http://doi.org/10.21555/top.v0i58.1088

Who's Thoughts Are These? Examining the *Top-Down* Approach to Attributions of Mental Agency

¿De quién son estos pensamientos? Examinando el modelo *top-down* de las atribuciones de agencia mental

Pablo López-Silva
Universidad de Valparaíso
Chile
pablo.lopez.silva@gmail.com
https://orcid.org/0000-0001-7457-7724

Recibido: 24 - 08 - 2018.
Aceptado: 23 -10 - 2018.

Abstract

An attribution of mental agency is defined as the act of assigning the initiation or authorship of a first personal phenomenal thought to a specific agent. This type of attribution is fundamental for the production of the belief that human beings are rational agents not only in behavioral terms, but also, in a cognitive manner. The top-down approach –one of the dominant theories in current literature– suggests that attributions of mental agency arise as retrospective rational explanations for the occurrence of phenomenal thoughts. Thus, the agency of a thought would not be contained in its most fundamental phenomenal structure, rather, it would be an imposed category. After introducing the most fundamental elements of the debate, this paper evaluates the top-down model in order to identify its argumentative strengths and weaknesses. It is concluded that, although this model possesses a number of merits, it cannot deal in a plausible way with some fundamental conceptual challenges.

*Key words***:** attributions of mental agency; rationality; cognitive phenomenology; thoughts.

Resumen

Las atribuciones de agencia mental se definen como el acto de asignar el inicio o autoría de un pensamiento fenoménico en primera persona a un agente específico. Este tipo de atribución es fundamental para la producción de la creencia de que los seres humanos somos agentes racionales no solamente en términos conductuales, sino también cognitivos. La teoría *top-down* –uno de los enfoques dominantes en la literatura actual– sugiere que las atribuciones de agencia mental emergen como explicaciones retrospectivas de carácter racional a la ocurrencia de pensamientos fenoménicos en primera persona. De esta forma, la agencia de un pensamiento no es parte de su contenido fenoménico inicial, sino que es una categoría impuesta por un análisis retrospectivo. Tras introducir algunos aspectos básicos sobre la discusión, el presente artículo ofrece una evaluación crítica del modelo *top-down* con el fin de identificar sus fortalezas y debilidades argumentativas. El artículo concluye que, si bien el modelo posee méritos importantes, éste no está en condiciones de responder a ciertos desafíos conceptuales fundamentales.

*Palabras clave***:** atribuciones de agencia; racionalidad; fenomenología cognitiva; pensamientos.

La vida sólo puede ser comprendida hacia atrás, pero únicamente puede ser vivida hacia delante

—Soren Kierkegaard

1. Fenomenalidad y formas de conciencia cognitiva[1]

Los *estados mentales conscientes* son caracterizados en la filosofía de la mente como aquellos en los cuáles existe 'un algo que es como' estar en ellos (Nagel, 1974; Jackson, 1982; Chalmers, 1996).[2] Si bien esta expresión es un tanto ambigua incluso en su forma original (*something that is like to be in those states*; Block, 1980), la idea detrás de ella es que los estados conscientes son aquellos que se experimentan –o se 'sienten'– de cierta forma específica; son estados que poseen propiedades cualitativas particulares (Zahavi, 2005). Por ejemplo, contrastemos la experiencia de beber una caipiriña con la experiencia de pensar cómo sería estar bebiéndola. En ambos casos, el objeto intencional de estos estados mentales es el mismo. Sin embargo, tales estados difieren en la actitud proposicional que los instancia y podemos notar, además, que ambos estados mentales 'se sienten' de forma distinta incluso teniendo el mismo objeto intencional (degustar P vs. pensar en degustar P). Diversos autores han denominado *fenomenalidad* a la propiedad que permite esta diferencia cualitativa; tal propiedad permitiría que un sujeto experimente de *cierta forma específica* un estado mental en su conciencia (cfr. Zahavi, 2005; Kriegel, 2015; Pitt, 2011).[3] Pues bien, nuestra conciencia de estados

[1] Agradecimientos: Quisiera agradecer a Joel Smith, Tim Bayne, Tom McClleland, Jöelle Proust y a los dos revisores ciegos provistos por la revista por sus comentarios provistos en versiones preliminares de este trabajo. Algunas ideas contenidas en este trabajo fueron presentadas en el Institut Jean Nicod (Francia), Universidad Alberto Hurtado (Chile), y la Universidad de Rikkyo (Japón). Financiamiento: La escritura final de este trabajo se realizó en el marco del Proyecto FONDECYT No 11160544 'La Arquitectura Agencial del Pensamiento Humano' otorgado por la Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) del Gobierno de Chile.

[2] Esta viene a ser una traducción del autor de la expresión *'something that is like'* célebremente propuesta por Nagel (1974).

