CARRUTHERS Y LA TRANSPARENCIA DE LA MENTE

MARTIN FRICKE

ABSTRACT. CARRUTHERS AND THE TRANSPARENCY OF MIND.
Self-knowledge presents a challenge for naturalistic theories of mind. Peter Carruthers's (2011) approach to this challenge is Rylean: He argues that we know our own propositional attitudes because we (unconsciously) interpret ourselves, just as we have to interpret others in order to know theirs'. An alternative approach, opposed by Carruthers, is to argue that we do have a special access to our own beliefs, but that this is a natural consequence of our reasoning capacity. This is the approach of transparency theories of self-knowledge, neatly encapsulated in Byrne's epistemic rule (BEL): If p, believe that you believe that p (Byrne 2005). In this paper, I examine an objection to Carruthers's theory in order to see whether it opens up space for a transparency theory of self-knowledge: Is it not the case that in order to interpret someone I have to have some direct access to what I believe (cf. Friedman and Petrashek 2009)?

KEY WORDS. Self-knowledge, privileged access, other minds, Peter Carruthers, Alex Byrne.

El autoconocimiento representa un reto para las teorías naturalistas de la mente. La razón es que el autoconocimiento parece ser especialmente seguro, aunque no parece mostrar aquellos rasgos que nos dan tal seguridad en el conocimiento de otras cosas. Descartes pensaba que cierto tipo de autoconocimiento era el conocimiento más seguro que se pueda tener y por ello trataba de fundar todo el resto del conocimiento en esta primera certeza. Hoy hemos algo perdido la fe en tal programa fundacionista. El naturalismo en la epistemología podría ser descrito como la idea de que el conocimiento científico, en especial el conocimiento de las ciencias naturales, es más certero, por lo menos en su totalidad, que cualquier fundación que los filósofos podrían proponer para él. El conocimiento científico no necesita tal fundación. Lo mejor que la filosofía puede hacer es tratar de

Instituto de Investigaciones Filosóficas y Centro Peninsular en Humanidades y Ciencias Sociales, Universidad Nacional Autónoma de México / martin_fricke@yahoo.com.uk

integrar las teorías que considera importantes en ese sistema de las ciencias naturales.

Se podría argumentar que eso es exactamente lo que la filosofía analítica trata de hacer cuando discute el autoconocimiento. El autoconocimiento que trata de las propias creencias, deseos y estados fenoménicos parece ser excepcionalmente seguro; tal vez no completamente infalible como Descartes pensaba, pero mucho más seguro que cualquier conocimiento ordinario o incluso científico del mundo. Al mismo tiempo, está claro que el autoconocimiento carece de los rasgos que caracterizan el conocimiento científico. No parece estar basado en la observación, la inferencia, experimentos, un gran cuerpo de conocimiento teórico, confirmación por pares, etc. Algunos creen que no está basado en nada y ciertamente parece ser mucho más directo que cualquier conocimiento científico. Si eso es así, entonces, en el contexto del naturalismo, el fenómeno del autoconocimiento requiere una explicación. Tal vez no le atribuyamos gran importancia a este conocimiento para el resto de nuestras teorías, tal como lo hizo Descartes, pero no hay duda de que se debe explicar cómo tal conocimiento es posible.

En este texto examinaré dos teorías del autoconocimiento que se acercan a este problema en maneras diferentes. La primera, propuesta por Peter Carruthers, desarrolla un tipo de teoría anteriormente defendido por Gilbert Ryle. Dice que adquirimos el conocimiento de nuestras propias actitudes proposicionales de la misma forma que adquirimos el conocimiento de las actitudes proposicionales de otras personas. Tenemos un módulo para leer la mente (*mindreading module*) que podemos aplicar tanto a otros como a nosotros mismos, y si lo aplicamos a nosotros mismos adquirimos un conocimiento de nuestras propias actitudes proposicionales. Se sigue que el autoconocimiento en realidad no es tan especial y diferente de otros conocimientos y sólo se nutre de una riqueza especial de datos, ya que estamos con nosotros mismos todo el día, juntando evidencia para posibles autoatribuciones, mientras que tenemos que trabajar con datos más limitados cuando se trata de otras personas.

