teorema
Vol. XXXIV/1, 2015, pp. 95-100
ISSN: 0210-1602
[BIBLID 0210-1602 (2015) 34:1; pp. 95-100]

SIMPOSIO SOBRE EL LIBRO/SYMPOSIUM ON THE BOOK

TRANSPARENT MINDS: A STUDY OF SELF-KNOWLEDGE

Resumen de *Transparent Minds*

Jordi Fernández

El proyecto de *Transparent Minds*: *A Study of Self-Knowledge* consiste en explicar nuestro conocimiento de nuestros estados mentales o, más específicamente, nuestro conocimiento de nuestras actitudes proposicionales. El libro se divide en dos partes. En la primera, que abarca los capítulos 1-3, se especifica de manera más profunda el problema del auto-conocimiento y se propone una explicación del auto-conocimiento para la creencia y el deseo. En la segunda parte, que abarca los capítulos 4-6, se presentan tres aplicaciones de esta explicación para iluminar, respectivamente, la paradoja de Moore, el delirio de inserción del pensamiento, y el auto engaño.

La explicación del auto-conocimiento que se propone en *Transparent Minds* pretende tomar en serio la famosa observación de Gareth Evans sobre la 'transparencia de la creencia'; la observación de que nos atribuimos creencias dirigiendo nuestra atención hacia afuera, hacia el mundo, y no hacia adentro (como sugiere la noción de introspección). En el capítulo 1 me alineo con aquellos teóricos del auto-conocimiento que piensan que obtenemos conocimiento de nuestras propias creencias y deseos mirando hacia afuera o, como me gusta expresarlo, mirando más allá de esas creencias y deseos. El capítulo 1 especifica también un cierto número de constricciones que debe respetar cualquier explicación del auto-conocimiento. El resultado principal

de esta discusión es que existe una interesante tensión entre dos tesis acerca del auto-conocimiento que, a primera vista, parecen bastante razonables. Una de ellas es que el auto-conocimiento constituye un logro cognitivo. La otra es que el auto-conocimiento es falible. Dicho de manera aproximada: la tensión consiste en que la primera tesis nos empuja hacia la idea de que tenemos que tener razones para atribuirnos creencias y deseos. Pero esos estados mentales que parecen ser candidatos plausibles para desempeñar el papel de razones para atribuirnos creencias y deseos, hacen que el auto-conocimiento sea infalible, algo que se da de bruces contra la segunda tesis. La tarea que emprende el resto del libro es la de construir una explicación 'transparente' del auto-conocimiento que, por un lado, explique lo que es epistémicamente peculiar de nuestras auto-atribuciones de creencias y deseos y que, por otro, resuelva la tensión que se acaba de mencionar.

En el capítulo 2, presento lo que llamo la explicación '*bypass*' del auto-conocimiento para la creencia. La afirmación principal de esta explicación es que nos atribuimos creencias sobre la base de nuestros fundamentos para esas creencias. Los fundamentos que un sujeto tiene para una creencia se estipula que son estados mentales que tienden a causar que ese sujeto tenga la creencia en cuestión. Así, si yo tiendo a creer que hace un día soleado cuando me parece que percibo el Sol, entonces mi experiencia perceptiva del Sol constituye mi fundamento para creer que hace un día soleado. La idea es que, cuando me atribuyo la creencia de que hace un día soleado, lo hago basándome en el mismo estado mental que constituye mi fundamento para creer que hace un día soleado, a saber: mi experiencia perceptiva del Sol. Esto explicaría por qué la auto-atribución que hago está justificada puesto que, normalmente, tendré la creencia de que hace un día soleado cuando me parece que percibo el Sol. Esto significa que el tipo de razón que aduciría para apoyar mi auto-atribución de la creencia de que hace un día soleado (a saber: que me parece que percibo el Sol) tiende a estar correlacionada con la verdad de esta auto-atribución. Explicaría también por qué cuando me atribuyo la creencia de que hace un día soleado, ya no es para mí una cuestión abierta si hace o no un día soleado: en virtud de esta atribución experimento algún tipo de presión para creer que hace un día soleado. Me refiero a este rasgo de nuestras auto-atribuciones de creencia como su 'asertividad'. El modelo *bypass* del auto-conocimiento para la creencia parece explicar la asertividad de nuestras auto-atribuciones de creencia. Sin embargo, la

idea de que un único estado mental pueda constituir tanto mis fundamentos para una creencia de primer orden, como la base para la auto-atribución de esa creencia hace que la explicación resulte vulnerable a un cierto número de objeciones que tienen que ver con la idea de que mi justificación para creer algo y mi justificación para creer que lo creo no van necesariamente de la mano. Esas objeciones se responden separando los hechos en virtud de los que un estado mental puede constituir una buena evidencia para una creencia de primer orden de los hechos en virtud de los cuales tal estado puede constituir una buena evidencia para una creencia de segundo orden.