[3] El término *fenomenalidad* se introduce con el fin de eliminar algunas ambigüedades asociadas al término 'fenomenología'. Se entiende, por

mentales fenoménicos –también denominada *conciencia fenoménica* (Nagel, 1974; Kriegel, 2015)– no posee como objetos intencionales solamente estados sensoriales, como aquellos derivados de los sentidos básicos (oler P, ver P, tocar P, escuchar, etc.); también incluye la conciencia de estados mentales cognitivos como los *pensamientos* (Strawson, 2003; Pitt, 2004; Montague, 2017). Durante los últimos años, diversos autores han sugerido que los pensamientos conscientes poseerían un tipo específico de *fenomenalidad* que los distinguiría de los estados de orden puramente sensorial (Strawson, 2003; Pitt, 2004; Bayne & Montague, 2011; Chudnoff, 2016; Montague, 2017, entre otros).[4]

Ahora bien, existen al menos tres formas en las cuales un sujeto puede devenir consciente de sus propios pensamientos. Cada una de estas modalidades de *conciencia cognitiva* desplegará una forma específica de relación entre sujeto y pensamiento que permitirá la integración del estado cognitivo nuevo a su vida mental general.[5] La primera modalidad implica devenir consciente de un pensamiento P de forma *subjetiva*, modalidad que emerge por el mero hecho de que P ha aparecido en el campo fenoménico privado del sujeto en cuestión. Esta forma de conciencia establece la relación más fundamental que se puede observar entre pensamiento y sujeto (Guillot, 2017; López-Silva, 2017a). En este caso, la mera disponibilidad fenoménica de un pensamiento es suficiente para reconocer una relación subjetiva entre éste y su poseedor –o *bearer*– (Gallagher, 2000, 2015; De Hann & De Bruin, 2010; Zahavi, 2005; Grünbaum & Zahavi, 2013). La segunda

consiguiente, que la fenomenalidad de una experiencia es la propiedad que permite que ésta despliegue una fenomenología específica que finalmente puede ser reportada mediante el uso del lenguaje natural.

[4] La definición de la naturaleza de esta fenomenalidad asociada a los estados puramente cognitivos es un debate abierto en la literatura actual (Bayne & Montague, 2011; Montague, 2017). Por una parte, algunos indican que existe una fenomenalidad exclusiva a los estados cognitivos que es distinta de las propiedades cualitativas asociadas al objeto intencional del estado cognitivo en cuestión (Strawson, 1994; Pitt, 2004; Montague, 2016). Por otra parte, algunos indican que, de existir tal fenomenalidad, ésta sólo se derivaría de las propiedades cualitativas asociadas con el objeto intencional del estado cognitivo específico (Carruthers, 2005).

[5] Utilizo la expresión *conciencia cognitiva* con el fin de diferenciar este tipo específico de conciencia de otras modalidades tales como la conciencia corporal, la conciencia propioceptiva, entre muchas otras.

modalidad implica devenir consciente de un pensamiento como algo que *es mío* o que *me pertenece* (De Hann & De Bruin, 2010; Zahavi, 2005; López-Silva, 2016; Grünbaum & Zahavi, 2013). Este tipo de relación no sólo reconoce que un pensamiento es dado en mi campo fenoménico subjetivo, sino que, además, es dado *como si fuese mío* (Gallagher, 2000, 2015). Algunos autores indican que lo que permite la diferencia entre la primera y la segunda modalidad de conciencia cognitiva es que aquellos pensamientos finalmente incluidos en la segunda modalidad poseerían una propiedad cualitativa extra denominada *sensación de propiedad* (*sense of mineness* o *sense of ownership*).[6] Actualmente existe un debate abierto respecto de la existencia y función de tal propiedad por lo que profundizar en el rol que desempeña ésta en la instanciación de un tipo específico de conciencia cognitiva sería apresurado y, además, nos alejaría del punto central del presente trabajo.[7] Finalmente, un sujeto puede devenir consciente de un pensamiento P como su autor o iniciador (Frith, 1992; Stephens & Graham, 1994; Gallagher, 2004, 2015, entre muchos otros). Diversos autores indican que este tipo de modalidad de conciencia cognitiva es permitida por la existencia de una *atribución de agencia mental*, la cual es definida como el acto de asignar la autoría o inicio voluntario de un pensamiento consciente en primera persona a un agente específico (Frith, 1992; Graham & Stephens, 1994; Campbell, 1999, 2002; Proust, 2009; Vosgerau & Voss, 2014). La idea es que este tipo de conciencia sería posible tras una asignación de autoría específica a un pensamiento. Ahora bien, no es obvio que tal atribución sea siempre autorreferente; diversos autores indican que reportes de delirios psicóticos permiten observar casos en los cuáles tal atribución puede ser externalizada (Mullins & Spence, 2003). Pacientes que sufren de este síntoma informan experiencias como las siguientes:

> I didn't hear these words as literal sounds, as through the houses were talking and I were hearing them; instead, the words just came into my head–they were

[6] El término 'sensación de propiedad' hace referencia a una cualidad fenoménica de ciertos pensamientos que permite que sean experimentados *como si fuesen míos*, por lo que su presencia establece una distinción fenomenológica fundamental en la forma de auto-reportar los pensamientos propios. Para una mayor discusión sobre este concepto en español, cfr. López-Silva (2014, 2017b).

[7] Para una buena discusión sobre este asunto, cfr. McClelland (2017).

> ideas I was having. Yet I instinctively knew they were not my ideas. They belonged to the houses, and the houses had put them in my head (Saks, 2007, p. 29).
>
> Thoughts are put into my mind like 'Kill God'. It's just like my mind working, but it isn't. They come from this chap, Chris. They are his thoughts (Frith 1992, p. 66).