La segunda teoría es la que propone Alex Byrne, la cual a su vez se inspira en una observación famosa de Gareth Evans, según la cual "contesto la pregunta de si creo que *p* poniendo en marcha el proceso (cualquiera que éste sea) mediante el cual respondo a la pregunta de si *p*" (Evans, 1982: 225). Las teorías que parten de esta nota han sido llamadas "teorías de la transparencia" (*transparency theories*) del autoconocimiento porque consideran la pregunta de si creo que *p* como "transparente" a la pregunta de si *p*. Byrne trata de resumir tal transparencia en la siguiente regla epistémica:

(CRE) Si *p*, cree que crees que *p* (Byrne, 2005: 95).

Al seguir esta regla, incluso sólo *tratando* de seguir esta regla, pero equivocándose con los hechos (es decir, no es cierto que *p*), se dan autoadscripciones verdaderas de creencias. La regla no depende de alguna percepción, sino sólo de capacidades muy básicas inferenciales: proceder de "*p*" a "creo que *p*". Byrne piensa que eso explica la seguridad especial del autoconocimiento. No hay mucho que puede salir mal. Además de explicar este "acceso privilegiado", se supone que CRE también explica nuestro "acceso peculiar" a las propias creencias. Si procedemos de "*p*" a "*él* cree que *p*", la probabilidad de llegar a una adscripción falsa de creencia es mucho más alta. Es evidente, entonces, que el tipo de acceso descrito por CRE es peculiar a la *propia* mente. No hay un tipo equivalente de acceso a la mente de otras personas. Resulta, pues, que en contraste con la teoría ryleana de Carruthers sobre el autoconocimiento, Byrne intenta mostrar que sí tenemos un acceso especial a la propia mente, un acceso no sólo especial por la cantidad de datos a partir de los cuales hacemos inferencias sobre nosotros mismos, sino también por el *método* a través del cual nos conocemos a nosotros mismos, un método que sólo es aplicable a nosotros mismos. Se podría decir que Byrne responde al reto que representa el naturalismo frente al fenómeno del autoconocimiento, no negando el carácter especial del autoconocimiento, sino mostrando que este carácter especial es una consecuencia de los poderes normales del razonamiento combinado con una simple regla epistémica.

Prefiero la teoría de Byrne. Cómo la de Carruthers no es compatible con ésta, examinaré, en lo que queda de este texto, la forma en que este autor justifica su teoría. Luego discutiré una posible objeción en contra de su propuesta y cómo sugiere responder a la objeción. El objetivo es ver si en esta discusión se abre un espacio para la teoría de Byrne.

La originalidad de la propuesta de Carruthers no es, por supuesto, su ryleanismo, sino la manera en que lo defiende con la ayuda de las ciencias cognitivas contemporáneas. Central para esta defensa es una teoría modular, según la cual la estructura de la mente conforma a una "arquitectura de transmisión global" (*global broadcast architecture*). La idea es que la mente consiste de diferentes sistemas especializados, organizados alrededor de un espacio común de trabajo. Los sistemas no pueden comunicar directamente el uno con el otro, sino sólo a través de mensajes que se transmiten globalmente en el espacio de trabajo, y llegan así a ser "conscientes por acceso" (*access conscious*). Este arreglo se asemeja a un grupo de especialistas en diferentes áreas de la ciencia, agrupados alrededor de un pizarrón, que sólo pueden comunicarse entre ellos escribiendo mensajes en el pizarrón. El aspecto crucial, en nuestro contexto, de este espacio común de trabajo es que sólo se pueden transmitir mensajes *sensoriales*, por ejemplo, estados perceptuales, imágenes o ejemplos de habla interna (*inner speech*). Nuestras decisiones, juicios, creencias, intenciones u otras

actitudes proposicionales no pueden ser transmitidos como tal. Primero deben ser expresados en estados sensoriales tales como imágenes o habla. Ahora bien, uno de los sistemas que constituyen la mente es una facultad para leer la mente (*mindreading faculty*) que se utiliza para atribuir estados mentales a otras personas (que incluyen se puede suponer, a animales no humanos). Según Carruthers, esta facultad para leer la mente de otras personas también tiene la tarea de proveernos con el conocimiento de los *propios* estados mentales. Para este fin tiene que usar la información que recibe a través del espacio común de trabajo de la mente. No tiene un acceso directo a las actitudes proposicionales del sujeto porque éstas no se transmiten globalmente, excepto después de ser transformados en datos sensoriales. Incluso cuando se presentan sensorialmente, los datos —por ejemplo, cuando oímos lo que alguien, tal vez nosotros mismos, dice —todavía deben ser *interpretados* por la facultad para leer la mente para determinar qué actitud proposicional se expresa en ellos.

