es angustioso, pero no cree que esto sea algo constitutivo del autoengaño. Dado que es posible que un sujeto se autoengañe con respecto a algo que

aún no ha generado en el sujeto ningún tipo de experiencia ni de sentimiento y, por tanto, no le angustie aún, esto no es una condición necesaria. [Mele (2001), p. 56].

En los casos extremos suele haber un salto entre la escasa evidencia en favor de aquello que deseamos que sea el caso y la enorme evidencia de aquello que no deseamos y nos causa malestar. Acudir a un proyecto *intencional* para abrazar la creencia con escaso apoyo nos hace caer de nuevo en el problema de la autoaniquilación del proyecto de engaño. Pero ¿y una intención inconsciente? ¿Sería más plausible que le explicación ofrecida por el modelo FTL? [Mele (2001), pp. 67-68]

Según Mele, la teoría FTL es capaz de explicar el autoengaño sin acudir a supuestos extravagantes o exóticos, incluso en casos tan extremos como el del marido tan crédulo que, aunque a una noche ve con sus propios ojos a hombre salir del cuarto de su mujer, ella le convence posteriormente de que lo ha imaginado todo. De acuerdo con una interpretación acorde con FTL el umbral para la hipótesis dolorosa (no deseada), es muy alto. Sin embargo, en un primer momento, la fuerte evidencia llega a sobrepasar este umbral, haciendo que el sujeto crea que su esposa le es infiel. Pero por otro lado, el umbral para aceptar que está equivocado y ella es fiel, es bajo. Además, los costes asociados a la potencial creencia errónea de que su esposa le sea infiel son muy altos, así que abandona esta creencia con poca evidencia; es tan poca la evidencia que necesita para abrazar la nueva creencia, que la adopta tras ser convencido por su mujer de que alucina a los amantes.

Amelie O. Rorty ofrece también un presunto ejemplo de autoengaño con creencias verdaderas *inconscientes*. En su ejemplo, una mujer tiene la creencia inconsciente y verdadera de que tiene cáncer. Debido a esta creencia inconsciente escribe cartas afectuosas a sus allegados, redacta su testamento, etc. Según Rorty, la mujer se autoengaña al afirmar sinceramente que no tiene cáncer, y la muestra de que tiene la creencia inconsciente (y contradictoria con la creencia consciente que afirma) de que tiene cáncer, es que realiza todas esas actividades que no realizaría si no sintiese que su muerte es cercana.

Mele en cambio sostiene que es más plausible interpretar esto como una creencia consciente de padecer la enfermedad unida al deseo de que los demás no se enteren de su estado de salud; esto explicaría perfectamente su comportamiento aparentemente contradictorio sin ningún tipo de supuesto extraño [Mele (2001), p. 71]. No obstante, también señala que no cree que sea imposible interpretar este caso como un caso de autoengaño bajo el modelo de FTL. Según él, no es necesario que la creencia que tiene la mujer de que padece cáncer sea inconsciente. Bastaría con que creyese que, de acuerdo con la evidencia que le ofrece el médico, hay *una posibilidad* de padecer la enfermedad (creencia consciente), pero después de todo no creyese en modo alguno (ni inconsciente ni conscientemente) que efectivamente la padece. Este estado, unido al deseo de no tener cáncer, el deseo de vivir, etc. puede conducirle a redactar cartas o dejar por escrito su voluntad. En este caso, el autoengaño consistiría en creer que *sólo* hay *una* posibilidad de tener cáncer.

Puede que algunos objeten que la explicación que hace uso de la creencia inconsciente es tan válida como la alternativa de FTL. Sin embargo, a menos que tengamos base empírica que nos lo confirme, no hay razón para preferir esta explicación basada en el inconsciente a la que defendería FTL y aludiría a una creencia de que hay probabilidades (aunque muy pocas) de que *p* sea el caso, junto con el deseo de que no lo sea. De hecho, en no pocas ocasiones actuamos en virtud de probabilidades aun siendo muy bajas, como cuando contratamos seguros de vivienda ante posibles incendios o cuando tomamos el paraguas al salir de casa pese a que las predicciones meteorológicas estimaban la probabilidad de lluvia en un 30%. [Mele (2001), p. 72]

Los principales desafíos o críticas que se le plantean a la teoría deflacionaria de Mele son los siguientes:

En primer lugar, los defensores de la explicación intencional exigen que Mele distinga entre pensamiento desiderativo (*wishful thinking*) y autoengaño. La respuesta de Mele es la siguiente: si el pensamiento desiderativo no es lo mismo que la creencia desiderativa (*wishful believing*), no hará falta decir que él se refiere a creencia desiderativa cuando habla de autoengaño. Si se trata de lo mismo, habrá que explicar la diferencia entre ambos en la cantidad de evidencia en contra. Al igual que Mark Johnston, Mele cree que quien se autoengaña tiene mucha más evidencia en contra que quien presenta pensamiento desiderativo.

Sin embargo, hay un núcleo de verdad en esta crítica que se ha dirigido en varias ocasiones tanto a Mele o Johnston como a otros teóricos no-

intencionalistas: en tanto que Mele considera que el caso prototípico de autoengaño consiste en que algunos agentes acaben por creer que *p* debido a la adulteración motivacional de la evidencia (producida a su vez por el deseo de que *p* sea el caso) [Mele (2001), p. 61], esto se encuentra, a nuestro juicio, a medio camino entre pensamiento desiderativo y el error evaluativo.

La otra crítica a la que se ha enfrentado la teoría de Alfred Mele ha sido el dilema ofrecido por William Talbott y José Luis Bermúdez, conocido como el *problema selectivo*:

> El autoengaño es paradigmáticamente selectivo […] Hay situaciones de todo tipo en las que, pese a que deseamos fuertemente que sea el caso que *p*, no estamos de ningún modo sesgados en favor de la creencia de que *p*. ¿Cómo vamos a distinguir estos casos de las situaciones en las que nuestros deseos dan lugar a un sesgo motivacional? Llamo a esto el *problema selectivo*.[80]

Mele acepta que el hecho de que ante historias o casos similares en algunos haya autoengaño y en otros no, sólo puede ser explicado por algún tipo de diferencia. Lo que no acepta es que ésta haya de ser necesariamente una diferencia en la intención (de autoengañarse o hacer más fácil para el sujeto el engaño). En su opinión, la teoría FTL puede dar una explicación satisfactoria de esta diferencia de *output* en función de los distintos umbrales y costes asociados.

---

[80] «Self-deception is paradigmatically selective […] There are all sorts of situations I nwhich, however strongly we desire it to be the case that pm we are not in any way biased in favour of the belief that p. How are we to distinguish these from situations in which our desires result in motivational bias? I will call this the selectivity problem.» [Bermúdez (2000), p. 317]

Bermúdez rebate indicando que no se trata de dos sujetos con distintos umbrales, sino del mismo sujeto y, además, *ex hypothesi* tanto costes, como umbrales, deseos etc. son iguales. Alfred Mele ofrece dos respuestas:

a) reconoce que FTL no podría dar una respuesta, pero afirma que en todo caso, no hay ninguna teoría que pueda dar respuesta a la etiología completa del autoengaño [Mele (2001), p. 65]

   Como ya hemos señalado, Mele no niega que en algún caso la intención juegue un papel en el autoengaño; sin embargo, este tipo de situaciones es tan poco común que no da cuenta de los casos paradigmáticos. [Mele (2001), pp. 16, 67 y 116 (nota 12)]

b) concede, por mor del argumento, que la respuesta es la intención; sin embargo, Mele señala que en muchas ocasiones intentamos cosas que no conseguimos realizar (al margen de aquellas otras que abandonamos o no llegamos a ejecutar). Por tanto, los intencionalistas se enfrentan también a su propio problema selectivo: si la intención va a ser la clave de la explicación, ¿por qué a veces intentamos engañarnos y lo conseguimos y otras veces *pese a intentarlo* no lo conseguimos? Si apelan a que sólo han de tomarse en cuenta los intentos exitosos, se traslada la pregunta: ¿por que algunos intentos son exitosos y otros no? Podrán entonces tratar de robustecer la historia explicativa acudiendo a otros elementos, pero irónicamente, estos elementos serán tan útiles en apoyo de la teoría intencionalista, como para la teoría motivacional FTL y, sin embargo, ésta última no necesita acudir a intenciones.

Mele reserva todo un capítulo para el análisis del autoengaño "retorcido" [Mele (20001), cap. 5]. En los casos más comunes de autoengaño (*autoengaño directo*) el sujeto, ante una realidad dolorosa o fuertemente desagradable, trata de acomodar la realidad a sus deseos. Sin embargo, en los casos de autoengaño retorcido el sujeto se engaña al inducirse una creencia que va no sólo en contra de la evidencia de la que dispone, sino en contra además de su propio deseo. El ejemplo que pone Mele es el del esposo inseguro y celoso que cree que su esposa le es infiel pese a que posee sólo una evidencia muy débil —frente a la gran evidencia que tiene de que su esposa le es fiel. Según Mele esto se debería interpretar en términos del coste que supondría un error para el sujeto. El sujeto trata de evitar los errores (como decía la tesis FTL) y estos errores tendrían un distinto coste en función de las consecuencias que acarreasen. En el caso del esposo sumamente celoso e inseguro, un posible error al creer que su esposa le es fiel sería tan costoso para él, que el umbral para la evidencia de fidelidad sube enormemente, mientras el umbral para la evidencia favorable a la infidelidad baja considerablemente.

Self-Deception Unmasked, obra que puede considerarse el compendio de las tesis de Alfred Mele sobre el autoengaño, tiene dos virtudes principalmente, como son la de distinguir entre *autoengaño directo* y *retorcido* haciendo un pormenorizado y novedoso análisis del *autoengaño retorcido*, y la de ofrecer una panorámica de las distintas posturas de los principales autores ante la problemática del autoengaño haciendo un esfuerzo por someterlas a crítica y ensayar una explicación alternativa y convincente. Sin embargo, este esfuerzo nos parece a veces insuficiente, pues la

caracterización del autoengaño que propone Mele resulta demasiado simple en varios aspectos. Por un lado, parece que sólo se preocupa de explicar cómo el sujeto llega a autoengañarse —sin detenerse demasiado en las causas que puedan llevarle a ello— y por otro lado, no contempla cómo el sujeto sostiene esa creencia, es decir, qué ha de hacer para mantener el engaño impidiendo que florezca la verdad. Mele descuida el hecho de que la creencia contraevidencial que adquiere el sujeto estará continuamente amenazada por la memoria y la nueva evidencia. El autoengañado habrá de estar sesgando continuamente la evidencia, y desde la teoría de Mele este continuo soslayo de la evidencia amenazadora parece difícil de explicar, sobre todo si es un proceso involuntario e inconsciente por parte del sujeto, pues parecería que el autoengaño está condenado al fracaso en un escaso margen de tiempo (al no ser el sujeto consciente de estar autoengañándose parece que no se guardaría mucho de seguir sesgando la nueva evidencia). Además parece que al sujeto le basta con sesgar la evidencia que atañe directamente a la creencia de que *p*. Sin embargo, nos parece que las creencias se disponen en un entramado, y que habría que adulterar más evidencia y engañarse con respecto a otras creencias que se implican y apoyan entre sí.

Por otro lado, la inconsciencia por parte del sujeto de estar engañándose genera otro problema, pues acerca el autoengaño peligrosamente al simple error o ceguera intelectual. En muchas ocasiones nuestro deseo de que algo ocurra hace que no evaluemos bien la evidencia disponible, pero esto no puede ser tratado como un caso de autoengaño. Entre otras razones porque el autoengaño no tiene nada que ver con el

error en la evaluación de la evidencia, sino con la adulteración y el apoyo de una creencia que en principio era contraevidencial: si el sujeto yerra en la evaluación de la evidencia pero se forma una creencia tomando esta evidencia como base, esta creencia se habrá formado de un modo legítimo, aun cuando resulte falsa por tener una base errónea.

III.3 - Enfoques escépticos y eliminativistas

El tercer gran grupo de teorías aúna todas aquellas que plantean serias dudas acerca de que un fenómeno como el autoengaño sea posible. No aceptan ni los postulados ni las conclusiones tanto de los intencionalistas como de los no-intencionalistas. Los enfoques intencionalistas les parecen más adecuados desde el punto de vista semántico por exigir la *intención* como elemento esencial y tratar de recoger así la tensión entre las dos creencias, pero consideran que conducen o bien a paradojas irresolubles y estados mentales contradictorios e imposibles, o bien a la suposición de entidades mentales innecesarias y más que probablemente inexistentes. Por tanto, aunque en principio se muestren como enfoques más adecuados, estos autores creen que en realidad son conceptualmente imposibles y, por ende, empíricamente imposibles. Por el contrario, estiman que los no-intencionalistas son empíricamente verosímiles, pero creen que no son semánticamente ajustados, ya que al no exigir intención de engañarse por parte del sujeto, parecen disolverse en otro tipo de situaciones similares no problemáticas, como el pensamiento desiderativo, la ceguera intelectual y, en general, la influencia de las pasiones en la cognición, etc., pero no se ajustan a lo que se presume que comporta el autoengaño.

No obstante, dentro de este grupo de escépticos hay también distintos grados: desde los eliminativistas, que niegan el autoengaño en su sentido literal, pues designaría un proceso o estado contradictorio y por tanto imposible, hasta aquellos que, observando que ninguna propuesta logra

solventar satisfactoriamente las paradojas, simplemente no se sienten convencidos por ninguna, aunque no cierran la posibilidad a que algún nuevo enfoque logre explicar el asunto. Otra postura intermedia es la que sostienen quienes aceptan que dado que los sujetos adscriben comúnmente autoengaño, su existencia es en cierto modo un hecho, aunque se reduzca meramente a una categoría socialmente aceptada; en concreto, consideran que el valor de la noción de autoengaño no es tanto conceptual cuanto social: la adscripción de autoengaño tiene diversas funciones, como la de restarle culpa a un sujeto sin eliminar su responsabilidad, o controlar a ciertos sujetos.

Otros, aun aceptando que el autoengaño parece contradictorio en el sentido literal, consideran que no podemos obviar el hecho de que hay situaciones que nos resultan a primera vista difíciles de interpretar. En estos casos, el único modo de avanzar parece ser mediante la reducción del número de casos de difícil interpretación, reexplicándolos como otro tipo de actitudes cognitivas más suaves y, por tanto, menos problemáticas. Veremos a continuación algunas de estas propuestas.

III.3.1 - Una explicación diversificada

Jon Elster es, sin duda, un referente en el estudio de la racionalidad y sus problemas. Como es bien sabido, ha publicado numerosos ensayos, artículos y libros acerca del asunto a lo largo de su trayectoria académica, y goza de una notable popularidad y autoridad en este terreno.

Según Jon Elster, los criterios para hablar de creencias racionales han de tomar en cuenta la *relación causal* entre evidencia y creencia; pero comparar evidencia y creencia resulta insuficiente, porque a veces uno tiene una creencia que está en consonancia con la evidencia, pero no está sustentada en ella [Elster (1983), p. 217; cf. Davidson (1985), p. 112; (1993), p. 216]. Dice Elster:

> Todos hemos conocido personas repantigadas en la autosatisfacción, una autosatisfacción que al mismo tiempo parece justificada y no justificada: justificada, puesto que hay buenas razones para estar satisfechas consigo mismas y no justificada, porque sentimos que podrían estar igualmente contentas si esas razones desaparecieran. [Elster (1983), p. 217]

En este sentido, se hace necesario el estudio de las conexiones causales y evidenciales en tanto que el autoengaño plantea problemas a nuestra concepción de la racionalidad. Según Elster, en un extremo se encuentra la afirmación de que el autoengaño es imposible porque entrañaría que la misma persona a la vez conozca y no conozca la misma cosa y al mismo respecto, esto es, tendría creencias contradictorias. Sin embargo, Elster considera que el autoengaño ha de ser claramente distinguido del mantenimiento simultáneo de creencias contradictorias.

> No se trata de que el autoengaño implique alimentar simultáneamente creencias contradictorias, lo cual haría del autoengaño algo imposible. [Elster (1983), p. 214; cf. (1979), p. 289].

En el otro extremo Sartre, al sostener que al hombre le resulta imposible ser auténtico ante sí mismo, convierte toda acción humana en mala fe o autoengaño, y disuelve así tanto su problematicidad como su importancia. La omnipresencia y la necesidad del autoengaño es una tesis

demasiado radical. Entre ambos casos extremos se encuentra el sentido común, que nos dice que parece que los hombres se engañan a veces, pero no siempre.

En opinión de Elster, la resolución de este asunto representa la prueba fundamental que ha de pasar toda teoría de la naturaleza humana; pero lo cierto es que ninguna teoría ha conseguido dar una respuesta satisfactoria. [Elster (1979), p. 286]. Como ya habían indicado Sartre o Fingarette, la teoría freudiana es incapaz de dar cuenta satisfactoriamente del autoengaño en términos de defensa y resistencia pese a los posteriores refinamientos. Pero tampoco Fingarette ha resuelto el problema, pues deja sin explicar cómo surge el yo unitario. Dice Fingarette: "El mundo es tal que puede hacer que los individuos se vuelvan personas", pero más adelante añade que esto es "una cuestión que no necesitamos debatir aquí" [Fingarette (1969), p. 108]. Elster sostiene, sin embargo, que el modelo del ego y la consciencia de Fingarette necesita una explicación de ese proceso si quiere que se tome en serio. Tampoco los enfoques desde la psicología cognitiva han captado la naturaleza del problema y, aunque realmente sea un problema de "cognición caliente" (*hot cognition*), el autoengaño tiene unas características distintivas que exigen un enfoque particular [Elster (1979), p. 287].

Elster considera que el autoengaño parece estar relacionado con la debilidad de la voluntad, ya como fenómeno cognoscitivo paralelo-similar, ya como condición de posibilidad [cf. Rorty (1972), p. 405; Fingarette (1969), pp. 76 y ss.; Davidson [(1985), p. 102]. Con respecto a ambos

fenómenos se han producido discusiones acerca de su posibilidad, preguntándose

> si representan nociones contradictorias acerca de fenómenos mentales, y no nociones acerca de fenómenos mentales contradictorios. [Elster (1979), p. 287][81]

Por otro lado, muchos autores utilizan las nociones de autoengaño y de pensamiento desiderativo como si fueran intercambiables [vid., por ejemplo, Kolakowski (1978), Vol. III, pp. 89, 116, 181; Levenson (1968), vol. I, pp. 59 y ss., 70], pero ambos fenómenos presentan características bien distintas y el autoengaño, además, conduce a paradojas que no surgen en el pensamiento desiderativo. Mientras el pensamiento desiderativo consiste para Elster en la "tendencia a formar creencias cuando y porque prefiero el estado del mundo en que estas creencias se hacen verdaderas por encima de los estados que las hacen falsas" y comporta un proceso causal que responde a un *impulso no consciente*, el autoengaño consiste en una creencia a voluntad, lo cual no es tanto un proceso causal como una elección deliberada que responde a un *deseo consciente* [Elster (1983), p. 213-214]. En el autoengaño hay un *proyecto intencional,* y es precisamente porque comporta un *proyecto de engañarse* por lo que surgen la paradojas. [Elster (1979), p. 288]

> El autoengañador intencional e instantáneamente adopta una nueva creencia pero de alguna manera logra ocultarse esta intención a sí mismo. ¿Cómo lograr esto? Tal se ha convertido en *el* problema del autoengaño. No propondré una solución, salvo para expresar la

[81] «Las atribuciones de autoengaño son atribuciones de incoherencia, no atribuciones incoherentes» [Rorty (1972), p. 395]

> esperanza de que la clase residual de casos que no pueden recibir un análisis más directo acabará por reducirse a un cascarón vacío [Elster (1979), p. 293].

En primer lugar, si el autoengaño comportase dos creencias contradictorias, esto lo convertiría en algo imposible. [Elster (1983), p. 214] Lo que hay que explicar es cómo el sujeto es capaz de *olvidar intencionalmente* aquello que "verdaderamente" cree para conseguir, tras este reto imposible, creer a voluntad posteriormente algo que considera que no tiene los fundamentos necesarios para ser creído. Esto parece "una hazaña más allá de la capacidad humana", ya que la capacidad de creer a voluntad, además de resultar conceptualmente imposible, tiene la paradójica consecuencia de que cuanto más trata uno de ejercerla, menos probabilidades de éxito tiene: es "como tratar de crear oscuridad por medio de la luz". Sin embargo, Elster señala que, al igual que sucede con otros fenómenos aparentemente incoherentes como la debilidad de la voluntad,

> la incoherencia de la noción parece evaporarse ante la masiva experiencia clínica, ficcional y cotidiana que atestigua la realidad del fenómeno [Elster (1983), p. 215].

Algunos casos de autoengaño podrán ser explicados mediante la teoría de la racionalidad imperfecta [Elster (1979), cap. II][82], pero algunos de

[82] Esta teoría hace referencia al modo indirecto en que los sujetos ejercitan su racionalidad. Un individuo que quiere —por ejemplo— dejar de fumar, no siempre toma en consideración los riesgos e inconvenientes, decide que ha de dejar de fumar y simplemente lo deja. En muchas ocasiones, utiliza métodos indirectos como no fumar en presencia de sus hijos, evitar pasar junto al estanco, acudir a locales para no fumadores, etc. Este modo de actuar no es irracional, pero tampoco muestra una racionalidad perfecta. El caso de Ulises, que se hace atar al mástil de su

ellos sólo podrían ser abordados añadiendo cláusulas adicionales y, finalmente, quedarían algunos casos irresolubles; el problema es que estos casos son los más interesantes.

El autoengaño está ligado, según Elster, a la formación del carácter y, en concreto, a la decisión de creer y la decisión de olvidar. Habría tres modos de llevar a cabo el proyecto:

1) De un modo directo y plenamente consciente. Esto, en opinión de Elster es, como ya mostró Bernard Williams (1973), *conceptualmente imposible*.

2) Indirectamente, mediante una *interrupción en la continuidad del yo*. El sujeto prepara el escenario posterior y de algún modo consigue disociar el yo anterior y posterior al engaño (p. ej., se emborracha o toma una píldora que le produce la amnesia necesaria para llevar a cabo el proyecto). Esto es más bien un engaño a uno mismo. [Elster (1979), pp. 292-293; (1983), p. 216; cf. Johnston (1988), pp. 76-78; Audi (1982), p. 143].

3) El caso intermedio: aquí se sitúan los casos difíciles. El sujeto adopta intencional y conscientemente una creencia, pero logra ocultarse esa intención.

Elster renuncia a ensayar una explicación omnicomprensiva del autoengaño. En su lugar, tratará de reducir el número de casos difíciles

embarcación para no sucumbir a los cantos de las sirenas y ordena que si en algún momento pide que lo desaten, ellos le aten más fuerte aún, es otro ejemplo de este modo de proceder, que además da título al libro en el que Elster desarrolla la teoría de la racionalidad imperfecta: *Ulises y las sirenas* [Elster (1979)].

mediante una explicación diversificada que reinterprete los distintos casos como fenómenos diferentes que se parecen sólo superficialmente.

En este sentido, algunos simplemente son tentativas no exitosas de autoengaño directo; otras veces llamamos autoengaño a lo que es en realidad un *intento fracasado de automodificación*; y en otras ocasiones, se trata de *una negativa a recabar las pruebas disponibles*. [Elster (1979), p. 294]

Con respecto al primer tipo, Elster advierte que intentar creer directamente y a voluntad aquello para lo que uno no tiene evidencia suficiente, no es más paradójico que el *intento* de alcanzar otros estados contradictorios, incoherentes o imposibles. Respecto del segundo tipo indica cómo en algunos casos decir algo lo convierte en verdad. No se trata tanto de creencias que se autocumplen, cuanto de autoconvencimiento. Por ejemplo, uno puede superar el miedo a fuerza de repetirse a sí mismo que no tiene miedo, y esto no constituye problema alguno; sin embargo, repetirse algo a uno mismo no garantiza el éxito y, así, algunos casos "rudamente tildados de autoengaño" se diluyen en situaciones en las que el sujeto trata de automodificar sin éxito su ser, haciendo creer (a él o los demás) algo que realmente no es el caso. Respecto del tercer tipo, Elster se refiere a algo similar a la caracterización que hace Fingarette del autoengaño como una negativa a confesar o explicitar los compromisos con el mundo. En este sentido, el autoengaño surgiría cuando uno se niega a reconocer una prueba amenazadora. El sujeto elige deliberadamente no adquirir las creencias de nivel inferior que darían sustancia a las de nivel superior, con lo que la creencia se hace más

tolerable. El dictador les dice a sus súbditos que no quiere conocer los detalles —aun cuando sabe que hay detalles desagradables que conocer— es un ejemplo de este tipo. Elster sospecha que la habilidad de millones de alemanes para no enterarse del exterminio de los judíos que veían desaparecer a diario puede tener que ver con este modelo. [Elster (1979), p. 295] Sin embargo, admite que este segundo tipo no constituye un ejemplo prototípico de autoengaño, en tanto que *no* podemos atribuir al sujeto el conocimiento de unos hechos que por otro lado no quiere conocer. Sabe que hay tales hechos, pero no sabe en qué consisten.

Ciertamente, en el parágrafo que hemos dedicado a la disonancia cognitiva (§I.4) veíamos cómo esta estrategia era denominada "exposición selectiva a la información". Según Leon Festinger, el sujeto trata de exponerse sólo a fuentes de información que tiendan a aumentar la consonancia o eviten que aumente la disonancia. Sin embargo, Elster advierte que Festinger no tuvo en cuenta en un principio una paradoja asociada a este tipo de manipulación: cuando el sujeto conoce los detalles, no puede evitarlos. Festinger se vio obligado a reconocer años después:

> Una vez que se ha dicho a la persona que la información que existe no apoya su decisión, se ha introducido ya evidencia adicional: en cierto sentido es imposible evitarla, pues sabe ya que existe. [Festinger (1964), p. 82]

Como vemos, Elster es bastante escéptico con respecto a la verdadera existencia de casos genuinos de autoengaño. Por un lado, niega tanto que el sujeto pueda poseer creencias contradictorias, como que uno pueda engañarse a sí mismo de modo directo.

Además, se pregunta por qué hemos de acudir al autoengaño cuando algo puede explicarse mediante el pensamiento desiderativo. En concreto, por qué no pensar que el creyente acude directamente a la creencia placentera (pensamiento desiderativo) en lugar de atribuirle autoengaño a través de un complejo proceso de cuatro pasos: 1) llegar a una creencia bien fundamentada, 2) decidir que no es agradable, 3) suprimirla, y 4) adherirse a otra más agradable.

> ¿Por qué ha de tener la fuerza repelente de una creencia desagradable privilegio explicativo sobre la fuerza de atracción de una creencia agradable? Sostengo que a falta de argumentos específicos de lo contrario, la racionalización de la esperanza[83] es una explicación más prudente que el autoengaño. [Elster (1983), p. 218]

Para el resto de casos que permanezcan *prima facie* problemáticos, tratará de reducir su número ofreciendo explicaciones alternativas que diluyan el problema. Es consciente de que no todos los casos quedan explicados con estas herramientas, pero cree que "sólo se logrará progreso mediante futuras reducciones" de los casos que plantean dificultades [Elster (1979), p. 297], con la "esperanza de que la clase residual de casos que no pueden recibir un análisis más directo acabará por reducirse a un cascarón vacío." [Elster (1979), p. 293]

[83] Puede resultar desorientador el hecho de que Enrique Lynch traduce a lo largo de todo el libro *Uvas amargas,* la noción de "*wishful thinking*" como "racionalización de la esperanza", en lugar de "pensamiento desiderativo". Nosotros hemos preferido este segundo giro por estar mucho más extendido en la literatura acerca del asunto.

III.3.2 - Un enfoque eliminativista

Según David Kipp, si el autoengaño se presenta como un fenómeno controvertido y problemático se debe en buena medida a que quienes han tratado ofrecer una explicación satisfactoria, han insistido en una caracterización o interpretación totalmente desafortunada del fenómeno.

En este sentido, aquellos que acusan al sujeto de ser víctima de autoengaño no se ponen de acuerdo en un punto crucial, a saber, si el sujeto realmente cree aquello que presenta como su creencia. Para Kipp hay tres posturas fundamentales con respecto a este punto, y de este modo los casos de autoengaño pueden ser clasificados como literalistas (*literalists*), melioristas (*ameliorists*) y eufemistas (*euphemists*).

Según los que defienden los casos melioristas estos sujetos realmente mantienen la creencia que dicen abrazar debido a que aquello que les motiva a inducirse esta creencia les impide ver la realidad con claridad. Estos sujetos estarían cegados por la pasión o perdidos por el deseo. Otros autores defienden casos literalistas, en los que el sujeto, empujado por algún motivo, mantiene y no mantiene la creencia que dice mantener debido a que el sujeto se persuade a sí mismo de que algo que sabe que es el caso, en realidad no lo es. Por último, los casos eufemistas serían aquellos en los que en individuo no cree aquello que dice creer, ya que ninguna estrategia puede extinguir su consciencia de la obviedad de la falsedad de esta creencia [Kipp (1980), p. 306]. El sujeto simplemente actúa como si creyese que *p*.

Según Kipp el enfoque que defiende los casos literalistas según los cuales es posible que el sujeto mantenga creencias contradictorias es descaminado ya que describe un estado mental imposible. En primer lugar Kipp afirma que esta tesis descansa sobre el supuesto de que un sujeto puede ser su propio "engañador", lo cual describe un estado de cosas imposible; de hecho Kipp dice no conocer ningún argumento que pueda escapar de esta acusación [Kipp (1980), p. 308].

No obstante, aunque para algunos el aspecto paradójico de la posición literalista descansa sobre el hecho de que parece que el sujeto es a la vez engañador y engañado, algunos literalistas se han defendido de esta paradoja argumentando que en realidad el sujeto comenzaría siendo sólo engañador y acabaría siendo sólo engañado. Sin embargo esto es para Kipp una cuestión retórica, pues no se entiende como un sujeto puede estar autoengañado sin estar engañado y cómo puede estar engañado sabiendo a la vez aquello que debe conocer como engañador. De hecho, no es más sencillo comprender que un sujeto comience siendo engañador y acabe siendo engañado que comprender que un sujeto sea a la vez engañador y engañado.

Además Kipp cree que no puede entenderse un proceso en el que un sujeto comience por ser meramente engañador para acabar siendo meramente engañado sin que ese proceso pase por una fase intermedia en la que el sujeto sea ambas cosas, y *esa* es la paradoja que precisamente trataban de evitar los literalistas.

Por tanto el autoengaño entendido de modo literalista no sería posible. Otra cosa bien distinta es que un sujeto finja mantener una creencia en la que en realidad no cree. Esto para Kipp es una descripción de hechos posibles, pero representaría ya un caso del llamado autoengaño eufemista, que en todo caso *no es una instancia de autoengaño genuina*. Las razones para que un sujeto actúe de este modo son siempre de cara a los demás, para no parecer penoso o irrisorio frente a ellos, para que no sepan que conoce la verdad o, simplemente, para no darles el gusto de que sepan que sufre por su consciencia de su fracaso existencial [Kipp (1980), p. 316]. Pero no sería un caso de autoengaño porque el sujeto no abraza creencias contradictorias ni es engañador de sí mismo.

Para Kipp aquellos individuos a los que se les atribuye autoengaño, ni mantienen la creencia que dicen mantener, ni la dejan de mantener, ni la mantienen y no la mantienen a la vez. En realidad debemos ver a aquellos que son caracterizados como autoengañados como personas que no están seguras de que sea el caso o no aquello que desean, y que fingen que sí lo es *no* en un intento de engañar a los demás, sino en un intento de hacerles tambalearse en su opinión y obtener así evidencia confirmatoria de ellos acerca de aquello que desean sea el caso.

En conclusión, para Kipp el autoengaño no es posible; el autoengaño entendido de modo fuerte, consistiría en que un sujeto mantuviese dos creencias contradictorias, que fuese a la vez engañador y engañado, y que diseñase además una estrategia para ello, todo lo cual es imposible. Es cierto que en algunas ocasiones en las que vemos que un amigo al que

admiramos y en el que confiamos se comporta de modo extraño y dice mantener algo que no parece razonable, nos es más fácil creer que está autoengañándose que hacernos a la idea de que está siendo deshonesto en aquello que dice creer. Pero entonces, si nos negamos a la posibilidad de que nos está engañando ¿no estaremos siendo nosotros mismos víctimas de autoengaño de tipo meliorista? ¿No estaremos cegados por nuestros sentimientos y deseos al evaluar lo que es el caso? El llamado autoengaño de tipo meliorista sí es posible, pero *tampoco es un caso de autoengaño genuino*; simplemente al estar cegados por nuestros sentimientos o deseos, no somos capaces de evaluar la realidad objetivamente y hay una especie de error al valorar la evidencia, pero no somos conscientes de ello; se trata de *ceguera intelectual*. No creemos una cosa y tratamos de inducirnos la contraria mediante una estrategia engañadora ni tampoco mantenemos en ningún momento dos creencias contradictorias a la vez.

III.3.3 - El autoengaño como constructo social

> El concepto de autoengaño es un constituyente de la etnopsicología de la cultura.
>
> [Gergen (1985), p. 236]

El enfoque de Gergen resulta interesante principalmente por dos motivos. En primer lugar, es una de las pocas voces disconformes con la idea de que el autoengaño es un *factum* psicológico. Gergen sostiene, por el contrario, que el autoengaño es un *factum* sociológico. Y aquí es donde entra en juego el segundo motivo: Gergen trata de profundizar y

esclarecer la razón sociológica por la que el concepto se mantiene funcionalmente activo.

Con respecto al primer asunto, Gergen tiene una idea de por qué el concepto de autoengaño tiene una presencia especial en el siglo XX. Según él, desde que irrumpiese con fuerza el psicoanálisis, los mecanismos de defensa freudianos han sido ampliamente aceptados como constituyentes fundamentales de la vida mental. En este sentido, no resulta extraño oír decir a otros (aunque pocas veces respecto de uno mismo) que "se engañan a sí mismos", "se mienten a sí mismos", "no encaran la verdad sobre sí mismos", o se involucran en otras formas cualesquiera de auto-disimulo. De modo paralelo, este discurso ha incrementado la accesibilidad de la gente común a otras formas descriptivas en las que la psique de uno está dividida contra sí misma, como por ejemplo la "falsa conciencia". De hecho, la labor de la terapia psicoanalítica estaría orientada en gran medida a reducir la magnitud del autoengaño [Gergen (1985), pp. 228-229].

Sin embargo, Gergen lamenta que no se haya hecho especial hincapié en justificar y argumentar aquello que garantiza el concepto de autoengaño. No se investiga rigurosamente si está fundamentado de algún modo sólido, si hay una base empírica adecuada para este concepto, esto es, si podemos constatar empíricamente la existencia de casos de autoengaño a partir de los que abstraer sus características fundamentales o sobre los que construir o basar el concepto y sus requisitos [Gergen (1985), p. 229]. En su lugar, el analista emplea este concepto como

> unas lentes interpretativas para determinar "lo que hay". Dadas las ambigüedades de significado en los enfoques normales que la gente hace de sus vidas, un analista debería hallar poca dificultad en "encontrar" que virtualmente todos los analizandos están autoengañados de un modo u otro [Gergen (1985), p. 229].

Los problemas teóricos y ambigüedades con respecto a los casos de estudio han provocado que distintos investigadores hayan intentado explorar el autoengaño en circunstancias controladas de modo más sistemático. Gergen cree que quizá el esfuerzo más ambicioso haya sido el realizado por Gur y Sackeim, aunque posteriormente se ha demostrado que su metodología es problemática. El problema estriba en que, como no podría ser de otro modo, los resultados de todos sus experimentos están sujetos a la interpretación de los datos, y estas interpretaciones ni están justificadas ni son inmediatas:

> […] ninguno de los elementos explicativos es transparentemente accesible a la observación; cada uno ha de ser inferido a partir de la actividad conductual. Sin embargo, no hay límite aparente (salvo los límites de la creatividad humana) para el número de interpretaciones que pueden hacerse de las bases psicológicas de cualquier acción dada [Gergen (1985), pp. 229-230].

Aun más: en la medida en que siempre interpretamos las acciones del sujeto realizando inferencias explicativas que están basadas en otras acciones suyas, y a su vez estas mismas acciones se justifican por otras inferencias, que necesitan nuevamente de otras acciones que las justifiquen, y así sucesivamente, resulta que finalmente todas las inferencias del "desconocido reino de la mente" (sea consciente o no) reciben apoyo de la red de inferencias asociadas al enfoque previo que uno

elige hacer. En este entramado, el concepto de autoengaño tiene "un interés irresistible, profundas ramificaciones, y una amplia utilidad social y profesional", que nos hace dudar acerca de si debería ser relegado al estatus de mero *mito folk*.

La investigación futura no podría, en principio, ofrecer un fundamento objetivo para el concepto, ya que hay dificultades conceptuales que impiden que el término adquiera credibilidad. En concreto, además de las paradojas ya comentadas, el autoengaño plantea problemas en tanto que defensa física (como veíamos en la crítica sartriana al concepto de defensa en Freud).