La forma general que poseen los reportes de delirios de inserción de pensamiento muestra sujetos conscientes de pensamientos en primera persona *como si fuesen* insertados en su mente por agentes externos. En el primer ejemplo, las ideas reportadas por el sujeto son referidas en primera persona e insertadas por las casas aledañas. En el segundo ejemplo, el paciente es consciente en primera persona de la idea 'Kill God' como si fuese insertada en su cabeza por Chris. Diversos autores han interpretado este fenómeno como el producto de una alteración en la forma en que los sujetos afectados realizan el acto de asignar agencia. La idea es que diversas alteraciones experienciales y cognitivas estimularían una externalización de la agencia de pensamientos con ciertas características específicas, lo que generaría la experiencia consciente de inserción. Este enfoque ha incluso llegado a ser denominado el *enfoque estándar para los delirios de inserción de pensamiento* (Stephens & Graham, 2000; Gallagher, 2000, 2015; Vosgerau & Voss, 2014). Ahora bien, el fenómeno atribucional no-patológico –i.e. una auto-atribución de agencia mental– implicaría que el inicio de un pensamiento consciente al cual posea acceso en primera persona es atribuido a uno mismo. Sin embargo, a pesar de que ambos tipos de atribución mental puedan ser diferenciadas, existe un debate abierto respecto de cómo explicar su producción, condiciones de posibilidad y naturaleza. Por ello, el debate se extiende a la forma en que tales modelos explicativos podrían ser ocupados para entender los casos no convencionales que podrían dar origen a casos como los delirios antes mencionados. Las siguientes secciones de este artículo examinan los méritos argumentativos de una de las teorías dominantes que intenta explicar la forma en que los seres humanos producen sus atribuciones de agencia mental en la literatura actual.

2. Atribuciones de agencia mental: el modelo *top-down*

Como asunto preliminar a nuestro objetivo principal, es importarte distinguir el debate sobre las normas epistémicas y prácticas que gobiernan la agencia racional del debate acerca de las normas y mecanismos que gobiernan el proceso de adscripción de agencia mental. Mientras el primer debate refiere a la discusión sobre la ontología y epistemología de la agencia racional, el segundo refiere a la forma en que un estado mental específico –*doxástico* si se desea– es producido por la mente humana y, por lo tanto, la forma en que la *sensación consciente de agencia racional* figuraría en la fenomenología cognitiva. Si bien ambas discusiones son distintas, éstas se mantienen profundamente enlazadas al ser posible relacionar una sensación específica de agencia con una forma de conocer nuestros propios pensamientos (cfr. O'Brien & Soteriou, 2009). Es más, el acto de atribuir el inicio o autoría de un pensamiento a un agente específico –para algunos, un acto de agencia racional– posibilitaría la representación de un tipo de *relación causal* entre agente y pensamiento por lo que este acto informaría la creencia de que los seres humanos poseen agencia racional y la creencia de que poseemos algún grado de control por sobre nuestra actividad cognitiva.

Durante los últimos años, el enfoque *top-down* se ha vuelto una de las alternativas más populares dentro del debate que intenta clarificar las reglas que gobiernan el proceso de producción de atribución de agencia mental (Graham & Stephens, 1994; Campbell, 1999; Stephens & Graham, 2000; Vosgerau & Voss, 2014, entre otros). La expresión *top-down* hace referencia a la dirección causal en la forma en que los diversos elementos implicados en el proceso atribucional se organizan según su naturaleza y su jerarquía en el proceso de producción del estado atribucional final. Por un lado, los elementos *top* pertenecen al dominio cognitivo; por otra parte, los elementos *down* refieren a los elementos fenoménicos derivados del tal procesamiento informacional (experiencia). Así, dentro de una teoría de tipo *top-down* la relación causal de producción de un estado mental va desde *arriba* (diversos elementos relacionados con el procesamiento de información de contenido y contextual de un pensamiento) hacia *abajo* (estado fenoménico final, sensaciones, etc.). Finalmente, la intuición tras este modelo es que una atribución de

agencia mental es un acto racional de la mente humana cuando ésta posee las razones suficientes para su ejecución (Vosgerau & Voss, 2014).

La teoría *top-down* sugiere que las atribuciones de agencial mental surgen como una explicación retrospectiva de carácter racional a la aparición de ciertos pensamientos en el flujo de la conciencia de un sujeto específico. Por lo tanto, la naturaleza del estado mental final producido por el acto atribucional sería fundamentalmente cognitiva e inferencial, ya que solamente surgiría luego de analizar introspectivamente ciertos pensamientos en términos causales (Stephens & Graham, 2000). Tal como se podría deducir, el modelo *top-down* también ha sido llamado 'explicacionista' dado que el contenido final del estado atribucional surge como una explicación causal de la ocurrencia de un pensamiento de primer orden. Ahora bien, *¿qué reglas gobiernan este tipo de explicación agencial?* El modelo sugiere que un sujeto primero tiene acceso al contenido de un pensamiento y, luego de acceder a este estado mental, el sujeto realiza una inferencia causal basada en estados doxásticos previos (creencias sobre el mundo y la realidad), conocimientos previos, ideas culturalmente heredadas, expectativas, potenciales predicciones, análisis retrospectivos de información fenoménica disponible y el contexto de la tarea en la cual el pensamiento en cuestión ha surgido. Luego de ponderar toda esta información a la luz del contenido del pensamiento en cuestión, el sujeto le da sentido a la ocurrencia de este pensamiento en términos causales, ya que este estado parece ser coherente con los estados intencionales subyacentes a la vida mental del sujeto. Sobre esto, los principales defensores del modelo claramente indican que: *[W]hether I take myself to be the agent of a mental episode depends upon whether I take the occurrence of this episode to be explicable in terms of my underlying intentional states* (Graham & Stephens, 1994, p.93).