Si todo eso es verdad, entonces la facultad para leer la mente no puede aplicar un procedimiento tal como la regla epistémica de Byrne (CRE). Para aplicar la regla "si *p*, cree que crees que *p*", primero tenemos que saber (o por lo menos pensar que sabemos) que *p*. En otras palabras, tenemos que tener acceso al contenido de la creencia de primer orden, es decir, a lo que se cree. Pero si Carruthers tiene razón, el sistema para leer la mente no tiene un acceso general a lo que el sujeto cree. Sólo tiene acceso a lo que se percibe o imagina de alguna manera y a lo que de tal forma se transmite globalmente en el espacio de trabajo. Así, si hubiera un procedimiento análogo a (CRE) para autoadscribir estados *perceptuales* o imágenes internas (*imagistic states*), éste sí podría ser aplicado por el sistema para leer la mente. De hecho, Carruthers cree que nuestro acceso a los propios estados perceptuales *es* transparente en este sentido y no depende de una interpretación (Carruthers, 2011: 72ss.).

Carruthers hace un considerable esfuerzo para presentar evidencia empírica y otras consideraciones a favor de su propuesta. Por ejemplo, argumenta que la arquitectura de la transmisión global es ideal para explicar la posibilidad de un desarrollo gradual de la mente en pasos incrementales, donde un sistema tras otro se agrega a través de una evolución natural. Este desarrollo también explica, por qué sucesos mentales no sensoriales, tales como juicios o decisiones, no pueden ser transmitidos globalmente —la arquitectura de la transmisión ya se había desarrollado antes de que tal rediseño de la arquitectura básica hubiera sido útil. Carruthers también discute ampliamente muchos casos en los cuales sujetos parecen confabular sus propias intenciones, deseos e, incluso, creencias. Por ejemplo, una persona hipnotizada que recibe una orden, frecuentemente, cuando se despierta cumple con la orden, digamos, de levantar algún libro de la mesa y ponerlo en el librero. Cuando se le

pregunta a la persona por qué hace eso, ella explica que no le gusta el desorden en la mesa y que decidió limpiarla —o algo semejante (cfr. Wegner, 2002). Carruthers interpreta tales casos como evidencia para la idea de que las autoatribuciones de actitudes proposicionales siempre se basan en autointerpretaciones no conscientes. Cuando confabulamos una intención que claramente no existió (¿ni existe?) nos interpretamos —¿de cuál otra forma podríamos llegar a la autoatribución? Como nos falta una información relevante (en este caso el dato de que estábamos hipnotizados), nuestra interpretación es errónea. Ya que no estamos conscientes de que sólo se trata de una interpretación, bien podría ser que *siempre* basamos nuestras autoatribuciones en interpretaciones, incluso cuando son verdaderas.

No tengo suficiente espacio para discutir estos argumentos aquí. Más bien quiero enfocarme en una objeción particular a la teoría de Carruthers sobre el conocimiento de otras mentes y en su respuesta a esta objeción, porque así se puede elucidar la relación de su propuesta con las teorías de la transparencia del autoconocimiento como la de Byrne. Como hemos visto, Carruthers dice que el sistema para leer la mente no tiene un acceso general a las creencias, intenciones, decisiones, etc. del sujeto. Más bien, como todos los demás sistemas, tiene que conformarse con la información que recibe a través de las transmisiones globales de los datos sensoriales (y una cantidad limitada de principios, datos, etc., específicamente necesarios para leer otras mentes). ¿Pero —esta es la objeción que varios comentaristas han hecho (Currie y Sterelny, 2000; Friedman y Petrashek, 2009; Lurz, 2009)— es posible interpretar la mente de otras personas sin tener un acceso general a las propias creencias? Parece que a menudo necesitamos información sobre el mundo que no está perceptualmente presente en el momento de la interpretación para atribuir estados mentales a otras personas. Las interpretamos no sólo en vista de lo que observamos ahora mismo, sino también en vista de lo que creemos sobre ellas y sobre el mundo en general.