> Aceptar un nivel desconocido (virtualmente incognoscible) en la mente, ya le compromete a uno con un romanticismo peligroso en el que los misterios de lo desconocido logran prominencia sobre lo que se conoce. Añadir un tercer nivel de funcionamiento, una consciencia de lo desconocido que no informa a lo conocido, parece amenazar aun más la inteligibilidad de la teoría. [Gergen (1985), pp. 232-233; cf. Sartre (1943), p. 102]

De hecho, Gergen es aún más explícito que Sartre acerca de las condiciones que impondría un proceso sub-agencial como el que se corresponde con la teoría freudiana a la entidad censora. Este proceso nos obligaría a suponer:

a) Un dispositivo sensible que permitiese que se le informase de los contenidos de lo inconsciente.

b) Un aparato conceptual que permitiese que estos contenidos fuesen clasificados (por ejemplo, como amenazantes o no amenazantes).

c) Un sistema de almacenamiento y recuperación que permita que estos contenidos sean recordados cuando sea necesario.

d) Un dispositivo sensible que permitiese que se le informase del estatus de la mente consciente (para que se dé cumplimiento a sus disposiciones).

e) Un aparato conceptual, un banco de memoria y un sistema de recuperación que permita comprender y ajustar los compromisos de la mente consciente.

f) Un dispositivo de procesamiento lógico o comparador que permita comprender qué contenidos de lo inconsciente son antitéticos a los compromisos de lo consciente.

g) Un centro de mando que active las maniobras de defensa.

h) Un dispositivo de retroalimentación que permita mantener control sobre las maniobras defensivas.

i) Un sistema de almacenaje y recuperación, necesario para las operaciones propias del sistema de mando.

Estos requerimientos le parecen a Gergen asombrosos y prohibitivos [Gergen (1985), p. 233]. Además, si ya incluso en el nivel más transparente de la vida mental —el de la vida consciente— es difícil establecer cuestiones aparentemente tan sencillas como cuándo finaliza un evento mental y comienza otro, en cuanto tratamos de adivinar qué es aquello que sucede en la vida inconsciente, las dificultades se vuelven inatacables.

Pero entonces, si el concepto resulta paradójico e incuso contradictorio, habrá que preguntarse por qué es de uso tan corriente, así como si debe erradicarse del vocabulario contemporáneo. ¿Es meramente una forma

estúpida de expresión de psicología folk que resulta desorientadora? Gergen cree que las distintas caracterizaciones de la mente no reciben su apoyo de la experiencia que los individuos tienen de su propia mente, sino de las convenciones culturales, de modo que en cada momento histórico y en cada cultura ha habido un modo de caracterizar los estados mentales. Así,

> […] el concepto de autoengaño es un constituyente de la etnopsicología de la cultura (sistema de creencias folk sobre la naturaleza del funcionamiento humano a nivel psicológico) [Gergen (1985), p. 236]

Sin embargo, la caracterización de este fenómeno de la etnopsicología tiene los problemas de justificación que Gergen señalaba anteriormente, por lo que la justificación parece reducirse únicamente a su uso y función social. [Gergen (1985), p. 237]

El concepto de autoengaño cumple dos funciones para Gergen:

1) exculpa al sujeto sin eliminar su responsabilidad, y
2) es un arma de control social.

Con respecto a la primera de las funciones, Gergen expone cómo la adscripción de un determinado estado mental conlleva implícitamente la atribución de responsabilidad moral. Según Gergen, hay tres momentos en la adscripción de estados mentales; en el momento en que al hablar de una acción humana se le atribuye a alguien un estado mental, simultáneamente se le atribuye una segunda función: no sólo se nombra o describe la acción, sino que también se fija el origen de tal estado bajo la

forma de una intención que proporciona la explicación para tal comportamiento. Esto lleva directamente a considerar al sujeto de la acción como un potencial blanco de elogio o culpabilidad. De este modo, decir de alguien que "agredió" a otro individuo, supone atribuirle la intención de agredir y, por tanto, hacerle además responsable de tal agresión. Las descripciones sirven como explicaciones, y éstas como criterios morales. [Gergen (1985), pp. 237-238]

Sin embargo, la gente desea evitar a menudo las sanciones morales y suele desviar la responsabilidad hacia otro lado, unas veces hacia otros y otras hacia el entorno u otras situaciones externas: "me obligaron a hacerlo", "las circunstancias escapaban a mi control", "fui una víctima de las circunstancias". A veces no resulta sencillo echarle la culpa al entorno o el propio sujeto es parte principal de ese entorno, por lo que desviar la responsabilidad en ese sentido sería poco efectivo; también puede ocurrir que el sujeto sienta simpatía por otro sujeto al que podría desviar la culpa y no desee cargarle con el castigo. En esos casos puede tratar de escamotear su responsabilidad afirmando algo así como "no puede evitarlo, estaba cegado por la ira" (la pasión o la tristeza), "lo necesitaba tanto que no pude pensar con claridad" o "mi deseo era tan intenso que no puede evitarlo". El problema es que en el mundo occidental las emociones no suelen exonerar de responsabilidad, y es aquí donde aparece el autoengaño y la función social que ejerce: permite considerar al individuo responsable de sus acciones, pero a la vez parcialmente inocente.

Gergen señala que este papel lo cumplieron en otros momentos históricos otros conceptos; así, en la antigua Grecia la locura momentánea se atribuía a la influencia de los dioses. La sociedad no era responsable de las acciones del sujeto, lo era él; pero la influencia de los dioses lo exculpaba en cierto modo. La misma función la ejerció en la cristiandad la figura de Satán.

Sin embargo, el proceso de laicización en occidente ha producido que en este momento las explicaciones que acudían a los dioses o a Satán hayan perdido fuerza, y el autoengaño ha venido a ocupar su lugar. Así, el individuo que se autoengaña es percibido como un sujeto que es arrastrado por fuerzas internas y necesita que le eduquen antes que le castiguen; hospitalizarle antes que encarcelarle: "responsable, pero perdonado" [Gergen (1985), pp. 238-40]

La segunda función social que cumple la noción de autoengaño es la de resultar un *potente arma de control.*

Como es bien sabido, el lenguaje es una formidable herramienta de poder. Gergen argumenta que debido al grado desproporcionado de voz que la sociedad moderna les da ahora a aquellos que parecen ostentar objetividad (científicos, médicos, etc.), los psicoterapeutas se han visto favorecidos por este prestigio heredado debido a su cercanía profesional con la medicina. La gravedad del asunto estriba en que si un psicoterapeuta le adscribe autoengaño a un paciente, éste tenderá a ver todas sus afirmaciones y compromisos como meras tretas que esconden la verdadera actitud perversa e inconsciente a la que por definición no puede

acceder por estar reprimida. Pero entonces, cualquiera que tenga el poder de adscribir autorizadamente autoengaño, tiene un potente elemento para cambiar patrones de conducta. En todo caso, el concepto de autoengaño es "un potente arma en el arsenal del control social". Gergen añade que si la sociedad se beneficia o sale perjudicada por esto será algo que se verá en el futuro [Gergen (1985), p. 241]

Como hemos visto, Gergen es eliminativista con respecto a la posibilidad del autoengaño en tanto que fenómeno psicológico.

En primer lugar, Gergen está de acuerdo con Bach en que el autoengaño no puede consistir meramente en equivocarse al considerar algo, o en que los pensamientos de uno contengan alguna inconsistencia. [Gergen (1985), p. 232; cf. Bach (1981), p. 352]. En segundo lugar, cree que no es posible que el sujeto crea que *p y no-p*. Y por último, cree que acudir a motivaciones inconscientes conlleva aceptar una serie de nociones explicativas tan oscuras que hacen del problema algo ininteligible.

No obstante, concede que tiene una enorme presencia como concepto social, y trata de explicar por qué no nos hemos deshecho de él si efectivamente no es adecuado para la descripción de nuestros estados mentales. En este sentido, cree que la justificación de la permanencia de este término folk se reduce a una doble función social. Por un lado exime de las culpas que se seguirían de acciones de las que somos claramente responsables, y por otro lado es una potente herramienta de control social.

Desde luego, hay otros estados mentales, como creencias, deseos, etc., que entre otras cosas, pueden ser utilizadas como herramientas de control social. El punto que subraya Gergen es que, dado que el autoengaño involucra —a diferencia de lo que ocurre con estos otros estados mentales— una imposibilidad conceptual, la *única razón* por la que seguimos haciendo uso de este concepto es porque resulta útil, en tanto que cumple una función social.

CUARTA PARTE

IV. ALGUNAS PROPUESTAS DESDE LA CIENCIA

IV.1 - Autoengaño y psicología

Uno de los aspectos más destacados del comportamiento social de las especies cuyo sistema social es más complejo, es el del engaño en sus distintas versiones. El disimulo, el fingimiento o la simulación son formas asociadas al engaño que pueden observarse no sólo en el hombre, sino en otras especies, sobre todo en primates. Sin duda muchas otras especies inferiores poseen capacidades que inducen al error tanto a congéneres como a otras especies depredadoras, tales como el camuflaje, la mimetización y otros mecanismos adaptativos. Sin embargo, no consideramos estos fenómenos como casos de engaño en tanto que no comportan intención por parte del organismo en cuestión de inducir a error. Más bien nos parece que aquellos individuos que han desarrollado un comportamiento de este tipo se han ido seleccionando, esto es, han sobrevivido frente a aquellos otros que no poseían esas capacidades y no eran por tanto capaces de despistar a los potenciales depredadores o competidores, y en esto consiste que ese mecanismo les haya conferido ventaja adaptativa.

Algunos autores como Whiten y Byrne [(1988); (1997)] denominan "inteligencia maquiavélica" a la inteligencia social de los primates basándose en los estudios que habían realizado otros psicólogos y primatólogos, principalmente Nicholas Humphrey. Con el término "inteligencia maquiavélica" se refieren a las complejas formas de manipulación y engaño que emplean los primates —principalmente los macacos rhesus, según indican los recientes estudios de Dario Maestripieri[84]— con el fin de establecer intrincadas "alianzas políticas" para mantener la cohesión social y, en último término, desarrollar estrategias sociales útiles para la supervivencia.

Sin embargo, es en el ser humano en quien este arte de engañar alcanza su mayor expresión. Todo el mundo ha engañado o ha tenido la experiencia de ser engañado. No obstante, pese a que el engaño es un fenómeno casi omnipresente en nuestras vidas, quizá no somos tan buenos detectores de mentiras e intentos de engaño. Algunas veces mentimos con la intención de no dañar a otro y en no pocas ocasiones la mentira es necesaria para el florecimiento de las relaciones sociales. Pero en este escenario plagado de mentiras y engaños, el ser humano tiende a creer lo que le dice su interlocutor. Este fenómeno de confianza en los testimonios es el que, a buen seguro, garantiza indirectamente el éxito de las mentiras y está causado, en mi opinión, por algo que está conectado con la eficacia del Principio de Cooperación griceano. La supervivencia y

[84] Maestripieri, Dario (2007), *Macachiavellian Intelligence: How Rhesus Macaques and Humans Have Conquered the World*, Chicago, University of Chicago Press.

el desarrollo del ser humano han sido posibles gracias a los lazos sociales, que a su vez dependen del éxito comunicativo. Como es bien sabido, Paul Grice fijó en la *cooperación* el punto esencial de la comunicación. Tanto el hablante como el oyente han de cooperar en el acto comunicativo para garantizar su éxito. En este proceso, no sólo el oyente espera cooperación por parte del hablante, sino que éste supone de antemano que el oyente espera dicha cooperación. Y es esta confianza mutua la que garantiza que tanto hablante como oyente se ciñan —no sólo tácita, sino inconscientemente— a unas pautas en las que el respeto a la veracidad es uno de los rasgos cruciales. De este modo, hay una suposición de veracidad *bajo la condición* de que es racional asumirla, siempre que no se tengan motivos para lo contrario. No obstante, aunque no es una suposición generalizada, si no se asumiese por lo común la veracidad de los testimonios, la comunicación quedaría, si no ya impedida, sí fuertemente dañada.

Sin embargo, irónicamente es esa confianza, presunción de veracidad y cooperación lo que da puede dar lugar a que, por un lado, algunos traten de sacar provecho de esta situación y, por otro lado, en algunas ocasiones la mentira cumpla una función social más efectiva. Así, no siempre decimos —ni podemos decir, por razones sociales y/o morales— lo que pensamos.

Llegados a este punto podría sugerirse que, puesto que nuestra incapacidad para la detección de mentiras se debe a un sesgo inconsciente de presunción de verdad, si nos hiciésemos conscientes de este prejuicio

cognitivo y en consecuencia lo eliminásemos, seríamos más capaces de detectar el engaño. Nada más lejos de la realidad.

IV.1.1 - Experimentos empíricos en psicología social

En primer lugar, ha de señalarse que existe la creencia común de que "la mentira tiene las patas muy cortas" o de que "se coge antes a un mentiroso que a un cojo". Este saber popular cristalizado en el refranero parece indicar que, por lo general, nos sentimos bastante capacitados para detectar mentiras o engaños. Sin embargo, Jaume Masip (2005), ha realizado un profundo análisis de los resultados obtenidos en numerosos estudios empíricos realizados por distinguidos psicólogos sociales[85] en los que trata de mostrar que, en primer lugar, nuestra capacidad para distinguir correctamente entre verdades y mentiras es limitada. Estos experimentos consistían en la presentación de declaraciones en vivo y grabaciones audiovisuales o auditivas que el observador había de determinar si eran veraces o no. Normalmente, la mitad de ellas lo eran, así que por mero azar uno podía alcanzar un 50% de aciertos. Todos los experimentos arrojaron un acierto medio de en torno al 53-56 %, lo cual supone un avance mínimo. [Masip (2005), pp. 81-82]

Lo curioso es que los sujetos creen acertar en torno al 72% (lo cual contrasta con los datos reales), e incluso es más llamativo el hecho de que el balance de la correlación entre las atribuciones de veracidad y las

[85] Principalmente, Kraut (1980), Vrij (2000) y Bond y DePaulo (2006).

declaraciones veraces sea aun más pobre (*r* = .04, correlación prácticamente nula). [Masip (2005), p. 84] Y estos datos no varían mucho si se realizan con profesionales cuyos oficios están relacionados con la detección de mentiras, como abogados, jueces, policías o psicólogos, en cuyo caso los resultados no superaron el 60%.

Otro dato que señala Masip es que los estudios empíricos revelan que los indicadores no-verbales para la detección de mentiras que suelen tomarse en cuenta, tales como taparse la boca, tocarse la nariz, frotarse un ojo o el cuello, tirar del cuello de la camisa, parpadear más, mover más el tronco o cambiar de postura no tienen relación significativa alguna. Tampoco resultó significativo el que se realizasen más pausas ni que el sujeto apartase la vista. Este último criterio de detección de mentiras resulta ser una creencia fuertemente arraigada: "mírame a los ojos y dime la verdad". Este estereotipo tiene incluso un alcance universal, y fue referido como una marca de la mentira en 51 de los 58 países en los que se realizó el estudio. Sin embargo, los resultados al respecto eran poco significativos o incluso contradictorios en algunos casos. Por otro lado, mientras se cree que quienes mienten mueven más las extremidades, los datos revelan que las mueven *menos*. El psicólogo social Aldert Vrij (2000) ha indicado que los únicos criterios que suelen manejarse y parecen acertados son los dos siguientes: la gente hace pausas más largas cuando miente y tiene un tono de voz algo más agudo. Es fácil suponer que las pausas más largas corresponden a la necesidad que tiene aquel que miente de tejer una fabulación coherente, y podemos especular que el tono más agudo puede deberse tanto a la inseguridad como a la necesidad de

resultar convincente (presumiblemente en muchas ocasiones pretenderá ser más cercano, y es sabido que los mensajes cariñosos y amables solemos emitirlos en un tono más agudo).

Otro aspecto que señalan los estudios empíricos es que ni siquiera estos criterios son uniformes, sino que dependen de situaciones contextuales. Esto equivale, en mi opinión, a afirmar que los criterios que son relativos al contexto dejan de ser criterios.[86] En cualquier caso, ni siquiera tras un entrenamiento en la detección de veracidad o engaño tomando en cuenta estos criterios se han hallado resultados más esperanzadores en la detección de mentiras. Se han establecido entrenamientos de tres tipos:

1. Retroalimentación: Decir a los sujetos en cuáles han acertado y en cuáles no para que extraigan conclusiones.

2. Informacional: Dar directamente a los sujetos las claves acerca de cuáles son los criterios adecuados y cuáles no para detectar mentiras y el porqué.

3. Atencional: Decirles en qué se han de fijar sin explicarles qué incidencia tiene en la veracidad o falsedad de la declaración. [Masip (2005), p. 87]

Tras el entrenamiento se consigue pasar de un 56% a un 58%. Si bien es cierto que parece que aún se están ensayando nuevos métodos, a

[86] Esto es, se asemejan mucho a los puntos ecuantes con los que se parcheaban los epiciclos que a su vez habían sido añadidos *ad hoc* a la cosmología ptolemaica para explicar en cada contexto las retrogradaciones pertinentes. El carácter improvisto y contextual de los puntos ecuantes en cada momento hacía que los datos posteriores, arreglados con respecto a ese punto colocado *ad hoc*, no tuviesen valor predictivo —por tanto científico— alguno.

diferencia de lo que opina el propio Masip, no somos muy optimistas con respecto al futuro en esta tarea. Incluso ni siquiera el polígrafo u otros métodos técnicos son totalmente eficaces en la detección de mentiras. Masip hace un balance de nuestras capacidades en la evaluación de la mentira que parece adecuado reproducir a modo de resumen. Sus conclusiones son las siguientes:

> (a) la capacidad del ser humano para discriminar entre verdades y mentiras es extremadamente limitada; esto es así incluso en grupos profesionales para quienes la detección del engaño es una tarea importante en su trabajo;
>
> (b) las personas no tenemos conciencia de lo correctos o incorrectos que son nuestros juicios de credibilidad;
>
> (c) tendemos a sobreestimar nuestra capacidad de identificar verdades y mentiras;
>
> (d) utilizamos claves equivocadas al hacer juicios de credibilidad;
>
> (e) las creencias populares sobre los indicadores del engaño son erróneas;
>
> (f) las creencias de los profesionales para quienes la detección del engaño es una tarea importante son también erróneas y similares a las de las otras personas;
>
> (g) no se ha demostrado que los indicadores conductuales que se mencionan en la mayoría de los libros "de autoayuda" permitan una adecuada discriminación entre verdades y mentiras;
>
> (h) existen muy pocas conductas que realmente permitan diferenciar entre verdades y mentiras;

> (i) al contrario de lo que se da a entender en muchos libros "de autoayuda" y de lo que sostiene la sabiduría popular, el significado y el poder de discriminación de las claves conductuales dependen de una serie de variables situacionales;
>
> (j) también al contrario de lo que afirman determinados libros dirigidos al gran público, aprender a discriminar entre verdades y mentiras es extremadamente difícil, como muestra la limitada eficacia de distintos programas de entrenamiento; y
>
> (k) en lugar de incrementar la precisión global, los entrenamientos al uso aumentan el sesgo a decir que las declaraciones son falsas. [Masip (2005), pp. 88-89]

Lo que trata de mostrar Masip es que, sea lo que sea aquello en lo que confiamos para descubrir mentiras, los experimentos muestran que no solemos tener éxito en esta empresa: no detectamos muchas más de las que detectaríamos por puro azar. En este sentido, ni los criterios mesurables que se suelen aducir son eficaces, ni tenemos algo así como una capacidad estética para captar marcas imponderables de las mentiras.

Estas dificultades en la detección de mentiras y otro tipo de engaños son heredadas por el autoengaño. De este modo, parece que toda atribución de autoengaño que aspire a ser adecuada no podrá fundamentarse en los criterios anteriormente descritos, los cuales son, como muestran los experimentos empíricos, más el reflejo de algunos prejuicios equivocados que un elemento de juicio idóneo.

En cualquier caso, lo problemático del autoengaño es que uno no sabe bien de qué modo habrían de realizarse los experimentos, en tanto que no puede ser que quien se autoengaña sea alguien que meramente finja.

Necesitamos asegurarnos de que pese a que cree aquello que dice, a la vez sabe, cree o piensa lo que de algún modo niega. Pero si realmente cree lo que afirma, ¿cómo encontrar un indicio de que por otro lado miente? Es más ¿miente? ¿Cómo detectar esa creencia que ha sido relegada al subconsciente, preconsciente, a un lado de la mente, a un subsistema o lo que se quiera?

Con respecto al asunto del autoengaño, los trabajos en psicología social más destacados, tanto desde el punto de vista de su valor intrínseco como desde el punto de vista sociológico o bibliográfico (ya que son los que más citas suscitan) son sin duda los efectuados por Ruben C. Gur y Harold A. Sackeim.

IV.1.2 - Autoengaño: ¿sólo un concepto? La búsqueda del fenómeno

Gur y Sackeim creen que a pesar del enorme interés que ha suscitado el problema del autoengaño y del considerable papel que le han asignado diversas áreas de estudio, no ha habido verdaderos intentos de demostrarlo empíricamente.

Esto se debe en gran parte a que dentro de la psicología ha habido una larga tradición a partir de los estudios de Wundt que, al margen de lo que postulaban otras teorías menos ortodoxas como el psicoanálisis, consideraba que el sujeto había de ser necesariamente consciente de sus cogniciones. De hecho, la principal razón por la que se han considerado paradójicos algunos conceptos como el autoengaño, reside en que se ha

asumido que la percepción implica consciencia de lo percibido [cf. Sartre (1943), p. 102]. Esto comienza a ponerse de verdad en duda en los años 70 y, aunque la existencia de una *no-consciencia selectiva motivada* ha continuado siendo un asunto particularmente controvertido, una vez que uno acepta que no es necesaria la asunción de reflexividad de la conciencia, los problemas inicialmente asociados a distintos fenómenos, como la percepción subliminal o el propio autoengaño, si bien no desaparecen por completo, sí resultan ya menos paradójicos. Esto casa, de acuerdo con Gur y Sackeim, con el uso común del término "autoengaño", según el cual quien se autoengaña "sabría todo el tiempo", "de algún modo", o "en el fondo" algo que contradice aquello que declara [Gur y Sackeim (1979), pp. 148-149].

De hecho, Gur y Sackeim van a adoptar como conjunto de condiciones necesarias y suficientes, la caracterización que ofreció Raphael Demos [cf. Demos (1960), p. 595], según la cual un sujeto se autoengaña si y sólo si:

1. mantiene dos creencias contradictorias
2. mantiene estas creencias simultáneamente
3. no es consciente de que mantiene una de ellas
4. el acto que determina cuál de ellas es objeto de consciencia y cuál no, es motivado [Gur y Sackeim (1979), p. 149]

En busca de un caso empírico que cumpliese estos requisitos, Gur y Sackeim diseñaron un experimento que constaba de dos partes; en la primera parte, se tomó un grupo de sujetos y se grabó su voz leyendo

unos párrafos de un texto[87]. Posteriormente, se les hizo escuchar unas cintas en las que aparecían distintas voces leyendo fragmentos y tenían que responder, en el menor tiempo posible, con qué grado de seguridad creían que la voz era suya o de algún otro. Mientras hacían el reconocimiento de las voces, estaban monitorizados con un polígrafo que medía los impulsos galvánicos de modo que eran registrados por unos electrodos colocados en las falanges medias del segundo y tercer dedo de la mano no dominante. En la segunda parte, se les pasó un cuestionario de personalidad en el que se evaluaba el grado de seguridad en sí mismos y de aversión a la autoconfrontación.

Según Gur y Sackeim, diversos trabajos experimentales previos mostraban que el sujeto tenía una mayor reacción galvánica al oír la propia voz, independientemente de que la reconociese como suya; además, el tiempo de reacción para discriminar una voz que oyese inmediatamente después de la suya era menor.

Esto fue lo que Gur y Sackeim encontraron; aunque algunos no cometieron errores, no todos los sujetos reconocieron su propia voz; unos hicieron falsos reconocimientos positivos de su propia voz; otros hicieron falsas negaciones, y otros cometieron ambos errores. Sin embargo, la respuesta galvánica de la piel (*Galvanic Skin Reponse*, GSR) era *casi siempre* acertada, de modo que incluso los sujetos que fallaban en sus atribuciones, seguían teniendo altas respuestas a su propia voz, y más bajas a las de los

[87] Concretamente, las páginas 124-125 de *The Structure of scientific revolutions,* de Thomas S. Khun (1962).

demás: parecía que su piel reconocía mejor las voces de lo que ellos mismos eran conscientes y declaraban. Por tanto, cabía afirmar que el sujeto mantenía dos creencias contradictorias, y la creencia verdadera no se manifestaba en su vida consciente; era relegada al subconsciente y sólo descubierta por medio del polígrafo. Ahora bien, el carácter motivacional de esta relegación no parecía tan sencillo de demostrar [Gur y Sackeim (1979), p. 161].

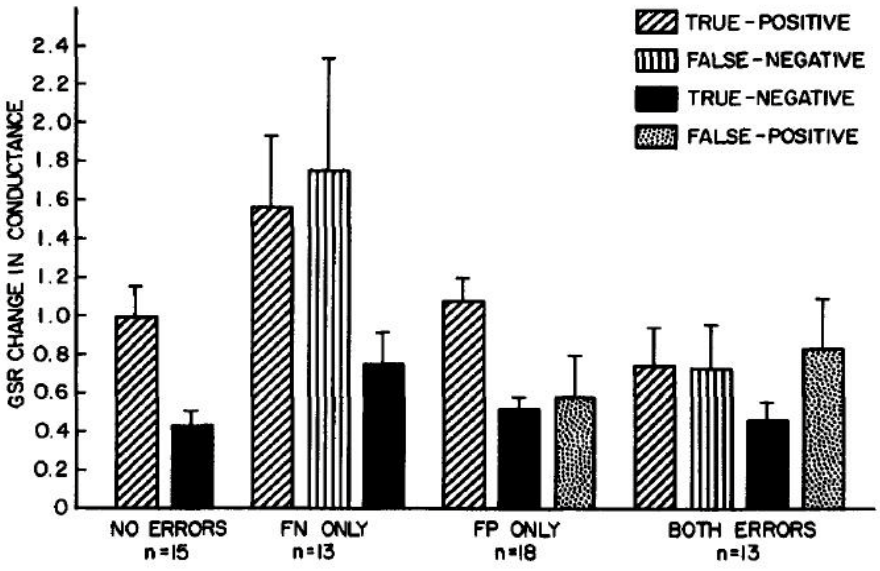

Fig. 1 Respuesta galvánica media (Galvanic Skin Response, GSR) medida en micromhos, ajustada en orden de intentos para los grupos que no cometieron errores (NO ERRORS), errores de falsos negativos (FN ONLY), errores de falsos positivos (FP ONLY) y falsos negativos y positivos (BOTH ERRORS) [Imagen extraída de Gur y Sackeim (1979), p. 156].

Más tarde comprobaron que, como habían predicho, en las audiciones que seguían a las de su propia voz, las respuestas eran siempre más rápidas, independientemente de que hiciesen bien o no la atribución

verbal; cuando por el contrario no se trataba de su voz, la respuesta siguiente no era tan rápida, independientemente de que se atribuyesen la voz como suya. Además, una vez que examinaron los cuestionarios, hallaron que aquellos sujetos que estaban menos a gusto con su imagen, tendían a cometer más falsas negaciones de su voz, mientras aquellos que estaban más a gusto, cometían más falsas afirmaciones. Es decir, los sujetos más inseguros con su imagen evitaban la autoconfrontación y negaban su propia voz, mientras los más narcisistas tendían a proyectarse, debido a que no sólo no evitaban la confrontación sino que la buscaban activamente. Además, casi todos los sujetos que hacían falsas negaciones no eran conscientes de haber cometido errores, y casi el 50% de los que hacían proyecciones, tampoco. Aunque Gur y Sackeim conceden que parte de estas proyecciones podría deberse a la falta de exposición previa a grabaciones de la propia voz, creían haber encontrado en la actitud de los sujetos ante la autoconfrontación el carácter motivacional del proceso.

En todo caso, los resultados del experimento mostraban, según Gur y Sackeim, que:

(a) varios individuos mostraban creencias contradictorias, ya que declaraban conscientemente una cosa, pero sus respuestas galvánicas mostraban que creían lo contrario.

(b) estas creencias, por la propia naturaleza del experimento (que registraba ambas respuestas, verbal y galvánica a la vez), eran simultáneas.

(c) el sujeto no era consciente de una de ellas; no sólo comete errores, sino que no es consciente ni de que los está cometiendo ni de que su piel sí reconoce las voces.

(d) el hecho de que los cuestionarios muestren que los sujetos que hacen más proyecciones son los que más contentos están consigo mismos, y que los que más niegan erróneamente su voz son quienes tienen más aversión a la autoconfrontación, parece indicar que los errores son motivados.

Por tanto, al satisfacer las 4 condiciones necesarias y suficientes, este experimento sería la demostración empírica del fenómeno del autoengaño. [Gur y Sackeim (1979), pp. 150-161]

Este tipo de hallazgos demostrarían colateralmente que debería atribuirse a la consciencia una falta de transparencia motivada y selectiva. Esto está relacionado con el concepto psicoanalítico de "represión", si bien Gur y Sackeim advierten que la *falta de transparencia motivada selectiva* es condición necesaria aunque no suficiente para la represión, ya que ésta requiere además que las creencias que no son objeto de consciencia estén almacenadas en el inconsciente. Dado que el inconsciente es un sistema de control funcionalmente independiente y capaz de influencia en la conducta, la represión no sólo implica que la consciencia no es transparente, sino también que no es unitaria. [Gur y Sackeim (1979), pp. 166-167]

El valor de estos experimentos es ciertamente controvertido. Alfred Mele criticó con dureza sus resultados, en tanto que no está nada claro que el sujeto mantenga dos creencias y, *a fortiori*, que mantenga creencias

contradictorias e, incluso, una de ellas de modo inconsciente. Como ya expusimos, Mele no encuentra problemático que el sujeto mantenga creencias contradictorias, pero sí cree que no es necesario suponer esto para dar cuenta de los casos más comunes de autoengaño. Lo que Mele cuestiona de los experimentos de Gur y Sackeim es que las respuestas fisiológicas en términos de saltos galvánicos sean muestra de creencias [Mele (1987b), p. 6; (1997), p. 96]. Quizá, añade Mele, el umbral para el reconocimiento de la propia voz sea menor que el de cualquier tipo de cognición (incluso la creencia inconsciente). Aún más, Mele indica que otro grupo de científicos [Douglas y Gibbins (1983); Gibbins y Douglas (1985)] ha hallado resultados similares para voces de conocidos, por lo que las fuertes reacciones galvánicas podrían indicar que, aunque el sujeto cree que no es su voz, ésta le resulta familiar [Mele (1997), p. 96].

A esta crítica, Sackeim y Gur responden, en primer lugar, que no se entiende cómo el sujeto puede desatender evidencia amenazante si el sujeto no *reconoce*, a algún nivel, esta evidencia como amenazante. Por esta razón, no hay modo de dar cuenta del carácter amenazador de la evidencia que se evita si uno no posee creencias contradictorias. El problema para ellos es que Mele tiene una visión demasiado estrecha de lo que ha de ser una creencia, ya que sólo considera como índices fiables de creencias las declaraciones del sujeto; esto provoca que Mele impida de antemano el estudio de los mecanismos evolutivos del autoengaño [Sackeim y Gur (1997), p. 125].

Mele apunta, sin embargo, que su objeción no era general y hacia todo test de carácter fisiológico; simplemente, cree que las pruebas fisiológicas que ofrecen Gur y Sackeim no son conclusivas y, por tanto, no pueden ser indicadores de creencias. Además, dice estar también comprometido con un estudio evolutivo del autoengaño [Mele (1997), p. 129].

IV.2 - Autoengaño y biología

En el terreno de la biología, se han realizado pocos estudios acerca de la presunta capacidad de los humanos para mentirse a sí mismos. Recientemente el genetista americano Dean Hamer ha publicado un libro que resultó ser un best-seller[88] acerca del gen VMAT2 (SLC18A2)[89] —que bautizó como el gen de Dios— y su presunta relación con el misticismo y la creencia en un poder trascendente. Según Hamer habría alguna relación entre la posición de este gen y la tendencia a formarse creencias de un determinado tipo. Aunque sus estudios no tratan directamente el autoengaño, sí tienen que ver en cierto modo con la formación de creencias, y resultarían muy significativos si consiguiesen demostrar que la formación de creencias no tiene que ver exclusivamente con la posesión de evidencia. Ciertamente, esto abriría la puerta a las creencias "genéticas"

[88] Dean H. Hamer (2004), *The God Gene: How Faith is Hardwired into our Genes,* New York, Doubleday.

[89] Este gen controla la producción de la proteína encargada del transporte de monoaminas, principalmente neurotransmisores como la dopamina, noradrenalina, serotonina e histamina.

—casi innatas— al margen de la evidencia y, de modo lateral, al autoengaño.

Aunque en un primer momento Hamer —director de la Unidad de Regulación de la Estructura Genética en el Instituto Nacional para el Cáncer de USA— estaba realizando unos ensayos para evaluar la propensión genética al tabaquismo en los que incluía preguntas acerca del sentimiento de transcendencia, parece que halló una correlación significativa entre la tendencia a la espiritualidad y la posición de este gen VMAT2. El estudio sobre la tendencia a la transcendencia tomaba tres parámetros:

1. olvido del yo (*self-forgetfulness*)
2. identificación transpersonal (*transpersonal identification*) y
3. misticismo (*mysticism*).

El primero de ellos hace referencia a la capacidad del sujeto para focalizarse en una actividad, aislándose y obviando por completo cualquier otra cosa. El segundo, a la tendencia de algunos sujetos a identificarse con una entidad mayor que el yo, sintiéndose, en la mayoría de las ocasiones, una parte inseparable y en cierto modo indistinguible del universo. Por último, el misticismo tendría que ver con la tendencia a una apertura mental que le lleva a uno a creer cosas cuya demostración resulta improbable, tales como la percepción extrasensorial.

Según Hamer, el sentimiento de transcendencia o tendencia a la espiritualidad estaría codificado genéticamente, sería heredable y resultaría

beneficioso. En concreto, la creencia en algún tipo de entidad transcendente o Dios resultaría beneficiosa evolutivamente porque produciría un estado de optimismo que haría a la gente más saludable y más proclive a tener descendencia.

Sin embargo, estos experimentos han sido duramente criticados desde varios frentes, principalmente por Carl Zimmer[90]. Zimmer ha señalado que sería interesante observar si los resultados presentados por Hamer soportarían el rigor de la revisión por pares ya que, en primer lugar, están basados en una muestra excesivamente pequeña (poco más de 1000 muestras de ADN) y, en segundo lugar, no pasaron el examen pertinente que consiste en ser publicados en una revista de prestigio científico. Sería igualmente interesante ver si otros científicos son capaces de repetir los resultados obtenidos, ya que Hamer presentó en 1993 unos resultados que indicaban una presunta relación significativa entre una región del cromosoma X y la homosexualidad que no pudieron ser replicados por otros colegas [Zimmer (2004), p. 111]. Finalmente, Zimmer cree que Hamer va demasiado lejos cuando afirma que el llamado gen de Dios es producto de la evolución; por un lado, Hamer no ha tenido en cuenta que la selección del gen ha podido producirse por otros efectos del mismo (por ejemplo, ese mismo gen también cumple la función de proteger al cerebro de neurotoxinas); por otro, no tendría por qué suponer ningún beneficio evolutivo, dado que parece que la selección evolutiva de algunos

[90] Carl Zimmer (2004), 'Faith-Boosting Genes. A search for the genetic basis of spirituality', *Scientific American*, vol. 291 (4), pp. 110-111.

elementos corresponden al azar, a la deriva genética. En todo caso, Zimmer cree que el libro de Hamer podría haber sido iluminador si se hubiese escrito tras 10 años de rigurosas investigaciones acerca de la correlación entre el gen VMAT2 y el sentimiento de transcendencia.

Aun siendo controvertidos los resultados hallados por Hamer, la idea general de que la tendencia a formarse ciertas creencias puede resultar adaptativa y, por tanto, evolutivamente favorecida, ha sido propuesta con notable repercusión por otro famoso científico, el biólogo evolutivo Robert Trivers. Bajo esta premisa, Robert Trivers ha estudiado tanto el carácter evolutivo del engaño como del autoengaño.

IV.2.1 - Autoengaño: un medio para hacer más eficiente el engaño a otros

Según Robert Trivers, todo engaño consiste en una imitación de la verdad, ya que, posiblemente, el engaño constituye siempre una forma de parasitismo de un sistema pre-existente destinado a la comunicación de información correcta. [Trivers (1985), p. 395].

En su estudio del engaño, Trivers comienza con una exposición de este fenómeno en el reino animal a través del análisis de los mecanismos de camuflaje y mimetismo, así como la relación existente entre las relaciones agresivas y el engaño. Tras la descripción por medio de distintos ejemplos de las ventajas adaptativas obtenidas por diferentes especies que practicarían el engaño mediante el camuflaje y mimetismo, Trivers señala que el engaño y la agresividad están íntimamente relacionados. En primer

lugar, algunos individuos tratan de engañar a su oponente intentando aparentar ser de mayor tamaño del que en realidad son, por ejemplo erizando el pelo (chimpancés) o cambiando su color de modo que alternan líneas blancas y negras horizontales (pececillo cabezón o *Pimephales promelas*); en segundo lugar, en algunos grupos donde ciertas características físicas son criterios del rango de sus miembros (como la oscuridad del plumaje del cuello y pecho de los gorriones de Harris *Zonotrichia querula*), cuando se detecta algún intento de engaño mediante la modificación de estas características, aquéllos que trataban de engañar son aislados o castigados de modo muy agresivo. [Trivers (1985), pp. 396-415]

Lo interesante para nuestro trabajo es que cuando aumenta la frecuencia de engaños, se intensifica la selección para su detección y, a medida que aumenta la detección, se intensifica la selección de engaños. De este modo, la selección natural mejora a la vez ambas capacidades, la de engañar y la de detectar engaños, por lo que favorecería un nuevo tipo de engaño: el autoengaño. El autoengaño resulta útil según Trivers porque vuelve inconsciente el engaño para quien lo practica, de modo que las sutiles señales que podría dejar al descubierto el mero engaño quedarían ocultas. [Trivers (1985), p. 395].