Un elemento fundamental dentro del modelo *top-down* es el establecimiento de una relación específica entre la psicología del sujeto y el tipo de explicación ofrecida a los pensamientos de los cuales éste deviene consciente, esto es, una relación entre la psicología del sujeto y la forma final del estado atribucional agencial. Los defensores de la teoría *top-down* indican que un sujeto aceptará que es el autor de cierto pensamiento específico (lo que finalmente informará el contenido de la atribución de agencia mental) sólo si el contenido de tal pensamiento es consistente con el tipo de contenido que el sujeto podría producir

dada la imagen y expectativas que posee de sí mismo (cfr. Graham & Stephens, 1994, p. 103).[8] Tal como señalan Graham & Stephens:

> [T]he subject's sense of agency regarding her thoughts likewise depends on her belief that these mental episodes are expressions of her intentional states. That is, whether the subject regards an episode of thinking occurring in her psychological history as something she does, as her mental action, depends on whether she finds its occurrence explicable in terms of her theory or story of her own underlying intentional states (Stephens & Graham, 2000, p. 162).[9]

Será fundamental para el análisis en la siguiente sección notar la existencia de dos reglas principales que determinan la racionalidad de la producción de una atribución de agencia mental: *consistencia* y *coherencia*. Los autores son claros en indicar que un pensamiento será auto-atribuido en términos agenciales si el contenido de éste es *coherente* con la autoimagen y expectativas que el sujeto posee de sí mismo y *consistente* con la información del contexto de aparición del pensamiento y los diversos estados intencionales subyacentes del sujeto. Graham & Stephens (1994) indican que una narrativa de carácter explicativo-causal servirá para darle sentido retrospectivamente a diversas conductas del sujeto y proveerá un marco de referencia para poder explicar la aparición de futuros pensamientos y conductas en la vida del sujeto.

Tal como podemos ver, la dirección causal de la producción de una atribución de agencia mental dentro del modelo *top-down* va desde el conjunto de conocimiento previo y auto-imagen (*top*-arriba) del sujeto, hacia la evaluación de un pensamiento con características fenoménicas específicas pero pasivas –desde un punto fenomenológico– a la luz de la

[8] Con el fin de no dejar duda sobre lo que es realmente dicho por los autores, a continuación se puede apreciar la cita original: 'the subject unproblematically accepts a thought as her action if, by her own lights, it accords with her intentional psychology–if roughly, it is the sort of thought [content] she would expect herself to think given her picture of herself' (Graham & Stephens, 1994, p. 103).

[9] Es importante señalar que los autores entienden el término 'sense of agency' como el producto de una atribución de agencial mental de carácter cognitivo, inferencial y retrospectivo (cfr. Gallagher, 2007).

pregunta por su agencia (*down*-abajo). Acá, el agente de un pensamiento surgiría como una operación de segundo orden de carácter evaluativa, inferencial y retrospectiva en tanto imposición de la categoría de 'agencia' a los pensamientos a los cuáles se tiene acceso en primera persona basado en cierta información privilegiada respecto de su ocurrencia. Ahora bien, una de las principales diferencias de este modelo con sus rivales (como por ejemplo, el modelo *bottom-up* de autores como Gallagher, 2000, 2007, 2012; Zahavi, 2005; Gallagher & Zahavi, 2008) es que, dentro del enfoque *top-down*, los pensamiento poseerían una fenomenología *pasiva* o *receptiva* respecto del elemento agencial i.e., los pensamientos no poseerían ninguna cualidad fenoménica agencial especial en sí mismos que propiciasen una adscripción de agencia. Dentro de esta teoría, la agencia no sería algo que pertenece experiencialmente al pensamiento, sino que es una forma en la cual se le da sentido a su ocurrencia experiencialmente pasiva. Ahondemos en este asunto en la siguiente sección.

3. Examinando el modelo *top-down*

Una de las fortalezas más evidentes del modelo *top-down* es que toda su estructura argumentativa parece ser consistente con una visión ampliamente aceptada dentro del debate sobre los componentes fenoménicos presentes en la experiencia del pensar. Si bien la exploración de la presencia de un componente fenoménico agencial en la fenomenología más fundamental de los pensamientos es una empresa esencial para la elaboración de una teoría de las atribuciones de agencia mental, este debate ha sido sistemáticamente ignorado en la literatura actual.[10] Con todo, algunos autores, como Strawson (2003), insisten en que la experiencia del pensar no posee un carácter activo homologable a la experiencia de una acción corporal, por ejemplo. Para entender este punto es necesario tener presente que la discusión sobre cómo explicar las atribuciones de agencia mental se origina en la expansión del debate acerca de las atribuciones de agencia motora al dominio de lo cognitivo (Proust, 2009). Es más, esto será la base de muchos

[10] Por ejemplo, ninguna de las publicaciones más recientes en el debate específico acerca de los diversos elementos fenoménicos relacionados al pensamiento contiene una exploración sistemática de este asunto. Cfr. Bayne & Montague (2011), Chudnoff (2017) y Breyer & Gutland (2015).

