La construcción de la sensación de agencia: un enfoque desde la neurociencia y la filosofía de la mente

Trabajo de Final de Máster

Estudiante: Ignacio Molinero Moles

Tutor: Fernando Martínez Manrique

1 DE JULIO DE 2025

MÁSTER INTERUNIVERSITARIO EN LÓGICA Y FILOSOFÍA DE LA CIENCIA
Universidad de Valladolid

Contenido

Abstract 2
Keywords 2
Resumen 2
Palabras clave 2
1. Introducción 3
2. Marco teórico y antecedentes filosóficos 4
Sustancialismo 4
Individualismo 5
Relacionismo 6
Impermanentismo 7
Conclusión del apartado 8
3. El eliminativismo y la agencia como ficción 9
4. La neurociencia de la agencia 12
Sensación de agencia y unión intencional 12
Sesgo de autoservicio 14
Modelos de agencia 14
Patologías relevantes 18
5. Discusión filosófica 19
Parecer y ser 19
Crítica al sujeto delgado como objeto 20
Desde la fenomenología 21
Reificación y ontología 21
Una contradicción 23
6. Conclusiones 23
7. Referencias 24

Abstract

This essay proposes a radical critique of the concept of agency from an interdisciplinary approach between neuroscience and philosophy of mind, adopting an eliminativist stance. It argues that agency is not an ontologically real property of the subject, but a perceptual and narrative construct generated by distributed brain mechanisms, conditioned by retrospective inferences and cognitive biases. Through the analysis of neuroscientific models and the study of pathologies, argues that the sense of agency is malleable, fallible and fabricated. Both the classical substantialist, individualist and relational conceptions are questioned, including Galen Strawson's position, which will require more detail than the others. The above positions are criticized —to a greater or lesser extent— for being ego-dualistic. It is concluded that even the most elementary forms of self-experience are based on opaque and confabulatory processes that do not justify a strong ontology of self or agent. This position has relevant implications for clinical ethics, moral responsibility, and theories of free will, and points toward a more fine-tuned and scientifically informed understanding of human subjectivity as a construct rather than as a substantial entity.

Keywords: Agency; Self; Confabulation; Ego-dualism.

Resumen

Este trabajo propone una crítica radical al concepto de agencia desde un enfoque interdisciplinar entre neurociencia y filosofía de la mente, adoptando una postura eliminativista. Se sostiene que la agencia no es una propiedad ontológicamente real del sujeto, sino una construcción perceptiva y narrativa generada por mecanismos cerebrales distribuidos, condicionada por inferencias retrospectivas y sesgos cognitivos. A través del análisis de modelos neurocientíficos y del estudio de patologías se argumenta que la sensación de agencia es maleable, falible y fabulada. Se cuestionan tanto las concepciones sustancialista clásica, individualista y relacionales, incluyendo la posición de Galen Strawson, que requerirá mayor detalle que las demás. Se critica a las posturas mencionadas —en mayor o menor medida— por ser dualistas del ego. Se concluye que incluso las formas más elementales de autoexperiencia se basan en procesos opacos y fabulatorios que no justifican una ontología fuerte del yo o del agente. Esta posición tiene implicaciones relevantes en ética clínica, responsabilidad moral y teorías del libre albedrío, y apunta hacia una comprensión más ajustada y científicamente informada de la subjetividad humana como construcción y no como entidad sustancial.

Palabras clave: Agencia; Yo; Fabulación; Dualismo del ego.

«[…] el «sujeto» no es nada dado, sino algo fingido-por-añadidura, introducido-por-detrás. ¿Definitivamente es necesario poner al intérprete detrás de la interpretación? Eso es ya poesía, hipótesis».

—Nietzsche, 2006, p. 17.

1. Introducción

El concepto de agencia ocupa un lugar central en numerosos debates contemporáneos, tanto en la filosofía como en las ciencias cognitivas. Desde la responsabilidad moral hasta el diseño de inteligencia artificial, la noción de un «agente» que actúa, elige o decide parece estar implícita en muchas explicaciones de conducta, decisión y experiencia. Sin embargo, esta ubicuidad contrasta con la vaguedad conceptual que lo rodea.