Trivers dedica un capítulo al autoengaño empujado en gran medida por el hecho de que frecuentemente se ha negado su posibilidad o existencia, centrándose en mostrar cómo el autoengaño favorece al engaño, cómo se ha demostrado experimentalmente en humanos y cuáles son los mecanismos que lo hacen posible [Trivers (1985), p. 396]. Según él, en

nuestra especie detectamos el estrés que acompaña al engaño por medio de señales como la mirada esquiva, las palmas sudorosas o la voz quebrada[91]. En este sentido, si quien se autoengaña consigue hacerse inconsciente de su propio engaño, logrará evitar los efectos del estrés y, consiguientemente, también ocultará mejor a los demás el engaño.

El lenguaje ha sido, por supuesto, una herramienta muy útil en el engaño, ya que permite practicar el engaño verbal de muy diversos modos y crear complejas redes de creencias que favorecen sesgos en beneficio del sujeto. Evidentemente, cuanto más hábil sea el sujeto en la tarea de ocultarse y ocultar a los demás estos sesgos, más difícil será detectarlos.

Sin embargo, Trivers advierte:

> Por supuesto, ha de ser ventajoso que la verdad quede registrada en algún lugar, así que es de esperar que los mecanismos del autoengaño permanezcan junto a los mecanismos para la correcta aprehensión de la realidad. La mente ha de estar estructurada de un modo muy complejo, dividida de modo reiterado en porciones públicas y privadas, con complicadas interacciones entre las subsecciones.[92]

[91] Ya hemos señalado en la sección precedente que estudios posteriores realizados por diversos psicólogos científicos, muestran que la mayor parte de estos índices son totalmente erróneos, y responden más bien a mitos infundados. [cf. Masip (2005), pp. 88-89; Vrij (2000)]

[92] «Of course it must be advantageous for the truth to be registered somewhere, so that mechanisms of self-deception are expected to reside side-by-side with mechanisms for the correct apprehension of reality. The mind must be structured in a very complex fashion, repeatedly split into public and private portions, with complicated interactions between subsections.» [Trivers (1985), p 416]

Este hecho es el que ha propiciado que, desde el punto de vista teórico, la noción de autoengaño haya "invitado a una reflexión vaga". [Trivers (1985), p. 416]. Han sido Ruben Gur y Harold Sackeim quienes, en opinión de Trivers, demostraron en un "brillante artículo" —que analizamos en la anterior sección— que pueden satisfacerse los 3 criterios[93] suficientes para el autoengaño:

(a) Creencias contradictorias simultáneas.

(b) Inconsciencia de la información verdadera y consciencia de la falsa.

(c) Carácter motivacional del proceso.

Los mecanismos que incluye el autoengaño según Trivers son los de beneficiación (*beneffectance*), exageración, ilusión de consistencia, percepción de las relaciones, defensa perceptual y vigilancia perceptual.

La beneficiación hace referencia a la tendencia a representarnos como beneficiosos y eficaces a la vez. De este modo, tendemos a exagerar nuestro rol cuando los resultados de una actividad son positivos y a negar el daño o responsabilidad en caso contrario (por ejemplo, cambiamos de un discurso en modo activo a un modo pasivo o a usar pasivas reflejas para escurrir el bulto). Además de negar o minimizar su responsabilidad, los sujetos suelen tender a reescribir su pasado de modo que los hechos del pasado resulten consistentes con lo que ahora hacen o defienden, lo

[93] Gur y Sackeim hablan de 4 criterios, pero se debe a que Trivers aúna el primer (creencias contradictorias) y el segundo (creencias simultáneas), en la clásula (a).

que les da una imagen de haber cometido pocos errores. En las relaciones personales, ambos individuos suelen estar de acuerdo en que hay un egoísta y un altruista, pero no se ponen de acuerdo en quién es cada uno: los individuos suelen verse a sí mismos como altruistas, mientras el otro u otros son egoístas.

Por otro lado, el sujeto tiende a evitar aquello que le produce estrés o angustia (defensa perceptual), y está siempre vigilante para ir recolectando las pruebas a favor de sus deseos o creencias como si fuera un abogado preparando su caso o un estudiante haciendo su tesis (vigilancia perceptual). Estos son los modos principales de autoengaño para Trivers [Trivers (1985), pp. 418-420]

QUINTA PARTE

V. LA FORMACIÓN DE CREENCIAS

El autoengaño le resulta sumamente familiar a cualquiera, al menos en el sentido de que las atribuciones —principalmente a otros— de tal estado no son en modo alguno infrecuentes. Como es natural, el hecho de que en muchas ocasiones tal atribución pueda deberse bien a un uso poco fino y confuso del término (donde la gente acostumbra a mezclarlo con otros fenómenos, sobre todo con la ceguera intelectual, el pensamiento desiderativo y la debilidad de la voluntad), bien simplemente a un error en la atribución de estados o actitudes mentales, no ha de empujarnos de antemano a negar y disolver apresuradamente el fenómeno.

No obstante, efectivamente nos parece que algo extraño ha de estar sucediendo cuando no somos capaces de explicar la actitud de un individuo —o la nuestra propia— de un modo coherente y racional. En este tipo de situaciones, a veces, atribuimos autoengaño.

En el capítulo introductorio realizamos el trabajo de exponer en qué *no* consistía el autoengaño, y se analizaron problemas y prejuicios cognitivos aparentemente similares con el objetivo de afinar la caracterización que propondremos ahora. No obstante, se hace necesaria, en primer lugar, la exposición de unas cuestiones y dificultades previas.

En este sentido, es evidente que en tanto que el autoengaño tiene relación con la adquisición de creencias, cualquier enfoque que pretenda ser adecuado habrá de estudiar cuidadosamente la naturaleza de la *creencia*. Como es bien sabido, la noción de creencia ha constituido uno de los asuntos más controvertidos a lo largo de la tradición filosófica, desde Platón hasta el presente. Tanto su naturaleza como la relación ontoepistémica con otros conceptos tales como los de *conocimiento* y *verdad* han sido objeto de interminables discusiones e intentos de clasificación. Sin embargo, el presente trabajo no puede permitirse realizar un estudio del devenir histórico del concepto —lo cual supondría una tarea hercúlea— ni ensayar siquiera un estudio a fondo de las implicaciones, objeciones y posibles soluciones de *un* enfoque particular. Más bien, trataré de exponer ciertos elementos y presupuestos necesarios para centrar la discusión.

Desgraciadamente, tan pronto como uno cree inocentemente haber dado cuenta de estos asuntos de un modo medianamente consistente, otro tipo de dificultades aparecen en el horizonte en forma de amenazadores nubarrones. Por esto, ha de tenerse en cuenta desde un primer momento una serie de cuestiones íntimamente relacionadas en una compleja maraña conceptual.

Por ejemplo, si uno acepta que la evidencia desempeña algún papel en la formación de creencias, la propia noción de "evidencia" —que se suele usar de una manera bastante neutra y acrítica— comporta, a nada que excavemos, interesantes problemas que afectan de manera crucial a

nuestra explicación del modo en el que se forman las creencias. Lo mismo ocurre con cuestiones como qué es y qué comporta la *racionalidad*, tanto en la formación de creencias como en su atribución; si la interpretación es una cuestión regional o ha de adoptarse un plan similar al de la *interpretación radical;* o cuáles son los límites del *Principio de caridad.* Todos ellos son asuntos que necesitan aclaración. Igualmente, en esta maraña tienen especial relevancia las nociones de *razón*, *causa*, *causa no-desviada*, *motivo*, *intención* o *propósito*, entre otras muchas.

No obstante, por si no fuera suficiente, hay una dificultad mayor aún; como señaló Schopenhauer, el problema de la *consciencia* es, quizá, "el nudo del mundo". La explicación del problema de la ntauraleza de la mente y la consciencia ha sido, especialmente desde la Edad Moderna, la piedra de toque de casi todo planteamiento filosófico. En los últimos años, sin embargo, algunos enfoques cientificistas, materialistas o eliminativistas han reivindicado la neurofisiología como el verdadero y único modo de atacar honestamente la cuestión. En este sentido, estos autores niegan la existencia de estados mentales como algo distinto de ciertos estados físico-químicos del cerebro. Éste no es el lugar tampoco para establecer una discusión acerca del eliminativismo, el funcionalismo o el emergentismo. Damos más bien por sentada la cuestión: los estados mentales son algo distinto de los procesos físico-químicos del cerebro, aunque evidentemente estén causados por ellos. La razón más importante —aunque no la única— es que, a nuestro juicio, es bastante evidente que la posesión y (auto)atribución de tales estados mentales es lo que nos hace ser el tipo de seres que somos y, si fuesen substituidos por algo, habría de

tener las mismas propiedades que tienen los llamados estados mentales. Un giro de este tipo constituiría, efectivamente, un mero cambio nominal y nunca substancial en la cuestión.

Ahora bien, aceptar que existen estados mentales sólo supone el inicio de una larga serie de problemas adyacentes. ¿En qué consiste entonces tener un determinado estado mental? Y la consciencia, ¿es una propiedad mental *reflexiva*? ¿Es la mente *transparente* y es, por tanto, el sujeto *infalible* con respecto a la autoatribución de estados mentales? Si no lo es, ¿tiene al menos un acceso privilegiado a sus propios estados mentales frente al que tienen otros? Si lo tiene, ¿qué significa que al no ser infalible a veces pueda equivocarse en la autoatribución? ¿Hay condiciones bajo las que es más sencillo equivocarse?; si no tiene tal *acceso privilegiado*, ¿a qué se debe? ¿Hemos de suponer que hay alguna región necesariamente inaccesible? Finalmente, si concedemos tal cosa, ¿qué virtudes y qué limitaciones explicativas tendría la existencia de un *inconsciente,* inaccesible por hipótesis?

Trataremos de fijar todas estas cuestiones en el siguiente capítulo con el fin de hacer una propuesta explicativa del autoengaño.

V.1 - La naturaleza de la creencia

> A wise man proportions his belief to the evidence.
>
> [David Hume (1748), SB 110]

> We cannot strictly be said to believe without evidence: what is so described is not belief but something else.
>
> [Henry H. Price (1973), p. 140]

Una creencia puede presentarse de dos modos: uno que podemos llamar *original* y otro *derivado*. En su sentido original, una creencia es la *ocurrencia* (*ocurrence*) de un estado mental que consiste en un *sentimiento* más o menos fuerte con respecto a la verdad de un determinado hecho, sentimiento provocado por la acumulación de evidencia significativa. Una creencia es, por tanto, un estado mental con contenido.

En su sentido derivado, la creencia tiene un carácter *disposicional*. Por supuesto, uno puede tener creencias aun cuando no estén presentes en su mente en todo momento, y ciertamente la mayoría de nuestras creencias son de este tipo. Así, uno no sólo tiene una creencia cuando tiene una ocurrencia de un determinado estado mental, sino también cuando bajo ciertas circunstancias lo tendría. Concretamente, cuando estuviese dispuesto a afirmar que *p* es el caso.

Por tanto, un sujeto tiene la creencia de que *p* es el caso, si y sólo si:

a) Tiene un estado mental consistente en un fuerte sentimiento de que *p* es el caso apoyado en evidencia que cree significativa y, por tanto, está dispuesto a afirmar sinceramente la verdad de *p*

O bien,

b) *Siempre que* pensase en la posibilidad de que *p* sea o no el caso, tendría un fuerte sentimiento de que *p* es el caso apoyado en evidencia que creyese significativa y, por tanto, estaría dispuesto a afirmar sinceramente la verdad de *p*.

Esta segunda caracterización de lo que es una creencia tiene consecuencias a primera vista extrañas, como el hecho de que parece que todos nosotros, hispanohablantes competentes, compartiríamos la creencia de que "Venus es más grande que una zapatilla", aun cuando seguramente nunca hemos pensado en establecer una relación de ningún tipo entre estos dos objetos hasta ahora. Sin embargo, da cuenta tanto de todas las creencias que hemos alcanzado y mantenemos aunque no pensemos en ellas en todo momento, como del uso común que hacemos del término y de nuestra capacidad *potencial* de alcanzar creencias.

Otro aspecto importante es que, aunque las creencias son en cierto modo un proceso automático y no-inferencial, pueden ser objeto de un análisis ulterior haciendo uso de distintos métodos, como la *teoría de la probabilidad subjetiva*.

Cuando sostengo que son automáticas y no-inferenciales, quiero decir —como trataré de explicar más adelante— que el sujeto no controla su

formación y que son algo que, al margen de que puedan estar equivocadas, uno *sabe* que tiene (o *puede saber que tiene*, si no son objeto de reflexión en ese momento) de un modo directo; o en otras palabras: no es necesario que uno haga inferencias para descubrir que las tiene. Esto no significa que no se puedan alcanzar creencias por medio de pasos inferenciales a partir de un conjunto de evidencia u otras creencias previas. Sin embargo, en los casos en que un sujeto no se da cuenta de las consecuencias que se siguen de los postulados o creencias que sostiene, no diremos que realmente estas inferencias se hallan entre el repertorio de sus creencias *implícitas* o *inconscientes;* diremos que el sujeto no cree estas consecuencias en modo alguno. Las creencias que el sujeto asume implícitamente quedan reducidas a las creencias disposicionales.

Una consecuencia que se sigue de esto, es que el sujeto puede tener creencias contradictorias sin percatarse de ello. Efectivamente, si uno no extrae las consecuencias de todas sus creencias, puede permanecer ignorante de la inconsistencia del conjunto; sus creencias no son contradictorias *per se*, pero implican cosas contradictorias y, por tanto, son inconsistentes. No obstante, esto no es un problema serio para ninguna teoría de la racionalidad modesta, ya que es evidente que los seres humanos no son omniscientes y cometen diversos errores interpretativos, inferenciales, etc.

Así, el sujeto no necesita hacer un análisis introspectivo extravagante para descubrir que tiene una creencia concreta; pero tampoco suele, por lo general, hacer un análisis técnico o matemático de la evidencia al formarse

una creencia. Sin embargo, como ya señalábamos, esto ni impide ni desaconseja la propuesta de modelos matemáticos o lógicos que nos permitan tratar de comprender mejor los procesos de formación de creencias. En este sentido, el proceso de formación de la creencia de que *p* es el caso puede caracterizarse como la *probabilidad subjetiva* que un sujeto le asigna en un momento *t* a que *p* sea el caso dada la evidencia E que posee. Bajo este análisis, un sujeto cree que *p* si y sólo si considera que la probabilidad subjetiva de que *p* sea el caso dado *E* (siendo *E* el conjunto de evidencia que este sujeto cree relevante para la verdad o falsedad de *p*) es *bastante alta.*

La pregunta que surge inmediatamente es qué puede significar que la probabilidad que asigna sea "bastante alta". Aquí hay dos cuestiones importantes que matizar. En primer lugar, una relacionada con los grados de creencia; en segundo lugar, una que tiene que ver con las llamadas "políticas de creencia" [Helm (1994)].

En cuanto a la primera de ellas alguien podría sostener que, en cierto sentido, la posesión de una creencia no es cuestión de grados: o se tiene o no se tiene; otro puede replicar que las creencias pueden ser más o menos fuertes. Pero estas posturas no tienen por qué estar en absoluto en oposición, en tanto que es plausible imaginar que quienquiera que abrace una creencia ha de tener al menos cierto grado de convicción de que *p* es el caso. Si lo alcanza, tendrá la creencia; si no lo alcanza, no. Pero esto no significa que una vez alcanzado tal umbral no pueda acumular más evidencia y convencerse más de que realmente *p* es el caso. En este

sentido, parece evidente que las creencias son graduales, al depender de la fuerza que tenga para el sujeto *el sentimiento* de que *p* es el caso.

Por esta razón ha habido varias propuestas para analizar las creencias bajo distintos modelos explicativos en términos de grados o de asignación de probabilidades subjetivas en un rango de 0 a 1, siendo 0 la creencia en la imposibilidad de que *p* sea el caso, y 1 la seguridad de que lo es. Los valores pueden ser desde tres (0/0,5/1) hasta infinitos, entendiendo los posibles valores entre 0 y 1 como un continuo.

En cuanto a la segunda cuestión, las *políticas de creencia* suponen a su vez dos asuntos importantes que han de investigarse:

a) qué tipo de cosas son las que uno acepta como evidencia de que *p* es el caso

b) qué grado de evidencia —si es que hay tal cosa— necesita uno para creer que *p* es el caso.

En función de la política adoptada, es posible que dos sujetos difieran en la aceptación o no de una creencia determinada sobre la misma base evidencial.

Con respecto al primer asunto, no es inconcebible que, por ejemplo, un sujeto considere que el testimonio de desconocidos no es fiable, y por tanto estime que no constituye evidencia de nada (pues evidencia no fiable es una contradicción en los términos). Otro quizá considere que las *Escrituras* son una fuente fiable, mientras un tercero puede juzgar que las *Escrituras* son meros mitos, a la vez que acepta como evidencia indubitable

la información ofrecida por los espacios informativos de la televisión. Tampoco es inconcebible que un cuarto sólo acepte como evidencia fiable aquello que puede ver, oír y tocar, y así sucesivamente.

El segundo asunto tiene que ver con el grado de evidencia que uno necesita para aceptar la creencia de que *p* es el caso. Sin duda, la respuesta a esta cuestión dependerá en buena parte del *carácter* del sujeto: unos serán más cautos y otros más osados. O, para decirlo en términos de James, unos aceptarán el principio de alcanzar verdades y otros el de minimizar errores. Cualquier elección entre estos principios tiene, a buen seguro, un factor *pasional o de carácter,* pero hay otro factor a tener en cuenta: la carga emocional de la cuestión relacionada con el objeto de creencia. De este modo, uno quizá eleva los niveles de escrutinio cuando la cuestión que está considerando tiene importantes implicaciones en su vida. Como vimos que señalaba Alfred Mele, los umbrales para la aceptación de una determinada creencia no son fijos, sino que dependen de los costes emocionales que se siguen de equivocarse en una cuestión importante.

En cualquier caso, parece que el grado de evidencia necesario para que un sujeto crea que *p* es el caso ha de ser significativo, a saber: *como mínimo*, suficiente para que el sujeto considere que *es más probable que p sea el caso que que no lo sea*. Esto sin perjuicio de que en algunas ocasiones el sujeto pueda elevar los niveles de escrutinio debido a los costes emocionales, y pueda demandar así una cantidad de evidencia *mayor* para aceptar una determinada creencia. Esta demanda de más evidencia no consistiría en una mera acumulación de datos insustanciales, sino en el acopio de datos

relevantes que redundasen en el aumento del grado de apoyo para la verdad de un estado de cosas concreto.

Lo que me interesa subrayar es que el grado de apoyo nunca podrá ser *menor*. Es decir, si un sujeto cree que es más probable que *p no* sea el caso, no podrá creer que *p* es el caso. En esto consiste, entre otras cosas, que las creencias aspiren o apunten a la verdad. Sin duda, las creencias pueden estar equivocadas, pero parece que el sujeto no puede ser consciente de que sus creencias no apuntan a la verdad sin abandonarlas en ese mismo momento.

Además, ha de tenerse en cuenta que la probabilidad subjetiva de que *p* sea el caso y la probabilidad subjetiva de que no lo sea no tienen por qué ser complementarias. Es decir, la suma de las probabilidades subjetivas asignadas no tiene por qué ser igual a 1. Esto se debe, obviamente, a que el individuo puede considerar —acertadamente o no— que carece de información relevante. Quizá le otorga una probabilidad subjetiva de 0,4 a que *p* sea el caso, y una probabilidad subjetiva de 0,3 a que no lo sea. De nuevo, el sujeto sólo puede creer que *p* es el caso si le asigna una probabilidad subjetiva *superior* a 0,5, dado el conjunto de evidencia *K* de que dispone el sujeto en un tiempo *t*.

No hace falta insistir en que este tipo de análisis son hipótesis explicativas posteriores, ya que quizá sólo sujetos que están muy prejuiciados (filósofos, psicólogos o científicos profesionales) hacen, en contadísimas ocasiones y bajo circunstancias muy particulares, tal asignación de valores. Por esta razón, aunque no niego que estas

explicaciones *a posteriori* puedan tener algún tipo de valor, resulta difícil ver cómo podrían corresponderse con aquello que hace un sujeto —si es que hace algo— cuando cree que *p* y, en todo caso, no parecen prometedoras para la cuestión que nos ocupa, a saber: indagar si es posible que un sujeto tenga creencias contradictorias.

Quizá pueda resultar desorientador que las creencias se hayan considerado un tipo de "acto mental" o "actitud proposicional". En mi opinión, el sujeto que cree algo no *hace* nada al creerlo. Ni calibrar o ponderar evidencia para asignar probabilidades, ni nada similar. Aún más: estoy plenamente de acuerdo con la idea humenana según la cual las creencias son algo que uno *padece*, antes que algo que uno *hace*. El sujeto que cree que *p* padece un sentimiento acerca de la verdad de *p* como resultado de la evidencia de que dispone: esta evidencia le inclina a considerar que (muy) probablemente *p* es el caso. De este modo, las creencias suponen un *asentimiento* a una determinada cognición, pero este asentimiento no es algo que el sujeto *haga* o decida hacer intencionalmente. No hay un acto mental por medio del cual uno adopte una creencia como quien se pone un sombrero.

En este sentido, los análisis en términos de grados de creencias y de análisis de atribución de probabilidad subjetiva o de grados de posibilidad quizá sean efectivos una vez que se fijan ciertos parámetros, pero en el examen de aquello en lo que consiste tener una creencia, me parecen totalmente inútiles por desorientadores y oscuros. Desorientadores porque tras conducir la discusión a otro terreno, no nos proporcionan

ningún avance y se quedan más bien en un mero replanteamiento de la cuestión. Oscuros, porque incluyen elementos difícilmente cuantificables, como grados de probabilidad subjetiva, pesos de evidencia, etc. Dicho de otro modo: cuando hablamos de sujetos concretos y sus creencias, ¿cómo puede el evaluador afirmar que este individuo le atribuye a la verdad de que *p* una probabilidad subjetiva de 0,75 más bien que una probabilidad de 0,8, 0,6 ó 0,72? ¿Quién tiene la balanza mágica en la que pesar el hecho de que una mujer llega tarde 8 días sin dar una explicación coherente, como evidencia de que es infiel?

Por lo demás, aunque pueda pensarse lo contrario en un primer momento, parece que acudir en estos asuntos a cantidades difusas no soluciona nada. Aunque uno pueda inclinarse a pensar que asignar valores difusos puede captar el modo difuso en que el sujeto pondera cierta la evidencia, no facilita nuestra investigación. Seguramente no resulta mucho más sencillo hacer valoraciones del peso "difuso" que podría tener una pieza de evidencia específica y, en todo caso, parece enrevesado y forzado suponer que el sujeto asigna probabilidades difusas a la verdad de que *p* es el caso tras ponderar difusamente el peso que una evidencia determinada tiene para la verdad de que *p* es el caso.

En resumen, las creencias son sentimientos —que pueden ser más o menos fuertes— acerca de la verdad de un hecho, y están causados y sustentados por evidencia que de algún modo consideramos significativa (es decir, relevante y suficientemente abundante). En la creencia de que *p*, la evidencia no sólo es causa, sino *razón*, para aceptar la verdad de que *p*,

esto es: no sólo la causa, sino que la explica. Un corolario de esto es que no todo sentimiento acerca de la verdad de *p* es una creencia; a veces podemos tener un sentimiento fuerte de que *p* es el caso en forma de corazonadas, intuiciones, deseos o esperanzas. La diferencia entre estos estados mentales y las creencias sería que ninguno de ellos está causado por la evidencia de la que dispone el sujeto (aun cuando disponga de ella) y menos aún, se explica por ella. Naturalmente, una actitud de este otro tipo puede estar incentivada por la posesión de cierta evidencia, y puede llegar a convertirse en una creencia. Sin embargo, sólo si el sujeto siente que la evidencia hace más probable que *p* sea el caso que que no lo sea, el sujeto posee una creencia.

Otra cuestión que ha de examinarse es en qué consiste tener *evidencia* de que algo es el caso.

V.2 - La evidencia

El concepto de evidencia es, a mi juicio, uno de los puntos más oscuros en la mayoría de los planteamientos generales acerca de la adquisición de creencias y su justificación y, por ende, del autoengaño. No obstante, no pretendo desarrollar aquí un estudio en profundidad de la cuestión, ni hacer una propuesta sumamente novedosa y reveladora. Antes bien, me conformo con señalar ciertas cuestiones que, aunque son en cierto modo comúnmente asumidas, muy a menudo parece que se olvidan y no se toman en cuenta o se simplifican de tal manera que desorientan por completo la investigación.

La más importante de estas cuestiones tiene que ver con la naturaleza de la evidencia. Al atacar la cuestión, la primera dificultad aparece cuando examinamos el significado corriente del término. "Evidencia" proviene directamente del vocablo latino *evidentia* y se refería en un principio a la propiedad que tiene aquello que se ve, que se capta directamente. Así, lo "evidente" no necesita demostración alguna, ya que se tiene a ello un acceso directo e infalible, por tener la propiedad de la evidencia.

Sin embargo, tanto en castellano como en otros idiomas (p. ej. en inglés, pero no así en francés), el término adquirió además otro significado y acabó designando también a todo aquello que puede ser aducido a favor de una hipótesis, aquello que, amontonándose, acabaría por hacer evidente algo: un *indicio* o *señal* de algo; destino similar ha tenido la palabra "prueba". Así, mientras "evidente"/"evident" conservan su sentido original, "evidencia"/"evidence" han pasado a significar tanto la propiedad que tiene aquello que es claro y distinto y no necesita de prueba, como "indicio" o "señal" para la demostración de que algo es el caso. Para deshacer esta ambigüedad, en inglés se usa a veces el término "evidentness", que designa únicamente la propiedad de ser evidente, pero en castellano carecemos de un término similar.

El problema es que el hecho de que la palabra tenga ese doble significado, hace que a veces se mezclen. Así, en algunas ocasiones algo que en realidad no es más que un indicio o señal de que algo es el caso, se toma como un hecho *evidente* que no ha de cuestionarse. Parece entonces,

que si algo es evidencia, ha de ser evidente para todos. Nada más lejos de la realidad: la evidencia es siempre evidencia *para alguien*.

Para aclarar esta cuestión, es necesario decir algo acerca de la verdadera naturaleza de la evidencia en tanto que indicios o señales que apoyan la verdad de un determinado estado de cosas. Sin entrar en debates epistemológicos profundos acerca de nuestro modo de conocer, propondré, con el objetivo de ser lo más claro posible de aquí en adelante, unos supuestos de partida.

Tengo una concepción realista con respecto a los hechos. De este modo, los hechos son lo que quiera que sea el caso, independientemente de que haya un sujeto que lo conozca o lo capte; independientemente del modo en que lo capte y de que pueda estar descaminado al hacerlo. Por supuesto, nuestro modo de captar los hechos del mundo, las proposiciones que los expresan y las proferencias de estas proposiciones también son hechos. Los seres humanos tenemos diversos modos de acceder al mundo y de obtener información acerca de lo que es el caso. Parece claro también que nuestro modo de captar el mundo es limitado, ya que es parcial y *falible*. Sin embargo, es *bastante fiable*. Es decir, aunque a veces cometemos errores, la información que captamos suele mostrarse como verdadera, al menos en el sentido de que

a) Compartimos instrumentos de medida (allí donde se puede medir)

b) parece difícil que por casualidad estemos bastante de acuerdo entre nosotros en los resultados obtenidos, y

c) estos resultados no parecen tampoco una mera alucinación colectiva, ya que se revelan útiles en la resolución de tareas y problemas cuando operamos con el mundo.

Por estas razones, tenemos la convicción de que la falibilidad es parcial, y el error masivo no parece posible. Además, un sujeto cuyas creencias fuesen mayoritariamente falsas resultaría ininteligible.

Sin embargo, el grado de fiabilidad de nuestra información acerca de lo que es el caso depende en buena parte de la naturaleza del objeto. Hay hechos de los cuales podemos obtener *evidencia ponderable*, mientras otros hechos sólo nos proporcionan *evidencia imponderable*.

Llamo *datos* a los resultados que obtenemos cuando tratamos de obtener evidencia ponderable de algo que es el caso. Son datos, por ejemplo, la temperatura de un cuerpo medida por un termómetro o la presión de un balón medida por un manómetro; pero también es un dato que mi perro está en el jardín jugando con mi vecino, que su mujer está mientras con el entrenador personal o que la bolsa ha bajado dos puntos. Por supuesto, estos datos pueden ser erróneos, en tanto que nuestros sentidos pueden engañarnos y los dispositivos diseñados para medir determinada propiedad pueden estar mal calibrados o estropeados. Lo que me interesa subrayar es que los resultados obtenidos son generalmente fiables en tanto que pueden ser intersubjetivamente compartidos y repetibles bajo las mismas condiciones.

Sin embargo, un dato es a menudo *insuficiente* sin una interpretación. No hace falta insistir en que esta interpretación no es una valoración arbitraria

ni una interpretación en el sentido adoptado por la tradición hermenéutica; más bien tiene el carácter de la interpretación tal y como es entendida en el experimento mental de la interpretación radical quineano o davidsoniano: el individuo que recoge datos, los interpreta de modo que resulten coherentes teniendo en cuenta un conjunto previo de hechos, creencias, deseos, proyectos, prejuicios, etc. que considera relevantes. Mi objetivo, por tanto, es simplemente señalar que los datos desnudos son "ciegos". Por ejemplo, puedo tomarme la temperatura corporal con varios termómetros y cerciorarme de que los datos que obtengo son fiables: tengo 38 ºC. Pero este dato no me dice mucho. Asumiendo que sepa de antemano cuál es la temperatura normal de mi cuerpo, me dice únicamente que mi temperatura es más alta de lo que se consideran los valores normales; pero nada más. Un médico podría, a la vista de más datos que yo también veo pero no presto atención, y él interpreta como síntomas, *interpretar* a su vez todos esos datos como *evidencia* de que tengo tal o cual enfermedad. En cambio, un sujeto sin conocimientos de medicina también podría interpretar ese dato como una mera fiebre sintomática de una gripe, y equivocarse gravemente. Tanto para el médico como para el lego, el dato (los 38 ºC) es el mismo, y para ambos también este dato supone *evidencia* de algo; sin embargo, el dato por sí solo no constituye evidencia de nada; es el dato interpretado lo que tomamos como evidencia. Del mismo modo, cuando en 1648 Pascal envió a su cuñado Florin Périer a medir en tres puntos de altitud diferente del Puy-de-Dôme el nivel que alcanzaba la columna de mercurio en un barómetro, observó varios datos: la columna era menor en la cima; era igual en varios

puntos de la cima que estaban a la misma altura aproximadamente, y a medida que descendía del volcán, la columna ascendía. Esos datos podrían haber sido evidencia de distintas hipótesis; sin embargo, Pascal conjeturó que suponía evidencia de que lo que presionaba el mercurio y hacía subir la columna era un mar de aire sobre nuestras cabezas; el espesor de este mar de aire había de ser lógicamente menor en la cima del volcán que en la base. La evidencia es siempre evidencia *para alguien*.

Sin duda, los datos recogidos por Florin Périer podían haber sido erróneos. Por ejemplo, podría haber calculado mal la altura a la que hacía las mediciones o los tubos podían haber estado mal sellados. Sin embargo, lo que es importante subrayar es que los datos, al margen de ser falibles, infradeterminan la teoría que ha de explicarlos y, por tanto, no son evidencia en sí mismos, sino que han de interpretarse bajo una hipótesis general para que constituyan evidencia de algo.

Esto afecta de modo especial a la evidencia obtenida de acciones humanas. Por supuesto, las acciones humanas son hechos del mundo, y, en tanto que hechos, podemos captarlos como datos; por ejemplo, podemos ver a la esposa de nuestro mejor amigo pasear con cierto individuo extraño para nosotros. Eso en principio no es más que un dato (falible además, ya que podemos haberla confundo con otra mujer), y no constituye evidencia de nada. En este sentido, sólo podemos decirle con honestidad a nuestro amigo: "Vi a tu mujer con otro hombre. No sé qué puede significar, pero que estaban dando un paseo juntos es un hecho". Es cuando tratamos de valorar ese dato y lo interpretamos bajo una

hipótesis particular, cuando pasa a constituir evidencia de algo, por ejemplo: "Tu mujer anda flirteando con otro hombre".

Por supuesto, es obvio también que algunos datos dejan más holgura interpretativa que otros. Si decido seguir a la pareja y veo que el paseo con el misterioso individuo acaba en casa de éste, el nuevo dato tiene ya menos explicaciones posibles que el simple paseo que había presenciado anteriormente. Sin embargo, hay aún cierta holgura: quizá nuestro misterioso hombre sea un familiar, o un amigo de la pareja que yo desconozco y que planea la fiesta de cumpleaños a mi amigo. Pero si acercándome a la casa descubro que están de algún modo enredados, los datos no ofrecen ahora margen alguno como evidencia de que esta mujer le está siendo infiel a mi mejor amigo (siempre y cuando —y esto también es importante— los términos de su relación no permitan las relaciones extra-maritales).

Lo que hace más problemática la interpretación cuando tiene relación con acciones humanas, es que a menudo incluye elementos tanto de evidencia ponderable como imponderable. Los datos a los que nos hemos referido hasta ahora corresponden al tipo de evidencia ponderable; es decir, existen para ellos —aunque nosotros no dispongamos de ellos— patrones interpretativos. Una vez que fijamos unos criterios y parámetros, los datos obtenidos son evidencia en cierto modo objetivada para la verdad de algo. Que no dispongamos de estos datos, bien porque en ese momento no tengamos a mano un instrumento para obtenerlos, porque

no sepamos cómo obtenerlos o porque ni siquiera sepamos que ahí hay un dato relevante que obtener, no quiere decir que éstos sean imponderables.

La evidencia imponderable tiene que ver en cambio con datos que *de iure* no tienen una explicación precisa o estipulada. Por ejemplo, una determinada expresión facial (de dolor, de placer, de desaprobación, de angustia, etc.). Siguiendo con nuestro ejemplo, quizá mientras presenciábamos el paseo, vimos ciertos gestos sutiles de cariño que resultan imponderables, pero que interpretamos como evidencia de que el cariño no era meramente fraternal.

De hecho, buena parte de lo que supone desarrollarse como individuo consiste en aprender a captar lo que este tipo de cosas significa, aún sin disponer de un manual de "geometrías de la cara" que nos lo explique; no seríamos la clase de seres que somos si no tuviéramos esta capacidad "estética" de tomar en cuenta evidencia imponderable.

Hasta aquí la primera dificultad con respecto a la evidencia: tanto la evidencia ponderable como la imponderable requieren la interpretación de unos datos que suelen presentar holgura e infradeterminan su explicación.

El segundo problema está relacionado con el anterior, aunque es distinto: supongamos que nuestra interpretación de los hechos es adecuada. Nuestro problema ahora surge una vez superadas las dificultades de la interpretación: ¿cómo saber qué peso tiene cada pieza explicativa dentro del conjunto de evidencia? Parece claro que el valor tiene que ver, por un lado, con la holgura explicativa que permita (a menor

holgura, mayor valor) y, por otro, con relaciones que mantiene con el resto de lo que tomemos como evidencia. Sin embargo, no nos es posible otorgar valores claros de este peso e, incluso como señalamos en la sección anterior, determinados prejuicios cognitivos pueden interferir en la valoración que hagamos de determinados datos. Así, los costes emocionales que se siguen de la interpretación de ciertos datos como evidencia en apoyo de algo emocionalmente perturbador para el sujeto, pueden interferir en dicha valoración.

En resumen, la evidencia de que *p* es el caso no consiste, por tanto, en una mera acumulación de datos acerca del mundo. Uno tiene que valorar estos datos e *interpretarlos como datos que apoyan la verdad de que* p *es el caso*. Esta valoración no es algo que se haga al margen de prejuicios, costes emocionales y otras consideraciones previas. La evidencia no es nunca, por tanto, un conjunto de datos claro, medible y objetivo.

Ésta es la razón por la que en el estudio de la acción humana, la teoría de la probabilidad que se ha usado en ciencias naturales no sirve, y se han buscado explicaciones alternativas, como la teoría de la posibilidad, que intenta tratar con esta imponderabilidad asignando valores difusos. Pero la dificultad no reside únicamente en que no se puedan dar valores claros o discretos; aunque tengan ciertas virtudes, ya expresé mi falta de esperanza con respecto a esta clase de enfoques, precisamente debido tanto a la imposibilidad de valorar cuánto apoya una determinada pieza de evidencia a la plausibilidad de que un sujeto intente, piense, desee, sospeche, haga o crea algo, como al carácter *esencialmente imponderable* de cierto tipo de

evidencia como una determinada expresión facial de aprobación, alegría, rechazo, o dolor.

V.3 - Creencias y voluntad

¿Puede la voluntad desempeñar algún papel, directo o indirecto, en la formación de nuestras creencias? ¿Puedo formarme creencias al margen de mis consideraciones acerca de los hechos, al margen de mi evidencia? ¿Puedo manipular la evidencia a voluntad? Y, en caso de ser posible, ¿debería hacerlo?