problemas asociados al debate general dentro del plano cognitivo. Dentro del dominio motor, la discusión se centra en la forma en que un sujeto termine auto-adscribiendo ciertos movimientos corporales, lo que permitiría la distinción fenoménica entre movimientos voluntarios e involuntarios (Frith, 2002). Pues bien, el enfoque más aceptado dentro de este contexto indica que la auto-atribución de agencia motora es propiciada por un elemento fenoménico específico que figura en la fenomenología más fundamental de los movimientos que después podrían ser categorizados como voluntarios, i.e. una sensación de agencia motora (Gallagher, 2000, 2007; Frith, 1992; Bayne & Pacherie, 2007; Bayne, 2011). Así, la sensación de ser el iniciador de cierto movimiento corporal específico estaría alojada en la misma estructura fenoménica de tal movimiento (Bayne & Pacherie, 2007). Pues bien, la misma idea no parece ser compartida cuando el debate se expande al dominio de lo cognitivo (Proust, 2009). Es más, la idea dominante sería que la experiencia de pensar no poseería ninguna *sensación de agencia* previa a ser explicada causalmente, por lo que cualquier experiencia de agencia sería posterior (Graham & Stephens, 2000). La experiencia de ser el agente de un pensamiento no figuraría en la fenomenología de un pensamiento. Sobre esto, De Hann & De Bruin (2010) han sugerido que, desde un punto de vista fenomenológico, el pensar implica una especie de filtro pasivo que permite dirigir el foco de atención de un pensamiento a otro, y que tales pensamientos no poseen ningún tipo de sensación de agencia cuando aparecen en el flujo de la conciencia de un individuo. Para usar una metáfora, los pensamientos aparecen en el flujo de la conciencia como los dientes de león se reparten en el aire; así, en el acto de seguir un tren del pensamiento específico, ponemos atención a unos pensamientos por sobre otros al igual que cuando observamos la trayectoria de algunos dientes de león por sobre otros en el aire. Ahora bien, el asunto relevante aquí es que, tal como observamos en la sección 2, el modelo *top-down* es capaz de integrar estas ideas al asumir que la fenomenología fundamental de los pensamientos se caracteriza por ser *pasiva* y posteriormente explicada en términos causales con base en diversos elementos contextuales y representacionales. A su vez, ésta será una de las principales ventajas del modelo *top-down* por sobre su mayor rival –la teoría *bottom-up*–, la cual, estableciendo un paralelismo entre el caso motor y el cognitivo, asume que las atribuciones de agencia mental surgen como la mera adopción del contenido representacional

de un pensamiento el cual contiene una sensación de agencia integrada en su estructura fenoménica (Gallagher, 2007, 2015).

Una segunda fortaleza del modelo *top-down* es que logra enlazar la psicología específica de un sujeto con el tipo de estado mental atribucional que éste es capaz de producir. Esto permite al modelo integrar la idea de cómo ciertas creencias culturales, expectativas y otros diversos elementos psicológicos podrían penetran el contenido fenoménico mismo de nuestros estados mentales conscientes y los modos de producción de éstos, lo cual añade una riqueza considerable a la propuesta. Con esto, el modelo *top-down* lograría ofrecer una respuesta a la pregunta acerca de cómo la agencia individual llega a figurar en nuestros estados mentales. Una constante queja dentro del mundo de la filosofía de la mente es que la mayoría de los modelos que explican diversos elementos de la mente humana parecen carecer de conexiones claras entre la historia psicológica de un sujeto y los componentes fenoménicos que éste logra o puede instanciar. En esto, la teoría *top-down* claramente propone un avance explicativo importante.

Ahora bien, a pesar de sus fortalezas, el modelo *top-down* enfrenta una serie de dificultades que parecen disminuir su poder explicativo. Un problema inicial nace de las dos reglas que guían la racionalidad tras la adscripción de agencia mental dentro del modelo. En éste se sugiere que un pensamiento P será auto-atribuido en términos agenciales por un sujeto S si el contenido de P es *coherente* con la autoimagen y expectativas que S posee de sí mismo y *consistente* con la información del contexto de aparición del pensamiento y los diversos estados intencionales subyacentes de S. Primero, el modelo no indica las condiciones mínimas en que la información subyacente determina la adscripción de agencia. No se indica la calidad o cantidad de información que debe tener para que pueda realmente informar una atribución de agencia. Segundo, en relación con lo anterior, no se clarifican los límites que el análisis retrospectivo debe tener en el proceso de rastreo informacional como para obtener *información suficientemente necesaria* para realizar una atribución de agencia mental bien informada. Sin estas dos condiciones, un análisis retrospectivo podría conducir al infinito. Ahora bien, es necesario notar que éstas no parecen ser críticas fundamentales, ya que una simple precisión del modelo podría solucionarlas. Sin embargo, existe un problema mucho más fundamental dentro de la propuesta.