En el capítulo 3 sugiero que el modelo *bypass* del auto-conocimiento puede extenderse fácilmente al caso del deseo. La principal afirmación del modelo, por lo que respecta al deseo, es que nos atribuimos deseos basándonos en los fundamentos que tenemos para esos deseos. Los fundamentos que tiene un sujeto para un deseo se estipula que son estados mentales que tienden a causar que ese sujeto tenga el deseo en cuestión. Así, si tiendo a querer beber cuando tengo la sensación de que estoy sediento, entonces mi sed constituye el fundamento que tengo para querer beber. La idea es entonces que, cuando me atribuyo el deseo de beber, lo hago basándome en el mismo estado mental que constituye el fundamento que tengo para querer beber, a saber: mi sed. Esto explicaría por qué mi auto-atribución está justificada puesto que, normalmente, tendré el deseo de beber cuando tengo la sensación de que estoy sediento. Esto significa que el tipo de razón que aduciría en apoyo de mi auto-atribución del deseo de beber (a saber: que tengo la sensación de que estoy sediento) tiende a estar correlacionada con la verdad de esa auto-atribución. Sin embargo, como en el caso de la creencia, la idea de que un único estado mental puede constituir tanto el fundamento que tengo para un deseo, como aquello en lo que se basa el que me atribuya ese deseo, plantea un cierto número de dificultades. Abordo esas dificultades, una vez más, separando los hechos in virtud de los cuales un estado mental puede constituir el fundamento que un sujeto tiene para un deseo, de los hechos en virtud de los cuales el mismo estado mental puede constituir una buena evidencia para su creencia de que tiene ese deseo. Considero que una virtud del modelo del *bypass* es que su explicación del auto-conocimiento para la creencia y su explicación del auto-conocimiento para el deseo estén tan estrechamente relacionadas.

El capítulo 4 contiene la primera aplicación del modelo del auto-conocimiento a otros problemas filosóficos. La paradoja de Moore, como afirmo, puede resolverse aplicando el modelo *bypass* del auto-conocimiento para la creencia. La paradoja de Moore es un puzle sobre las auto-adscripciones de creencia de los tipos 'P y no creo que P' y 'P y creo que no P'. El puzle consiste en que esas afirmaciones pueden ser verdaderas, pero parece irracional hacerlas y parece igualmente irracional creer sus contenidos. ¿De dónde viene esta intuición de irracionalidad? Sostengo que viene de que se detecta aquí que el sujeto que hace esas afirmaciones peca de negligencia epistémica, puesto que no está tomando en consideración los fundamentos que tiene para esas creencias. En el caso de 'P y no creo que P', el sujeto está afirmando que no cree que P de modo que, si se acepta que el modelo del *bypass* para el auto-conocimiento es correcto, no puede haber encontrado fundamentos para creer que P y, con todo, afirma que P. En el caso de 'P y creo que no P', el sujeto está afirmando que cree que no P; con lo que si se acepta que el modelo del *bypass* para el auto-conocimiento es correcto, tiene que haber encontrado fundamentos para creer que no P y, con todo, afirma que P. De este modo, la propuesta de que la negligencia epistémica explica la intuición de irracionalidad de la paradoja de Moore trata análogamente ambas formas de la paradoja, lo que parece ser una ventaja.