En primer lugar, parece claro que la experiencia cotidiana nos indica que creer algo directamente, por las buenas, y al margen —o en contra— de la evidencia de la que disponemos, no es posible. No es verosímil —dejando a un lado trastornos mentales— que si mis cuentas están casi en números rojos, consiga llegar a creer, porque sí, que tengo 3 millones de euros en el banco. Si esto es así, debemos investigar a qué se debe; esclarecer si es algo contingente o necesario. Después de dilucidar ese asunto, hemos de ver otro género de casos menos extremos, en los que la pregunta interesante sería ya más bien la siguiente: ¿es posible que la voluntad interfiera *de algún modo* —quizá indirecto, dando un rodeo— en la formación de nuestras creencias?

Muchos teóricos han tratado de responder a esta cuestión. Es de sobra conocido —y ya lo hemos indicado con anterioridad— que David Hume, por ejemplo, defendió que las creencias son algo que nos sucede y que no

controlamos en absoluto. Hume pensaba que en la medida en que nuestro conocimiento no es infalible, la confianza o la certeza que tenemos acerca de lo que suponen los datos empíricos que obtenemos de la naturaleza pueden oscilar. En estas circunstancias, "un hombre sabio proporciona su creencia a la evidencia"[94] [Hume (1748), p. 228].

A la vista de esta afirmación uno podría pensar que, en tanto que la sabiduría consiste en proporcionar nuestras creencias a la evidencia, parece que cabría no hacerlo. ¿Podemos, entonces, creer algo sin tomar en consideración lo que nos impone con más fuerza la evidencia? Según Hume, todo apunta a que esto no es posible. En otro pasaje aclara que las creencias están fuera de nuestro control:

> Se sigue, por tanto, que la diferencia entre *ficción* y *creencia* reside en algún sentimiento o sensación que se añade a la última, no a la primera, y que no depende de la voluntad ni puede manipularse a placer. Ha de ser suscitado por la naturaleza como todos los demás sentimientos y ha de surgir de una situación particular, en la cual la mente se encuentra colocada en una coyuntura especial. Cada vez que un objeto se presenta a la memoria o a los sentidos, inmediatamente, por la fuerza de la costumbre, lleva a la imaginación a concebir aquel objeto que normalmente le está unido. Y esta representación es acompañada por una sensación o sentimiento distinto de las divagaciones de la fantasía. En esto sólo consiste la naturaleza de la creencia, pues, como no hay cuestión de hecho en la que creamos tan firmemente como para que no podamos imaginar su contrario, no habría diferencia entre la representación aceptada y la que rechazamos si no hubiera un

[94] «A wise man, therefore, proportions his belief to the evidence.» [Hume (1748), SB 110]

> sentimiento que distinguiese la una de la otra [Hume (1748), pp. 164-165].[95]

Cabría interpretar, pues, que lo que a primera vista parece un consejo que nos invita a proporcionar nuestras creencias a la evidencia no implica, al menos en cierto sentido, que podamos dejar de hacerlo. Si el pasaje resulta confuso, se debe a que no especifica a qué se refiere con "evidencia": si se trata de la *evidencia potencial*, la que está al alcance y el individuo podría obtener, o bien a la evidencia de la que ya dispone realmente. Creo que es bastante claro —o es más caritativo pensar— que Hume se refiere a la primera de ellas, a la evidencia disponible. Así, se trata más bien de un consejo epistémico referido a la cautela con la que uno debe molestarse en adquirir la mayor cantidad de evidencia relevante posible. De este modo, Hume advierte que quien presta poca atención a lo que es el caso, obtiene una evidencia parcial y, por consiguiente, se forma una creencia errónea. El sabio sería quien atiende con cuidado al mundo, y adecúa su creencia a la evidencia relevante disponible.

---

[95] «It follows, therefore, that the difference between *fiction* and *belief* lies in some sentiment or feeling, which is annexed to the latter, not to the former, and which depends not on the will, nor can be commanded at pleasure. It must be excited by nature, like all other sentiments; and must arise from the particular situation, in which the mind is placed at any particular juncture. Whenever any object is presented to the memory or senses, it immediately, by the force of custom, carries the imagination to conceive that object, which is usually conjoined to it; and this conception is attended with a feeling or sentiment, different from the loose reveries of the fancy. In this consists the whole nature of belief. For as there is no matter of fact which we believe so firmly that we cannot conceive the contrary, there would be no difference between the conception assented to and that which is rejected, were it not for some sentiment which distinguishes the one from a the other.» [Hume (1748), SB 48]

Pero no todos los autores han defendido que las creencias estén totalmente fuera de nuestro control. William K. Clifford publicó en 1876 un famoso artículo acerca de la formación de creencias titulado "La ética de la creencia", en el que defendía vehementemente la tesis de que hemos de ejercitar algún tipo de control sobre nuestras creencias, siendo escrupulosos en la exigencia de que toda creencia esté basada en la evidencia de la que dispone el sujeto para que sea considerada como *moralmente legítima*.

> La pregunta sobre lo correcto o lo erróneo tiene que ver con el origen de su creencia, no con el asunto sobre el que ésta versa; no con qué creencia sea, sino con cómo la alcanzó; no sobre si resulta ser verdadera o falsa, sino sobre si tenía derecho a creer de acuerdo con la evidencia en cuestión tal como ésta se le presentaba.[96]

Según Clifford, sabemos que la evidencia es una marca de verdad y, además, tenemos una obligación hacia otros miembros de la sociedad[97] que nos exige evitar formarnos creencias en ausencia de evidencia significativa favorable [Clifford (1876), p. 184], por lo que "es malo

[96] «The question of right or wrong has to do with the origin of his belief, not the matter of it; not what it was, but how he got it; not whether it turned out to be true or false, but whether he had a right to believe on such evidence as was before him.» [Clifford (1876), p. 178]

[97] Esta idea nos recuerda a la demanda kantiana en "Sobre un supuesto derecho a mentir por filantropía", donde sostiene que tenemos la obligación de decir siempre la verdad a causa del pacto social. No podemos mentir porque pondríamos en peligro los vínculos que hemos aceptado tácitamente, y que incluyen en su misma base la mutua confianza entre los miembros. Cualquier mentira introduciría un elemento podrido en el conjunto que legitimaría cualquier otra mentira subsiguiente.

siempre, en todo lugar, y para todos, creer algo sobre evidencia insuficiente".[98]

En esta misma línea cientifista, Peirce sostuvo un año más tarde que, aunque el mejor método para fijar nuestras creencias es el científico, porque sólo él tiene en cuenta los datos empíricos que afectan o podrían afectar a cualquier hombre [Peirce (1877), pp. 253-54], no es imposible creer a voluntad. En este sentido, afirma que por medio del "método de la tenacidad" algunos pueden adquirir, a la fuerza y *al margen de la evidencia,* una creencia deseada, y tratar de mantenerla hasta las últimas consecuencias, pase lo que pase. Peirce advierte que esta actitud produce un "éxito brillante pero poco duradero", y

> un hombre […] debería considerar que, después de todo, desea que sus opiniones coincidan con los hechos, y no hay razón por la que los resultados de estos tres primeros métodos[99] deberían hacerlo. [Peirce (1877), p. 256][100]

En cualquier caso, ni Peirce ni Clifford proclaman que sea imposible creer a voluntad y al margen de la evidencia. Simplemente sostienen que

---

[98] «To sum up: it is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence.» [Clifford (1876), p. 186]

[99] Se refiere al método de la tenacidad, el de la autoridad y el *a priori*, como opuestos al método científico. Todos ellos serían falibles en la búsqueda de la verdad, y poco fiables si uno quiere formarse creencias adecuadas en tanto que ajustadas a la realidad.

[100] «A man […] should consider that, after all, he wishes his opinions to coincide with the fact, and that there is no reason why the results of those three first methods should do so.» [Peirce (1877), p. 256]

es inmoral: "evitar echar un vistazo al apoyo de cualquier creencia por miedo a que pueda estar podrido es tan inmoral como desventajoso."[101]

Como es de sobra conocido, William James respondió a Clifford veinte años más tarde en un artículo también famoso titulado "La voluntad de creer". Aunque, como el propio autor se vio obligado a reconocer tras múltiples malentendidos, el título pueda resultar desorientador, James simplemente quería defender la legitimidad de la fe voluntariamente adoptada. Admite que, obviamente, siempre que nos sea posible basar nuestras creencias sobre evidencia empírica, sería *irracional y ridículo* que la voluntad entrase en juego. Sin embargo, a veces estamos ante una opción genuina, esto es, una opción viva, forzosa e importante para la que por la naturaleza del asunto no tenemos evidencia que nos indique en qué creer. En estos casos,

> […] *nuestra naturaleza pasional no sólo legítimamente puede, sino que debe, optar entre propuestas […] pues decir, en tales circunstancias, "no decida usted, deje la cuestión abierta", es en sí mismo una decisión pasional —exactamente lo mismo que decidir sí o no— y comporta el mismo riesgo de perder la verdad.* [En cursivas en el original][102]

[101] «... to avoid looking into the support of any belief from a fear that it may turn out rotten is quite as immoral as it is disadvantageous.» [Peirce (1877), p. 257]

[102] «[…] *our passional nature not only lawfully may, but must, decide an option between propositions, […] for to say, under such circumstances, "Do not decide, but leave the question open," is itself a passional decision,—just like deciding yes or no,—and is attended with the same risk of losing the truth*.» [James (1896), p. 11]

De acuerdo con James, hay dos modos de atender nuestro deber en cuestiones de opinión: *Debemos conocer la verdad* y *debemos evitar el error*. Éstas son dos leyes materialmente diferentes y mientras unos prefieren la primera, otros la segunda. Ninguna de ellas es más racional: la elección entre ambas es una cuestión pasional. El punto de James es que Clifford escoge la segunda, pero al hacerlo, renuncia a la posibilidad de obtener algunas verdades.

Incluso, James admite que en situaciones en las que estemos en ausencia de evidencia y forzados a elegir entre dos opciones, debemos elegir el principio que nos instiga a evitar creer algo falso, *siempre y cuando la cuestión no sea importante* [James (1896), pp. 19-20]. Sin embargo,

> *Una regla de pensamiento que me impidiera absolutamente reconocer ciertos tipos de verdad si estos tipos estuvieran realmente ahí, sería una regla irracional.* [En cursivas en el original][103]

La tesis de James es, por tanto, que con respecto a cuestiones importantes y forzosas en las que por la naturaleza del asunto no sólo carecemos de evidencia, sino que no podríamos acceder de ningún modo a ningún cuerpo de evidencia, podemos y debemos decidir si creer o no. Lo que defiende no es tanto la creencia a voluntad en general, sino *el derecho a creer* en determinadas circunstancias.

[103] «[…] *a rule of thinking which would absolutely prevent me from acknowledging certain kinds of truth if those kinds of truth were really there, would be an irrational rule.*» [James (1896), p. 28]

Ésta es una cuestión controvertida que ha sido ampliamente discutida. Los que siguen a Clifford en el estricto y necesario requisito de evidencia para abrazar creencias son denominados evidencialistas; frente a ellos, los llamados no-evidencialistas simpatizan con James. No obstante, a pesar de que estas dos posturas han sido presentadas como prácticamente opuestas, de hecho no lo son tanto. James defiende —tal como hace Clifford— la importancia del método científico y de las verdades alcanzadas por este medio; también piensa que la evidencia desempeña un papel esencial en la formación de la mayoría de nuestras creencias. Incluso con respecto a aquellas hipótesis que no pueden decidirse (*de iure*, no *de facto*) por medio de la evidencia, James afirma que podemos renunciar a mantener una creencia si no versa sobre una cuestión importante. Es únicamente en las disputas genuinas, importantes, forzosas y no empíricas donde deberíamos tener *derecho* a decidir voluntariamente. Por su parte, ha de subrayarse que Clifford no juzga que sea imposible creer bajo tales condiciones, pero sí que resulta —no sólo por razones epistémicas sino también por cuestiones sociales— inmoral. En este sentido, se ha considerado que Peirce habría estado seguramente de acuerdo con Clifford, en la medida en que afirma —como hemos visto— que evitar investigar acerca del apoyo de cualquier creencia a causa del miedo de que pudiese estar podrido es tan inmoral como desventajoso [Peirce (1877), p. 257][104]. No está muy claro, sin embargo, si en cuestiones por principio no-empíricas como las que tiene en mente James (creencias religiosas,

[104] «[T]o avoid looking into the support of any belief from a fear that it may turn out rotten is quite as immoral as it is disadvantageous.»

principios filosóficos, etc.) es posible descubrir que su apoyo esté podrido; si a lo que se refiere Peirce es a que descubramos que no tienen apoyo sólido, sea de la clase que sea, Peirce considerará ciertamente la postura de James como inmoral; pero si a lo que se refiere es a que descubramos que está *en contra de la evidencia*, es obvio que, *ex hypothesi*, no puede ser así.

De hecho, Peirce ni sostiene que la evidencia sea necesaria para abrazar una creencia, ni considera que sea inmoral creer sin evidencia; como hemos expuesto ya, de los cuatro métodos que presenta para fijar creencias, tres de ellos no están necesariamente basados en la evidencia. Además, Peirce sostiene que

> La acción de pensar puede, dicho sea de paso, tener otros resultados. Puede servir para divertirnos, por ejemplo, y entre los *dilettanti* no es raro encontrar a aquellos que han pervertido tanto el pensamiento con el objeto de obtener placer que parece disgustarles pensar que las cuestiones sobre las que ejercitan con deleite pueden incluso zanjarse […] Esta disposición es la corrupción misma del pensamiento. [Peirce (1878), p. 80] [105]

[105] «[T]he action of thinking may incidentally have other results; it may serve to amuse us, for example, and among *dilettanti* it is not rare to find those who have so perverted thought to the purposes of pleasure that it seems to vex them to think that the questions upon which they delight to exercise it may ever get finally settled; and a positive discovery which takes a favorite subject out of the arena of literary debate is met with ill-concealed dislike. This disposition is the very debauchery of thought.» [Peirce (1878), p. 263] En un tono muy similar, Clifford había expresado con respecto a la facultad de creer: «Se la profana cuando se entrega a afirmaciones nunca puestas en tela de juicio ni sometidas a prueba, para el disfrute y placer privado del que cree». [Clifford (1876), p. 98] [«It is desecrated when given to unproved and unquestioned statements, for the solace and private pleasure of the believer.» Clifford (1876), p. 183]

Además, no parece considerar que creer sobre la base de una evidencia insuficiente sea siempre inmoral. En este sentido, decía un año antes:

> Una persona puede pasar por el mundo manteniendo sistemáticamente fuera de su vista todo lo que pudiera causar un cambio en sus opiniones, y si la única consecuencia fuera tener éxito —basando su método, como lo hace, en dos leyes psicológicas fundamentales— no veo qué puede decirse contra lo que hace. Sería una impertinencia egoísta objetar que su forma de proceder es irracional, ya que eso sólo equivale a decir que su método de asentar creencias no es el nuestro. Esa persona no se propone ser racional, y desde luego a menudo hablará con desprecio de la débil e ilusoria razón humana. Así que dejémosle pensar como le plazca. [Peirce (1877), pp. 46-47] [106]

Para ser honestos, Peirce seguramente no veía con buenos ojos la tesis de James acerca del papel de la voluntad en ausencia de evidencia, creer en Dios o cosas por el estilo, pero no parece congeniar tampoco con la tesis de Clifford de que no deberíamos formarnos jamás creencias en ausencia de suficiente apoyo empírico por razones sociales o morales. Más bien, lo que Peirce subraya es que sólo en el caso en que *descubriésemos que los hechos están en contra de nuestras creencias*, deberíamos abandonarlas, pues actuar de otro modo sería la depravación del pensamiento. El método científico es, simplemente, el mejor método para fijar creencias, porque en principio nos acerca más a la verdad al acudir directamente a los hechos.

---

[106] «A man may go through life, systematically keeping out of view all that might cause a change in his opinions, and if he only succeeds — basing his method, as he does, on two fundamental psychological laws — I do not see what can be said against his doing so. It would be an egotistical impertinence to object that his procedure is irrational, for that only amounts to saying that his method of settling belief is not ours. He does not propose to himself to be rational, and, indeed, will often talk with scorn of man's weak and illusive reason. So let him think as he pleases.» [Peirce (1877), pp. 249-50]

Otro celebérrimo argumento relacionado con la concepción que parece sostener James, es el que se conoce como la *apuesta de Pascal*. Pascal defendía que creer en Dios es racional incluso si no tenemos evidencia en absoluto de su existencia. ¿Por qué? Quizá haya una infinidad de opciones y muy pocas posibilidades de que Dios exista realmente; sin embargo, si apostamos por Dios y existe realmente, nuestra ganancia será infinita; y si, después de todo, no existe, no perderemos nada (y viceversa). Simplemente por este argumento, sería racional creer, deberíamos creer. Sin embargo, en este punto Pascal prevé un inoportuno problema para la creencia en función de razones probabilísticas o racionales: tal vez, incluso habiendo comprendido y aceptado la prueba racional de la conveniencia de creer, no podemos *hacerlo*. La respuesta de Pascal es la siguiente: no se preocupe. Si está racionalmente convencido, ha de ser su naturaleza pasional lo que le impide creer, de modo que debe trabajar ese aspecto. ¿Cómo hacerlo? Tome agua bendita, acuda a misa… verá como pronto se embrutecerá y creerá [Pascal (1670), pp. 126-127].

Como ya hemos dicho en otro lugar, esta idea recuerda a la famosa divisa Aristotélica, "el hábito engendra la virtud" [*EN* 1103a 31-1103b 2]. Creo que ni Clifford ni James aceptarían este argumento, bien por inmoral, bien por irracional. Tampoco es descabellada la interpretación de James Wernham [(1987), pp. 75-80], según la cual al invitarnos a tomar agua bendita, acudir a misa, etc., Pascal está tratando de presentar la hipótesis de Dios más viva para el no creyente, pues la razón por la que el sujeto no puede creer se halla en que la hipótesis está muerta para él. Sin embargo, hay algo que no acaba de encajar: no me parece que se trate

únicamente de vivificar la opción. Además de lograr que la hipótesis sea viva, uno debe vencer la tendencia a adoptar la política de creencias según la cual es siempre preferible evitar los errores (Clifford), para pasar a abrazar la otra que nos empuja a lanzarnos en busca de cierto tipo de verdades que por su naturaleza no gozan de apoyo empírico (James). Sólo de este modo podremos acceder a ciertas clases de verdad, y no se entiende muy bien cómo podemos llegar aquí sin una decisión personal y pasional, como quería James.

En todo caso, este tipo de interpretación acercaría las posturas de James y Pascal. De hecho, James presenta un ejemplo similar. A veces, en casos en los que (esta vez *de facto*, no *de iure*) no tenemos evidencia para creer algo que deseamos, la única manera de obtener esa evidencia es decidir creerlo de antemano, en ausencia de evidencia.

> *¿Te gusto o no?* — por ejemplo. El que sea o no así depende, en innumerables ocasiones, de si nos encontramos a medio camino, de si deseo asumir que debo gustarte y te muestro confianza y expectación. La fe previa por mi parte en la existencia de tu cariño es en tales casos lo que hace que surja tu cariño. Pero si permanezco quieto y me niego a mover un dedo hasta que tenga evidencia objetiva, hasta que tú hayas hecho algo adecuado, como dicen los absolutistas, *ad extorquendum assensum meum*, te apuesto que tu cariño nunca surgirá [James (1896), 23-24][107].

[107] «*Do you like me or not*? — for example. Whether you do or not depends, in countless instances, on whether I meet you half-way, am willing to assume that you must like me, and show you trust and expectation. The previous faith on my part in your liking's existence is in such cases what makes your liking come. But if I stand aloof, and refuse to budge an inch until I have objective evidence, until you shall have done something apt, as the absolutists say, ad *extorquendum assensum meum*, ten to one your liking never comes.»

James cree que, en estas ocasiones, uno ha de acercarse un poco para que el otro reconozca nuestra intención y a su vez se mueva, encontrándonos así a medio camino; pues si cada uno espera a que el otro haga todo el recorrido, ninguno de los dos se moverá un paso. Es lo que en el lenguaje cotidiano llamamos "mover ficha".

Sin negar que esto ocurra a veces, creo que en estos casos no decimos que tenemos la creencia de que le gustamos a la otra persona o de lo que sea. A lo sumo, uno tiene la esperanza de gustarle a alguien. Sin embargo, no cabe duda de que a veces actuamos guiados sólo por esperanzas o deseos. En casos como éste, confiamos en que haciendo ver que nos gusta, quizá ella se sienta segura de nuestro cariño y surja en ella un aprecio también hacia nosotros que de otro modo no hubiese surgido. Pero, insisto, no creo que en ningún momento tenga una creencia acerca de su cariño.

De hecho, Russell ya presentó una objeción en términos similares cuando indicó que a menudo actuamos sobre la base de una hipótesis sin creer en ella. Cuando nos enfrentamos con una elección forzosa para la que no tenemos suficiente evidencia, actuamos sobre la base de probabilidades sin abrazar creencia alguna. Según Russell, James confunde ambas actitudes.[108] Incluso, Clifford ya había hecho una observación semejante cuando, anticipándose a futuras críticas, señaló:

[108] «In the cases which William James has in mind, the option between rival hypotheses is, he says, a "forced" option; i.e. it is not avoidable: "lf I say, «Either accept this truth or go without it», I put on you a forced option, for there is no

> No hay peligro real de que tales consecuencias deriven jamás de un cuidado escrupuloso y el autocontrol en el caso de la creencia […]. Además, hay muchos casos en los que es nuestro deber actuar en función de las probabilidades, aunque la evidencia no sea tal como para justificar la creencia presente; porque es precisamente por medio de tal acción, y por la observación de sus frutos, como se obtiene la evidencia que puede justificar la creencia futura. De modo que no tenemos razón alguna para temer que un hábito de investigación concienzuda paralice las acciones de nuestra vida cotidiana.[109]

No obstante, lo que me interesa subrayar es que el desencuentro en este punto entre la postura de James y Clifford no se debe tanto a una cuestión esencial cuanto a una cuestión verbal. En efecto, tengo la sensación de que lo que distingue a estos autores no es tanto que uno sostenga que se

---

standing place outside of the alternative." This statement appears to us to be contrary to many of the plainest facts of daily life. If, in walking along a country road, I come to a fork where there is no signpost and no passer-by, I have, from the point of view of action, a "forced" option. I must take one road or other if I am to have any chance of reaching my destination; and I may have no evidence whatever as to which is the right road, I then *act* on one or other of the two possible hypotheses, until I find some one of whom I can ask the way. But I do not *believe* either hypothesis. My action is either right or wrong, but my belief is neither, since I do not entertain either of the two possible beliefs. The pragmatist assumption that I believe the road I have chosen to be the right one is erroneous. To infer belief from action, in the crude way involved in the assumption that we must "either accept this truth or go without it", is to ignore the plain fact that our actions are constantly based upon probabilities, and that, in all such cases, we neither accept a truth nor go without it, but entertain it as an hypothesis.» [Russell (1909), p. 264]

[109] «There is no practical danger that such consequences will ever follow from scrupulous care and self-control in the matter of belief. […] Moreover there are many cases in which it is our duty to act upon probabilities, although the evidence is not such as to justify present belief; because it is precisely by such action, and by observation of its fruits, that evidence is got which may justify future belief. So that we have no reason to fear lest a habit of conscientious inquiry should paralyze the actions of our daily life.» [Clifford (1876), pp. 188-189]

puede creer al margen de evidencia y el otro no. Con respecto a las cuestiones que Clifford tiene en mente (creencias factuales), ambos afirman que aunque es posible creer al margen de la evidencia, en principio es inmoral (Clifford) o irracional y ridículo (James). Lo que empaña la cuestión es que Clifford condena toda creencia sin evidencia suficiente; creo que Clifford no estaba pensando en modo alguno en las creencias religiosas, porque para Clifford en ese ámbito no podría haber creencias por principio: la definición de creencia que Clifford maneja no contempla que pueda haber una creencia sin algún tipo de evidencia.

James, por su parte, reacciona y se defiende contra un fantasma: cree que la postura de Clifford atenta contra el derecho a disponer de una guía para la conducta allí donde no hay evidencia.

Por tanto, la disputa se plantea en realidad con respecto a si en ciertas cuestiones en las que en *de iure* carecemos de evidencia por su naturaleza, podemos tener guías para la conducta y si éstas han de llamarse creencias o no. La cuestión, por tanto, se reduce a en qué consiste tener una creencia, y por tanto es verbal. Aunque Clifford seguramente creía que algo que nos sirve de guía para la acción allí donde no hay evidencia no puede ser otra cosa que una esperanza o un deseo, James no encuentra ninguna diferencia pragmática entre una actitud de este tipo y una creencia: ambas son actitudes mentales con respecto a un determinado hecho que suponen reglas para la acción y que pueden mostrarse más o menos adecuadas en función de los resultados. En tanto que James no

aprecia diferencia pragmática alguna entre ambas actitudes, ambas son para él creencias; de entre ellas, unas disponen de evidencia y otras no.

UN ENFOQUE CONCEPTUAL

Como hemos visto, sin duda James estaría de acuerdo con Clifford en que allí donde hubiese evidencia, la voluntad no debería asumir ningún papel. No obstante, sus posturas se pueden denominar como evidencialista y no-evidencialista: Clifford es evidencialista porque considera que la voluntad *nunca debería* desempeñar ningún papel por ser inmoral formarnos nuestras creencias sin apoyo en la evidencia disponible; James asume que la voluntad ha de desempeñar *necesariamente* un papel crucial en determinados casos especiales donde no hay otro elemento al que asirse y es necesario elegir, por lo que su enfoque es no-evidencialista.

Frente a estas posturas, Bernard Williams ha negado que la voluntad cumpla algún papel no por razones morales o racionales, sino *conceptuales*. En su famoso artículo titulado "Deciding to believe", sostiene que *creer a voluntad* no es *ni posible bajo ciertas condiciones muy especiales* (como muchos han querido ver en James) *ni posible pero inmoral* (como piensa Clifford); Williams afirma que creer por las buenas es *conceptualmente imposible*.

Ha de notarse que no está teorizando —al menos no directamente— acerca de creencias religiosas, filosóficas o morales, sino sobre casos de

creencias fácticas más simples.[110] Más bien, está pensando en el tipo de creencia que alguien tiene cuando sencillamente cree que está lloviendo o que la sustancia que tiene en frente es sal. Además, cuando habla de creencias, éstas han de entenderse como estados mentales (no como contenidos) [Williams (1973), p. 136].

Según Williams, las creencias tienen cinco atributos:

1. Las creencias aspiran o apuntan a la verdad.
2. La expresión más simple y básica de una creencia es la aserción.
3. Aunque la aserción es la expresión más simple y básica de una creencia, no es condición necesaria ni —subraya— suficiente.
4. No toda creencia está basada en evidencia.
5. Una creencia es una noción explicativa.

A su vez, que las creencias *aspiran a la verdad* significa tres cosas:

a) A diferencia de otros estados mentales, su contenido es veritativo-condicional.

b) Creer que *p* es lo mismo que creer que *p es verdadero,* y

c) Decir "pienso *p*" implica, en general, la afirmación de que *p* es verdadero (lo cual, como señala Williams, está relacionado con la paradoja de Moore) [Williams (1973), pp. 136-137].

[110] No resulta claro si, a pesar de tomar en consideración sólo creencias comunes para los propósitos de la explicación, estima que las conclusiones alcanzadas pueden extenderse al resto de las creencias. Si no es así, ignoro la razón que Williams tiene en mente para excluir las creencias morales o religiosas pero, obviamente, en tal caso la observación de Williams no afectaría del mismo modo a las tesis de James, en la medida en que James reserva el papel de la voluntad precisamente para casos no-factuales.

La manera más directa de expresar la creencia de que *p* es la aserción de que *p*, y no "pienso que *p*", cuya función es, según Williams, más bien especial.[111] [Williams (1973), pp. 137-138]. La aserción, sin embargo, ni es

[111] «The most elementary and straightforward expression of the belief that it is raining is to say 'it is raining', not to say 'I believe that it's raining'. 'I believe that it's raining' does a special rather job.» [Williams (1973), p. 137]

Esta afirmación de Williams parece, cuando menos, contraria al sentido común y, posiblemente, equivocada. Si bien es verdad que podemos expresar —y de hecho lo hacemos a menudo— nuestra creencia de que *p* a través de la afirmación de que *p*, no es verdad que siempre suceda así, ni siquiera lo es en la mayoría de los casos. Es más, en absoluto es excepcional decir "*creo* que *p*", "*me parece* que *p*" o "*tengo la impresión de* que *p*" cuando uno tiene la creencia de que *p*. En la mayoría de las ocasiones, sólo cuando tengo un tipo muy particular de creencias —convicciones— me permito afirmar que *p* directamente. En estos casos, aunque no dudo aquello que mantengo, no puedo decir que tenga conocimiento, en tanto mi consideración es falible. El enfoque de Williams descansa sobre el hecho de que no distingue entre creencia, convicción y conocimiento o, cuando menos, los mezcla. Sospecho que cuando sostiene que la aserción "pienso que *p*" es derivada y extraña, tiene en mente cosas que conocemos —o de las que estamos convencidos— y no cosas que meramente *creemos*. El ejemplo del propio Williams que citamos al comienzo de esta nota es instructivo a este respecto (aunque tenga el resultado contrario al que él pretende). Normalmente decimos "está lloviendo", cierto; pero la razón no es que la aserción sea una forma primaria de expresión de una creencia; más bien, la razón es que generalmente cuando decimos esto, lo *sabemos o tenemos la convicción*. Si, por ejemplo, simplemente tuviéramos un recuerdo de haber oído media hora antes un suave murmullo de lluvia, podríamos decir, dudando si aún está lloviendo: "*creo* que está lloviendo". Otro ejemplo que alega Williams en su defensa pero que parece darnos de nuevo la razón, es como sigue: «Si alguien me dice: "¿Dónde está la estación?" Y respondo: "Creo que está tres manzanas más abajo, a la derecha", el interlocutor tendrá un poco menos de confianza en mi expresión que si simplemente digo: "Se encuentra tres manzanas más abajo, a la derecha"». [Williams (1973), p. 183] [«If somebody says to me: Where it is the railroad station? And I say 'I believe that it's three blocks down and to the right' he will have slightly less confidence in my utterances than if I just say 'It's three blocks down and to the right'.» Williams (1973), p. 138]. Obviamente, cuando creemos algo tenemos menos seguridad que cuando lo sabemos. Pero, por desgracia, sólo en escasas ocasiones la evidencia de la que disponemos es conclusiva. En efecto, en la medida en que las creencias son graduales, pueden ser muy débiles; esto se refleja a menudo en

condición necesaria de la creencia (hay muchas creencias que no hacemos ni haremos explícitas) ni es suficiente (pues puede ser insincera) [Williams (1973), p. 140].

Lo que puede llamar fuertemente la atención es que Williams admite en 4) que hay creencias que no están basadas en evidencia. Pero entonces, si afirma que no se puede creer a voluntad y que hay creencias que no están basadas en evidencia, ¿cómo las obtenemos? Es importante señalar que Williams entiende por 'evidencia' para una creencia, otras creencias que apoyen la primera.[112] En este sentido, hay creencias que están causadas directamente por datos perceptivos, los cuales no constituyen —en

nuestras afirmaciones o aserciones. Por tanto, el argumento que Williams presenta parece débil y, más bien, cabría defender lo contrario: a veces, uno expresa una creencia (incluso una creencia débil) enmascarada bajo la forma de una aserción para alcanzar una "seguridad robada", para engañar a otros, etc.

[112] Ésta es otra de las tesis controvertidas de Williams. Su argumento es como sigue: si toda creencia necesita evidencia para su justificación (considerando que "evidencia" significa "otras creencias"), entonces jamás podríamos detener o si quiera comenzar la justificación (argumento *ad infinitum*) [Williams (1973), p. 143]. Williams rechaza, por tanto, la explicación coherentista de la justificación inclinado hacia una postura fundacionalista: cree que a la base justificativa ha de haber otras creencias no basadas en creencias: las creencias perceptivas fundadas en datos sensoriales. Pero dejando a un lado la crítica poderosa que podría hacerse a este tipo de explicaciones fundacionalistas [vid. Haack (1993)], parece claro que estos datos sensoriales son buena parte de la experiencia que conforma lo que muchos otros llamamos "evidencia". Por tanto, parece arbitrario reservar el término "evidencia" para las creencias que apoyan otras creencias. Después de todo, como hemos mostrado, la interesante afirmación de Williams de que algunas creencias requieren evidencia y otras no (no por ser auto-evidentes en un sentido racionalista, sino porque están basadas en datos sensoriales directos) no difiere de aquella otra que dice que "toda creencia requiere evidencia". Bajo mi punto de vista, por ende, toda creencia demanda evidencia —sea perceptual, basada en la memoria o doxástica— para su formación y justificación. Mi postura podría etiquetarse como *fundherentista*, en términos de Haack (1993).

términos de Williams— evidencia. Sin embargo, no es esto lo que quiere subrayar Williams cuando afirma que hay creencias que no están basadas en la evidencia. Lo que le interesa recalcar tiene más que ver con la relación causal entre dos (o más) creencias. Según Williams, esta relación puede ser racional o meramente causal, es decir, a veces una creencia de que *p* puede *causar* otra creencia de que *q*, sin que pueda aducirse la creencia de que *p* como apoyo para la creencia de que *q*. Dicho de otro modo: uno puede creer que *q* porque cree que *p*, no siendo la verdad de *p* una *razón* para la verdad de *q*. Diremos entonces que "A cree que *q* porque cree que *p*", pero no estaremos legitimados para decir "*q* porque *p*" ni "la creencia de que *p* implica la creencia en *q*" en la medida en que, aunque hay una conexión entre creencias, no es una conexión racional [Williams (1973), pp. 141-2].

Por otro lado, la creencia es una noción explicativa, es decir, sólo es posible explicar lo que hace un hombre acudiendo a lo que cree. No obstante, las creencias no son suficientes para explicar la acción; además de creencia y acción necesitamos un tercer componente: el proyecto.[113] [Williams (1973), p. 144].

[113] «Un ejemplo típico de esto es el siguiente: veo a un hombre que marcha con paso firme y decidido sobre cierto puente. Decimos que esto muestra su creencia de que el puente es seguro, pero, desde luego, esto es sólo relativo a un proyecto que es muy razonable suponer que tiene, a saber, evitar ahogarse. Si se tratara de una persona que, sorprendentemente, tuviera el proyecto de caerse al río, entonces, el que caminase con paso firme sobre este puente no manifestaría necesariamente la creencia de que el puente era seguro» [Williams (1973), p. 191]. [«A standard example of this is: I see a man walking with a determined and heavy step onto a certain bridge. We say that it shows that the bridge is safe, but this, of course is only

Tras establecer las condiciones para caracterizar cualquier creencia, Williams propone un experimento conceptual en el que examina la posibilidad de crear una máquina que pudiese formarse creencias. Concluye que la mayor dificultad para su creación sería que esta máquina no podría satisfacer la tercera condición, es decir, que no podría ser insincera.[114] Williams sostiene que la voluntad desempeña un papel

relative to a project which it is very reasonable to assume that he has, namely to avoid getting drowned. If this were a man who surprisingly had the project of falling in the river, then his walking with firm step onto this bridge would not necessary manifest the belief that the bridge was safe.» Williams (1973), p. 144]. Supongo que no hay aquí gran diferencia entre "proyecto" e "intención".

[114] Williams distingue entre creencias y estados-C, que serían "creencias empobrecidas", "*impoverished beliefs*" (ya que no satisfacen la cláusula 3) [Williams (1973), pp. 145-7]. Considero que esta tesis de Williams es desacertada. Incluso garantizando que pudiéramos adscribir "creencias" a las máquinas —soy bastante reacio a admitir esto—, las razones que Williams propone para distinguir estados-C y creencias me parecen demasiado débiles. De modo que el hecho de que una máquina no pueda ser hipócrita no es lo que demuestra que no pueda tener creencias; en todo caso, sólo indica que podemos confiar en sus aserciones (a condición de que empleen métodos fiables para captar estímulos, realizar procesos inferenciales, etc., y no estén desajustadas o estropeadas). En el caso de los seres humanos, la voluntariedad para decir (o no decir) lo que uno piensa impide la adscripción directa de cualquier creencia particular, y, por tanto, no podemos confiar en sus aserciones. Pero entonces, los estados-C alegados no son un estado mental diferente. Como mucho, las máquinas carecen de la capacidad de mentir, lo cual no afecta o modifica al estado "mental" en sí mismo, sino a su expresión.

Otra tesis problemática en la que no podemos detenernos tanto cuanto sería necesario es la que sigue. Williams dice acerca del posible conocimiento de una máquina: «Por lo que respecta a los estados-C, la máquina podría encontrarse en falsos estados-C; en estados-C verdaderos a los que habría accedido accidental o fortuitamente; en estados-C verdaderos y a los que no habría accedido accidentalmente, esto es, estados-C que fueran verdaderos y que se produjesen de maneras relacionadas con el hecho de que fueran verdaderos, y a estos últimos los podríamos llamar conocimiento». [«With regard to the B-states, there could be false B-states that the machines was in; accidentally or randomly true B-states; and non-

esencial con respecto a las creencias: *voluntariamente expresamos o no* aquello que creemos. No obstante, a pesar de que la voluntad está relacionada con la decisión de decir o no decir, no está directamente relacionada con la decisión de creer. En este aspecto, la caracterización de Williams se asemeja a la descripción de la creencia ofrecida por Hume, según la cual la creencia es algo que nos pasa, un fenómeno pasivo. Ahora bien, Williams se separa de Hume con respecto a la naturaleza del fenómeno, pues mientras Hume considera que este rasgo de las creencias es *contingente*, Williams cree que es conceptualmente imposible que fuese de otra manera.