A menudo se invoca el término «agencia» sin ofrecer una definición clara o consistente, lo que genera un campo fértil para equívocos, reificaciones y presupuestos no examinados. De hecho, en ocasiones se reconoce explícitamente que no se dará una definición (Ferrero, 2022, p. 3) y se continúa el desarrollo del trabajo a partir de aspectos tangenciales o circunstanciales de la agencia. Esto es problemático, ya que conduce a la elevación injustificada del estatus ontológico del agente.[1]

En este trabajo adoptaré una metodología interdisciplinar y un enfoque de corte eliminativista para evaluar críticamente el concepto de «agencia». Este trabajo defiende que lo que llamamos «agencia» pertenece al reino de la apariencia —y no al de la realidad— desde la neurociencia y la filosofía de la mente. Por tanto, mi tesis es que no existe tal cosa como un agente real, sino una sensación de agencia maleable, falible y fabulada, que debe entenderse como una ficción útil. Los teóricos de la agencia y del «yo» que caen en reificaciones del «yo» serán llamados en este trabajo «dualistas del ego» y serán mis principales rivales.[2]

El concepto de agencia ha sido tradicionalmente esquivo (Emirbayer y Mische, 1998, p. 962; Ferrero, 2022, p. 3). Una definición operativa de agencia podría ser la capacidad de «hacer que las cosas sucedan», «marcar la diferencia» o «provocar algún tipo de cambio» (Ferrero, 2022, p. 2). Y el agente podría ser «aquello que actúa». Esta concepción primordial del agente ya nos condiciona a considerarlo como un ente real, único y unificado. Nuestra fenomenología sirve de sustento a esta creencia.

Sin embargo, este trabajo parte de una hipótesis crítica: que lo que llamamos agencia no es una propiedad ontológicamente robusta de ningún sujeto, sino una construcción

[1] Soy consciente de lo arduo que es dar una definición de agencia y no es mi deseo criticar con esto ningún autor que no dé una definición.

[2] Encontramos dualismos del ego mediante una concepción popular tradicional del alma, la psique o la mente, o bien sea mediante un enfoque materialista más contemporáneo. Aunque la crítica no vaya a centrarse en ellos, encontramos términos como «ejecutivo central» (Baddeley y Hitch, 1974), «sistema atencional supervisor» (Norman y Shallice, 1986), «homúnculo» o «fantasma en la máquina» (Ryle, 1949), «intérprete» (Gazzaniga, 2012), etc., que llevan a un enmascaramiento del dualismo implícito en el pensamiento acerca del «yo». Los últimos —debo destacar— que son conscientes de esto y lo hacen a propósito para mostrar el error conceptual.

perceptiva y narrativa derivada de patrones predictivos y atribuciones *post hoc*. En realidad, lo que ocurre es que un conjunto de eventos produce una apariencia. De aquí surgen dos mecanismos complementarios: 1) La apariencia de que ocurre un acto genera en nosotros la inferencia de que debe existir un agente que lo ha causado. Y 2) No solo nos viene la apariencia de un acto, sino que también disponemos de la sensación de ser un agente unificado. Estos mecanismos no son mutuamente exclusivos, sino que se retroalimentan. No obstante, esta construcción puede debilitarse. Hay casos donde la sensación de agencia se deteriora por lesiones cerebrales, enfermedades o síndromes, entre otros. La inferencia de que hay un agente tras los actos pierde fuerza o desaparece por completo. Este hecho evidencia que la sensación y la inferencia no remiten a un agente real, sino a una ficción cognitiva funcional.

Para desarrollar el análisis propuesto, el trabajo se estructurará en cinco secciones principales. En primer lugar, examinaré las concepciones tradicionales de la agencia que incurren en lo que denomino el dualismo del ego, es decir, la presuposición implícita de un «yo» o «agente» sustancial o unificado. En segundo lugar, presentaré las posiciones eliminativistas, tanto filosóficas como científicas, que cuestionan la existencia ontológica del agente y que servirán de soporte a la tesis defendida. A continuación, analizaré modelos y evidencias procedentes de la neurociencia cognitiva —incluyendo investigaciones sobre la sensación de agencia, los procesos de atribución y los errores en la autopercepción— que refuerzan la interpretación de la agencia como una construcción ilusoria. La cuarta sección se dedicará a una discusión filosófica más amplia sobre las implicaciones ontológicas y epistémicas de esta posición, así como a la comparación crítica con enfoques alternativos. Centraré mi foco en la posición de Strawson dado que plantea una de las versiones más sutiles del dualismo del ego que critico. Finalmente, en las conclusiones, evaluaré las posibles objeciones a la tesis propuesta, propondré líneas de investigación futuras y destacaré el potencial explicativo del marco eliminativista en el estudio de la agencia.

2. Marco teórico y antecedentes filosóficos

Sustancialismo

En este apartado voy a comentar diferentes posturas acerca del agente: el sustancialismo de Descartes; el individualismo de Hobbes, Locke y Kant; el relacionismo de Emirbayer y Mische y el SESMET[3] de Strawson. Estas posiciones servirán de marco para introducir la postura del eliminativismo y la crítica principal a las posturas anteriores que llevo a cabo en este trabajo: el dualismo del ego.