El capítulo 6 ofrece una explicación de una forma particular de auto-engaño que es paralela a la solución de la paradoja de Moore que acabamos de considerar. La forma relevante de auto-engaño es aquella en la que el sujeto se comporta de una manera que sugiere que cree que P, pero afirma que no cree que P. Por ejemplo, Jack puede afirmar que no cree que esté enfermo a la vez que evita claramente visitar a un médico o tomar consejo médico de cualquier género. Mi tesis es que el modelo *bypass* del autoconocimiento puede explicar nuestra intuición de que este sujeto merece ser censurado, en tanto que víctima de un auto-engaño de una manera análoga a como explicamos nuestra intuición de que es irracional hacer una aserción de la paradoja de Moore del tipo 'P y no creo que P'. Jack afirma que no cree que esté enfermo; suponiendo que el modelo *bypass* del auto-conocimiento sea correcto, entonces Jack no puede haber encontrado fundamento alguno para creer que está enfermo. Pero, con todo, su conducta indica que él cree que está enfermo. Así pues, se ha formado una creencia para la que no tiene fundamentos. Hay, desde luego, explicaciones al-

ternativas de esta variedad de auto-engaño, y variedades alternativas de auto-engaño que han de explicarse. En el capítulo 6 subrayo las virtudes de esta explicación en relación con otras explicaciones del auto-engaño y hago un esfuerzo para incluir otras variedades de este fenómeno. Al final, sin embargo, la conclusión es que la explicación no puede cubrir todos los casos que incluiríamos bajo el rótulo 'auto-engaño' y planteo la cuestión de si el auto-engaño podría no ser, por así decirlo, un género natural.

El capítulo 5 es un poco diferente. En los capítulos 4 y 6 se pone énfasis sobre la idea de que nuestro fundamento para una cierta creencia (o deseo), así como la base para nuestra auto-atribución de tal creencia o deseo, es un único estado mental. Esta es la idea clave de la explicación propuesta para el auto-engaño y de la solución propuesta para la paradoja de Moore. Por contraste, en el capítulo 5 el énfasis se coloca en una idea diferente, a saber: la idea de que nuestras auto-atribuciones de creencias son asertivas. Esta idea se usa en una explicación del delirio de inserción de pensamiento. Los pacientes que lo sufren están bajo la impresión de que tienen pensamientos que no son suyos. Mi propuesta es que tales pacientes experimentan esos pensamientos como meros ítems de información, respecto de cuya verdad no se sienten comprometidos. Son capaces de atribuirse esos pensamientos puesto que los pueden encontrar en sus propias mentes. Pero no experimentan presión alguna para suscribirlos en virtud del hecho de que ellos mismos se los atribuyen. Y esta es la razón por la que afirman que los pensamientos relevantes no son suyos. Si esto es correcto, y suponiendo que el modelo *bypass* del auto-conocimiento explica por qué nuestras auto-atribuciones de creencias son asertivas, la razón por la que los pacientes afectados por el delirio de inserción de pensamientos no experimentan sus auto-atribuciones de pensamientos como asertivas parece ser que es que no pueden atribuirse esos pensamientos a sí mismos por medio del *bypass*.

El resultado es que el modelo *bypass* del autoconocimiento explica nuestra posición de privilegio epistémico con respecto a nuestras propias creencias y deseos y se puede aplicar de manera interesante a otros problemas filosóficos. El modelo identifica las razones para nuestras auto-atribuciones de creencias y deseos, es decir: los fundamentos de que disponemos para esas creencias y deseos. Y deja lugar para la posibilidad de error en nuestras auto-atribuciones de creencias y deseos, puesto que el hecho de que tengamos fundamentos para una creencia o un de-

seo, no lleva necesariamente a que tengamos esa creencia o ese deseo. De este modo, el modelo acomoda tanto la tesis de que el autoconocimiento es falible, como la de que implica un logro cognitivo.

School of Humanities
University of Adelaide
Adelaide SA 5005, Australia
E-mail: jorge.fernandez@adelaide.edu.au

AGRADECIMIENTOS

Trabajo realizado en el marco del proyecto de investigación "Autoconocimiento, expresión y transparencia" (FFI2012-38908-C02-02), financiado por el Ministerio de Economía y Competitividad del Gobierno de España

teorema
Vol. XXXIV/1, 2015, pp. 101-105
ISSN: 0210-1602
[BIBLID 0210-1602 (2015) 34:1; pp. 101-105]

Précis of *Transparent Minds*

Jordi Fernández

The project in *Transparent Minds: A Study of Self-Knowledge* is the project of explaining our knowledge of our own mental states or, specifically, our knowledge of our propositional attitudes. The book is divided in two parts. In the first part, which comprises chapters 1, 2 and 3, the problem of self-knowledge is specified further, and an account of self-knowledge for belief and desire is proposed. In the second part, which comprises chapters 4, 5 and 6, three applications of this account are drawn to illuminate, respectively, Moore's paradox, the thought insertion delusion, and self-deception.