Williams señala que hay procesos en los que la voluntad no toma parte —por ejemplo, uno no se sonroja a voluntad. Pero incluso aunque nadie pudiera sonrojarse a voluntad, esto sería un hecho contingente, en tanto que no es inconcebible sonrojarse a voluntad. O, al menos, cabe imaginar un modo de sonrojarse "dando un rodeo" (*taking a detour*), quizá

accidentally true B-states, that is B-states which were connected with the fact that they were true, and these last we could call Knowledge.» Williams (1973), p. 147]. Sin embargo, esta tesis acerca del conocimiento depende de la *situación del examinador* que Williams había rechazado previamente. Pienso que el conocimiento es reflexivo, es decir, que el sujeto que conoce, conoce que conoce o no hay conocimiento alguno. Así, pues, una máquina que estuviese en estados-C, no siendo ni consciente ni capaz de saber si éstos son verdaderos o falsos (es decir, no sabiendo si se basan en datos que han sido captados de modo adecuado o si se han debido a una avería), no podría tener un estado mental semejante al conocimiento, incluso si su creencia fuese verdadera y adecuada a los hechos. La justificación que exige el conocimiento no depende únicamente de que haya una relación adecuada con los hechos, sino de que el sujeto sepa que la relación es adecuada.

En resumen, Williams no sólo es incapaz de distinguir entre estados-C y creencias, sino que la relación entre creencia y conocimiento que ofrece es controvertida.

imaginando una situación embarazosa. Por el contrario, Williams piensa que es la propia naturaleza de las creencias lo que impide a la voluntad tomar parte en el proceso de adquirir una creencia.

> No es un hecho contingente que no pueda hacerme, así por las buenas, creer algo, tal como es contingente el hecho de que no pueda, así por las buenas, ruborizarme. ¿Por qué es esto? Una razón está ligada a la característica, por parte de las creencias, de que apuntan a la verdad [Williams (1973), p. 196].[115]

Si uno pudiese creer a voluntad, debería saber que puede hacerlo, lo cual significa que sabe que puede formarse una creencia con independencia de lo que la evidencia nos dice acerca de la realidad. Pero no es posible al mismo tiempo pensar que mi creencia aspira a la verdad y que no representa la realidad [Williams (1973), p. 148]. Así, aunque la idea de creer algo *por las buenas* es muy extraña, Williams se pregunta si la voluntad puede desempeñar algún papel en la decisión de abrazar una creencia: ¿somos capaces de obtener una creencia (o de rechazarla) por medio de caminos indirectos?

Williams admite que sería concebible que la voluntad desempeñase un papel para producir una creencia por medio de factores no directamente conectados con la verdad, tales como la hipnosis o ciertas drogas

[115] «It is not a contingent fact that I cannot bring it about, just like that, that I believe something, as it is a contingent fact that I cannot bring it about, just like that, that I'm blushing. Why is this? One reason is connected with the characteristic of beliefs that they aim at truth.» [Williams (1973), p. 148]

[Williams (1973), p. 149].[116] Pero estos métodos no siempre son efectivos. La razón principal es que hay dos tipos de motivos para querer creer algo: *motivos centrados en la verdad* y *motivos no centrados en la verdad*. Supóngase, dice Williams, que un hombre tiene una fuerte evidencia para creer que su hijo se ha ahogado en el mar. Este hombre desea con fuerza creer que su hijo está vivo; pero entonces, ¿por qué no acudir a la hipnosis o las drogas? Por medio de estos métodos quizá logre abrazar la creencia que desea. No obstante, ¿quiere realmente abrazar la creencia de que su hijo está vivo? Williams indica que parece más apropiado pensar que "desea creer que su hijo está vivo" habría de interpretarse como "desea que la creencia sea verdadera", es decir, lo que quiere es que su hijo *esté* realmente vivo. Esto es lo que Williams denomina *motivos centrados en la verdad* [Williams (1973), p. 150]. Por consiguiente, en tales casos creer a voluntad sería inútil.

Por el contrario, en algunas situaciones particulares "desear creer que su hijo está vivo" puede interpretarse de otro modo. Nuestro hombre podría tener un motivo no centrado en la verdad; sin duda querría, por supuesto, que su hijo estuviese vivo, pero, después de todo, sabe que *no puede cambiar la realidad*. Sin embargo, la encuentra intolerable y por esta razón desea abrazar la creencia, sea o no verdadera, y decide así hacer uso voluntariamente de mecanismos tales como la hipnosis o las drogas. Según Williams, esto no sería inconsistente, pero sí profundamente irracional; un proyecto que la mayoría de nosotros rechazaríamos

[116] Aunque Williams concede que no sólo es posible sino también común que la gente se engañe a sí misma, llegando a creer cosas que no son verdaderas, declina la discusión acerca del tema del autoengaño.

[Williams (1973), p. 150]. Es más, este tipo de proyecto entraña otros problemas: la naturaleza holista de nuestra estructura doxástica hace que la eliminación de una creencia requiera también la destrucción de todas las creencias que la implican, y un proyecto de este tipo podría llegar a implicar destrucción total de la realidad, llevando a la paranoia [Williams (1973), p. 151].

Williams indica que en caso de que el sujeto desease salvar algunas de sus creencias de tal destrucción masiva, podríamos estar hablando de autoengaño. En estos casos, el sujeto necesitaría reconocer la evidencia que resulta contraria y conflictiva a la creencia deseada, precisamente para distinguirla y evitarla eficazmente. Si el sujeto ha de creer o no aquello que *sabe que es conflictivo* es lo que constituye el problema del autoengaño, pero, en todo caso, implicaría un proyecto distinto a los casos puros de inducción de creencias que Williams analiza.

Me parece, sin embargo, que de las tesis de Williams se sigue naturalmente que el autoengaño no es posible. Por mi parte, considero —con Hume y Williams— que las creencias son algo que nos pasa y que, en principio, no controlamos en absoluto. A diferencia de Hume, coincido con Bernard Williams en sostener que esta característica no es contingente ni psicológica, sino necesaria y conceptual. Sin embargo, no concedo que podamos —salvo en casos muy exóticos como, por ejemplo, el de las pastillas que crean amnesia retroactiva— adquirir o retener una creencia en ausencia de evidencia dando un rodeo.

Carecemos de control sobre la formación de nuestras creencias; las creencias son una respuesta automática, un sentimiento involuntario respecto de la verdad de algo al acumular evidencia a su favor.

Las razones para sostener esto son las siguientes:

1. *Las creencias apuntan a la verdad.* Por tanto, la creencia de que *p* incluye de algún modo la aserción de que *p* es el caso. Si somos conscientes de que actuamos tomando en consideración algo que hemos aceptado por un acto de voluntad, necesariamente sabemos que no tiene por qué casar con la realidad y, en consecuencia, *no tenemos el sentimiento de que sea el caso*: no lo creemos. Así, la voluntad no puede desempeñar ningún papel epistémico en la formación de creencias: éste es un proceso necesariamente involuntario bajo su propia definición.

2. *Las creencias requieren evidencia*: como dije anteriormente, toda creencia consiste en *un sentimiento acerca de la verdad de algo apoyado por un conjunto de evidencia que el sujeto considera relevante y significativa, sentimiento que le dispone para afirmar o declarar con sinceridad la verdad de ese algo*. Es una consecuencia directa de que las creencias aspiren a la verdad, que las creencias requieran evidencia.

3. Esta evidencia no tiene por qué ser adecuada, verdadera u objetiva; *el peso de esta evidencia ha de ser subjetivamente significativo,* pues es el modo en que el sujeto pondera la evidencia y el grado de apoyo que él estima que ofrece su evidencia para la verdad de la proposición

> en cuestión, lo único que resulta relevante en la formación de su creencia. Obviamente, el grado de disposición para afirmar la verdad de tal proposición puede variar, dependiendo *no* de que la evidencia *sea* más o menos concluyente, sino de cómo considere el individuo que es. Esto, por supuesto, no garantiza la verdad de la creencia, incluso cuando crea algo con certeza.

Se sigue que la autoinducción de creencias directa es conceptualmente imposible, no sólo en los casos en que la evidencia tiene un peso significativo, sino también en los casos en que cabría la posibilidad de creer a voluntad en los términos de James, es decir, allí donde no hay, ni puede haber, evidencia empírica. Quizá sea necesario recordar que la ausencia de creencias no necesariamente impide la acción. Como es bien sabido, toda creencia puede convertirse en una regla para la acción, pero no toda regla para la acción consiste en una creencia. La fe, los deseos, las esperanzas, las sospechas o los miedos son estados cognitivos respecto de la verdad de algo y, en sentido amplio, pueden ejercer también como *motivo* para la acción.

Por lo demás, coincido con los principios pragmatistas acerca del papel de las creencias como reglas para la acción, aunque —como indicaron Peirce y James— no toda creencia conduzca inmediatamente a la acción.

No obstante, aun cuando no podemos creer algo a voluntad, lo que sí podemos hacer es preparar *evidencia ficticia y consonante con el estado deseado* y *auto-provocarnos el olvido tanto de la antigua evidencia disonante, como del hecho de que la nueva evidencia es preparada, falsa*. No encuentro ningún modo de

explicar esto mediante un proceso introspectivo o en cierto sentido natural; sin embargo, es concebible que, por ejemplo, tras cometer un asesinato en la habitación de un hotel, uno prepare el escenario de modo que oculte todas las pistas que lo incriminan, y a continuación logre olvidar lo sucedido por medio de abundante alcohol o algún tipo de potente píldora que cause amnesia retroactiva, interrumpiendo así la consciencia y la unidad del yo. Aún así, lo que habríamos conseguido controlar es la orientación de la evidencia, confiando en que surja en nosotros *de modo involuntario* el sentimiento hacia el estado de cosas que deseábamos que fuera el caso; de esta forma, dejamos que sea el proceso incontrolable y automático el que, al despertar, haga el resto del trabajo. Es, por tanto, un control indirecto y en casos muy especiales. Por lo demás, estrictamente hablando, este género de casos exóticos no representa una prueba de que haya creencias en ausencia de evidencia; más bien, son la prueba de que no podemos inducir una creencia sin evidencia que la sustente: por ello hemos de "arreglarla". Disponemos de evidencia, aunque ciertamente sea de un tipo especial. Lo que puede despistarnos es que esta evidencia no nos parece en absoluto fiable, pero esto no es relevante: es aquello que subjetivamente considera el sujeto acerca de la evidencia y su fiabilidad lo que nos interesa. Y precisamente el experimento mental contempla que éste consigue olvidar tanto el carácter ficticio de la evidencia cuanto que ese olvido se lo ha auto-provocado. Lo que imposibilitaba la autoinducción de creencias era la tensión entre la aspiración a la verdad y *la consciencia* de que la evidencia que el sujeto deseaba obtener sería sesgada y autoprovocada. Al desaparecer la

consciencia de ello, desaparece también la tensión. Por desgracia, la desaparición de esta tensión es sólo temporal: la coherencia de la red doxástico-evidencial nos conducirá, quizá, a preguntarnos por el apoyo de otra creencia relacionada con la inducida. O nos enfrentará a nueva evidencia conflictiva con la evidencia "arreglada" y la creencia inducida. Por tanto, este tipo de estados son inherentemente inestables o metaestables y condenados, a buen seguro, a una vida corta.

Una consecuencia que parecería seguirse de la tesis que mantengo es que no debemos preocuparnos en absoluto por la formación de nuestras creencias. No se trata de afirmar que no podamos fracasar al captar la realidad y en la aspiración de que nuestras creencias sean verdaderas: evidentemente, podemos errar, pues nuestros métodos de adquisición de evidencia (nuestros sentidos, recuerdos y capacidades inferenciales) son falibles. Pero, en tanto que (1) estos errores son involuntarios, (2) las creencias son respuestas cuasi-automáticas a una evidencia relevante, y (3) la voluntad o la intención no tienen nada que hacer en el proceso de adquisición o mantenimiento de las creencias, se sigue que no podemos ser negligentes en este proceso. Entonces, ¿estamos totalmente exentos de responsabilidad? En absoluto.

Aunque el asunto de la consideración de las consecuencias morales del autoengaño y la ética de la creencia no son parte de nuestro trabajo, en tanto que estudio puramente conceptual del fenómeno del autoengaño, diremos sólo algunas cosas a modo de apuntes en relación con este asunto.

En primer lugar, me parece que la responsabilidad exige capacidad de control. Dejando a un lado los casos exóticos en los que "arreglamos" evidencia ficticia, debemos desplazar la atribución de responsabilidad desde la formación de creencias (que escapa a nuestro control) a otro lugar. La responsabilidad se encontraría en

a) El modo de adquisición de evidencia

b) Las declaraciones de nuestras creencias

c) Las acciones que se siguen de las creencias

A veces un individuo puede decidir dejar de buscar evidencia porque teme hallar elementos dañinos, pero en estos casos no se forma una creencia sesgada; como dice Peirce, más bien duda ya acerca de la cuestión y, por esta razón, no tiene ninguna creencia en absoluto.[117] Cuando hay serias dudas no hay verdadera creencia, porque las creencias son un estado en cierto modo estable; para decirlo con Ortega, en las creencias *se está*. Es cierto que, en cualquier caso, las creencias *no* son la única regla posible para la acción,[118] y un individuo puede actuar sobre la base de sospechas,

[117] «[…] cuando ven que cualquier creencia suya está determinada por cualquier circunstancia ajena a los hechos, no sólo se limitarán a partir de ese momento a admitir de palabra que esa creencia es dudosa, sino que experimentarán una duda real acerca de ella, de modo que deja de ser una creencia». «[…] when they see that any belief of theirs is determined by any circumstance extraneous to the facts, will from that moment not merely admit in words that that belief is doubtful, but will experience a real doubt of it, so that it ceases to be a belief.» [Peirce (1877), p. 253]

[118] Aunque, como se ha dicho, siempre que podamos actuar sobre la base de creencias será mejor, precisamente porque éstas, al aspirar a la verdad, se apoyan en evidencia y tratan de aprehender la realidad, en algunas situaciones resulta muy útil actuar por fe, o en función de esperanzas, sospechas, deseos (sin creencia previa)

esperanzas o deseos; estas cogniciones no apuntan a la verdad y, por esta razón, es más plausible que estén desajustadas con la realidad. Esa responsabilidad recae sobre el sujeto.

Otro aspecto que también influye en el modo de adquisición de evidencia es la *política de creencia* que adoptemos; no se trata sólo de si aceptamos el principio de evitar errores o el de buscar verdades; se trata también de qué aceptamos como evidencia; por ejemplo, si aceptamos como evidencia testimonial ciertos argumentos de autoridad, debemos ser cuidadosos en quién consideramos como autoridad. Pero también hemos de poner atención a los umbrales de aceptación de evidencia y a cómo nuestras emociones influyen en ellos, en la ponderación de la evidencia y, de modo indirecto, en la rectitud de nuestros juicios. Sin obviar que ciertas pasiones pueden resultar incontrolables y cegar al individuo, podemos exigir al menos cierto propósito de autocontrol. Esta responsabilidad es también del sujeto, y tiene una enorme importancia en la adquisición de la evidencia que posteriormente genera la creencia.

Por otro lado, es algo trivial e inocuo que un sujeto pueda ser hipócrita o insincero y negarse a declarar sus creencias. Sin embargo, cuando en un asunto importante alguien sabe que podría buscar (más) evidencia *relevante* pero decide no hacerlo, y *afirma sin reservas algo que no le está permitido afirmar* o actúa *fingiendo conocer la verdad*, es completamente responsable de ello e, incluso, culpable. En la medida de nuestras posibilidades, debemos buscar

para buscar evidencia favorable y resolver un problema. Ninguna de estas disposiciones son creencias, pero pueden resultar útiles en ausencia de las mismas.

toda la evidencia que consideremos relevante para la verdad del asunto que nos ocupa, y guiar nuestras acciones en función de la fuerza que trasmita la evidencia obtenida a nuestras creencias.

En resumen, la autoinducción de creencias parece una tarea sobrehumana. De forma voluntaria y directa es conceptualmente imposible adquirir una creencia, porque si el sujeto es consciente tanto del proceso de adulteración como de que la evidencia está sesgada, es consciente también de que el estado mental que pretende abrazar no apunta a la verdad, y es la consciencia de este tipo de cosas lo que le impide formarse la creencia. Y de forma indirecta, "dando un rodeo", presenta problemas particulares; hemos examinado en la primera parte de nuestro trabajo múltiples intentos de explicaciones conceptuales de este fenómeno, pero todas ellas resultan insatisfactorias principalmente por dos razones: o bien hacen uso de *explicaciones blindadas*, como la apelación a la noción del inconsciente, o bien acuden a explicaciones *ad hoc* retorcidas que implican entidades mentales exóticas, tales como los subsistemas o la división de la mente —que otras explicaciones alternativas no requieren—, perdiendo desde un punto de vista empírico la encarnación en seres reales y causando en el plano puramente conceptual, además, acusaciones persistentes de homuncularidad. Otras explicaciones indirectas que acuden a drogas o hipnosis son problemáticas porque, por un lado, pueden llevar a la paranoia y a estados demasiado inestables como para permitir un éxito duradero, debido a los conflictos en la red doxástica y con futura evidencia; por otro lado, no los encuentro satisfactorios principalmente porque me parecen demasiado exóticos y ocasionales como para constituir

el tipo de situación que buscamos cuando hablamos del familiar fenómeno denominado "autoegaño".

Una vez expuestas en líneas generales mis consideraciones acerca de la naturaleza de la creencia, la evidencia (ponderable e imponderable), de cómo se forma la creencia y del papel que tiene en todo ello la voluntad, pasaré a considerar qué puede querer decir que un sujeto se autoengañe, y cuál podría ser la naturaleza del fenómeno.

VI. ANÁLISIS CONCEPTUAL DEL AUTOENGAÑO

Es ciertamente difícil encontrar consenso entre distintos teóricos acerca de lo que habría de suponer entrar en un estado que pudiésemos denominar, sin miedo a ser imprecisos, autoengaño. Como hemos visto, la cantidad de explicaciones alternativas es abrumadora. Esto, no obstante, no es algo excesivamente sorprendente; sólo demuestra que estamos ante un verdadero problema filosófico.

En contraste con esta dificultad, la gente que atribuye autoengaño no suele ser tan dubitativa. Ni necesita hacer un examen cuidadoso de evidencia, creencias, conductas, etc., ni tiene encendidos debates internos o públicos acerca del concepto antes de aplicarlo. No obstante, parece que la gente no lo aplica aleatoriamente. Sin duda, en muchas ocasiones, la gente puede hacer uso del término "autoengaño" de un modo multívoco o equívoco, para referirse a distintos fenómenos que, en principio, no son excesivamente problemáticos para una teoría de la racionalidad modesta. Así, un individuo podría hablar de autoengaño cuando se le presentasen casos de pensamiento desiderativo, ceguera intelectual, debilidad de la voluntad, falsa conciencia, mala fe, disonancia cognitiva u otros prejuicios cognitivos, sin tener muy clara la diferencia. O podría considerar el autoengaño como el género del cual todos esos fenómenos son distintas

especies. Varios no plantean un serio desafío para nuestra racionalidad; otros quizá lo hagan, pero tampoco serían casos genuinos de autoengaño, aunque puedan verse posibilitados o reforzados por éste.

La pregunta que nos parece filosóficamente interesante es si, en todo caso, podemos encontrar algo distinto de estos otros prejuicios cognitivos que se corresponda con lo que debería cubrir nuestro concepto. Hay dos cuestiones relevantes aquí que no deben mezclarse o confundirse:

1) Dilucidar en qué consistiría desde un punto de vista conceptual, que un sujeto iniciase un *proceso* de autoengaño o permaneciese en un *estado* de autoengaño.

2) Examinar las condiciones bajo las cuáles es racional suponer que alguien está autoengañado.

La primera de las cuestiones es de tipo onto-epistemológico, mientras la segunda de ellas es más pragmática. En este capítulo nos dedicaremos exclusivamente a la cuestión de la naturaleza del autoengaño y en el siguiente veremos la cuestión de las atribuciones de autoengaño.

Podemos comenzar preguntándonos si el autoengaño ha de ser, de algún modo, algo así como un tipo de *engaño*, una especie de engaño *autoinducido*. No diremos "engaño causado por uno mismo" porque reservamos este concepto para los casos en los que sería concebible que uno decidiese engañarse y, por medio de caminos indirectos como las drogas o la hipnosis, lograse alcanzar una creencia deseada. Por un lado, ya hemos visto las características especiales y problemas que tendría algo así. Por otro lado, este tipo de casos no son ni familiares (como parece

serlo el autoengaño) ni excesivamente problemáticos. Por razones similares, también excluimos de nuestro planteamiento conceptual las descripciones de estados mentales correspondientes a distintos trastornos mentales, como la esquizofrenia o la doble personalidad. Si el autoengaño supone un reto para cualquier teoría de la naturaleza humana o de la racionalidad, es porque parece algo tan omnipresente y natural, como inexplicable sobre fundamentos racionales.

También deberíamos estar de acuerdo en que el autoengaño no puede consistir en un estado alcanzado sin pretenderlo uno. Davidson vio esto muy bien cuando dijo que

> […] hacer algo intencionalmente con la consecuencia de que el sujeto de la acción resulte engañado no constituye, sin más, autoengaño, pues de otro modo una persona se autoengañaría al leer una noticia falsa en un periódico. [Davidson (1985), p. 110]

Todas las teorías, tanto las intencionales como —curiosamente— las no-intencionales, recogen esto de un modo u otro. De hecho, las teorías no-intencionales hacen lo posible por evitar acudir a elementos intencionales con el fin de eludir las paradojas del autoengaño, pero acaban introduciendo elementos como "propósitos" o "conductas deshonestas" que parecen *intenciones disfrazadas*. Mele dice acudir sólo a motivos, pero no es más que un espejismo: sostiene que uno puede sesgar intencionalmente la evidencia sin que por ello haya tratado de engañarse intencionalmente. Veremos esto más adelante.

En cualquier caso, ninguna propuesta suele eliminar por completo el ingrediente intencional, pues en tal caso su caracterización correría el

riesgo de acercarse demasiado al simple error o ceguera intelectual causados por fuertes emociones o angustias. Al menos en este aspecto el autoengaño se asemeja al engaño: exige algún tipo de *intención*.

Otra cuestión debatida es si el autoengaño ha de incluir, como el engaño interpersonal típico, creencias en conflicto o contradictorias. Del mismo modo que en el caso de la intención, aunque en el engaño interpersonal se acepta con naturalidad y no se discute este punto, las paradojas a las que conduce este requisito en el caso del engaño a uno mismo provoca que se haya tratado de evitar o replantear el asunto.

Se dice que uno ha de "contemplar de algún modo" cosas contradictorias, mirar por el rabillo del ojo de la mente una cosa y tener presente otra, enviar a una esquina de la mente lo que es doloroso, observar *in foro interno*, etc. Todas estas metáforas o imágenes representan formas más o menos satisfactorias de replantear la cuestión: hemos de aceptar que quien se autoengaña ha de contemplar de algún modo dos creencias contradictorias.

Ya hemos visto que uno puede tener *de modo trivial* creencias contradictorias. Es evidente que nuestros sentidos, nuestra memoria y nuestras capacidades inferenciales son falibles, por lo que nuestra racionalidad es limitada; sin duda, *no* somos conscientes de algunas consecuencias de nuestras creencias, y esto abre la puerta a posibles incoherencias entre nuestras cogniciones. Pero esto no nos convierte en seres carentes de razón, del mismo modo que un mero error al efectuar una suma no nos convierte en seres que no saben sumar. Sabemos sumar,

entre otras cosas, porque podríamos detectar nuestros errores. De modo similar, sólo supondría un problema de racionalidad que conociésemos estas incoherencias y no las rechazásemos. El autoengaño es paradójico porque parece exigir el reconocimiento —a algún nivel— de este tipo de incoherencias, en tanto que el sujeto desea adquirir una creencia placentera que está en conflicto tanto con la evidencia de la que dispone como con la creencia a la que le empuja esta evidencia. Aún más, parece obvio que la única razón para que el sujeto desee adquirir esta creencia contraevidencial, es que la evidencia contraria o la creencia le causan dolor. Así, como apuntaba Davidson, el problema del autoengaño se resume en explicar cómo una creencia causa otra contradictoria con ella y la sustenta, sin que el sujeto rechace ninguna de las dos.

Como he dicho en la sección dedicada a la formación de creencias, acudir en este caso a una explicación mediante el análisis de teorías de la posibilidad y grados de creencias no solventa nada. Únicamente reformula en clave lógico-matemática un problema que no le es posible manejar: o bien lo disuelve de modo ingenuo admitiendo la posibilidad de creencias con escasa evidencia y, con ello, de creencias contradictorias en distinto grado (uno puede creer que *p,* y creer que *no-p en un grado muy bajo*, dado que no está seguro de que *p*) o bien se topa con la misma dificultad: si para creer que *p* es *necesario* que uno estime que su evidencia total indica que es *más plausible que p* sea el caso que que no lo sea, entonces parece difícil explicar cómo podría creer que *no-p*, aunque sea consciente de que hay alguna evidencia que apoya la posibilidad de *no-p*.

De este modo, parece que en los casos de autoengaño, el sujeto dispone de una evidencia que le produce una creencia, un fuerte sentimiento acerca de la verdad de *p*. Sin embargo, esta creencia, junto con la evidencia que la apoya, le causa un dolor, angustia o malestar al sujeto, de modo que desea abrazar otra creencia placentera y olvidar aquello que le causa dolor. Por este motivo, inicia un proceso intencional que concluye cuando el sujeto abraza la nueva creencia; lo irracional es que

(1) Su evidencia total no apoya esa creencia placentera.

(2) Tiene una creencia basada en la evidencia total que es contradictoria con la autoinducida.

(3) La creencia autoinducida es, además, causada y sustentada por la creencia evidencial.

(4) No rechaza ni una creencia ni otra.

Dado (1) y (3), la creencia evidencial causa la creencia deseada, pero no puede ser nunca *razón* de ella.

Además, este proceso da lugar a un estado inestable; además de mantener ocultos tanto los mecanismos que permiten el engaño, como la angustia y la evidencia contraria, la evidencia disponi*ble* y contraria a la creencia inducida siempre será amenazante. El sujeto ha de ser capaz de evitar nueva evidencia contradictoria. Por esta razón, si el sujeto desea que el proyecto no naufrague en el momento de botarlo, necesita fuel y un timón que lo hagan avanzar evitando los escollos en forma de evidencia contraria. Por esta razón, no parece posible que el sujeto "olvide" la

evidencia contraria y la creencia original: perdería de vista aquello que le obliga a sesgar continuamente la evidencia. No obstante, todo apunta a que este estado sería inestable por definición y abocado seguramente al fracaso. Quizá, sólo un estado transitorio en la esperanza de que lleguen tiempos mejores. Algo así como pedir confianza a crédito.

Una vez establecidas algunas de las condiciones que debería cumplir un proceso o estado de autoengaño, debemos preguntarnos: ¿es conceptualmente posible un proceso y un estado de este tipo? ¿Hay algún impedimento conceptual que convierta en contradictorio un proyecto o estado de este tipo?

En primer lugar, parece que hay consenso con respecto a la opinión según la cual el autoengaño no es un estado en el que se pueda entrar de un modo directo. Esto de se debe principalmente a una poderosa razón:

Uno no puede abrigar directamente *dos creencias contradictorias,* pues sabría que al menos una no puede apuntar a la verdad. Parece que ni siquiera podría adquirir *una* creencia a voluntad, al margen de lo que apoya su evidencia. Esto se debe a que el sujeto sería consciente —como hemos visto que advirtió Williams (1973)— de que su estado cognitivo no tiene más visos de ser verdadero que falso. No habría nada que le condujese a asentir que es más probable que *p* sea verdadero que falso, y es esto en lo que consiste una creencia. Esto es básicamente lo que quiere decir que las creencias apuntan a la verdad.

No parece prometedor, tampoco, acudir en esta ocasión a distintos tipos de creencias, más o menos fuertes. Aunque las creencias pueden admitir grados, sólo hay creencias cuando uno considera que dispone de un conjunto significativo de evidencia relevante (esto no es algo objetivamente mesurable; tiene que ver con una *apreciación* que el sujeto hace y que, aunque suele ser fiable en la mayoría de los casos, puede ser errónea) y esta evidencia le produce el sentimiento de que es más probable que *p* sea el caso que que no lo sea. Desde luego, la creencia puede ser falsa por diversos motivos: porque se ha formado sobre *poca* evidencia o evidencia *no relevante* (es decir, no era un conjunto significativo), porque era evidencia errónea, porque ha realizado inferencias infelices, etc. Admite grados, porque el rango de creencias abarca desde el grado de asentimiento hasta el convencimiento.

Del mismo modo, hay estados cognitivos con respecto a la verdad de algo en los que, *grosso modo* y en términos de probabilidad subjetiva, el valor que le damos a la verdad de algo es menor de 0,5. A veces, por ejemplo, uno tiene un débil sentimiento de que algo puede ser el caso. Es decir, considera que dispone, después de todo, de poca evidencia; o, dicho de otro modo, estima que *carece de un conjunto significativo de evidencia*. Por tanto, considera que *su* evidencia de que *p* no es garantía en absoluto de que *p* sea verdad y que *no le justifica para considerar que así sea*, pero, a la vez, siente que la evidencia de la que dispone hace *p* más posible que *no-p*. De cualquier manera, en tanto que tiene algún indicio de que posiblemente *p* sea el caso, el sujeto tiene un sentimiento hacia la verdad de *p*, pero no lo cree, lo sospecha; no asiente a que *p*, sino a que es posible que *p*. Esta

sospecha puede llevarlo a buscar más evidencia que la confirme o disconfirme. Si la confirma, de modo que llega a considerar que su evidencia es ya significativa y le parece que es más posible que *p* sea el caso que que no lo sea, entonces tendrá una creencia.

En otras ocasiones, uno tiene un deseo, esperanza o fe. A diferencia de otros estados cognitivos como el conocimiento, la creencia y la sospecha, que son de tipo epistemológico, el deseo, la esperanza y la fe no están *directamente* relacionados con la evidencia, sino que con cuestiones valorativas o axiológicas esenciales de manera que, en estos casos, uno *prefiere* que *p* sea el caso que que no lo sea. No obstante, aunque no necesitan esencialmente de evidencia, es cierto que, a veces, pueden verse *afectados* por la evidencia. Así, la evidencia favorable al estado de cosas que uno desea, puede darle esperanzas o fe renovadas. Y al contrario, la carencia de evidencia de cariño por tu parte, puede acrecentar mi deseo de tal cariño. En cualquier caso, lo que quisiera subrayar es que la evidencia no es necesaria ni para tener esperanza ni para tener un deseo, a diferencia de lo que ocurre en las sospechas, creencias, convicciones o conocimientos.

Planteo esta tosca clasificación para aclarar que, en el intento de evitar la paradoja de las creencias contradictorias que encierra el autoengaño, no parece un camino prometedor ver el fenómeno como el mantenimiento de una creencia por un lado, y una esperanza, deseo, fe, sospecha, etc. por el otro. En primer lugar, porque los estados que no dependen esencialmente de evidencia (deseos, esperanzas, fe) son trivialmente

compatibles con cualquier creencia. Así, uno puede tener la creencia de que hay injusticias en el mundo, y tener fe en la "Justicia Divina". Si nos parece que hay injusticias, o bien es que no entendemos qué ocurre, o bien Dios nos pone a prueba para fortalecer nuestra fe, o bien Dios nos lo compensará en una vida futura. En principio, la fe siempre es compatible con cualquier creencia factual porque está *empíricamente blindada*. Evidentemente, una creencia de tipo religioso o teológico puede ser incompatible con otra aunque ambas estén blindadas empíricamentes; las razones en tales casos para rechazar la inconsistencia son conceptuales. En todo caso, hemos estipulado que las creencias religiosas, si carecen de evidencia, son más bien convicciones o artículos de fe. Dos artículos de fe pueden ser incompatibles; un artículo de fe y una creencia en sentido estricto, *siempre* se pueden compatibilizar.

Pensar en el autoengaño como en una cuestión de una creencia de que *p* y la sospecha de que *no-p* no mejora mucho las cosas. Pese a que las sospechas requieren menos evidencia, al pensar en creencias y sospechas contradictorias podríamos estar queriendo decir que, o bien

a) la sospecha imposibilita la creencia, pues la creencia es un estado estable (aquello que calma una duda) que queda desestabilizado al entrar en juego la duda en forma de sospecha. [Peirce (1877), p. 253] Pero en este caso, surgirían las mismas paradojas, ya que —aunque en menor medida— las sospechas también apuntarían a la verdad.

O bien

b) la sospecha es compatible con la creencia. En este caso, la compatibilidad muestra una mera trivialidad: cuando tenemos una creencia, sabemos que podemos estar equivocados. En ese sentido, el hecho de que la evidencia total —*pero no toda*— apunta a la verdad de *p,* puede dar lugar a la sospecha de que podamos estar equivocados y *no-p* sea el caso.

Si en lugar de acudir a deseos, esperanzas o sospechas, acudimos a convicciones y conocimientos, el resultado es igualmente infructuoso: en este caso surgen las mismas paradojas que ante creencias contradictorias, pues estos estados apuntan igualmente a la verdad. De hecho, el autoengaño es a menudo definido en términos de conocimiento de que *p* y creencia de que *no-p*. No hay razón para oponerse a que pudiera haber casos de autoengaño consistentes en que un sujeto está convencido de que *p* y a la vez cree que *no-p*. Lo que parece claro es que en todas las reformulaciones surge el mismo problema y bajo ninguna de estas descripciones alternativas resulta posible creer a voluntad o engañarse de modo directo.

Únicamente hemos examinado esta cuestión del autoengaño directo por mor de la completitud de nuestra investigación, pero no le dedicaremos más espacio, ya que ninguna explicación del autoengaño propone tal cosa.

Todas las explicaciones conceptuales suelen concebir el autoengaño como un modo indirecto de adquisición de creencias placenteras en condiciones muy especiales, y las estrategias para explicar el fenómeno de modo coherente pasan por suponer

a) Que el sujeto consigue abrazar la creencia deseada porque previamente de modo intencionado consigue hacer inconsciente la evidencia dolorosa y la creencia apoyada en ella.

O bien,

b) El sujeto tiene una creencia y evidencia dolorosa que desearía olvidar. Esto provoca que comience a sesgar la evidencia y obtenga nueva evidencia adulterada que sustenta ahora una creencia adulterada y contradictoria con la primera. Tiene acceso a ambas creencias, pero lo importante es que, de un modo aún por explicar, consigue no poner en conjunción ambas creencias, ni creencia deseada y evidencia contraria.

Es dudoso si han de incluirse algunas explicaciones no-intencionales dentro del grupo (a). Por ejemplo, las explicaciones de Mark Johnston o Annette Barnes hablan de *propósito* de autoengaño. En algunas ocasiones se refieren al "propósito" como a algo que cumple una función y, por tanto, podría pensarse que es un fenómeno no-intencional que *casualmente* resulta beneficioso. Así, el autoengaño tendría una función, pero no habría modo de explicar por qué unas veces se desencadena el proceso y otras no; en otras ocasiones parece que el propósito es una pseudo-intención o "proceso subintencional", concepto que resulta oscuro y no se entiende del todo sin aceptar un preconsciente o un inconsciente, lo cual veremos a continuación que tiene problemas particulares.[119]

[119] Quizá quiera decir que se produce por debajo del umbral de la consciencia. Sin embargo, sigue sin entenderse qué quiere decir que captamos, por debajo del nivel de la consciencia, cierta evidencia como amenazadora. Al menos Johnston acude directamente al concepto de represión freudiano.

Por otro lado, dejo fuera de la discusión explicaciones que acudan a nociones puramente no-intencionales ya que, como he dicho anteriormente, tales procesos serían indistinguibles del mero error. Por ejemplo, un sujeto, en una situación dolorosa, angustiosa o de tensión, puede carecer de la calma necesaria para evaluar con detenimiento la evidencia y, consecuentemente, equivocarse en la valoración y formarse una creencia errónea. Estos casos no nos parecen instancias de autoengaño y, aunque puedan tener su interés para la psicología, distan mucho de ser filosóficamente relevantes.

Examinemos, pues, las dos alternativas que hemos propuesto y que parecen, a primera vista, las únicas capaces de ofrecer una explicación conceptual satisfactoria acerca del fenómeno:

La propuesta (a) pasa por que el sujeto consiga, siempre de modo intencionado, hacer inconsciente la evidencia dolorosa y la creencia apoyada en ella, para después abrazar la creencia contraria.