Las reglas de *consistencia* y *coherencia* simplemente no parecen determinar la racionalidad de una atribución de agencia mental, y

por racionalidad aquí quiero decir específicamente las razones que un sujeto específico posee para realizar un acto mental de esa naturaleza. Simplemente no es cierto que los humanos terminan atribuyendo agencialmente solamente los pensamientos que son coherentes con su autoimagen y expectativas y consistentes con la información del contexto de aparición del pensamiento en cuestión. Muchas veces, los humanos auto-atribuyen de forma agencial pensamientos que son altamente inconsistentes con sus expectativas y auto-imagen e inconsistentes con el contexto de cierta tarea cognitiva específica que se encuentran realizando.[11] Sobre la primera condición (coherencia), podemos observar que muchas veces los humanos auto-atribuyen agencialmente pensamientos que son altamente conflictivos dado que su contenido es inconsistente con su auto-imagen y expectativas. Pensemos en el caso de los pensamientos moralmente reprochables. Este tipo de pensamiento se caracteriza por poseer un carácter conflictivo para el sujeto que lo posee que nace exactamente de la incoherencia entre el contenido del pensamiento y lo que el sujeto piensa de sí mismo. Es más, si no existiese auto-atribución de agencia, tal conflicto sería ininteligible. Es la auto-atribución de agencia la que posibilita el conflicto en la mente del sujeto al integrar a su vida mental un pensamiento que es incoherente con su auto-imagen. En este caso, un sujeto piensa un contenido reprochable y tal contenido no es consistente con la auto-imagen del sujeto; a pesar de esto, éste termina auto-atribuyendo el pensamiento en cuestión. El enfoque *top-down* no parece poder explicar este tipo de casos que son característicos de la vida mental humana. Existe otro caso que podría fortalecer este punto. Pensemos en los pensamientos obsesivos que acompañan un cuadro de *Transtorno Obsesivo Compulso.* Este tipo de

[11] Tal como lo ha señalado uno de los revisores, sería interesante analizar la forma en que la pertenencia a ciertos grupos con creencias culturales específicas podría influencia o determinar la racionalidad del acto agencial. Por ejemplo, el trabajo de autocategorización de Tajfel & Turner (1979), junto con el de Siegel (2016) podrían explicar cómo tales creencias penetrarían no sólo las formas de interpretar información cognitiva sino también los procesos que producen ciertos estados mentales. Intuitivamente, el concepto de agencia mental es una noción individualista en nuestra cultura occidental. Sin embargo, en otras culturas la agencia (subjetiva) no tiene por qué ser necesariamente individual, y podría ser compartida subjetivamente. Si bien ésta es una excelente línea de investigación, es algo en lo que no puedo profundizar en este escrito por razones de extensión y alcance.

pensamientos está caracterizado por tener contenidos egodistónicos, displacenteros y hasta agresivos. Tal como lo reportan los pacientes, muchas veces tales pensamientos están abiertamente en contra de su auto-imagen (Julien *et al*., 2007). Sin embargo, tales pensamientos no son externalizados y permanecen auto-atribuidos de forma agencia (lo que hace el conflicto que reportan los sujetos posible). Como podemos observar, el enfoque *top-down* no parecería ser capaz de explicar este tipo de casos tampoco.

Ahora, sobre la segunda condición (consistencia), es necesario recurrir a otro tipo de pensamiento que caracteriza la vida mental normal. El modelo *top-down* no parece estar tampoco en buen pie para explicar la auto-atribución de agencia mental que caracteriza a los pensamientos repentinos o *unbidden thoughts*. Según Frankfurt (1976), este tipo de pensamientos se caracteriza por aparecer en nuestra mente de forma súbita, sorpresiva y descontextualizada. Las características fenoménicas de este tipo de pensamiento son explicadas por la incapacidad del sujeto de rastrear las causas específicas de su contenido en el contexto de su aparición (Gallagher, 2015). Sin embargo, a pesar de tal incapacidad, este tipo de pensamiento termina siendo auto-atribuido de forma agencial, lo cual parece ir en contra de lo propuesto por el enfoque *top-down.* Simplemente no es verdad que un sujeto solamente auto-atribuye agencialmente aquellos pensamientos de los cuáles logra rastrear sus causas de aparición. Este problema fundamental dentro del modelo *top-down* parece surgir al imponer un requerimiento demasiado exigente al acto atribucional. Tal como Martin & Pacherie (2013) indican, es normal que los humanos posean dificultades para rastrear la información contextual-causal de un pensamiento. Es más, muchas veces nosotros no estamos en posición de realizar esto de forma clara. Sin embargo, el modelo *top-down* no logra explica cómo, a pesar de esta inhabilidad común, los seres humanos términos auto-atribuyendo tales pensamientos en términos agenciales.[12] El caso de los pensamientos obsesivos en TOC parece reafirmar este punto. Además, de poseer un contenido egodistónico, los pacientes que reportan este tipo de

[12] Aquí una potencial respuesta de los defensores del modelo *top-down* sería indicar que son las dos condiciones (coherencia y consistencia) conjuntamente las que posibilitan la atribución de agencia. Sin embargo, esta respuesta no ayudaría en nada a la defensa de la posición porque los casos anteriores tampoco podrían ser explicados con base en tal idea.

pensamientos no parecen ser capaces de rastrear la producción de éstos basados en información contextual. Sin embargo, a pesar de esto, tales pensamientos siguen siendo auto-atribuidos de forma agencial.