The account of self-knowledge proposed in *Transparent Minds* is meant to take seriously Gareth Evans's famous observation about the 'transparency of belief'; the observation that we self-attribute beliefs by focusing our attention outwards, upon the world, and not inwards (as the notion of introspection suggests). In chapter 1, I align myself with those theorists of self-knowledge who think that we obtain knowledge of our own beliefs and desires by looking outwards or, as I put it, by looking past those beliefs and desires. Chapter 1 also specifies a number of constraints that any account of self-knowledge should respect. The main outcome of that discussion is that there is an interesting tension between two views about self-knowledge that, at first glance, seem reasonable enough. One of them is that self-knowledge constitutes a cognitive achievement. The other one is that self-knowledge is fallible. Roughly, the tension is that the first view pushes us towards the idea that we must have reasons for our self-attributions of beliefs and desires. But those mental states which seem to be plausible candidates for the role of reasons for our self-attributions of beliefs and desires make self-knowledge infallible, which flies in the face of the second view. The task for the rest of the book is that of

building a 'transparent' account of self-knowledge that, on the one hand, explains what is epistemically distinctive about our self-attributions of beliefs and desires and, on the other hand, resolves the just-mentioned tension.

In chapter 2, I put forward what I call the 'bypass' account of self-knowledge for belief. The main tenet of this account is that we self-attribute beliefs on the basis of our grounds for those beliefs. A subject's grounds for a belief are stipulated to be mental states which tend to cause that subject to have the belief in question. Thus, if I tend to believe that it is sunny when I seem to perceive the sun, then my perceptual experience of the sun constitutes my grounds for believing that it is sunny. The thought is that, when I self-attribute the belief that it is sunny, I do it based on the very same mental state which constitutes my grounds for believing that it is sunny, namely, my perceptual experience of the sun. This would explain why my self-attribution is justified since, normally, I will have the belief that it is sunny when I seem to perceive the sun. Which means that the type of reason that I would produce in support of my self-attribution of the belief that it is sunny (namely, that I seem to perceive the sun) tends to correlate with the truth of that self-attribution. It would also explain why, when I self-attribute the belief that it is sunny, the question of whether it is sunny or not is no longer open for me: I thereby experience some pressure to believe that it is sunny. I refer to this feature of our self-attributions of belief as their 'assertiveness.' The bypass model of self-knowledge for belief seems to explain the assertiveness of our self-attributions of belief. However, the idea that a single mental state can constitute both my grounds for a first-order belief and the basis of my self-attribution of that belief makes the account vulnerable to a number of objections; objections having to do with the idea that my justification for believing something and my justification for believing that I believe it do not necessarily come hand in hand. These objections are handled by separating the facts in virtue of which a mental state can constitute good evidence for a first-order belief from the facts in virtue of which it can constitute good evidence for a second-order belief.

In chapter 3, I suggest that the bypass model of self-knowledge can be easily extended to the case of desire. The main tenet of the model, with regards to desire, is that we self-attribute desires on the basis of our grounds for those desires. A subject's grounds for a desire

are stipulated to be mental states which tend to cause that subject to have the desire in question. Thus, if I tend to want to drink when I feel thirsty, then my thirst constitutes my grounds for wanting to drink. The thought is that, when I self-attribute the desire to drink, I do it based on the very same mental state which constitutes my grounds for wanting to drink, namely, my thirst. This would explain why my self-attribution is justified since, normally, I will have the desire to drink when I feel thirsty. Which means that the type of reason that I would produce in support of my self-attribution of the desire to drink (namely, that I feel thirsty) tends to correlate with the truth of that self-attribution. As in the case of belief, however, the idea that a single mental state can constitute both my grounds for a desire and the basis of my self-attribution of that desire raises a number of difficulties. I approach those difficulties, once again, by separating the facts in virtue of which a mental state can constitute a subject's grounds for a desire from the facts in virtue of which the very same mental state can constitute good evidence for her belief that she has that desire. I take it to be a virtue of the bypass model that its explanation of self-knowledge for belief and its explanation of self-knowledge for desire are so closely related.