Estas propuestas acuden siempre a la noción de inconsciente freudiano, aunque sea en versiones convenientemente actualizadas o remodeladas. Así Audi, Pears, Bermúdez o Johnston creen que hay intenciones inconscientes. Fingarette no habla de intenciones, sino de proceso a propósito (*purposefulness*), lo cual viene a ser lo mismo. Además, señala que era precisamente el postulado de transparencia de la consciencia lo que le ponía en serios aprietos a Sartre cuando trataba de explicar la mala fe como mentira a uno mismo. Fingarette niega tal transparencia. También Talbott habla de "divisiones inocentes" y de la no-transparencia de

algunos procesos de la consciencia. La función de un postulado de no-transparencia es, en estos casos, en todo equivalente al inconsciente: impide al sujeto tener acceso a sus propios estados mentales (es decir, considera que con respecto a algunos elementos, perdemos la autoridad).

Todos estos enfoques creen que el único modo de dar cuenta del autoengaño, pasa por encapsular o aislar la evidencia y las creencias que resultan dañinas en la vida consciente del sujeto. Con este movimiento, consiguen que para el sujeto quede únicamente disponible evidencia consonante con el estado de cosas que desean que sea el caso. Sin embargo, dado que el sujeto necesita sustento para la nueva creencia, y una búsqueda sin ningún tipo de filtro resultaría fatal, también contemplan que el sujeto adultere los procesos de adquisición de evidencia. Nuevamente, el hecho de que las creencias *apuntan a la verdad* les conduce a suponer que la intención de adulterar estos procesos ha de ser desterrada de la vida consciente del individuo y convenientemente aislada.

Unos suponen que el destino de todas estas cogniciones disonantes e intenciones de adulteración, puede ser algo similar al inconsciente freudiano (Audi o Bermúdez). Otros suponen que este lugar ha de ser un subsistema, pero, aunque sea un lugar apartado de la consciencia, no puede ser hogar del caos y el desorden: ha de tener algunas características, como consistencia interna (Pears). Por esta razón, suponen que los subsistemas incluyen todo aquello que necesitan para su consistencia, a saber: todo aquello del sistema principal que no genera inconsistencia en el subsistema. Y a la inversa: el sistema principal incluye todo menos lo

que generaría inconsistencia, es decir, aquello que provocó la segregación del subsistema y su aislamiento del sistema principal. El dibujo consiste en conjuntos —sistema y subsistemas— solapados.

No queda claro si este aislamiento es algo que intencional y conscientemente hace el sujeto — ¿de qué otro modo podría hacerlo? Además, la exigencia de coherencia interna del subsistema que reclama Pears parece un mero capricho y, probablemente, nos forzaría a suponer la existencia de un subsistema para cada objeto de autoengaño u olvido (porque la evidencia y creencias que causan la segregación no tienen por qué ser consonantes, y cada subsistema si ha de serlo). En cualquier caso, no veo la necesidad de suponer que haya consistencia interna en los subsistemas. Está claro que la necesidad de consistencia en el sistema principal es algo dependiente de que en él rige la consciencia: tanto de las cogniciones que forman parte de él, como de nuestra tendencia a la verdad. Si somos conscientes de que dos cogniciones se contradicen, o bien no sostenemos ninguna de ellas hasta encontrar más evidencia, o bien rechazamos una de ellas. Pero ¿qué nos fuerza a suponer consistencia en un subsistema de cuyos elementos, por hipótesis, no podemos ser conscientes? ¿Por qué no verlo como un cajón de sastre? No encuentro ninguna razón que impida la inconsistencia de un subsistema tal.

En cualquier caso, no son estas dificultades las que nos invitan a abandonar una explicación de este tipo. Mele, por ejemplo, sostiene que sin duda es concebible pensar en intenciones ocultas "freudianas" y conjeturar que tengan efecto en nuestro modo de actuar. Pero dado que

no tenemos ni justificación empírica ni necesidad conceptual de hacer uso de ellas para dar cuenta del fenómeno, es *injustificada* como explicación de los casos estándar de autoengaño [Mele (2001), p. 17]. Según Mele, son "exotismos mentales" innecesarios.

Sin embargo, tengo la impresión de que Mele sólo se resiste a aceptar este tipo de explicaciones porque su deseo no es explicar el autoengaño desde un punto de vista *conceptual*, sino ofrecer una explicación coherente y verosímil de los casos comunes y, por tanto, empíricos. De hecho, Mele aclara que tal y como él ve el autoengaño, es una *noción explicativa*, no una verdad conceptual [Mele (2001), p. 10].

No obstante, aun admitiendo que es cierto que estas hipótesis están empíricamente blindadas porque acuden al inconsiente y los contenidos del inconsciente son por definición inaccesibles, creo que no podemos rechazarlas como *explicación puramente conceptual* acerca de la posibilidad del autoengaño.

Más bien, lo que me impide aceptar este tipo de propuestas es que sospecho que cometen la *falacia homuncular* [cf. Sartre (1943), p. 102; Fingarette (1969), p. 78; Rorty (1988), p. 19; Johnston (1988), p. 65]. Suponen que el yo está dividido en distintos sistemas —sean de la naturaleza que sean— pero a su vez, no explican cómo algo pasa a formar parte del subsistema. No me refiero a que no ofrecen una explicación psicológica de cómo pueda producirse el proceso, sino a que conceptualmente resulta difícil de entender.

Si aquello que se aísla del sistema principal es algo que en un primer momento fue captado como disonante o contradictorio con los elementos del sistema principal, ¿por qué el sistema principal no lo rechazó directamente? ¿Por qué rechaza unas cosas y otras no? ¿Es el yo consciente el que intencionalmente segrega cogniciones fuera del sistema principal y las envía al inconsciente? Sea cual sea la explicación elegida, hay algo invariable: el sistema principal, Yo, o como quiera denominarse al sistema consciente, ha de concebir tales cogniciones *como algo que ha de ser reprimido*. Pero entonces, no sólo sabrá que hay algo relevante que está ocultando, sino que sabrá también que aquello que forma parte de su vida consciente es parcial, sesgado; desconoce parte de la información y no puede acceder a ella (porque la ha hecho inconsciente), pero sabe que aquella de la que dispone *no es toda la información*.

Sin duda, los defensores de este tipo de explicación pueden argumentar que el sistema principal podría reconocer esta dificultad y hacer inconsciente el propio proceso de censura, de represión de la evidencia y creencia dolorosas (o de segregación de subsistema o, en cualquier caso, el proceso por el que se suponga que escinde o segrega parte del sistema principal). De nuevo, la objeción más notable es que para iniciar un proceso de ocultación de la censura porque supone una dificultad para llevar adelante el engaño, el sujeto ha de ser consciente de la censura *como tal dificultad*. Se produce así un regreso *al infinito* de cogniciones disonantes que se perciben como amenazadoras y se reprimen.

Quizá la única escapatoria viable implica suponer que es algún tipo de elemento subpersonal —y no el yo consciente— lo que decide qué puede acceder a la consciencia y qué no. Pero es entonces cuando damos el paso fatal que nos conduce a la homuncularidad: hemos admitido una consciencia independiente dentro de la consciencia, que discrimina lo que es aceptable de lo que no, pero no hemos resuelto el problema, sino que lo hemos reproducido a menor escala. De igual modo, tendríamos que explicar como es que esa consciencia capta la inconsistencia sin verse obligada a rechazar una de las dos cogniciones, si no queremos volver a suponer una represión que genere otro subsistema dentro del subsistema, y así sucesivamente.

No se me ocurre ningún buen argumento para evitar esta dificultad. Johnston creyó encontrar un modo de evadir esta falacia acudiendo a procesos subintencionales. Sin embargo, no se entiende muy bien qué puede ser un proceso subintencional: si es un proceso intencional disfrazado, los problemas siguen ahí. Si es un proceso que resulta funcional pero es inconsciente y no intencional, tiene un problema aún mayor: al eliminar la intención, vuelve a hacer del autoengaño un error indistinguible del pensamiento desiderativo o incluso de la mera ceguera intelectual. De hecho, como ya vimos, Johnston cree que el autoengaño es una especie de pensamiento desiderativo, pero con esto se renuncia a su especificidad e importancia.

Quizá sería apropiado que desistiésemos de cualquier explicación que acuda a procesos inconscientes, no ya porque esté empíricamente blindada, sino porque parece *conceptualmente* insostenible.

La propuesta b) nos dice que el sujeto tiene un conjunto de evidencia y creencia que encuentra particularmente insoportables. Esto provoca que comience a sesgar la evidencia y obtenga nueva evidencia que sustenta ahora una creencia adulterada y contradictoria con la primera. Lo que distingue a este enfoque es que el sujeto tiene en principio acceso a ambas creencias, pero, de un modo aún por explicar, consigue no poner en conjunción ni las creencias, ni creencia deseada y evidencia contraria. Ésta es la propuesta de Davidson.

Como hemos visto, el resto de enfoques o bien evitan atribuir intención al sujeto (no-intencionalistas), o bien suponen que el único modo de evitar la incoherencia del estado y la consciencia del plan para engañarse, es postular que habría una parcela de la mente inaccesible para el sujeto, bien porque algunos procesos no son transparentes, bien porque hay subsistemas con contenidos desterrados de la consciencia y, por tanto, inconscientes.

Davidson cree que el autoengaño no puede consistir en un mero error al evaluar la evidencia. El autoengaño ha de incluir *debilidad de la justificación*, es decir, uno ha de formarse una creencia en contra de lo que le indica su evidencia total. Sobra decir que un sujeto puede tener evidencia, pero no advertir *de qué* es evidencia. Aun más: puede advertir de qué es evidencia, pero no percatarse de que la evidencia total apuntaba abrumadoramente

en una dirección. Davidson indica que, sin duda, todos estos casos son posibles, e incluso "hay un sinnúmero de preguntas que la tortuga puede hacerle a Aquiles" acerca de las posibles razones por las que uno podría no cometer debilidad de la justificación. [Davidson (1985), p. 104]

Lo que está claro es que el autoengaño incluye debilidad de justificación y que lo que causa la violación del *requisito de evidencia total* y hace posible esta debilidad, es la creencia evidencial que le producía un dolor insoportable. Sin dejar de percibir la toda su evidencia, se forma una creencia sobre *sólo una parte* de lo que considera como evidencia relevante [Davidson (1985), 106-107]. Lo que hace que el autoengaño sea algo más que mera debilidad de la voluntad, es que ésta es *autoinducida* por el sujeto.

Así, Davidson piensa que aunque el autoengaño no pueda ser totalmente caracterizado bajo el modelo del engaño interpersonal, seguramente algunas de sus características han de estar presentes: la consideración simultánea de creencias incompatibles y la intención de abrazar una creencia en contra de la evidencia total de la que dispone, han de ser algunas de sus características esenciales. Lo importante es que Davidson dice que es posible creer un conjunto de proposiciones contradictorias entre sí, siempre y cuando la contradicción no sea obvia. En los casos de autoengaño, la existencia de creencias obviamente antagónicas hace necesaria la suposición de unos límites que las separen, pero *no ve ninguna razón por la que alguno de los territorios hubiese de estar vedado a la consciencia.* Lo que sí está claro es que el sujeto no podría contemplar el todo sin borrar los límites, descubrir la inconsistencia y rechazar alguna de

las creencias. [Davidson (1985), p. 211] Por tanto, de algún modo, el autoengaño ha de suponer una consciencia no completamente integrada.

Aunque el propio Davidson reconoce una deuda con otros enfoques como el de Pears o Bach, también indica que su propia explicación muestra de modo más claro el estado irracional inherente al autoengaño. Si Davidson no separa las creencias acudiendo al supuesto de la inconsciencia, es porque cree que disolvería el problema de irracionalidad: generaría un baúl donde todo cabe, que no sabemos cómo rayos llega ahí y no podemos inspeccionar. El problema es que se pierde de vista la tensión que a todos nos parece inherente al autoengaño, tanto entre las distintas creencias como entre evidencia total y creencia contraevidencial.

Además Davidson señala que no se entendería cómo desde el inconsciente, la evidencia y creencia displacenteras siguen sustentando la nueva creencia. Para Davidson es central subrayar que la creencia original es *causa* de la nueva creencia, sin ser *razón* para ella. Sin embargo, también reconoce que su postura hace más difícil comprender cómo pueda producirse esto psicológicamente. No obstante, cree que sería un error tratar de ofrecer una explicación psicológica.

La postura de Davidson, aunque logra esquivar algunos problemas de otras teorías intencionales a la vez que muestra el elemento irracional que nos parece que debería reflejar el autoengaño y otros disuelven, tampoco nos parece absolutamente satisfactoria. La causa de esta insatisfacción no reside en que, como él mismo dice, ninguna teoría vaya a ser capaz de dar cuenta de todos los casos porque estos sean de muy distinto tipo y grado

[Davidson (1993), pp. 221, 230]. Nuestra reticencia se debe, de nuevo, únicamente a cuestiones conceptuales.

En primer lugar, uno de los reproches que Davidson le hace a otras teorías como la de Bach, es que no incluyen lo que él llama *debilidad de la justificación,* es decir, niegan que sea posible que un sujeto crea algo en contra de su evidencia (vid. §I.6) [Davidson (1985), p. 210]. No obstante, creo que se podría devolver la pelota al tejado de Davidson. Esta característica de las creencias no sólo la encontramos en Kant Bach, sino también en David Hume, Bernard Williams, Henry Price o William Alston, entre otros, y no parece en absoluto descabellada. Davidson tampoco ofrece un argumento para defender que sea posible creer en contra de la evidencia, al menos, no en contra de aquella de la que uno dispone —y no de la disponi*ble*—, que es a la que Davidson se refiere. Por mi parte, mi concepción de aquello que es una creencia supone que ante un conjunto significativo de evidencia de que *p* es el caso, uno no puede sino sentir que *p* es más probablemente verdadero que falso y, consiguientemente, asentir a que *p*.

No obstante esta característica de las creencias no impide el autoengaño: simplemente exigiría que el sujeto adulterase su conjunto total de evidencia e intentase esconderse cierto cuerpo de evidencia dolorosa de modo que consiguiese que sólo la evidencia agradable le resultara accesible a la consciencia. Pero entonces, un sujeto que creyese bajo estas condiciones, lo haría apoyado en *toda la evidencia* que tras el

proceso de represión le es accesible a la consciencia (sólo la consonante y agradable) y, por tanto, dejaría de mostrar debilidad de justificación.

Es previsible que Davidson se haya mostrado reacio a hacer este movimiento, principalmente por dos razones: en primer lugar, porque reintroduciría todos los problemas homunculares que aquejan a las teorías que acuden al inconsciente y, en segundo lugar, porque desharía la irracionalidad que él desea reflejar. Para Davidson es esencial mostrar que quien se autoengaña cree algo *en contra de su evidencia total*, y si el sujeto de algún modo no es consciente de algo, no es posible dar cuenta de la debilidad de la justificación. En cierto modo, Davidson se encuentra en una situación similar a la que se le presentaba cuando pretendía rebatir la concepción aristotélica de la debilidad de la voluntad: dado que no quería perder el elemento de irracionalidad e inconsistencia, negaba que la *akrasia* fuese un caso de mera ceguera intelectual, como se seguía de la exposición de Aristóteles. En aquel caso, se aferró a que el sujeto toma en algunas ocasiones en consideración un conjunto de evidencia *prima facie* total. El problema es que en resumidas cuentas, la razón por la que el conjunto de evidencia era sólo *prima facie* total, y no era el conjunto total hechas todas las consideraciones, era que el sujeto tenía un motivo para no contemplar toda la evidencia. Es decir, acababa por reintroducir el elemento emocional que nos nublaría el juicio, y por tanto, no conseguía separarse de la explicación aristotélica.

De modo similar, al tratar de explicar el autoengaño como un proceso irracional, Davidson se encuentra con que suponer que el sujeto no

contempla toda la evidencia por la influencia de las emociones es algo muy común, pero es mera ceguera intelectual. El elemento de debilidad de la justificación es indispensable, pero obliga o bien a

a) Afirmar que el sujeto se forma su creencia sobre un conjunto de evidencia sólo *prima facie* total (es decir, parcial),

O bien

b) recurrir a algún tipo de *olvido intencionado* o *represión* que oculte parte de la evidencia.

Como ya hemos indicado, Davidson rechaza explicar el autoengaño por medio de mecanismos de represión o del inconsciente. Cree que esto facilita la comprensión conceptual del *proceso*, pero desharía el problema de la irracionalidad del *estado*. Por eso, supone que todo ha de poder ser accesible al sujeto, aunque ciertamente no sea posible que contemple *a la vez* ambas creencias: es "una conciencia no totalmente integrada". Sin entrar a valorar si su postulado de una línea divisoria es *ad hoc* o no, parece que Davidson no logra encontrar un modo de explicar cómo sería conceptualmente posible que un sujeto tomase en consideración la evidencia sólo *prima facie* total, sin verse abocado a suponer o bien un mecanismo no intencional (como la ceguera intelectual) o bien un mecanismo como el subconsciente. Lo que no parece posible, en todo caso, es que el sujeto decida voluntaria y conscientemente, tomar en consideración sólo una parte de la evidencia sin dar lugar a que no pueda creer nada en absoluto, pues las creencias aspiran a la verdad. Davidson ha explicado cómo sería conceptualmente posible un *estado* en el que un

sujeto mantuviese creencias que son contradictorias pero, como él mismo reconoce, no es capaz de explicar a la vez conceptualmente el proceso que produciría esto. Si se trata de explicar el proceso, se disuelve la irracionalidad del estado y si se da cuenta de la irracionalidad del estado, no se entiende cómo pueda concebirse el proceso que dé lugar a ese estado.

Pero el autoengaño, si ha de ser posible —al menos conceptualmente— ha de ser *un proceso que dé lugar a un estado*. Parece que una teoría que no sea capaz de dar cuenta *a la vez de ambas cosas*, no puede resultar satisfactoria.

Hemos visto que las explicaciones no-intencionales o bien eran insuficientes (por reducirse a algún tipo de error causado por la pasión, angustia, dolor, etc.) o bien introducían elementos "intencionales disfrazados". Los enfoques intencionalistas sí recogen la tensión, pero precisamente esta tensión daba lugar a las paradojas estática y dinámica (las creencias contradictorias y el conocimiento de la propia estrategia de engaño) y, para tratar de solucionar esto, unas veces recurren al inconsciente u otros primos conceptuales (no-transparencia, represión, olvido motivado, etc.) y otras suponen una línea divisoria entre las creencias que las mantenga separadas y dé lugar a una consciencia no completamente integrada. Los primeros incluían nociones oscuras, blindadas empíricamente, y que nos conducen a un regreso *ad infinitum* o a la falacia homuncular. Los segundos evitan estos problemas, pero hacen más difícil de explicar el proceso y, finalmente, al no conseguir explicar

satisfactoriamente por qué el sujeto no integra toda su evidencia, o bien recurren a la ceguera intelectual o a algún tipo de olvido intencional similar a la represión, con lo que no han conseguido ganar nada. Estamos de acuerdo con Bach cuando afirma que está claro que el autoengaño no es ni un caso de ceguera intelectual, pues "quien se autoengaña ve todo demasiado bien, ni un caso de pensamiento sesgado, ya que cuando acusamos a alguien de estar prejuiciado o sesgado, simplemente queremos decir que está influenciado por sus sentimientos" [Bach (1981), pp. 351-352].

Como dije al principio, uno puede hablar de autoengaño para referirse de un modo confuso a fenómenos no excesivamente problemáticos, como la ceguera intelectual o la mala fe entre otros, pero el autoengaño, en tanto que engaño, sólo parece concebible acudiendo a una caracterización intencional que incluya de algún modo la consideración de creencias contradictorias. Desgraciadamente, parece que se torna conceptualmente muy difícilmente resoluble cuando sacamos las consecuencias de aquello que requeriría un enfoque de este tipo, se hagan los arreglos que se hagan.

No he considerado, sin embargo, enfoques intencionales más modestos. Bajo una teoría de esta clase, no se exige la consideración de creencias contradictorias, ni subsistemas, ni procesos inconscientes. El autoengaño podría ser concebido como un estado en el cual uno tiene una creencia displacentera, y simplemente decide obviarla o no hacerle mucho caso. En un caso de este tipo, no se trata *por hipótesis* de que el sujeto deje de tener la creencia, ni siquiera que la olvide; precisamente la tiene

presente, de manera que trata de sobrellevar sus consecuencias del mejor modo posible, evitando —en la medida que le se lo permitan sus habilidades, conocimientos previos y circunstancias— exponerse a *más* evidencia dolorosa. Podemos conjeturar que una actitud así podría encerrar bien el deseo de que las cosas volviesen a su cauce, bien el deseo de no actuar precipitadamente de un modo del que después pueda arrepentirse; también podría ser una muestra de una disposición a cambiar las cosas, una negación a reconocer un error del pasado o un intento de evadir la responsabilidad sobre una determinada acción.

Quizá alguien desee decir, ante la verosimilitud de cualquiera de estas hipótesis explicativas, que el hecho de que este sujeto no desee enfrentarse a la realidad es una muestra de que se engaña. Y, desde luego, es perfectamente posible que su actitud responda a un proyecto intencional. Sin embargo, no parece que un individuo pudiese, en estas condiciones, llegar a olvidar la creencia dolorosa, aun ni siquiera como un efecto indeseado de sus acciones intencionales de mirar hacia otro lado. Como se ha subrayado en muchas ocasiones, tratar de olvidar algo o mirar hacia otro lado para no verlo, tiene como consecuencia un mayor afianzamiento y presencia de esa cognición. Es como el insomne que trata de dormir intentado dejar la mente en blanco y no pensar en que no puede dormir: no conseguirá pegar ojo en toda la noche salvo, quizá, en caso de que, cansado, olvide tal intención.

En cualquier caso, no me parece adecuado llamar a esto autoengaño debido a que no consigue olvidar la creencia dolorosa o angustiosa, ni

tampoco puede creer lo contrario; a lo sumo es, como señalara Elster, un intento infructuoso de engañarse a uno mismo, no más paradójico que el intento de alcanzar cualquier otro estado incoherente, o imposible [Elster (1979), p. 294]. También es posible ver su actitud como un intento de engaño a otros, bien para hacer que éstos se tambaleen en sus creencias y dejen de ofrecerle evidencia dolorosa, bien para no sufrir la humillación de que sepan que es consciente de algo (una infidelidad, por ejemplo), etc.

Desde luego, uno puede estar de acuerdo en que todos estos casos son posibles y, aun así, preguntarse si no puede haber otros casos concebibles en los que *genuinamente* el sujeto no trate de engañar a otros, sino a sí mismo y de algún modo lo consiga. Aquí se abren tres posibilidades:

a) Lo intenta y no lo consigue, lo cual no mostraría autoengaño, sino intento de autoengaño.

b) Lo intenta y lo consigue. Ya hemos visto que no hay modo de comprender conceptualmente cómo un sujeto puede lograr esto, ni de modo directo, ni indirecto.

c) No lo intenta, pero lo consigue. Examinaremos a continuación esta vía.

Los enfoques de este tipo son los que han tenido un mayor auge en los últimos años. Aceptando que las paradojas a las que nos conduce un modelo intencional son insalvables, Mele y otros han tratado de ver de qué modo sería conceptualmente posible el autoengaño sin acudir a una intención de engañarse.

Hemos indicado ya que una teoría que elimine por completo la intención como ingrediente, se ve directamente forzada a admitir que no es capaz de distinguir el mero error del autoengaño. Mele es consciente de esto, y trata de buscar una explicación alternativa.

Su estrategia es hacer de la acción intencional un proyecto "no transitivo". Si yo sé que A implica B, y hago intencionalmente A, esto no quiere decir que haga intencionalmente B. Sin duda, esta intuición es a menudo correcta: si yo sé que llevando el almuerzo a la facultad unos gamberros me asaltarán y aun así decido llevar el almuerzo, esto no implica que decida que me asalten los gamberros.

Mele sostiene que, en casos en los que la evidencia es insoportable para un sujeto, éste puede, y a veces decide, apartar intencionalmente la vista de esta evidencia, sesgarla, focalizar su atención sobre evidencia placentera, etc. Estas maniobras pueden dar lugar a que el sujeto esté engañado. El individuo no ha intentado engañarse, aunque las actividades que ha llevado a cabo intencionalmente le hayan conducido a tal engaño.

Sin embargo, el problema de una explicación de este tipo es que Mele no menciona que, aunque yo no tenga la intención de que me asalten, *ex hypothesi* sí soy consciente de que mi acción tendrá esas consecuencias. Por supuesto, la consciencia de las consecuencias que acarreará o puede acarrear algo que yo decida hacer, no me hace responsable de ello, ya que no implica que tenga la voluntad de que estas consecuencias sean efectivas. No obstante, la dificultad está, en el caso de la formación de creencias, en que uno es consciente de que su intención de adulterar los

métodos de adquisición de evidencia traerá como consecuencia una evidencia adulterada, es decir, una evidencia que no apunta a la verdad, que no es indicativa de lo que es el caso. La naturaleza de la creencia como estado mental que apunta a la verdad, impide conceptualmente que yo sea consciente de que mi evidencia está adulterada y, a la vez, la tome como garante de la verdad de *p*. Ésta es la razón por la que el enfoque de Mele nos parece insatisfactorio.

Apelar a que el proceso de adulteración de la evidencia es involuntario y provocado por una fuerte pasión que nos nubla el juicio sólo empeora las cosas. Estoy totalmente de acuerdo en que el temor y la angustia tienen a veces una influencia decisiva en nuestro modo de adquirir evidencia. Como vimos en §V.2, la evidencia no es una pasta de datos que el sujeto capta objetivamente. Más bien, supone a veces una ponderación de los datos que obtenemos, y otras veces, una interpretación de las expresiones y conductas de otros sujetos que, más que con una habilidad hermenéutica, tiene que ver con una capacidad estética que consiste en una apreciación de algo que *no* es ponderable. Esta interpretación es, qué duda cabe, falible. Entre otras cuestiones, porque nuestras emociones o pasiones pueden influir en el modo en que buscamos evidencia. Por ejemplo, podemos buscar evidencia en una dirección y no en otra; en un lugar y no en otro; buscando confirmar alguna intuición o corazonada, etc. Uno de los aspectos más importantes que el propio Mele indica, es que las emociones y sentimientos pueden influir en los umbrales de aceptación de evidencia. Así, los sujetos tienden a exigir menos evidencia para aquello que confirme lo que desean, que para rechazarlo. Uno eleva los niveles de

escrutinio cuando lo que ve no le gusta: necesita más evidencia para aceptar la verdad de ese hecho.

No vamos a detenernos en esta cuestión, pues es la psicología empírica quien puede y debe dirimir estos asuntos. Lo que me interesaría subrayar son dos cuestiones. En la explicación conceptual que estamos evaluando ahora, suponemos que es necesario que

1. Las emociones y sentimientos tienen influencia en nuestras políticas de creencia. Afectan al lugar en que buscamos evidencia, al modo en que interpretamos algo —pues los sentimientos nos pueden llegar a cegar—, a los niveles de escrutinio y umbrales de aceptación, e incluso pueden paralizarnos y dar lugar así a un cese de acumulación de evidencia no pretendido.

2. Esta influencia no es intencional.

Sin duda, no es contradictorio pensar en una situación así, e incluso, me parece que son muy frecuentes. Pero ¿qué diferencia conceptual hay en los casos más suaves con el mero error y, en los casos más extremos, con la ceguera intelectual?

Las creencias que abrazamos bajo una influencia emocional pueden ser falsas, desajustadas. En ese sentido uno puede decir, si quiere también, que estamos engañados. A veces se dice: "me engañé al pensar que podía confiar en ti", sin implicar por ello que se haya hecho un esfuerzo consciente por alcanzar la confianza en esa otra persona. Éste es un uso, en todo caso, derivado, como lo es también el uso: "si no estoy engañado,

las llaves están sobre el escritorio". En este caso no hay ni intención ni pasiones de por medio, y "engañado" simplemente significa "equivocado". Qué duda cabe que podemos estar equivocados en muchas ocasiones y que, en algunas de ellas, han sido nuestras emociones las que nos han cegado y conducido a esa equivocación; pero no me parece que ninguna de estas situaciones plantee un reto explicativo o conceptual serio similar a aquél al que nos enfrentábamos al comenzar este trabajo.

Después de este análisis, en el que he intentado someter a examen a todas las posturas formalmente posibles, no he encontrado ninguna alternativa que resulte totalmente aceptable. No niego de antemano que alguna teoría concreta que pueda llegar a dar cuenta, de modo satisfactorio, a la vez del estado y del proceso del autoengaño sin reducirlo a ceguera intelectual o pensamiento desiderativo. Sin embargo, no veo muy bien cómo puede llevarse adelante esta tarea sin acudir a nociones blindadas empíricamente.

VII. LA ATRIBUCIÓN DE AUTOENGAÑO

En este capítulo me propongo estudiar el autoengaño desde otro ángulo. Si en el anterior capítulo evaluábamos las distintas explicaciones conceptuales de aquello en lo que habría de consistir que un sujeto se autoengañase, tanto del proyecto que debería emprender como del estado —posiblemente inestable— al que llegaría, en este capítulo vamos a intentar examinar por qué los individuos atribuyen autoengaño a otros —aunque normalmente no a sí mismos, salvo en tiempos pasados—; qué les empuja a hacer tal atribución de irracionalidad y qué pueden estar diciendo al hacerla.

No es necesario insistir en que, efectivamente, incluso en el supuesto de que tuviésemos un acceso privilegiado a nuestros estados cognitivos, esto no implicaría que no pudiésemos tener cogniciones inconsistentes entre sí. Ya hemos señalado que, principalmente porque somos seres racionalmente limitados, un sujeto no puede hacer infinitas inferencias, ni siquiera un número muy elevado de ellas. Igualmente, no todos los individuos tienen sus capacidades inferenciales desarrolladas por igual y, en todo caso, cabe el error en el proceso inferencial. Además, estas

inferencias dependen casi siempre de modo crucial de recuerdos, datos de los sentidos, etc., que pueden estar equivocados.

Esta limitación puede dar lugar a que un sujeto tenga cogniciones inconsistentes de modo no problemático, precisamente porque le pasan desapercibidas. De hecho sabemos que las redes doxásticas de los individuos suelen incluir —si es que no lo hacen siempre— contradicciones internas. Este tipo de inconsistencias son detecta*bles* y no plantean una seria dificultad, ni para el sujeto ni para el intérprete. Si el sujeto descubriese, o le hiciésemos ver, que tiene cogniciones inconsistentes, en principio ha de imaginarse que rechazaría alguna de ellas o, cuando menos, no sabría ya qué pensar.

¿Cómo se produce la atribución de estados mentales a otros? Bueno, parece que queda fuera de toda duda que nuestro acceso a los estados mentales de los demás no puede ser en modo alguno directo. Estén o no los significados en las cabezas, parece natural pensar que al menos lo están *en parte*, y que nuestro modo de saber lo que otro piensa, es siempre una *inferencia a la mejor explicación* de su conducta, verbal y no verbal, tomando en consideración ciertos elementos, como otras conductas suyas, lo que hizo en el pasado, etc. Que hagamos una inferencia a la mejor explicación quiere decir, simple y llanamente, que al menos en principio, no debemos atribuirle estados mentales extravagantes o inconsistencias: esto es lo que Davidson llamó, como vimos en §III.1.3.2, el Principio de Caridad. Este principio es condición de posibilidad de la interpretación, ya que sólo

si las creencias que le atribuimos no son extravagantes, inconsistentes, etc., podremos comprender lo que dice. Además,

> […] puesto que una creencia no puede mantener su identidad al perder sus relaciones con otras creencias, no es posible que la misma proposición sirva para interpretar actitudes particulares de dos personas distintas y guarde al mismo tiempo, con las demás actitudes de una de ellas, relaciones muy diferentes de las que guarda con las de la otra. [Davidson (1985), pp. 105-106]

Por esta razón, cuando no somos capaces de interpretar lo que hace un sujeto sin tener que atribuirle dos creencias que son contradictorias, corremos siempre el riesgo de que toda su conducta se haga ininteligible para nosotros. Hay una tensión esencial en estos casos entre dos elementos: por un lado, *ex hypothesi* no entendemos nada si no le atribuimos dos creencias contradictorias[120], y deshacer esa inconsistencia nos empujaría a atribuirle creencias extravagantes y quién sabe qué otras inconsistencias a él y a los demás. Por ello, lo más caritativo es atribuirle esas creencias. Pero por otro lado, el Principio de Caridad establece que la inconsistencia o incoherencia conduce a la ininteligibilidad; esta dificultad es conocida como la paradoja de la irracionalidad:

> [S]i explicamos algo irracional demasiado bien, alcanzamos una forma encubierta de racionalidad, y si simplemente asignamos incoherencia de un modo sospechosamente demasiado sencillo, simplemente estaremos poniendo en entredicho nuestra capacidad para diagnosticar

[120] Por supuesto, le atribuimos es que cree que *p* y que cree que *no-p*, y nunca le atribuimos la creencia en la conjunción de dos hechos contradictorios, esto es, que cree que *p* y *no-p*.

> irracionalidad, al alejarnos del trasfondo de racionalidad necesario para justificar cualquier diagnóstico.[121]

Ya vimos que la salida a este dilema está, según Davidson, en ver esto como una cuestión de grado: mientras las diferencias entre nuestras creencias y las del sujeto y las incoherencias sean suficientemente pequeñas, no supondrán un gran problema a la hora de predecir y explicar su comportamiento, pero cuando se hacen demasiado grandes, nos resulta del todo incomprensible [Davidson (1982), pp. 183-184].

Ciertamente, el experimento mental que imagina Davidson *no* versa sobre casos empíricos en los que uno no tenga *de facto* otro modo de interpretar la evidencia; se trata de un caso hipotético en el que *de iure no tenemos otro modo* de interpretar lo que hace el sujeto que adscribiéndole creencias contradictorias. Por tanto, Davidson no está tomando en cuenta aquellas situaciones en las que el intérprete pudiera llegar a conocer de algún modo algo relevante que le disuadiera de su error en la atribución de inconsistencia.

Además, la tarea de interpretación no es, por así decirlo, individual; uno ha de interpretar a los demás como un todo relacional (esto es deudor, en parte, de la aceptación de que el significado depende parcialmente de cuestiones sociales y externas y, en otra parte, de la tesis general del holismo). Así, no sólo no puedo atribuir a un individuo un significado

[121] «[I]f we explain it too well, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, we merely compromise our ability to diagnose irrationality by withdrawing the background of rationality needed to justify any diagnosis at all.» [Davidson (1982), p. 184]

distinto para cada uso que haga de una misma palabra, sino que, en la medida de lo posible, los usos de los distintos hablantes de una comunidad han de coincidir.

Por otra parte, el proceso de interpretación es progresivo, una construcción, en donde en determinadas ocasiones podríamos renunciar a entender algún aspecto particular. No obstante, en ciertos casos se produce una tensión porque no podemos renunciar a la interpretación de ciertas conductas sin vernos abocados a no entender nada. El experimento mental nos dice que es *concebible* que haya casos (no casos aislados, sino en conjunto) en los que sea crucial para el significado global no renunciar a explicar la conducta similar de varios sujetos y, por tanto, hayamos de atribuirles determinados estados mentales; a la vez, no podemos interpretar sus conductas sin adscribirles cogniciones inconsistentes y, además, el modo *más caritativo* de interpretar a estos sujetos es adscribirles una inconsistencia particular.

Davidson sostiene que en estos casos, uno está forzado a interpretar si desea entender, y el modo menos dañino para el conjunto de atribuciones a todos los sujetos, es atribuir dos creencias contradictorias al individuo en cuestión. Debido a que el Principio de Caridad me impide atribuir inconsistencias al sujeto, he se suponer además, que estas creencias se hayan de algún modo separadas, y el sujeto no las contempla *a la vez*.

No voy a poner en duda el valor pragmático de la labor de interpretación. También admito que quizá sea el mejor modo, si no el único, de entender a los demás. Por supuesto es falible, pero esto no es

tampoco en principio una gran dificultad desde el momento que aceptamos que la labor interpretativa es continua y corregible: puede detectar fallos y consecuentemente reasignar intenciones, actitudes, etc.

La dificultad que yo veo, no tiene que ver con su valor *pragmático*, sino *epistémico*. Naturalmente, la interpretación ha de pretender captar los estados mentales de los individuos, y no ser un mero cuento de hadas consistente. En este sentido, hay algo que me hace sospechar de la atribución de creencias contradictorias: la razón por la que atribuimos en los casos de autoengaño creencias contradictorias a un sujeto es porque si deshiciésemos esta incoherencia no podríamos evitar atribuir creencias extravagantes o más inconsistencias a él o a los demás. Esto, sin embargo, no es lo que parece que ocurre cuando ese sujeto se autoengaña, ya que, como el propio Davidson admite, el autoengaño es un caso en el que el sujeto se autoinduce la incoherencia *porque lo que resultaría* coherente *le causa un dolor o malestar insoportable*. Por tanto, el sujeto no se autoinduce creencias inconsistentes porque en otro caso tendría que aceptar más incoherencias. Si en nuestro proceso interpretativo lo que nos fuerza a atribuir autoengaño es la evitación de mayores incoherencias, eso debería hacernos pensar que algo no va bien en nuestra interpretación, y que ese algo *no* está directamente relacionado con los estados mentales del sujeto en cuestión. Ésta me parece una dificultad importante para la teoría de la interpretación radical.

No obstante, dejando esto a un lado, estoy de acuerdo con Davidson en que si queremos entender a los demás, no podemos renunciar a

interpretar partes cruciales de la conducta de algunos individuos. Coincido también en que esta tarea interpretativa ha de realizarse de un modo holista, no sólo con respecto al conjunto doxástico y la semántica del individuo, sino dentro de una comunidad de hablantes. Es obvio también, que no hay ninguna razón por la que podamos o debamos asignar a la ligera incoherencias a otros, ya que resultaría fatal para la intención comunicativa suponer que los demás son *arracionales*. Hemos de suponer que los individuos rechazarían una incoherencia si la percibiesen de modo claro, lo cual ciertamente no impide que puedan tener fallos de racionalidad, y sus estados contengan incoherencias internas que les sean desconocidas.