Una dificultad final para el modelo surge al intentar aplicar su estructura explicativa a los casos de alteraciones de agencia mental que caracterizan fenómenos como los delirios de inserción de pensamiento (ver sección 1). Es común dentro de la filosofía de la mente recurrir a este tipo de casos para evaluar el poder heurístico que una teoría explicativa posee. De esta forma, una teoría que logre dar sentido a las alteraciones del fenómeno que intenta explicar podría ser elegida por sobre sus rivales (Gallagher, 2013). Ahora bien, incluso si lograse explicar las atribuciones de agencia mental en su instanciación normal (algo que no parece ocurrir), el modelo no logra explicar las alteraciones de este proceso. Pensemos en el caso específico de los delirios de inserción de pensamiento. Los defensores de este modelo indican que pacientes que sufren de este síntoma simplemente no son capaces de rastrear las causas de ciertos pensamientos, lo que los llevaría a concluir que si tales pensamientos no son de ellos (no ha sido creado por ellos), necesariamente deben ser de alguien más (debe haber sido creado por alguien más) (Martin & Pacherie, 2013). Finalmente, esto precipitaría la externalización del contenido del pensamiento en cuestión (Synofzik, Vosgerau & Newen, 2008; Martin & Pacherie, 2013; Vosgerau & Voss, 2014). El problema principal con esta explicación es que la inhabilidad de un sujeto S para rastrear las causas de un pensamiento P no explica necesariamente el tipo de atribución externa que caracteriza los delirios de inserción de pensamiento; no explica el elemento más importante del síntoma. Esta inhabilidad sólo explicaría el hecho de que S experimenta P como sorpresivo o 'descontextualizado', lo cual no es diferente de la forma en que experimentados varios de los pensamientos que comúnmente tenemos día a día, por ejemplo, los pensamientos repentinos que comentábamos anteriormente. Muchas veces uno mismo no logra rastrear las causas específicas de un pensamiento P que aparece repentinamente en nuestro flujo de la conciencia, sin embargo, uno no termina externalizando tales pensamientos. De esta forma, la explicación que el modelo *top-down* ofrece para los delirios de inserción de pensamiento se transforma en la misma explicación que ofrece para los pensamientos repentinos, lo que finalmente implica que el modelo realmente no logra explicar las diferencias fenomenológicas y estructurales que diferencian ambos fenómenos. Ahora bien, el modelo

top-down podría indicar que la inhabilidad de rastrear información causal de un pensamiento actuaría en conjunción con otros factores para generar el tipo de fenómeno que observamos en los delirios de inserción de pensamiento. Sin embargo, este tipo de propuesta no ha sido explorada actualmente por los defensores del modelo, por lo que en su estado actual, éste no parece estar en condiciones de explicar las alteraciones que son posible de observa en el proceso de atribución de agencia mental.

4. Conclusiones

Este trabajo ha intentado examinar las fortalezas y debilidades del modelo *top-down* de las atribuciones de agencia mental. Contrastando con sus rivales, este modelo ofrece una alternativa que parece ser coherente con la fenomenología más fundamental de nuestros pensamientos a la luz de la pregunta por su carácter agencial. A su vez, el modelo logra establecer plausiblemente una relación entre la psicología del sujeto y el tipo de estado mental que éste es capaz de generar en medio de tareas cognitivas específicas. Sin embargo, a pesar de sus fortalezas, el modelo posee importantes debilidades. La más fundamental tiene que ver con los dos requerimientos que definirían la racionalidad del proceso de atribución mental (consistencia y coherencia). El primer problema con tales requerimientos tiene que ver con que no han sido bien definidos realmente y, por lo tanto, no es claro cómo deberían ser analizados ni evaluados. Nuestro análisis ha intentado realizar esta tarea a la luz de casos paradigmáticos de pensamientos, lo que nos lleva al segundo problema. Muchos de los pensamientos que terminan siendo auto-atribuidos de forma agencial no parecen cumplir tales requerimientos; por lo tanto, su valor argumentativo ha de quedar más que en duda. Finalmente, el modelo tampoco parece estar en buen pie para explicar las potenciales alteraciones de los procesos de atribución de agencia mental que llevarían a la producción de delirios cognitivos tales como la inserción y robo de pensamiento. Todo esto no significa que el modelo deba ser completamente desechado, sino que las nuevas alternativas de tipo *top-down* deberían tener en cuenta las debilidades señaladas en este trabajo para así mejorar el modelo y sacar provecho de sus fortalezas. Con todo, el debate sobre la racionalidad de las atribuciones de agencia mental parece mantenerse abierto y a la espera de futuros desarrollos.

Referencias

Bayne, T. (2011). The Sense of Agency. En F. Macpherson (ed.), *The senses.* (pp. 355–374). Oxford: Oxford University Press.

Bayne, T. & Montague, M. (2011). *Cognitive Phenomenology.* Oxford: Oxford University Press.

Bayne, T. & Pacherie, E. (2007). Narrators and Comparators: The Architecture of Agentive Self-Awareness. *Synthese*, 159, 475–491.

Breyer T. & Gutland, C. (2015). *The Phenomenology of Thinking.* Londres: Routledge.

Block, N. (1980). Troubles with Functionalism. En N. Block (ed.), *Readings in the Philosophy of Psychology.* (pp. 268–305). Cambridge: Harvard University Press.

Campbell, J. (1999). Schizophrenia, The Space of Reasons, and Thinking as a Motor Process. *The Monist*, 82(4), 609– 625.

___ (2002). The Ownership of Thoughts. *Philosophy, Psychiatry & Psychology*, 9(1), 35–39.

Carruthers, P. (2005). Conscious Experience Versus Conscious Thought. En P. Carruthers (ed.), *Consciousness: Essays from a Higher-Order Perspective.* (pp. 134-151). Oxford: OUP.

Chalmers, D. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

Chudnoff, E. (2016). *Cognitive Phenomenology*. Londres: Routlegde.