Chapter 4 contains the first application of the model of self-knowledge to other philosophical problems. Moore's paradox, I claim, can be solved by applying the bypass model of self-knowledge for belief. Moore's paradox is a puzzle about self-ascriptions of belief of the types 'P and I do not believe that P' and 'P and I believe that not-P'. The puzzle is that those claims can be true, but it seems irrational to make them and it seems equally irrational to believe their contents. Where does this irrationality intuition come from? I contend that it comes from sensing that the subject who makes those claims is epistemically negligent in that she is disregarding her grounds for belief. In the case of 'P and I do not believe that P', she is claiming that she does not believe that P so, provided that the bypass model of self-knowledge is correct, she must have found no grounds for believing that P. And yet, she claims that P. In the case of 'P and I believe that not-P', she is claiming that she believes that not-P so, provided that the bypass model of self-knowledge is correct, she must have found grounds for believing that not-P. And yet, she claims that P. Thus, the proposal that epistemic negligence explains the irrationality intuition

in Moore's paradox treats both forms of the paradox analogously, which seems to be a virtue of it.

Chapter 6 offers an account of a particular form of self-deception that is parallel to the solution to Moore's paradox that we have just considered. The relevant form of self-deception is one in which the subject behaves in a way that suggests that she believes that P, but she claims not to believe it. For instance, Jack may claim that he does not believe that he is sick while he clearly avoids seeing the doctor or getting medical advice of any kind. My contention is that the bypass model of self-knowledge can explain our intuition that this subject is blameworthy for his self-deception in an analogous way to that in which it explains our intuition that it is irrational to make a Moore-paradoxical assertion of the type 'P and I do not believe that P.' Jack claims not to believe that he is sick. Assuming that the bypass model of self-knowledge is correct, then Jack must have found no grounds for believing that he is sick. And yet, his behavior indicates that he believes that he is sick. So he has formed a belief for which he has no grounds. There are, of course, alternative explanations for this variety of self-deception, and alternative varieties of self-deception to be explained. In chapter 6, I highlight the virtues of this account over alternative accounts of self-deception, and I make an effort to broaden the account so as to include other varieties of self-deception. Ultimately, though, the conclusion is that the account cannot cover all those cases which we would include under the title of 'self-deception', and I raise the question of whether self-deception might not be, so to speak, a natural kind.

Chapter 5 is a little different. The emphasis in chapters 4 and 6 is on the idea that a single mental state constitutes our grounds for a certain belief (or desire) as well as the basis for our self-attribution of it. That is the key idea in the proposed explanation of self-deception and the proposed solution to Moore's paradox. In chapter 5, by contrast, the emphasis is on a different idea, namely, the idea that our self-attributions of beliefs are assertive. This idea is used in an account of the thought insertion delusion. Patients who suffer this delusion are under the impression that they have thoughts that are not theirs. My proposal is that these patients experience those thoughts as mere pieces of information, the truth of which they do not feel committed to. They are able to attribute those thoughts to themselves, since they can find them in their own minds. But they do not experience the pressure

to endorse them in virtue of self-attributing them. And this is why they claim that the relevant thoughts are not theirs. If this is right, and provided that the bypass model of self-knowledge explains why our self-attributions of beliefs are assertive, the reason why thought insertion patients do not experience their self-attributions of thoughts as being assertive seems to be that they cannot self-attribute those thoughts through bypass.

The outcome is that the bypass model of self-knowledge explains our privileged epistemic position with respect to our own beliefs and desires, and it provides some interesting applications to other philosophical problems. The model identifies reasons for our self-attributions of beliefs and desires, namely, our grounds for those beliefs and desires. And it allows for the possibility of error in our self-attributions of beliefs and desires because our having grounds for a belief, or a desire, does not necessarily lead to our having that belief, or that desire. Thus, the model accommodates both the view that self-knowledge is fallible and the view that it involves a cognitive achievement.

School of Humanities
University of Adelaide
Adelaide SA 5005, Australia
E-mail: jorge.fernandez@adelaide.edu.au

ACKNOWLEGMENTS

Work funded by the Spanish government, as part of the research project "Self-knowledge, Expression and Transparency" (ref. FFI2012-38908-C02-02).