He aquí un ejemplo de Friedman y Petrashek: "Louise es una especialista en la historia británica, y así *sabe* que la batalla de Hastings sucedió en 1066" (2009: 146). Atribuimos tal conocimiento (una actitud proposicional) a Louise porque creemos que la batalla de Hastings sucedió en 1066, que Louise es una especialista en la historia británica y que especialistas en la historia británica saben tales cosas. Leer la mente de Louise depende, en este caso, de un acceso a nuestras propias creencias. Es imaginable que los tres (supuestos) hechos en cuestión se presenten sensorialmente al sujeto. Por ejemplo, el sujeto podría *leer* sobre ellos, así como el lector de estas líneas lo está haciendo ahora. Sin embargo, aunque eso podría suceder, parece claro que tales datos sensoriales no son *necesarios* para atribuir el conocimiento a Louise. Parece que la atribución del conocimiento podría

proceder directamente con base en nuestras creencias, sin un intermediario sensorial. Si eso es así, parece que tenemos un contraejemplo a la teoría de Carruthers.

De hecho, podría ser un principio general de leer la mente de otras personas el que, *ceteris paribus*, primero atribuimos a otros las mismas creencias que tenemos nosotros mismos. Si acepto que *p* es verdad, entonces, sin razones para lo contrario, debería atribuir la creencia de que *p* a otras personas también. Resulta, entonces, que hay una regla similar a (CRE) para la atribución de creencias a otras personas:

(CRE-3) Si *p*, cree que Fred cree *p* (Byrne, 2005: 96).

Aunque esta regla no es, por supuesto, igual de útil como (CRE) para producir verdaderas adscripciones de creencias, es por lo menos un buen punto de partida para aquellos que quieren saber qué es lo que creen otras personas.

Ahora bien, si estos argumentos son correctos, entonces, al contrario de lo que se dijo antes, parece que nuestro sistema para leer otras mentes, o algún otro mecanismo, sí tiene un acceso no sensorial a las propias creencias, en el sentido de que tiene acceso a qué es lo que creemos. Eso significa que no debería ser necesario utilizar el método ryleano para autoadscribir creencias. Si el sistema puede atribuir la creencia de que la batalla de Hastings sucedió en 1066 a Louise, razonando de los (supuestos) hechos de que la batalla sucedió en este año y de que Louise es una especialista en la historia británica, entonces también debería ser capaz de aplicar una regla epistémica tal como (CRE). Debería ser capaz, en otras palabras, de razonar del (supuesto) hecho (es decir de la creencia) de que Louise es una especialista en la historia británica directamente a la creencia de que yo *creo* que Louise es una especialista en la historia británica. Todo lo que se necesita para tal razonamiento es una regla epistémica como (CRE).

La respuesta de Carruthers a esta objeción viene en tres partes. Primero, concede que reglas como (CRE) y (CRE-3) pueden ser utilizadas por nosotros, pero dice que no se aplican por el sistema para leer la mente y que resultan en autoatribuciones puramente verbales. Segundo, concede que el sistema para leer la mente puede tener acceso a todas la creencias del sujeto, pero sólo indirectamente a través del espacio global de trabajo y operando en una manera lenta y reflexiva del tipo "sistema 2". Tercero, mantiene que cuando el sistema opera de una manera automática o "en línea" ("sistema 1") sólo tiene acceso a información sensorial.

A partir de la segunda y tercera parte de su respuesta, está claro que Carruthers piensa que el sistema para leer otras mentes nunca utiliza reglas como (CRE) o (CRE-3). Mantiene que el sistema para leer otras mentes no tiene acceso directo alguno a las creencias propias del sujeto. Cuando se lee la mente de otra persona de una manera automática o "en línea" (la

tercera parte de la respuesta de Carruthers), el sistema principalmente interpreta información sensorial actual y no tiene acceso a las creencias almacenadas o a otras actitudes proposicionales del sujeto. Sin embargo, el sistema tiene otra manera de operar que es más lenta y reflexiva. En esta manera lenta de operar el sistema para leer otras mentes sí puede tener acceso a todas las actitudes proposicionales del sujeto. Puede tenerlo si pone consultas en el espacio común de la mente. "Todo el conjunto de sistemas consumidores luego se pone a trabajar, haciendo inferencias y razonando de la manera normal, accediendo a cualquiera de las creencias del sujeto a las que normalmente accedería. Los resultados se depositan nuevamente en el espacio global de trabajo. [...] Aquí el proceso completo, colectivamente, tiene acceso a todas las creencias del agente" (Carruthers, 2011: 238). En este modo reflexivo, el sistema para leer otras mentes y tiene acceso a todas las creencias del sujeto, pero sólo indirectamente a través del espacio global de trabajo. En tanto cualquier información que viene del espacio de trabajo es sensorial, todavía debe ser interpretada para proveer información sobre las actitudes proposicionales del sujeto.