De Haan, S. & De Bruin, L. (2010). Reconstructing the Minimal Self, or How to Make Sense of Agency and Ownership. *Phenomenology and the Cognitive Sciences*, 9, 373–396.

Frankfurt, H. (1976). Identification and Externality. En A. O. Rorty (ed.), *The identities of persons.* (pp. 239–251). Berkeley: University of California Press.

Frith, C. (1992). *The Cognitive Neuropsychology of Schizophrenia.* Hillsdale: Lawrence Erlbaum.

Gallagher, S. (2000). Philosophical Conceptions of the Self: Implications for Cognitive Science. *Trends in Cognitive Sciences*, 4(1), 14–21.

___ (2007). The Natural Philosophy of Agency. *Philosophy Compass,* 2(2), 347-357.

___ (2013). *Phenomenology*. Londres: Palgrave McMillan.

___ (2015). Relations Between Agency and Ownership in the Case of Schizophrenic Thought Insertion and Delusions of Control. T*he Review of Philosophy and Psychology*, 6 (4), 865-879.

Gallagher, S. & Zahavi, D. (2008). *The Phenomenological Mind*. Londres: Routledge.

Graham, G. & Stephens, G. L. (1994). Mind and Mine. En G. Graham & G. L. Stephens (eds.), *Philosophical psychopathology.* (pp. 91-109). Cambridge: MIT Press.

Grünbaum, T. & Zahavi, D. (2013). Varieties of Self-Awareness. In K. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini & T. Thornton (eds.), *The Oxford Handbook of Philosophy and Psychiatry.* (pp. 221-239). Oxford: OUP.

Guillot, M. (2016). I Me Mine: on a Confusion Concerning the Subjective Character of Experience. *The Review of Philosophy and Psychology* [Online First]: DOI 10.1007/s13164-016-0313-4

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly*, 32, 127–136.

Julien, D., O'Connor, K. & Aardema, F. (2007). Intrusive Thoughts, Obsessions, and Appraisals in Obsessive-Compulsive Disorder: A Critical Review. *Clinical Psychology Review*, 27(3), 366-383.

Kriegel, U. (2015). *The Varieties of Consciousness*. New York: OUP.

López-Silva, P. (2014). Sensación de propiedad de la experiencia consciente y trastornos mentales: clarificaciones en torno al examen de anomalías subjetivas de la experiencia (EASE). *Gaceta de Psiquiatría Universitaria,* 10(3), 285-286.

___ (2016). Schizophrenia and the Place of Egodystonic States in the Aetiology of Thought Insertion. *The Review of Philosophy & Psychology*, 7(3), 577-594.

___ (2017a). Me and I are Not Friends, Just Acquaintances: On Thought Insertion and Self-Awareness. *The Review of Philosophy & Psychology.* [Online first]: DOI https://doi.org/10.1007/s13164-017-0366-z

___ (2017b). Conciencia fenoménica y mismidad. *Gaceta de Psiquiatría Universitaria,* 13(1), 17-20.

Martin, J. M. & Pacherie, E. (2013). Out of Nowhere: Thought Insertion, Ownership and Context-Integration. *Consciousness and Cognition*, 22(1), 111-12.

McClelland, T. (2017). Four Impediments to the Case for Mineness. En M. Guillot & M. García-Carpintero (eds.), *The Sense of Mineness*. Oxford University Press. https://www.academia.edu/35840865/Four_Impediments_to_the_Case_for_Mineness

Montague, M. (2017). *The Given*. Oxford: OUP.

Mullins, S. & Spence, S. (2003). Re-Examining Thought Insertion. *British Journal of Psychiatry*, 182, 293-329.

Nagel, T. (1974). What Is It Like to Be a Bat? *Philosophical Review*, LXXXIII, 435-50.

O'Brien, L. & Soteriou, M. (2009). *Mental Actions*. UK: OUP.

Pitt, D. (2011). Introspection, Phenomenality, and the Availability of Intentional Content. En T. Bayne & M. Montague (eds.), *Cognitive Phenomenology.* (pp. 141-173). Oxford: OUP.

Proust, J. (2009). Is there a Sense of Agency for Thoughts? En L. O'Brien & M. Soteriou (eds.), *Mental Actions.* (pp. 253-279). UK: OUP.

Saks, E. R. (2007). *The Centre Cannot Hold. My Journey Through Madness.* Nueva York: Hyperion.

Siegel, S. (2016). *The Rationality of Perception*. Oxford: OUP

Stephens, G. L. & Graham, G. (2000). *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts.* Cambridge: MIT Press.

Strawson, G. (2003). Mental Ballistics or the Involuntariness of Spontaneity. *Proceedings of the Aristotelian Society*, 103, 227-256.

Synofzik, M., Vosgerau, G. & Newen, A. (2008). Beyond the Comparator Model: A Multifactorial Two-step Account of Agency. *Consciousness and Cognition*, 17, 219 - 39.

Tajfel, H. & Turner, J. C. (1979). An Integrative Theory of Intergroup Conflict. En W. G. Austin & S. Worchel (eds.), *The Social Psychology of Intergroup Relations.* (pp. 33-47). Monterrey: Brooks-Cole.

Vosgerau, G. & Voss, M. (2014). Authorship and Control over Thoughts. *Mind & Language*, 29(5), 534-565.

Zahavi, D. (2005). *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge: MIT Press.