La otra dificultad en el experimento mental de Davidson que quisiera apuntar, es que parece que el propio experimento contiene una contradicción formal que problematiza todo de antemano: por un lado, supone que hemos de interpretar consistentemente; por otro lado, supone que los sujetos no son infalibles en la adquisición de creencias. Si fueran infalibles, parece que sus creencias serían verdaderas, adecuadas y consistentes, y no habría excesivos problemas interpretativos. Pero dado que suponemos que son falibles y, por tanto, pueden tener errores sensoriales e inferenciales, esto nos coloca ya en una situación de previsibles inconsistencias. No obstante, estas inconsistencias han de ser *detectables* en la conducta, o no tendrán ninguna importancia para nuestra tarea interpretativa. El problema es que, de que sean en principio detectables, no se sigue que sean todas ellas detectadas. Esta anomalía, en principio inocua —pues concedemos de antemano que la interpretación

es un proceso cuyas atribuciones son falibles y corregibles—, puede ser quizá la causante de todas nuestras dificultades.

Llamemos *inconsistencia trivial* a aquella que le pasa desapercibida al propio sujeto y se produce por su falibilidad sensorial e inferencial. La inconsistencia no-trivial sería aquella de la que, de algún modo, es consciente el sujeto, como es por hipótesis el autoengaño. Tal vez las inconsistencias triviales que *pueden* pasarle desapercibidas a quien realiza la interpretación porque no tiene modo de preverlas (aunque pueda llegar a detectarlas) debido a que son precisamente ilógicas o fallos dentro de la propia razón, sean las que nos empujen posteriormente a atribuir inconsistencias no-triviales. Es decir, no tenemos modo de saber si hemos detectado todas las inconsistencias triviales que son detectables, y por ello no tenemos modo de distinguir entre inconsistencias no-triviales y triviales. Pero dado que cuando interpretamos a veces no podemos parar el proceso interpretativo, y las inconsistencias triviales no detectadas introducen elementos podridos en la red de creencias, posteriormente pueden dar lugar a atribuciones de autoengaño ilegítimas.

Ya sabíamos que nuestras atribuciones son falibles, lo cual también afecta a las atribuciones de autoengaño. Este argumento anterior no demuestra —ni pretende hacerlo—, que las atribuciones de autoengaño sean erróneas. Lo que sí podría ayudar a aclarar es, quizá, por qué razón *podrían estar equivocadas* todas ellas: lo que nos forzaría a la atribución de creencias contradictorias sería un error previo en detección de inconsistencias triviales.

Tenemos dos pistas para suponer que en estos casos el error puede ser *siempre* del intérprete. La primera nos la daba el hecho de que *mientras* la razón por la que el sujeto intentaría autoinducirse una creencia es que desearía evitar el dolor que produce lo que infiere de modo *consistente*, la razón por la que le adscribimos autoengaño nosotros es para evitar una *inconsistencia mayor*. El sujeto nunca se autoengaña para evitar una mayor inconsistencia. La segunda nos dice que esa mayor inconsistencia que pretende evitar el intérprete quizá sólo se deba a que no hemos detectado aún una inconsistencia trivial previa.

Éstas son las dos principales razones por las que considero que ningún experimento mental puede demostrar la existencia de casos que puedan ser interpretados como casos genuinos de autoengaño, al margen de que los haya o no. En este sentido, las atribuciones de autoengaño podrían reducirse a significar esto: "no me es posible interpretar la conducta de este sujeto; algo extraño debe estar pasando". La falta de datos o el error por parte del intérprete son siempre una mejor explicación.

VIII. CONCLUSIONES

Cuando comenzamos este trabajo, explicar cómo podría producirse el autoengaño parecía una empresa formidable. Aunque estamos totalmente familiarizados con el término, nada más intentar hincar el diente al concepto, aparecían las primeras dificultades. Al observarlo bajo el modelo interpersonal o de engaño a otros surgían diversos problemas, entre los cuales, las paradojas estática y dinámica eran los más espinosos. Según la primera, cuando se produce un engaño del tipo que sea, el engañador sabe o cree que *p*, e induce al engañado a creer que *no-p*. En algún momento, ambos tienen creencias cuyos contenidos son contradictorios, pero esto es problemático cuando se trata de uno y el mismo sujeto. Según la segunda, para lograr el engaño, el engañador ha de diseñar una estrategia más o menos compleja de la que el engañado no puede tener noticia. El autoengaño reflexivo o dirigido a uno mismo parecía involucrar que un individuo creyese *p* y *no-p* a la vez y, además, que de algún modo fuese capaz de esconderse su propio proyecto de engaño, siendo conocedor de los detalles como verdugo y desconocedor como víctima.

A partir de la segunda mitad del siglo pasado, este asunto suscitó un amplísimo debate que nos ha proporcionado un corpus bibliográfico inabarcable. La puesta en cuestión de la autoridad de primera persona y del método de la introspección como fuente de evidencia, tanto por parte del psicoanálisis como del conductismo, reavivaron la disputa acerca de cómo un sujeto puede engañarse a sí mismo. El inconsciente freudiano, por ejemplo, parecía posibilitar que un sujeto tuviese cogniciones incoherentes o inconsistentes y realizase maniobras defensivas que le fuesen desconocidas. Hemos visto tanto estas ideas, como la crítica sartriana a su posibilidad conceptual en *El Ser y la Nada* (traducido al inglés en 1956), y cómo parece que fueron estas disputas las que encendieron definitivamente el debate, que a partir de los años 60 se ha convertido en interminable.

Aunque hay consenso con respecto a la imposibilidad de creer algo por las buenas, se ha tratado de argumentar cómo podríamos inducirnos creencias de modo indirecto. Hay tres posturas principales: intencionalistas, no-intencionalistas, y escépticos o eliminativistas. Dejando a un lado estos últimos, la disputa se centra en si debemos aceptar que este proceso indirecto ha de llevarse a cabo con intención por parte del sujeto o no. Los primeros defienden que un engaño sin intención no puede ser nunca un engaño; creen que este punto no debe estar en cuestión, y que el debate debería consistir en tratar de alcanzar la respuesta más satisfactoria posible a las paradojas que suscita la intención de engaño dirigida a uno mismo y el estado irracional que provoca. Los segundos creen que bajo este modelo las paradojas resultan irresolubles, y renuncian

a una explicación de este tipo. No buscan una inferencia a la mejor explicación, sino una argumentación indirecta que evite las paradojas.

He intentado señalar por qué ninguna de estas posturas parece satisfactoria. Las intencionalistas ciertamente captan mejor el carácter semántico del término, pero resultan más controvertidas tanto empírica como psicológicamente. Sin embargo, la razón por la que no parecen aceptables no es ésta. El verdadero motivo es que la mayoría acude a nociones *conceptualmente problemáticas* como las intenciones inconscientes, los subsistemas o la exposición selectiva a la información y (des)focalización voluntaria de la atención. Como ya vimos con detenimiento, varios problemas aquejan a cada una de estas alternativas: en primer lugar, las *intenciones inconscientes* de las que hablan Audi, Pears o Bermúdez, resultan demasiado exóticas, ya que una intención nos parece algo que uno ha de reconocer de modo no-inferencial. No obstante, si a uno no le persuade este argumento, aun concediendo que existiesen intenciones de esta naturaleza, resulta difícil explicar cómo tienen algún efecto sobre la conducta consciente. Como decía Sartre, al negar la unidad consciente de lo psíquico, "nos vemos obligados a sobrentender por doquiera una unidad mágica que vincula los fenómenos a distancia y por encima de los obstáculos" [Sartre (1943), p. 103]. Y una dificultad final aparentemente insalvable: toda explicación que recurre a lo inconsciente parece incurrir en la *falacia homuncular,* al reproducir dentro de la consciencia una consciencia que distinga previamente entre lo que ha de ser reprimido y lo que no.

De hecho, casi todas las explicaciones intencionales terminan ahogando o mitigando la intencionalidad al acudir a subsistemas y supuestos de no-transparencia de la consciencia. Estas formas de explicación acaban por hacer de la intención algo desconocido para el sujeto, con lo cual resultan oscuras y no escapan tampoco de la falacia homuncular.

No ofrecen mejor panorama las teorías intencionales que hablan de exposición selectiva a la información por medio de una focalización voluntaria en pensamientos y evidencia agradable o evitación voluntaria de pensamientos y evidencia dolorosa. El principal escollo es que, aunque hay sin duda procesos *involuntarios* por debajo del nivel de la consciencia (por ejemplo, en el proceso de percepción), parece que una vez que sabemos que algo es relevante, no podemos evitarlo. Como se vio obligado a reconocer Festinger, su teoría de la disonancia cognitiva no resolvía un problema que no había previsto: "Una vez que se ha dicho a la persona que la información que existe no apoya su decisión, se ha introducido ya evidencia adicional: en cierto sentido es imposible evitarla, pues sabe ya que existe". [Festinger (1964), p. 82] El problema es que tratar de evitar pensar algo, tiene como consecuencia que no podamos quitárnoslo de la cabeza: es como tratar de crear oscuridad por medio de la luz. Y si para evitar este resultado se acude a *intenciones no conscientes* de focalización y selección de la evidencia, caemos en los mismos problemas que las explicaciones previas.

Por tanto, todas estas teorías acaban por ocultar a la consciencia la intención que, a la vez que nos ofrecía la tensión requerida en el

autoengaño, generaba las paradojas indeseadas. Davidson ha indicado que esta tendencia se ha debido, en parte, a que todas ellas tratan de exponer las *condiciones de éxito* del autoengaño, disolviendo con ello la irracionalidad y la tensión que se le supone al proceso. Su principal interés ha sido, por el contrario, el de subrayar la irracionalidad.

Como veíamos, su enfoque asume la tensión y la irracionalidad que *prima facie* parecen connaturales al autoengaño. Para ello, exige que el sujeto *tenga creencias contradictorias*, que una *cause y sustente* a la otra (sin ser, evidentemente, razón para ella), que el sujeto crea algo en contra de *su* evidencia (debilidad de la justificación) y, a la vez, niega que algo de esto haya de ser inaccesible a la consciencia. El individuo no puede, no obstante, observar el todo sin apreciar la incoherencia y así rechazar alguna de las creencias, por lo que Davidson supone (por razones conceptuales, no empíricas) que ha de haber algún límite que separe las creencias en dos territorios distintos.

Traté de mostrar por qué esta explicación no me parece tampoco satisfactoria. Como hemos señalado, Hume, Alston o Williams han defendido persuasivamente que la naturaleza de la creencia impide que ésta se forme en contra de la evidencia. Davidson podría reformular su teoría de modo que el sujeto olvidase cierta evidencia o no la tomase en cuenta; seguramente, se negaría a conceder que este proceso supusiese un proyecto no intencional, pues entonces estaríamos hablando de ceguera intelectual por las pasiones, el temor o la angustia. Pero si acude a un

olvido intencional, no encontraríamos mucha diferencia con la explicación de Pears o Bach de las que, como él mismo reconoce, es gran deudor.

Davidson sostiene que las teorías de Pears y Bach son muy similares a la suya, aunque su interés sea distinto. Ellos no logran dar cuenta de lo que nos parece más interesante del fenómeno: el estado de irracionalidad. Al intentar resolver las paradojas, disuelven el problema. Pero la teoría de Davidson hace difícil de comprender psicológicamente cómo puede producirse el fenómeno. ¿Por qué? Porque se resiste a admitir que el proceso sea inconsciente, ya que esto diluiría la inconsistencia, y con ello la irracionalidad. Pero desde un enfoque que apele a la consciencia, la explicación del proceso parece inalcanzable.

El problema, pues, para las teorías intencionalistas, es que no pueden dar cuenta a la vez del estado y el proceso. Al dar cuenta conceptualmente del estado (Davidson), se ha de renunciar a comprender el proceso. Al explicar conceptualmente el proceso (Bach, Pears, Audi, Bermúdez, Rorty, etc.), se disuelve el estado de irracionalidad.

También he tratado de mostrar por qué las teorías no-intencionalistas no suponen ni siquiera una alternativa explicativa. Entre estas teorías, algunas no son más que "teorías intencionalistas disfradazas", pues introducen nociones explicativas —como la de propósito (*purpose*) o "conducta deshonesta para con uno mismo"— que resultan oscuras, si no ininteligibles, a menos que se comprenden bajo el modelo intencional. Obviamente, heredarían todos los problemas de las anteriores.

En otros casos, cuando argumentan que el autoengaño tiene un propósito, no se refieren a que tengan un objetivo o fin, sino a que involucra procesos "subintencionales" que consisten en mecanismos que cumplen una función (la reducción de angustia, por ejemplo) pero no son intencionales.

Este tipo de explicaciones, al igual que las que argumentan que el individuo tiene *motivos* que le llevan a sesgar involuntariamente la evidencia, disuelven el problema de irracionalidad al hacer del autoengaño un fenómeno muy cercano a la ceguera intelectual o al mero error. Por supuesto, las emociones, angustias o temores tienen relevancia en nuestro modo de ver el mundo, en nuestra captación de la evidencia y en la formación de nuestras creencias. Pero de aquí no se sigue que nos engañemos a nosotros mismos siempre que estemos bajo la influencia de nuestras pasiones. El elemento *intencional* parece indispensable.

Mele es, quizá, el campeón de la postura no intencional. Consciente de que considerar que el sesgo de evidencia es algo no intencional disuelve el problema del autoengaño, Mele sostiene que aún es posible mantener el carácter intencional de la adulteración de la evidencia sin estar obligado a aceptar que el sujeto se engañe intencionalmente. Su estrategia es mostrar que hacer A intencionalmente sabiendo que B se sigue de A, no implica hacer B intencionalmente. Sin embargo, aunque esto es cierto para algunos contextos, esto no es posible en el caso del autoengaño, ya que por la naturaleza de las creencias, no es posible sesgar la evidencia voluntaria y conscientemente, y abrazar la creencia que apoya esa

evidencia adulterada. Las creencias apuntan a la verdad, y sesgar intencionalmente la evidencia conduce a la consciencia de que la evidencia no es garante de la verdad.

Así pues, ni intencionalistas ni no-intencionalistas ofrecen una propuesta conceptualmente satisfactoria. Tampoco la psicología, que ha pretendido aislar empíricamente el fenómeno, ha ofrecido resultados interesantes. Quizá el mayor problema de estos intentos es que no consiguen encontrar un método para descubrir creencias que resulte superior al testimonio del propio individuo. Asumiendo que éste puede ser insincero, algunos psicólogos sociales como Gur y Sackeim han propuesto distintos experimentos en los que creen encontrar sujetos que mantienen creencias contradictorias porque les resulta dolorosa la autoconfrontación. El problema es que consideran como índice de creencias las respuestas galvánicas fuertes ante determinados estímulos, concretamente, ante la propia voz. Algunos sujetos negaban reconocer su voz, pero tenían fuertes respuestas galvánicas. La dificultad estriba en que, aun concediendo que fueran sinceros en sus declaraciones, no está claro que las respuestas galvánicas sean un índice fiable de creencias; quizá hay un reconocimiento de algunas cosas por debajo del nivel de la consciencia, lo cual impediría afirmar que el sujeto cree cosas contradictorias. Por otro lado, algunos experimentos han demostrado que los individuos tienen también fuertes respuestas galvánicas ante voces de conocidos (y no sólo para la propia). En resumen, todos estos experimentos dependen crucialmente de criterios que están más allá de lo que dictan las condiciones experimentales: captar si un sujeto tiene una creencia o no, no

tiene que ver con el descubrimiento de marcas empíricas localizables y cuantificables.

Robert Trivers ha argumentado que el autoengaño es un producto evolutivo resultante de la necesidad de mejorar nuestras capacidades para el engaño, debido a que también aumentan las capacidades para su detección. En principio, no hay nada incoherente en este tipo de explicaciones evolutivas. Uno puede señalar que la evolución no es apoyo de nada, en tanto que ofrece apoyo a todo cuanto existe por el mero hecho de *ser*. De cualquier forma, parece inobjetable que el autoengaño nos permitiría, de ser posible, ser más eficaces en la tarea de engañar a otro; pero Trivers no demuestra que sea posible. Simplemente alude a los experimentos de Gur y Sackeim como una demostración del fenómeno, pero ya hemos visto que sus resultados son más que cuestionables.

Por todas estas razones, con respecto a las teorías presentadas en este trabajo, podría suscribir unas palabras que ya citamos de Stanley Paluch:

> Es cualquier cosa menos obvio que pudiera decirse de cualquiera de ellas que refleje casos reales de autoengaño. De hecho, no es ni mucho menos claro que haya casos estrictos de autoengaño que reflejar. No hay duda acerca del hecho de que hablamos de gente que se engaña a sí misma, pero lo que no está claro es que haya instancia alguna en la que este modo de hablar sea el mejor o no pueda ser reemplazado por otras descripciones que no incluyan la noción de engaño en absoluto.[122]

[122] «It is anything but obvious that any of them could be said to mirror actual cases of self-deception. Indeed, it is anything but obvious that there are strict cases of self-deception to be mirrored. There is no doubt about the fact that we talk about people deceiving themselves, but what is not clear is that there are any instances where this way of speaking is best or could not be replaced by other

Frederick Siegler también nos ponía en alerta respecto a la posibilidad de confundir el hecho de que un sujeto se comporte de modo aparentemente contradictorio con el hecho de que posea creencias contradictorias.

> […] en el mejor de los casos, encontramos evidencia para "Jones creía p y ahora cree no p" donde no *nos* parece que haya nada que justifique el cambio de creencia. No encontramos evidencia para "Jones cree p y no p", ni para "Jones cree no p como resultado de su propio intento procedimiento para inducirse una creencia que creía falsa" [en cursivas en el original][123]

En este sentido, nos parecen más prudentes las estrategias que, como la de Jon Elster, consideran que quizá sea más adecuado reducir progresivamente el número de casos problemáticos ofreciendo explicaciones alternativas distintas, con la esperanza de que finalmente el número de casos difíciles tienda a cero.

¿Por qué, sin embargo, estamos tan familiarizados con el término y se hacen tantas atribuciones de autoengaño? Es evidente que algo tiene que estar sucediendo, que nos parece irracional e incomprensible, para que acudamos al autoengaño como explicación. He argumentado que muy

descriptions which do not involve the notion of deception at all.» [Paluch (1967), p. 277]

[123] «…at best, we find evidence for "Jones believed p and now he believes not p" where there seems to *us* to be nothing to justify the change in belief. We could not find evidence for "Jones believes p and not p", or for "Jones believes not p as a result of his own procedure intended to induce a belief which he believed to be false.» [Siegler (1963), p. 42]

probablemente el autoengaño tiene más que ver con una atribución descaminada de estados, que con estados mismos.

Se argumenta que es concebible que el sujeto que trata de comprender a los demás y no puede renunciar a entender ciertas conductas sin que todo se vuelva ininteligible, no tenga más remedio que suponer que un determinado individuo se autoengaña, por resultar menos costoso en términos de atribución de inconsistencias al conjunto de individuos. Sin embargo, la atribución de autoengaño supone una atribución de inconsistencias no-triviales (es decir, en cierto modo directas para el sujeto) que tengo la impresión de que no suele tener un fundamento sólido; aun más, en mi opinión, estas atribuciones son muy posiblemente la consecuencia de no haber tomado previamente en consideración determinadas inconsistencias triviales (esto es, generadas por la limitación en las capacidades del sujeto, tanto sensoriales, como cognitivas e inferenciales) que todo sujeto tiene. Está claro que cualquier tipo de inconsistencia relevante ha de ser detecta*ble* en el proceso de interpretación si tiene influencia en la conducta, pero por la propia naturaleza procesual y corregible de la tarea de interpretación, es concebible que no hayan sido *aún* detectadas. Sabemos que todas nuestras atribuciones son falibles, lo cual también afecta a las atribuciones de autoengaño y, aunque esto no demuestra que las atribuciones de autoengaño sean siempre erróneas, sí podría ayudar a aclarar, quizá, por qué razón *podrían estar equivocadas*: lo que nos forzaría a la atribución de creencias contradictorias sería un error previo en detección de inconsistencias triviales. Tenemos dos pistas para suponer que el error es

siempre del intérprete. La primera nos la daba el hecho de que mientras la razón por la que el sujeto intentaría autoinducirse una creencia es que desearía evitar el dolor que produce lo que infiere de modo *consistente*, la razón por la que le adscribimos autoengaño nosotros es para evitar una *inconsistencia mayor*. El sujeto nunca se autoengaña para evitar una mayor inconsistencia. La segunda nos dice que quizá esa mayor inconsistencia sólo se debe a que no hemos detectado una inconsistencia trivial previa. En cualquier caso, que la necesidad de continuar interpretando nos obligue a suponer una inconsistencia e irracionalidad como la que nos presenta el autoengaño, quizá nos esté indicando que el proyecto de interpretación radical no es del todo adecuado y, debería abandonar la pretensión de ser global y aceptar que, posiblemente, debe conformarse con ser local y más modesto.

No obstante, quienes aún crean que el autoengaño es, después de todo, un fenómeno real y cotidiano, quizá sientan que mi postura supone algún paso importante en falso. Quizá sean discutibles mi concepción de lo que es una creencia, mi intuición acerca de la autoridad de primera persona o mi postura con respecto al significado o al proyecto de interpretación radical. En cualquier caso, he querido subrayar que no se trata de una cuestión meramente verbal, y que los problemas que impiden una sencilla caracterización del fenómeno tampoco son de tipo empírico, sino conceptual.

BIBLIOGRAFÍA

ALSTON, WILLIAM P. (1989) *Epistemic Justification*, Ithaca, Cornell University Press.

— (1996), *A Realist Conception of Truth*, Ithaca, Cornell University Press.

AMES, ROGER T. Y DISSANAYAKE, WIMAL (eds.) (1996), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press.

AMES, ROGER T. (1996), 'The Classical Chinese Self and Hypocrisy', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 219-240.

ANSCOMBE, GERTRUDE ELIZABETH MARGARET (1957), *Intention*, Oxford, Basil Blackwell. [Edición en castellano: *Intención*, trad. de Ana Isabel Stellino, Barcelona, Paidós/I.C.E.-U.A.B., 1991].

ANSPACH, MARK R. (1998), 'Madness and the Divided Self: Esquirol, Sartre, Bateson', en *Self-Deception and Paradoxes of Rationality*, Jean-Pierre Dupuy (ed.), Stanford (CA), CSLI Publications, 1998, pp. 59-85.

AQUINO, TOMAS DE (1265-1273), *Summa de Teología*, Madrid, Biblioteca de autores cristianos, 2001.

ARISTÓTELES, *Poética*, ed. Trilingüe, trad. Valentín García Yebra, 1992.

— *Retórica*, trad. de Quintín Racionero Carmona, Madrid, Gredos (1990).

— *Ética a Nicómaco,* trad. de Julio Pallí Bonet, Madrid, Gredos (1985).

ASCH, SOLOMON E. (1951), 'Effects of Group Pressure Upon the Modification and Distortion of Judgements', en Harold Guetzkow (ed.), *Groups, Leadership, and Men*, pp. 177-190.

AUDI, ROBERT (1985), 'Self-Deception and Rationality', en Mike Martin (ed.) *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas, pp. 169-194.

— (1988), 'Self Deception, Rationalization, and Reasons for Acting', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 92-120.

— (1997a), 'Self Deception, Rationalization, and the Ethics of Belief An Essay in Moral Psychology', en R. Audi, *Moral Knowledge and Ethical Character*, New York, Oxford University Press, pp. 131-156.

— (1997b), 'Self-Deception vs. Self-Caused Deception: A Comment On Professor Mele', *Behavioral and Brain Sciences*, 20, p. 104.

BACH, KENT (1981), 'An Analysis of Self-Deception', *Philosophy and Phenomenological Research*, vol. 41 (3), pp. 351-370.

— (1985), 'More on Self-Deception: Reply to Hellman', *Philosophy and Phenomenological Research*, vol. 45 (4), pp. 611-614.

— (1997), 'Thinking And Believing In Self-Deception', *Behavioral and Brain Sciences*, vol. 20, p. 105.

— (1998), 'Apparent Paradoxes of Self-Deception and Decision', en *Self-Deception and Paradoxes of Rationality*, Jean-Pierre Dupuy (ed.), Stanford (CA), CSLI Publications, 1998, pp. 163-189.

BAIER, ANNETE C. (1996), 'The Vital But Dangerous Art of Ignoring: Selective Attention and Self Deception', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 53-72.

BARNES, ANNETTE (1997), *Seeing through self-deception*, Cambridge, Cambridge University Press.

BARON, MARCIA (1988), 'What Is Wrong with Self-Deception', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 431-449.

BEDFORD NAYLOR, MARGERY (1985), 'Voluntary Belief', *Philosophy and Phenomenological Research*, vol. 45 (3), pp. 427-436.

BEM, DARYL J. (1967), 'Self-perception: an alternative interpretation of cognitive dissonance phenomena', *Psychological Review*, vol. 74, pp. 183-200.

BERKICH, DON (2007), 'A Puzzle About Akrasia', *Teorema*, vol. XXVI (3), pp. 59-71.

BERMÚDEZ, JOSÉ L. (1997), 'Defending intentionalist Accounts of Self-Deception', *Behavioral and Brain Sciences*, vol. 20, pp. 107-108.

— (2000), 'Self-deception, intentions and contradictory beliefs', *Analysis*, vol. 60 (4), pp. 309-319.

BEYER, LAWRENCE (1998), 'Keeping Self Deception in Perspective', en *Self-Deception and Paradoxes of Rationality*, Jean-Pierre Dupuy (ed.), Stanford (CA), CSLI Publications, 1998, pp. 87-111.

BITTNER, RÜDIGER (1988), 'Understanding a Self-Deceiver', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 535-551.

BOND, CHARLES F. JR. Y DEPAULO, BELLA M. (2006), 'Accuracy of Deception Judgments', *Personality and Social Psychology Review*, vol. 10 (3), pp. 214-234.

BORDES SOLANA, MONTSERRAT (2001), 'Motivated Irrationality: the Case of Self-Deception', *Crítica: revista hispanoamericana de filosofía*, vol. 33 (97), pp. 3-32.

BRATMAN, MICHAEL (1987), *Intention, Plans, and Practical Reason*, Cambridge, Harvard University Press.

BROWN, JONATHON D. Y DUTTON, KEITH A. (1995), 'Truth and Consequences: The Costs and Benefits of Accurate Self-Knowledge', *Personality and Social Psychology Bulletin*, vol. 21, pp. 1288-1296.

BUTLER, JOSEPH (1726), 'Upon Self-Deceit', en *Human Nature and Other Sermons*, London, Cassell & Company, 1887.

CANFIELD, JOHN V. Y GUSTAVSON, DON F. (1962), 'Self-Deception', *Analysis*, vol. 23 (2), pp. 32-36.

CARNAP, RUDOLF (1950), *Logical Foundations of Probability,* Chicago, University of Chicago Press.

CARROLL, LEWIS (1895), 'What the Tortoise Said To Achilles', *Mind*, New Series, vol. IV (14), pp 278-280.

CHAMPLIN, THOMAS STEPHEN (1976), 'Double Deception', *Mind*, New Series, vol. 85 (337), pp. 100-102.

— (1988), *Reflexive Paradoxes*, London, Routledge.

— (1994), 'Deceit, Deception and the Self Deceiver', *Philosophical Investigations*, vol. 17 (1), pp 53-58.

CHANOWITZ, BENZION Y LANGER, ELLEN J. (1985), 'Self-Protection and Self-Deception', en Mike Martin (ed.), *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas, pp. 117-135.

CHURCHLAND, PAUL M. (1984), *Matter and Consciousness*, Cambridge (MA), MIT Press. [Edición en castellano: *Materia y conciencia*, trad. de Margarita N. Mizraji, Barcelona, Gedisa, 1992].

CLIFFORD, WILLIAM KINGDOM (1876), 'The Ethics of Belief', *Contemporary Review*, vol. 29, pp. 289-308. Reeditado en *Lectures and Essays,* London, Macmillan, 1879, pp. 177-211. [Edición en castellano: 'La ética de la creencia', en *La Voluntad de Creer. Un debate sobre la ética de la creencia,* trad. de Lorena Villamil, Madrid, Tecnos, 2003, pp. 91-134].

COLL MÁRMOL, JESÚS A. (2007), 'Autoengaño y responsabilidad', *Teorema*, vol. XXVI (3), pp. 145-159.

COOK, J. THOMAS (1987), 'Deciding to Believe Without Self-Deception' *The Journal of Philosophy*, vol. 84 (8), pp. 441-446.

CORREIA, VASCO (2007), 'Une conception émotionnaliste de la self-deception', *Teorema*, vol. XXVI (3), pp. 31-44.

CUA, ANTONIO S. (1996), 'A Confucian Perspective on Self-Deception', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 177-199.

DARWALL, STEPHEN L. (1988), 'Self-Deception, Autonomy, and Moral Constitution', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 407-430.

DAVIDSON, DONALD (1970), 'How Is Weakness of Will Possible?', en *Moral Concepts*, Joel Feinberg (ed.), Oxford, Oxford University Press. Reeditado en *Essays on Actions and Events*, Oxford, Clarendon Press, 1980, pp. 21–42. [Edición en castellano: '¿Cómo es posible la debilidad de la voluntad?', en *Ensayos sobre acciones y sucesos*, trad. coordinada por Olbeth Hansberg, Barcelona, UNAM/Crítica, 1995, pp. 37-62].

— (1974), ' On the Very Idea of a Conceptual Scheme'*, Proceedings and Addresses of the American Philosophical Association*, 47, pp. 5-20. Reeditado en *Inquiries into Truth and Interpretation, Oxford*, Clarendon Press, 1984, pp. 183-198. [Edición en castellano: 'De la idea misma de un esquema conceptual', en *De la verdad y de la interpretación*, Barcelona, Gedisa, 1990, pp. 189-203].

— (1982), 'Paradoxes of Irrationality', en Wollheim, R. y Hopkins, J. (Eds.), *Philosophical Essays on Freud,* Cambridge, Cambridge University Press*,* (1988), pp. 289-305. Reeditado en *Problems of Rationality*, Oxford, Clarendon Press, 2004, pp. 169-187.

— (1984a), 'First Person Authority', *Dialectica*, vol. 38, pp. 101-111. Reeditado en *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, 2001, 3-14. [Edición en castellano: *Subjetivo, Intersubjetivo, Objetivo*, trad. de Olga Fernández Prat, Madrid, Cátedra, 2003, pp. 25-40].

— (1984b), *Inquiries into Truth and Interpretation,* Oxford, Oxford University Press. [Edición en castellano: *De la verdad y de la interpretación*, trad. de Guido Filippi, Barcelona, Gedisa, 1990].

— (1985), 'Deception and Division', en Jon Elster (ed.), *The Multiple Self*, Cambridge University Press, 1986, pp. 79-92. Reeditado en *Problems of Rationality*, Oxford, Clarendon Press, 2004, pp. 199-212. [Edición en castellano: 'Engaño y División', en *Mente, Mundo y Acción*, edición de Carlos Moya Espí, Barcelona, Paidós, 1992, pp. 99-117]

— (1987), 'Knowing One's Own Mind', *Proceedings and Addresses of the American Philosophical Association*, pp. 441-458. Reeditado en *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, 2001, 15-38. [Edición en castellano: *Subjetivo, Intersubjetivo, Objetivo*, trad. de Olga Fernández Prat, Madrid, Cátedra, 2003, pp. 41-71].

— (1993), 'Who is Fooled?', en *Self-Deception and Paradoxes of Rationality*, Jean-Pierre Dupuy (ed.), Stanford (CA), CSLI Publications, 1998, pp. 1-18. Reeditado en *Problems of Rationality*, Oxford, Clarendon Press, 2004, pp. 213-230.

DE SOUSA, RONALD B. (1978), 'Self-Deceptive Emotions', *Journal of Philosophy*, vol. 75, pp. 684-697.

— (1988), 'Emotion and Self-Deception', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 324-341.

DELEUZE, GILLES (1968), *Différence et répétition*, Paris, PUF. [Edición en castellano: *Diferencia y Repetición*, trad. de María Silvia Delpy y Hugo Beccacece, Buenos Aires, Amorrortu, 2002].

DEMOS, RAPHAEL (1960), 'Lying to oneself', *Journal of Philosophy*, vol. 57, pp. 588-595.

DEUTSCH, ELIOT (1996), 'A Comparative Study', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 315-326.

DEWEESE-BOYD, IAN (2007), 'Taking Care: Self-Deception, Culpability and Control', *Teorema*, vol. XXVI (3), pp. 161-176.

DEWEY, JOHN (1929), *The Quest for Certainty: A Study of the Relation of Knowledge and Action,* New York, Minton, Balch and Co. [Edición en castellano: *La búsqueda de la certeza: Un estudio de la relación entre el conocimiento y la acción*, trad. de Eugenio Imaz, México, Fondo de cultura económica, 1952].

DISSANAYAKE, WIMAL (1996), 'Self-Deception and Cultural Contextualization: Reflections on Two Indian Novels', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 327-347.

DOUGLAS, WILLIAM Y GIBBINS, KEITH (1983), 'Inadequacy of voice recognition as a demonstration of self-deception', Journal of Personality and Social Psychology, vol. 44 (3), pp. 589-592.

DUPUY, JEAN-PIERRE (1998b), 'Rationality and Self-Deception', en *Self-Deception and Paradoxes of Rationality*, Jean-Pierre Dupuy (ed.), Stanford (CA), CSLI Publications, 1998, pp. 113-150.

— (ed.) (1998a), *Self-Deception and Paradoxes of Rationality*, Stanford (CA), CSLI Publications.

ELSTER, JON (1979), *Ulysses and the Sirens: Studies in Rationality and Irrationality*, New York/Cambridge: Cambridge University Press. [Edición en castellano: *Ulises y las sirenas: estudios sobre racionalidad e irracionalidad*, trad. de Juan José Utrilla, México, Fondo de Cultura Económica, 1989/1997].

— (1983), *Sour Grapes: Studies in the Subversion of Rationality,* Cambridge, Cambridge University Press. [Edición en castellano: *Uvas amargas. Sobre la subversión de la racionalidad,* trad. de Enrique Lynch, Barcelona, Península/Ideas, 1988].

— (Ed.) (1986), *The Multiple Self*, Cambridge, Cambridge University Press.

— (1989), *Solomonic Judgments: Studies in the Limitations of Rationality,* Cambridge, Cambridge University Press. [Edición en castellano: *Juicios salomónicos: Las limitaciones de la racionalidad como principio de decisión*, trad. de Carlos Gardini, Barcelona, Gedisa, 1991]

ERWIN, EDWARD (1988), 'Psychoanalysis and Self-Deception', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 228-245.

FEREJOHN, JOHN (1998), 'Cooperation and Time', en *Self-Deception and Paradoxes of Rationality*, Jean-Pierre Dupuy (ed.), Stanford (CA), CSLI Publications, 1998, pp. 151-61.

FESTINGER, LEON (1957), *A Theory of Cognitive Dissonance*, Stanford (CA), Stanford University Press. [Edición en castellano: *Teoría de la disonancia cognoscitiva*, trad. de José Enrique Martín Daza, Madrid, Instituto de Estudios Políticos, 1975]

— (1964), *Conflict, decision, and dissonance*. Stanford (CA), Stanford University Press.

FESTINGER, LEON Y CARLSMITH, JAMES M. (1959), 'Cognitive consequences of forced compliance', *Journal of Abnormal and Social Psychology,* vol. 58, pp. 203-210.

FINGARETTE, HERBERT (1950), '«Unconscious Behavior» and Allied Concepts. A New Approach to their Empirical Interpretation', *The Journal of Philosophy*, vol. 47 (18), pp. 509-520.

— (1969), *Self-Deception*, Londres, Routledge and Kegan Paul. Reeditado en University of California Press, 2000 [Cito por la edición del 2000].

— (1985), 'Alcoholism and Self-Deception', en Mike Martin (ed.), *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas, pp. 52-67.

— (1998), 'Self-Deception Needs No Explaining', *The Philosophical Quarterly*, vol. 48 (192), pp. 289-301.

FINKELSTEIN, DAVID H. (2003), *Expression and the Inner*, Cambridge (MA), Harvard University Press.

FLEMING, PATRICK (2005), 'Hume on Weakness of Will', en *British Journal for the History of Philosophy* (en prensa). También en: <http://www.geocities.com/flemingphilosophy/shorthume.doc>

FODOR, JERRY A. (1987), *Psychosemantics: The Problem of Meaning in the Philosophy of Mind,* Cambridge (MA), The MIT Press. [Edición en castellano: *Psicosemántica. El Problema del Significado en la Filosofía de la Mente*, trad. de Óscar Luis González Castán, Madrid, Tecnos, 1994]

FOSS, JEFFREY (1980), 'Rethinking Self-deception', *American Philosophical Quarterly*, vol. 17, pp. 237-243.

— (1997), 'How Many Beliefs Can Dance on the Head of the Self-Deceived?', *Behavioral and Brain Sciences*, 2 vol. 0, pp. 111-112.