La primera parte de la respuesta de Carruthers a la objeción es la más interesante en nuestro contexto. Carruthers concede que de hecho usamos reglas tales como (CRE) y (CRE-3) para atribuir creencias. Sólo que estas reglas no se implementan por el sistema para leer otras mentes, sino por los sistemas ejecutivos y para la producción de lenguaje; y el resultado no es un autoconocimiento verdadero (o un conocimiento verdadero de las creencias de otra persona). Más bien, Carruthers parece pensar que se trata de una atribución meramente verbal que podemos hacer en respuesta a una cuestión verbal:

> Si mi tarea es decir cuál ciudad —alguien cree— es la capital del Reino Unido, por ejemplo, inmediatamente responderé "Londres" sin saber algo más sobre esta persona. [...] los sistemas ejecutivos y para la producción de lenguaje cooperan (y en parte compiten) entre ellos, buscando en la memoria del atribuidor y dando el resultado en la forma de un reporte metarrepresentacional —"Creo que/Ella cree que P"— donde la forma del reporte puede ser copiada de la forma de la pregunta inicial (Carruthers, 2011: 237).

Esta explicación de cómo podemos llegar a hacer "un reporte metarrepresentacional" parece ser bastante similar a la teoría de la transparencia de Byrne. La diferencia crucial es que, según Carruthers, el reporte no expresa un autoconocimiento (en caso de tener la forma "creo que *p*"), ni un conocimiento de las creencias de otra persona (en caso de tener la forma "ella cree que *p*"). Más bien, el prefijo (en el caso de un reporte de primera persona) es "sólo una manera de hablar (*a mere manner of speech*) o una

forma de cortesía (para no parecer demasiado seguro o definitivo)" (Carruthers, 2011: 86).

Lo curioso en esta idea es que no parece hacer una distinción entre atribuciones de creencia a uno mismo y atribuciones a otras personas. ¿Cuál es la diferencia, en la mente del hablante, entre "creo que *p*" y "él cree que *p*", si el prefijo es sólo una manera de hablar? Parece que según Carruthers no hay ninguna. Más bien, tenemos que interpretar nuestros propios reportes verbales para saber sobre quién estamos hablando. Más extraño todavía, incluso cuando decimos "*él* cree que *p*," en realidad no expresamos una creencia sobre otra persona, sino sólo la creencia de que *p*.

Para concluir, ¿dónde deja esta discusión a las teorías de la transparencia del autoconocimiento? Si tomamos Carruthers en serio, hay un lugar para las autoadscripciones transparentes de creencia: se encuentra en nuestras respuestas verbales inmediatas y, presuntamente, no reflexivas a preguntas sobre nuestras creencias. Aquí directamente accedemos a nuestras creencias de primer orden y sólo verbalmente las prefijamos con "creo que [...]", así aplicando el procedimiento de Evans. Si Carruthers tiene razón, el resultado no es un autoconocimiento, sino sólo una manera de hablar. Es poco probable que eso satisfaga a los teóricos de la transparencia como Byrne. También tiene la consecuencia extraña de que parece que a menudo hablamos de creencias sin saber si son nuestras o de alguien más.

REFERENCIAS

Byrne, Alex (2005), "Introspection", *Philosophical Topics* 33: 79-104.
Carruthers, Peter (2011), *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.
Currie Gregory y Sterelny, Kim (2000), "How to think about the modularity of mind-reading", *Philosophical Quarterly* 50: 145-160.
Evans, Gareth (1982), *The Varieties of Reference*. Oxford: Clarendon.
Friedman, Ori y Petrashek, Adam R. (2009), "Non-interpretive metacognition for true beliefs", *Behavioral and Brain Sciences* 32: 146-147.
Lurz, Robert W. (2009), "Feigning introspective blindness for thought", *Behavioral and Brain Sciences* 32: 153-154.
Ryle, Gilbert (1949), *The Concept of Mind*. London: Hutchinson.
Wegner, Daniel M. (2002), *The Illusion of Conscious Will*. Boston: MIT Press.