FREUD, ANNA (1936), *Das Ich und die Abwehrmechanismen,* Viena/Leipzig, Internationaler Psychoanalytischer Verlag. [Edición en castellano: *El yo y los mecanismos de defensa,* trad. de Yvonne P. de Cárcamo y Celes E. Cárcamo, en *Obras escogidas*, Barcelona, RBA, 2006, pp. 23-142].

FREUD, SIGMUND (1886/1939), *Gesammelte Werke*, Hamburg, S. Fischer Verlag, 18 vols., 1972. [Edición en castellano: *Obras completas,* Ordenamiento, comentarios y notas de James Strachey, con la colaboración de Anna Freud, asistidos por Alix Stratchey y Alan Tyson (*Standard Edition*). Traducción directa del alemán por José Luis Etcheverry, Buenos Aires, Amorrortu editores, 24 vols., 1974-1985].

— (1894), 'Die Abwehr-Neuropsychosen (Versuch einer psychologischen Theorie der aquirierten Hysterie, vieler Phobien und Zwangsvorstellungen und gewisser halluzinatorischer Psychosen)',

Neurologisches Zentralblatt, vol. 13 (10), pp. 362-364, y (11), pp. 402-409. Reeditado en *Gesammelte Werke*, vol. 1, pp. 59-74 [Edición en castellano: 'Las neuropsicosis de defensa (Ensayo de una teoría psicológica de la histeria adquirida, de muchas fobias y representaciones obsesivas, y de ciertas psicosis alucinatorias)', en *Obras completas*, vol. 3, pp. 47-61].

— (1900), *Die Traumdeutung*, Leipzig/Viena, Franz Deuticke. Reeditado en *Gesammelte Werke*, vols. 2-3. [Edición en castellano: *La interpretación de los sueños,* en *Obras completas*, vols. 4-5, pp. 17-608].

— (1901), 'Zur Psychopathologie des Alltagslebens. Über Vergessen, Versprechen, Vergreifen, Aberglaube und Irrtum', *Monatsschrift für Psychiatrie und Neurologie*, vol. 10 (1), pp. 1-32, y (2), pp. 95-143. Reeditado en forma de libro en 1904, Berlín, Karger. También en *Gesammelte Werke*, vol. 4. [Edición en castellano: *Psicopatología de la vida cotidiana. Sobre el olvido, los deslices en el habla, el trastrocar las cosas confundido, la superstición y el error*, en *Obras completas*, vol. 6].

— (1912), 'A Note on the Unconscious in Psycho-Analysis', *Proceedings of the Society for Psychical Research*, vol. 26, parte 66º, pp. 312-318 (original en inglés). Reeditado en *Gesammelte Werke*, vol. 8, pp. 430-439. [Edición en castellano: 'Nota sobre el concepto de lo inconsciente en psicoanálisis', en *Obras completas*, vol. 12, pp. 271-277].

— (1915a) 'Die Verdrängung', *Internationale Zeitschrift. für ärztliche Psychoanalyse*, vol. 3 (3), pp. 129-138. Reeditado en *Gesammelte Werke*, vol. 10, pp. 248-261. [Edición en castellano: 'La represión', en *Obras completas*, vol. 14, pp. 141-152].

— (1915b) 'Das Unbewusste', *Internationale Zeitschrift. für ärztliche Psychoanalyse*, vol. 3 (4), pp. 189-203, y (5), pp. 257-269. Reeditado en *Gesammelte Werke*, vol. 10, pp. 264-303. [Edición en castellano: 'Lo inconsciente', en *Obras completas*, vol. 14, pp. 161-201].

— (1920), *Jenseits des Lustprinzips*, Leipzig/Viena/Zurich, Internationaler Psychoanalytischer Verlag. Reeditado en *Gesammelte Werke*, vol. 13, pp. 3-69 [Edición en castellano: *Más allá del principio de placer,* en *Obras completas*, vol. 18, pp. 7-62].

— (1923), *Das Ich und das Es*, Leipzig/Viena/Zurich, Internationaler Psychoanalytischer Verlag. Reeditado en *Gesammelte Werke*, vol. 13, pp. 237-289. [Edición en castellano: *El yo y el ello*, en *Obras completas*, vol. 19, pp. 13-59].

— (1926), *Hemmung, Symptom und Angst*, Leipzig/Viena/Zurich, Internationaler Psychoanalytischer Verlag. Reeditado en *Gesammelte Werke*, vol. 18, pp. 113-205. [Edición en castellano: *Inhibición, síntoma y angustia*, en *Obras completas*, vol. 20, pp. 71-161].

— (1932), *Neue Folge der Vorlesungen zur Einführung in die Psychoanalyse*, Viena, Internationaler Psychoanalytischer Verlag (publicado en 1933). Reeditado en *Gesammelte Werke*, vol. 15. [Edición en castellano: *Nuevas conferencias de introducción al psicoanálisis*, en *Obras completas*, vol. 22, pp. 5-168].

— (1938), 'Die Ichspaltung im Abwehrvorgang', *Internationale Zeitschrift für Psychoanalyse - Imago*, vol. 25 (3/4), pp. 241-244 (publicado en 1940).

Reeditado en *Gesammelte Werke*, vol. 17, pp. 56-62 [Edición en castellano: 'La escisión del yo en el proceso defensivo', en *Obras completas*, vol. 23, pp. 275-278].

FRIEDRICH, JAMES (1993), 'Primary Error Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena', *Psychological Review*, vol. 100, pp. 298-319.

GARDINER, PATRICK (1970), 'Error, faith, and self-deception', *Proceedings of the Aristotelian Society*, vol. 70, pp. 221-243.

GAZZANIGA, MICHAEL S. (1985), *The Social Brain: Discovering the Networks of the Mind*, New York, Basic Books. [Edición en castellano: *El cerebro social*, trad. de Carlos Frade Blas, Madrid, Alianza, 1993]

GERGEN, KENNETH J. (1985), 'The Ethnopsychology of Self-Deception', en Mike Martin (ed.), *Self-deception and self-understanding*, Lawrence (KS), University Press of Kansas, pp. 228-43.

GIBBINS, KEITH Y DOUGLAS, WILLIAM (1985), 'Voice recognition and self-deception: a reply to Sackeim and Gur', Journal of Personality and Social Psychology, vol. 48 (5), pp. 1369–1372.

GILBERT DANIEL T. Y COOPER, JOEL (1985), 'Social Phychological Strategies of Self-Deception', en Mike Martin (ed.), *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas, pp. 75-94.

GLÜER, KATHRIN (2003), 'Is There Such a Thing as Weakness of the Will?', en *A philosophical smorgasbord: essays on action, truth and other things in honour of Fredrick Stoutland*, K. Segerberg, R. Sliwinski, (eds.), Uppsala, Uppsala University, pp. 65-83.

GOMILA, ANTONI (2007), 'El retorno de la Represión', *Teorema*, vol. XXVI (3), pp. 97-111.

GOSLING, JUSTIN (1990), *Weakness of the Will*, London, Routledge.

GUR, RUBEN C Y SACKEIM, HAROLD A. (1979), 'Self-Deception: A Concept in Search of a Phenomenon', *Journal of Personality and Social Psychology*, vol. 37 (2), pp. 147-169.

HAACK, SUSAN (1993), *Evidence and Inquiry. Towards Reconstruction in Epistemology*, Cambridge (MA), Blackwell Publishers. [Edición en castellano: *Evidencia e Investigación. Hacia una reconstrucción en epistemología*, trad. de María Ángeles Martínez García, Madrid, Tecnos, 1997].

HAIGHT, MARY R. (1980), *A Study of Self-Deception*, New Jersey, Humanities Press.

— (1985), 'Tales From a Black Box', en Mike Martin (ed.), *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas, pp. 244-260.

HALL, DAVID L. (1996), 'Our Names Are Legion for We Are Many: On the Academics of Self-Deception', en Roger T. Ames and Wimal

Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 241-261.

HAMER, DEAN H. (2004), *The God Gene: How Faith is Hardwired into our Genes,* New York, Doubleday.

HAMLYN, DAVID W. (1971), 'Self-Deception', *Proceedings of the Aristotelian Society*, vol. 45 (suplemento), pp. 45-60.

HARE, RICHARD M. (1963), *Freedom and Reason*, Clarendon Press, Oxford.

HARMAN, GILBERT (1976), 'Practical Reasoning', *Review of Metaphysics*, vol. 79, 431-463. Reeditado en Alfred Mele (ed.), *The Philosophy of Action*, Oxford, Oxford University Press, 1997, pp. 149-177.

HARRÉ, ROM (1988), 'The Social Context of Self-Deception', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 364-379.

HAYES, RICHARD P. (1996), 'Ritual, Self-Deception, and Make-Believe: A Classical Buddhist Perspective', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 349-363.

HELLMAN, NATHAN (1983), 'Bach on Self-Deception**',** *Philosophy and Phenomenological Research*, vol. 44(1), pp. 113-120.

HELM, PAUL (1994), *Belief Policies*, Cambridge, Cambridge University Press.

HEMPEL, CARL G. (1962), 'Deductive nomological vs. statistical explanation', en Herbert Feigl and Grover Maxwell (eds), *Minnesota Studies in the Philosophy of Science*, Vol. 3. University of Minnesota Press, Minneapolis, pp. 98–169.

— (1965), *Aspects of Scientific Explanation,* New York, The Free Press.

HERMES, CHARLES M. (2007), 'How Peer Evaluations Can Be Deceptive', *Teorema*, vol. XXVI (3), pp. 123-130.

HERNÁNDEZ BORGES, MARÍA ROSARIO (2007), 'La etiología del autoengaño: ¿pretendo engañarme o me engañan mis mecanismos?', *Teorema*, vol. XXVI (3), pp. 19-30.

HIGGINS, KATHLEEN M. (1996), 'Bad Faith and Kitsch as Models for Self-Deception', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 123-141.

HOLTON, RICHARD (1999), 'Intention and Weakness of Will', *Journal of Philosophy*, vol. 96 (5), pp. 241-262.

— (2001), 'What is the role of the self in self-deception?', *Proceedings of the Aristotelian Society*, vol. 101, pp. 53-69.

— (2003), 'How Is Strength of Will Possible?', en Sarah Stroud y Christine Tappolet (eds.), *Weakness of Will and Practical Irrationality*, Oxford, Clarendon Press, pp. 39-67.

HUME, DAVID (1739-40), *A Treatise on Human Nature*, L. A. Selby-Bigge (ed.), Oxford at the Clarendon Press, 1888. [Edición en castellano: *Tratado de la naturaleza humana*, trad. de Félix Duque, Madrid, Tecnos, 2002].

— (1748), *An enquiry concerning the human understanding*, en L. A. Selby-Bigge (ed.), *Enquiries concerning the human understanding and concerning the principles of morals*, texto revisado y notas de P. H. Nidditch, Oxford at the Clarendon Press, 1902. [Edición en castellano: *Investigación sobre el conocimiento humano. Investigación sobre los principios de la moral,* trad. de Jaime Salas y Gerardo López Sastre, Madrid, Tecnos, 2007].

— (1751), *An enquiry concerning the principles of morals*, en L. A. Selby-Bigge (ed.), *Enquiries concerning the human understanding and concerning the principles of morals*, texto revisado y notas de P. H. Nidditch, Oxford at the Clarendon Press, 1902. [Edición en castellano: *Investigación sobre el conocimiento humaon. Investigación sobre los principios de la moral,* trad. de Jaime Salas y Gerardo López Sastre, Madrid, Tecnos, 2007].

JAMES, WILLIAM (1890), *The Principles of Psychology*, 2 vols., New York, Henry Holt and Company. [Edición en castellano: *Principios de Psicología*, trad. de Agustín Bárcena, México, Fondo de Cultura Económica, 1989].

— (1896), 'The Will to Believe', *New World*, vol. 5, pp. 327-34. Reeditado en The *Will to Believe and Other Essays*, New York, Longmans, Green, and Co., 1897. [Edición en castellano: 'La voluntad de creer', en *La voluntad de creer. Un debate sobre la ética de la creencia*, trad. de Lorena Villamil, Madrid, Tecnos, 2003, pp. 135-180].

— (1907), *Pragmatism. A New Name for Some Old Ways of Thinking. Popular Lectures on Philosophy*, New York, Longmans, Green & Co. [Edición en castellano: *Pragmatismo. Un nuevo nombre para viejas formas de pensar*, trad. de Ramón del Castillo, Madrid, Alianza, 2007].

— (1909), *The Meaning of Truth, A Sequel to 'Pragmatism'*, New York, Longmans, Green & Co. [Edición en castellano: *El significado de la verdad*, trad. de Luis Rodríguez Aranda, Madrid, Aguilar, 1966].

JANIS, IRVING L. Y KING, BERT T. (1954), 'The influence of role-playing on opinion change', *Journal of abnormal and Social Psychology,* vol. 49 (2), pp. 211-218.

JOHNSTON, MARK (1988), 'Self-Deception and the Nature of Mind', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 63-91.

KANT, IMMANUEL (1781/1787), *Kritik der reinen Vernunft*, en *Kant's gesammelte Schriften,* 29 vols., Berlin, Preussischen/Deutschen Akademie der Wissenschaften, 1902-1983, vols. III-IV. [Edición en castellano: *Crítica de la razón pura*, Madrid, Alfaguara, trad. de Pedro Ribas, 1998].

— (1797), *Die Metaphysik der Sitten,* en *Kant's gesammelte Schriften,* 29 vols., Berlin, Preussischen/Deutschen Akademie der Wissenschaften, 1902-1983, vol. VI. [Edición en castellano: *La metafísica de las costumbres*, trad. de Adela Cortina y Jesús Conill, Madrid, *Tecnos*, 1989].

— (1798), *Anthropologie in pragmatischer Hinsicht*, en *Kant's gesammelte Schriften,* 29 vols., Berlin, Preussischen/Deutschen Akademie der Wissenschaften,

1902-1983, Vol. VII. [Edición en castellano: *Antropología en sentido pragmático*, trad. de José Gaos, Madrid, *Alianza Editorial*, 2004].

KELMAN, HERBERT C. (1953), 'Attitude change as a function of response restriction', *Human Relations*, vol. 6 (3), pp. 185-214.

KING, BERT T. Y JANIS, IRVING L. (1956), 'Comparison of the effectiveness of improvised versus non-improvised role-playing in producing opinion change', *Human Relations*, vol. 9 (2), pp. 177-186.

KING-FARLOW, J. Y BOSLEY, RICHARD (1985), 'Self-Formation and the Mean (Programmatic Remarks on Self-Deception)', en Mike Martin (ed.) *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas, pp. 195-220.

KING-FARLOW, JOHN (1963), 'Self-Deceivers and Sartrian Seducers', *Analysis*, vol. 23 (6), pp. 131-136.

KIPP, DAVID (1980), 'On Self-Deception', *The Philosophical Quarterly*, vol. 30, pp. 305-317.

— (1985), 'Self-Deception, Inauthenticity, and Weakness of the Will', en Mike Martin (ed.) *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas, pp. 261-283.

KNIGHT, MARTHA (1988), 'Cognitive and Motivational Bases of Self-Deception: Commentary on Mele's *Irrationality'*, *Philosophical Psychology*, vol. 1(2), pp. 179-188.

KOHLENBACH, MARGRET (1988), 'Error or Self-Deception? The Case of Eduard in Goethe's *Elective Affinities*', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 515-534.

KRAUT, ROBERT E. (1980), 'Humans as lie detectors', *Journal of Communication*, vol. 30, pp. 209-216.

KUPPERMAN, JOEL J. (1996), 'Falsity, Psychic Indefiniteness, and Self-Knowledge', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 161-176.

KUNDA, ZIVA (1990), 'The Case for Motivated Reasoning', *Psychological Bulletin*, vol. 108, pp. 480-498.

LACAN, JACQUES (1966), *Écrits*, Paris, Éditions du Seuil. [Edición en castellano: *Escritos*, trad. de Tomás Segovia, Madrid, Siglo XXI, 1984, 2 tomos].

— (1978), *Le Séminaire de Jacques Lacan, Livre 2: Le moi dans la théorie de Freud et dans la technique de la psychanalyse*, París, Editions du Seuil. [Edición en castellano: *Seminario 2. El Yo en la Teoría de Freud y en la Técnica Psicoanalítica*, trad. de Irene Agoff, Barcelona, Paidós, 1983].

— (1998), *Le Séminaire de Jacques Lacan, Livre V: Les formations de l'inconscient*, París, Éditions du Seuil, 1998. [Edición en castellano: *Seminario 5. Las formaciones del inconsciente*, trad. de Enric Berenguer, Barcelona, Paidós, 1999].

LAFLEUR, WILLIAM R. (1996), 'A Half-Dressed Emperor: Societal Self Deception and Recent "Japanokritik" in America', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 263-285.

LAPLANCHE, JEAN Y PONTALIS, JEAN BERNARD (1967), *Vocabulaire de la psychanalyse*, París, PUF/Quadrige. [Edición en castellano: *Diccionario de Psicoanálisis,* trad. de Fernando Gimeno Cervantes, Barcelona, Paidós, 1996].

LAZAR, ARIELA (1997), 'Self-deception and the desire to believe', *Behavioral and Brain Sciences*, vol. 20, pp. 119-120.

— (1998), 'Division and Deception: Davidson on Being Self-Deceived', en *Self-Deception and Paradoxes of Rationality*, Jean-Pierre Dupuy (ed.), Stanford (CA), CSLI Publications,1998, pp. 19-36.

— (1999), 'Deceiving Oneself or Self-Deceived?', *Mind*, vol. 108 (430), pp. 265-290.

LEVY, NEIL (2003), 'Self-Deception and Responsibility for Addiction', *Journal of Applied Philosophy*, vol. 20 (2), pp. 133-142.

— (2004), 'Self-Deception and Moral Responsibility', *Ratio* (new series), vol. XVII, pp. 294-311.

— (2008), 'Self-Deception Without Thought Experiments', en J. Fernandez and T. Bayne (eds), *Delusions, Self-Deception and Affective Influences on Belief-formation*, New York, Psychology Press.

LINEHAN, ELISABETH (1982), 'Ignorance, Self-deception, and Moral Responsibility', *Journal of Value Inquiry*, vol. 16, pp. 101-115.

LOCKIE, ROBERT (2003), 'Depth Psychology and Self-Deception', *Philosophical Psychology*, vol. 16 (1), pp. 127-148.

MARTIN, MIKE (ed.) (1985), *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas.

— (1979), 'Self-Deception, Self-Pretence, and Emotional Detachment' *Mind*, New Series, vol. 88 (351), pp. 441-446.

— (1986), *Self-Deception and Morality*, Lawrence (KS), University Press of Kansas.

MARTÍNEZ MANRIQUE, FERNANDO (2007), 'Attributions of Self-Deception', *Teorema*, vol. XXVI (3), pp. 131-143.

MASIP, JAUME (2005), '¿Se pilla antes a un mentiroso a que a un cojo? Sabiduría popular frente a conocimiento científico sobre la detección no-verbal del engaño', *Papeles del Psicólogo*, vol. 26, pp. 78-91.

MCLAUGHLIN, BRIAN P. & RORTY, AMÉLIE O. (eds.) (1988), *Perspectives on Self-Deception*, University of California Press, Berkeley.

MCLAUGHLIN, BRIAN P. (1988), 'Exploring the Possibility of Self-Deception in Belief', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 29-62.

— (1996), 'On the very possibility of Self-Deception', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 31-51.

MELE, ALFRED R. (1982), 'Self-Deception, Action, and Will: Comments', *Erkenntnis*, vol. 18, pp. 159-164.

— (1983), 'Self-Deception', *Philosophical Quarterly*, vol. 33, pp. 365-377.

— (1987a), *Irrationality. An Essay on Akrasia, Self-Deception and Self Control*, Oxford, Oxford University Press.

— (1987b), 'Recent work on Self-Deception', *American Philosophical Quarterly*, vol. 24, pp. 1-17.

— (1997), 'Real Self-Deception', *Behavioral and Brain Sciences*, vol. 20 (1), pp. 91-102/127-136.

— (1998), 'Two Paradoxes Of Self-Deception', en *Self-Deception and Paradoxes of Rationality*, Jean-Pierre Dupuy (ed.), Stanford (CA), CSLI Publications, 1998, pp. 37-58.

— (1999), 'Twisted self-deception', *Philosophical Psychology*, vol. 12, pp. 117-137.

— (2001), *Self-Deception Unmasked*, Cambridge, Harvard University Press.

— (2003), 'Emotion and Desire in Self-deception', en Anthony Hatzimoysis (ed.), *Philosophy and the Emotions*, Cambridge, Cambridge University Press, 2003, pp. 163-179.

MIDGLEY, MARY (2003), *The Myths We Live By*, London/New York, Routledge.

MILO, RONALD D. (1984), *Immorality*, Princeton, Princeton University Press.

MIRI, MRINAL (1974), 'Self-Deception', *Philosophy and Phenomenological Research*, vol. 34 (4), pp. 576-585.

MORTON, ADAM (1988), 'Partisanship', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 170-182.

NEVILLE, ROBERT CUMMINGS (1996), 'A Confucian Construction os a Self-Deceivable Self', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 201-217.

NICHOLSON, ANNA (2007), 'Cognitive Bias, Intentionality and Self-Deception', *Teorema,* vol. XXVI (3), pp. 45-58.

NIETZSCHE, FRIEDRICH (1873), *Über Wahrheit und Lüge im außermoralischen Sinn*. [Edición en catellano: *Sobre verdad y mentira,* Madrid, Tecnos, 2007].

NISBETT, RICHARD Y ROSS, LEE (1980], *Human Inference: Strategies Social Judgment*, Englewood Cliffs, NJ, Prentice-Hall.

NUSSBAUM, MARTHA (1988), 'Love's Knowledge', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 487-514.

ORTEGA Y GASSET, JOSÉ (1940), *Ideas y Creencias*, Madrid, Alianza, 2001.

PALMER, ANTHONY (1979), 'Characterising Self-deception**',** *Mind*, New Series, vol. 88 (349), pp. 45-58.

PALUCH, STANLEY (1967), 'Self-deception', *Inquiry*, vol. 10, pp. 268-278.

PARKES, GRAHAM (1996), 'Facing The Self With Masks: Perspectives on the Personal from Nietzsche and the Japanese', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 287-313.

PASCAL, BLAISE (1670), *Pensées de M. Pascal sur la religion et sur quelques autres sujets, qui ont été trouvés après sa mort parmy ses papiers*, Paris*,* Guillaume Desprez. [Edición en castellano: *Pensamientos*, trad. de Xavier Zubiri, Espasa Calpe, Madrid, 1995]

PATTEN, DAVID (2003), 'How do we deceive ourselves?', *Philosophical Psychology*, vol. 16 (2), pp. 229-246.

PAULHUS, DELROY L., Y SUEDFELD, PETER (1988), 'A dynamic complexity model of self-deception', en Joan S. Lockard, y Delroy L. Paulhus (eds.), *Self-deception: An adaptive mechanism?*, Englewood Cliffs (NJ), Prentice-Hall, pp. 133-145.

PEARS, DAVID (1982), 'Motivated irrationality, Freudian theory and cognitive dissonance', en Wollheim, R. y Hopkins, J. (Eds.), *Philosophical Essays on Freud,* Cambridge, Cambridge University Press*,* (1988), pp. 264-288.

— (1984), *Motivated irrationality*, Oxford, Oxford University Press.

— (1986), 'The Goals and Strategies Of Self-Deception', en Elster, J. (Ed.) *The Multiple Self*, Cambridge, Cambridge University Press, pp. 59-77.

— (1991), 'Self-Deceptive Belief-Formation', *Synthese*, vol. 89 (3), pp. 392-405.

PEIRCE, CHARLES SANDERS (1877), 'The Fixation of Belief', *Popular Science Monthly*, vol. 12, pp. 1-15. Reeditado en *Writings of Charles S. Peirce. A chronological Edition, Volume 3, 1872-1878*, Bloomington, Indiana University Press, pp. 242-257. ['La fijación de la creencia', en *La fijación de la creencia. Cómo aclarar nuestras ideas*, Oviedo, KRK Ediciones, 2007, pp. 27-64].

— (1878), 'How to Make Our Ideas Clear', *Popular Science Monthly*, 12, pp. 286-302. Reeditado en *Writings of Charles S. Peirce. A chronological Edition, Volume 3, 1872-1878*, Bloomington, Indiana University Press, pp. 257-276. ['Cómo aclarar nuestras ideas, en *La fijación de la creencia. Cómo aclarar nuestras ideas*, Oviedo, KRK Ediciones, 2007, pp. 67-110].

PENELHUM, TERENCE (1964), 'Pleasure and Falsity', *American Philosophical Quarterly*, vol. 1, pp. 81-91. Reeditado en *Philosophy of Mind*, Stuart Hampshire (ed.), New York, Harper & Row, 1966, pp. 242-266.

PERRING, CHRISTIAN (1997), 'Direct, Fully Intentional Self-Deception Is Also Real', *Behavioral and Brain Sciences*, vol. 20, pp. 123-124.

PETERMAN, JAMES (1983), 'Self-deception and the problem of avoidance', *Southern Journal of Philosophy*, 21, pp. 565-574.

PIHLSTRÖM, SAMI (2007), 'Transcendental Self-Deception', *Teorema,* vol. XXVI (3), pp. 177-189.

PIPER, ADRIAN M. S. (1988), 'Pseudorationality', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 297-323.

PLATÓN, *Diálogos*, 9 vols., Madrid, Gredos.

PRICE, HENRY H. (1973), *Belief. The Gifford Lectures delivered at the University of Aberdeen in 1960*, London, George Allen & Unwin Ltd.

— (1973), *Perception*, London, Methuen.

QUATTRONE, GEORGE A., & TVERSKY, AMOS (1984), 'Causal versus diagnostic contingencies: On self-deception and on the voter's illusion', *Journal of Personality and Social Psychology*, vol. 46 (2), pp. 237-248.

QUINE, WILLARD V. Y ULLIAN, JOSEPH S. (1978), *The web of belief*, New York, Random House.

— (1956), 'Quantifiers and Propositional Attitudes'*, The Journal of Philosophy*, 53 (5), pp. 177-187. [Edición en castellano: 'Cuantificadores y actitudes

proposicionales', en Thomas Moro Simpson (comp.), *Semántica filosófica: problemas y discusiones*, Buenos Aires, Siglo XXI, 1993, pp. 217-230]

REY, GEORGES (1988), 'Toward a Computational Account of *Akrasia* and Self-Deception', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 264-296.

RORTY, AMÉLIE O. (1988), 'The Deceptive Self: Liars, Layers and Lairs', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 11-28.

— (1996), 'User-Friendly Self-Deception: a Traveler's Manual', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 73-89.

RORTY, RICHARD (1991), *Objectivity, Relativism, and Truth: Philosophical Papers, Volume 1*, Cambridge, Cambridge University Press. [Edición en castellano: *Objetivismo, relativismo y verdad. Escritos filosóficos 1*, trad. de Jorge Vigil Rubio, Barcelona, Paidós, 1996].

RUDDICK, WILLIAM (1988), 'Social Self-Deception', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 380-389.

RUSSELL, BERTRAND (1909), 'Pragmatism', *The Edinburgh Review*, 209, pp. 363-388 (sin firmar). Reeditado en *Philosophical Essays*, Longmans, Green, 1910; también en *The Collected Papers of Bertrand Russell, Vol. 6:*

Logical and Philosophical Papers, 1909-13, pp. 260-284. [Edición en castellano: 'Pragmatismo', en *Ensayos Filosóficos*, trad. de Juan Ramón Capella, Madrid, Alianza, pp. 95-134]

RYLE, GILBERT (1949), *The concept of Mind*, New York, Barnes & Noble. [Edición en castellano: *El concepto de lo mental*, trad. de Eduardo Rabossi, Barcelona, Paidós, 2005].

SACKEIM, HAROLD (1983), 'Self-deception, self-esteem, and depression: the adaptive value of lying to oneself', en J. Masling (ed.) *Empirical studies of psychoanalytic theories*, vol. 1., Hillsdale (NJ), Analytic Press.

— (1988), 'Self-deception: a synthesis', en Joan S. Lockard & Delroy L. Paulhus (eds), *Self-deception: an adaptive mechanism?*, Englewood Cliffs (NJ), Prentice Hall.

SACKEIM, HAROLD A., Y GUR, RUBEN C. (1979), 'Self-deception, other-deception, and self-reported psychopathology', *Journal of Consulting and Clinical Psychology*, vol. 47 (1), pp. 213-215.

— (1985) 'Voice recognition and the ontological status of self-deception', *Journal of Personality and Social Psychology*, vol. 48 (5), pp. 1365-1368.

— (1978), 'Self-deception, self-confrontation, and consciousness', en G. E. Schwartz, y D. Shapiro (eds.), *Consciousness and self-regulation: Advances in theory and research*, vol. 2, New York, Plenum Press, pp. 139-197.

— (1997), 'Flavors of Self-deception: Ontology and Epidemiology', *Behavioral and Brain Sciences*, vol. 20 (1), pp. 125-126.

SAHDRA, BALJINDER Y THAGARD, PAUL (2003), 'Self-deception and Emotional Coherence', *Mind and Machines*, vol. 13, pp. 213-231.

SANFORD, DAVID H. (1988), 'Self-Deception as Rationalization', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 157-169.

SARTRE, JEAN PAUL (1943), *L'Être et le Néant. Essai d'ontologie phénoménologique*, Paris, Gallimard. [Edición en castellano: *El ser y la Nada. Ensayo de Ontología y Fenomenología*, trad. de Juan Valmar, Buenos Aires, Losada, 2005, pp. 95-125].

SAUSSURE, FERDINAND DE (1916), *Cours de linguistique générale*, Paris, Payot. [Edición en castellano: *Curso de lingüística general*, trad. de Amado Alonso, Buenos Aires, Losada, 1945].

SCHMITT, FREDERICK F. (1988), 'Epistemic Dimensions of Self-Deception', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 183-204.

SCOTT-KAKURES, DION (1996), 'Self-Deception and Internal Irrationality', *Philosophy and Phenomenological Research*, vol. 56 (1), pp. 31-56.

SEARLE, JOHN (1980), 'Mind, brains and programs', *Behavioral and Brain Sciences*, vol. 3 (3), pp. 417-457.

— (1983), *Intentionality*, Cambridge (MA), Cambridge University Press.

SELLARS, WILFRID (1956), 'Empiricism and the Philosophy of Mind', en Herbert Feigl and Michael Scriven (eds.), *Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of*

Psychology and Psychoanalysis, Minneapolis, University of Minnesota Press, pp. 253-329. [Edición en castellano:'El empirismo y la filosofía de lo mental', en *Ciencia, percepción y realidad*, trad. de Víctor Sánchez de Zavala, Madrid, Tecnos, 1971, pp. 139-209]

SIEGLER, FREDERICK A. (1963), 'Self-deception', *Australasian Journal of Philosophy*, vol. 41, pp. 29-43.

SLOAN, TOD S. (1987), *Deciding. Self-Deception in Life Choices*, New York, Methuen.

SMITH, ADAM (1759), 'Of the Nature of Self-deceit, and of the Origin and Use of general Rules', en *The Theory of Moral Sentiments,* Parte III, Cap. IV. [Edición en castellano: *La teoría de los sentimientos morales*, trad. de Carlos Rodríguez Braum, Madrid, Alianza, 2004]

SMITH, BARBARA H. (1996), 'Unloading the Self-Refutation Charge', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 143-160.

SNYDER, C. RICHARD (1985), 'Collaborative Companions: The Relationship of Self-Deception and Excuse Making', en Mike Martin (ed.), *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas, pp. 35-51.

SOLOMON, ROBERT C. (1996), 'Self, Deception and Self Deception in Philosophy', en Roger T. Ames and Wimal Dissanayake (eds.), *Self and Deception, a cross-cultural philosophical enquiry*, Albany (NY), State University of New York Press, pp. 91-121.

SOMMER, VOLKER (1992), *Lob der Lüge. Täuschung und Selbstbetrug bei Tier und Mensch*, München, C.H. Beck. [Edición en castellano: *Elogio de la mentira. Engaño y autoengaño en hombres y otros animales*, trad. de Oliver Strunk, Galaxia Gutenberg-Círculo de Lectores, 2005].

SORENSEN, ROY A. (1985), 'Self-Deception and Scattered Events', *Mind*, vol. 94 (373), pp. 64-69.

STICH, STEPHEN (1983), *From folk psychology to cognoscitive science: the case against belief*, Cambridge, MIT Press.

STURM, THOMAS (2007), 'Self-Deception, Rationality and the Self', *Teorema,* vol. XXVI (3), pp. 73-95.

SZABADOS, BÉLA (1973), 'Wishful Thinking and Self-Deception', *Analysis*, vol. 33, pp. 201-205.

— (1974a), 'Self-Deception', *Canadian Journal of Philosophy*, vol. 4 (1), pp. 51-68.

— (1974b), 'The Morality of Self-Deception', *Dialogue*, vol. 13, pp. 25-34.

— (1974c), 'Rorty on Belief and Self-Deception', *Inquiry*, vol. 17, pp. 464-473.

— (1985), 'The Self, Its Passions and Self-Deception', en Mike Martin (ed.) *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas, pp. 143-168.

TALBOTT, WILLIAM J. (1995), 'Intentional self-deception in a single coherent self', *Philosophy and Phenomenological Research*, vol. 55 (1), pp. 27-74.

THOMAS, ALAN (2007), 'Practical Irrationality, Reflexivity and Sartre's Regress Argument', *Teorema,* vol. XXVI (3), pp. 113-121.

TOV-RUACH, LEILA (1988), 'Freud on Unconscious Affects, Mourning, and the Erotic Mind', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 246-263.

TRIVERS, ROBERT (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*, vol. 907, pp. 114-131.

TROPE, YAACOV Y LIBERMAN, AKIVA (1996), 'Social Hypothesis Testing: Cognitive and Motivational Mechanisms', en E. Tory Higgins y Arie W. Kruglanski (eds.), *Social Psychology: Handbook of Basic Principles*, New York, Guilford Press, pp. 239-270.

TVERSKY, AMOS Y KAHNEMAN, DANIEL (1974), 'Judgement under uncertainty: Heuristics and biases', *Science*, vol. 185, pp. 1124-1130.

— (1981), 'The framing of decisions and psychology of choice', *Science*, vol. 211, pp. 453-458.

VALDÉS VILLANUEVA, LUIS M. (2003), 'El derecho de creer', en *La voluntad de creer*, Madrid, Tecnos, 2003.

VAN FRAASSEN, BAS C. (1988), 'The Peculiar Effects of Love and Desire', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-*

Deception, Berkeley (CA), University of California Press, 1988, pp. 123-156.

VAN LEEUWEN, NEIL (2007), 'The Spandrels of Self-Deception: Prospects for a Biological Theory of a Mental Phenomenon', *Philosophical Psychology*, vol. 20 (3), pp. 329-348.

VELARDE LOMBRAÑA, JULIÁN (2005), 'Incertidumbre y grados de creencia', *Teorema*, vol. XXIV (2), pp 27-41.

VRIJ, ALDERT (2000), *Detecting Lies and Deceit: The Psychology of Lying and the Implications tor Professional Practice*, Wiley, Chichester.

WERNHAM, JAMES C. S. (1987), *James's Will-To-Believe Doctrine: A Heretical View*, Kingston (Ontario), McGill-Queen's University Press.

WHITE, ALAN R. (1964), *Attention*, Oxford, Basil Blackwell.

WHITE, STEPHEN L. (1988), 'Self-Deception and Responsibility for the Self', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 450-484.

WHITEN, ANDREW Y BYRNE, RICHARD W. (eds.) (1988), *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*, New York, Oxford University Press.

— (eds) (1997), *Machiavellian intelligence II: Extensions and Evaluations*, Cambridge (MA), Cambridge University Press.

WILLIAMS, BERNARD (1973), 'Deciding to Believe', en *Problems of the Self*, Cambridge, Cambridge University Press, pp. 136-151. [Edición en castellano: 'Decidirse a creer', en *Problemas del Yo*, trad. de José M. G. Holguera, México, UNAM, 1986, pp. 181-209].

— (1996), 'Truth, Politics, and Self-Deception', in *Social Research*, vol. 63 (3), pp. 603-617.

— (2002), *Truth and Truthfulness. A Essay in Genealogy*, Princeton, NJ, Princeton University Press.

WILSHIRE, BRUCE (1988), 'Mimetic Engulfment and Self-Deception', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 390-404.

WILSON, TIMOTHY D. (1985), 'Self-Deception without Repression: Limits on Access to Mental States', en Mike Martin (ed.), *Self-Deception and Self-Understanding. New Essays in Philosophy and Psychology*, Lawrence (KS), University Press of Kansas, pp. 95-116.

WITTGENSTEIN, LUDWIG (1953), *Philosophische Untersuchungen,* ed. bilingüe alemán/ingles, Gertrude Elizabeth Margaret Anscombe y Rush Rhees (eds.), texto alemán y trad. inglesa G.E.M. Anscombe, Oxford, Blackwell. [Edición en castellano: *Investigaciones Filosóficas,* ed. bilingüe alemán/castellano, trad. de Alfonso García Suárez y Ulises Moulines, Crítica/UNAM, Barcelona, 2002].

WOLLHEIM, RICHARD Y HOPKINS, JAMES (eds.) (1988), *Philosophical Essays on Freud, C*ambridge, Cambridge University Press.

WOOD, ALLEN W. (1988), 'Ideology, False Consciousness, and Social Illusion', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 345-363.

— (1988), 'Self-Deception and Bad Faith', en Brian P. McLaughlin and Amélie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley (CA), University of California Press, 1988, pp. 207-227.