Descartes presenta su dualismo sustancialista con claridad en las Meditaciones metafísicas (1641/1999), donde distingue entre la *res cogitans* (sustancia pensante) y la *res extensa* (sustancia material). Para él, el «yo» o el alma es una sustancia inmaterial que

[3] En adelante SESMET y SESMETs serán escritos como *sesmet* y *sesmets*.

piensa, duda y existe con independencia del cuerpo. Esta concepción fundamenta la idea de un agente esencial, previo a la acción, dotado de unidad, continuidad y poder causal: "Soy una cosa que piensa, es decir, que duda, afirma, niega, conoce pocas cosas, ignora otras muchas…" (Descartes, 1641/1999, en la primera página de la meditación tercera).

Esta concepción inaugura una forma de pensar la agencia como propiedad de una entidad sustancial unificada, a la que se atribuyen capacidades mentales como la voluntad o la decisión. El agente no se constituye en la experiencia, sino que la precede ontológicamente: es su causa y fundamento. Este modelo influirá en muchos desarrollos posteriores, incluso en contextos no dualistas, al conservar la idea de que hay un «yo» que actúa, decide y posee estados mentales. Para un enfoque eliminativista como el que aquí se defiende, esta es la forma más explícita y metafísicamente fuerte del dualismo del ego: el dualismo de sustancias.

Individualismo

Aunque Hobbes, Locke y Kant no sostienen un dualismo sustancial como el cartesiano, su concepción del agente sigue anclada en una visión individualista que atribuye a la persona una capacidad interna, racional y autónoma de acción. En el caso de Hobbes, en *Leviatán* (1651), el agente es concebido como un sujeto mecanicista, movido por deseos y pasiones, pero que posee voluntad y puede deliberar entre medios para alcanzar fines. Aunque niega el sustancialismo explícitamente: «Por tanto, si un hombre me habla de un rectángulo redondo; o de accidentes del pan en el queso; o de substancias inmateriales; o de un sujeto libre (…) yo no diré que está en un error, sino que sus palabras carecen de significación» (Hobbes, 1651/2005, pp. 34 y 35) podemos ver en múltiples ocasiones expresiones como «*creo en él, yo confío en él, yo tengo fe en él, yo me apoyo en él»* (p. 53) o «*yo doy, yo otorgo, yo he dado, yo he otorgado, yo quiero que esto sea tuyo;* (…) *yo daré, yo otorgaré*» (p. 110), en cuyo uso denota su concepto del agente como ente necesario y válido.[4] Aunque niega el libre albedrío y es materialista, Hobbes preserva una noción de sujeto individual como unidad de toma de decisiones, centrado en el cálculo racional de intereses.

En *Ensayo sobre el entendimiento humano* (1690), Locke vincula la identidad personal a la conciencia reflexiva que acompaña al pensamiento: el «yo» es quien se percibe a sí mismo como idéntico en el tiempo, en virtud de su conciencia (Locke, 1690/1956, Libro II, cap. XXVII)[5]. Esta conciencia unificadora sirve de base para atribuir responsabilidad

[4] Dice también: «En la *deliberación,* el último apetito o aversión inmediatamente próximo a la acción o a la omisión correspondiente, es lo que llamamos VOLUNTAD, acto (y no facultad) de *querer*» (p. 48). Aquí —y más adelante con su análisis de las pasiones— vemos como trata la deliberación como algo interno y unitario.

[5] Entre otras, dice: «Si pudiéramos suponer cualquier espíritu totalmente privado del recuerdo o de la conciencia de sus acciones pasadas, lo mismo que encontramos que nuestra mente siempre lo está respecto a gran parte de nuestras acciones anteriores, y algunas veces respecto a todas, entonces la unión o la separación de unas sustancias semejantes espirituales no provocaría más cambios de identidad personal que lo que lo hace la unión o separación de cualquier partícula de materia. Cualquier sustancia virtualmente unida al ser presente pensante es una parte de ese mismo sí mismo que ahora es; y cualquier cosa unida a él por un tener conciencia de sus acciones anteriores, también forma parte del sí mismo, que es el mismo entonces y ahora» (Libro II, cap. XXVII, §25).

moral y agencia. Kant, por su parte, en la *Crítica de la razón práctica* (1788), define al agente como un ser racional autónomo capaz de actuar según principios morales autoimpuestos. La agencia kantiana se funda en la libertad práctica: «La unión de la causalidad como libertad con ella como mecanismo natural, la primera mediante la ley moral y la segunda mediante la ley natural —y precisamente en un mismo sujeto: el hombre— es imposible sin representar éste en relación con la primera como ente en sí y en la segunda como fenómeno, lo primero en la conciencia pura, lo último en la empírica» (Kant, 1788/2003, p. 6).[6]

Aunque cada uno de estos autores tiene su peculiaridad, comparten una noción del agente como unidad individual racional, capaz de deliberar, proyectarse y asumir responsabilidad. Desde el punto de vista aquí adoptado, este modelo sigue siendo sustancialista en sentido funcional, ya que presupone la existencia de un «yo» homuncular, no dinámico, interno y autor único de sus actos.

Relacionismo

El enfoque sustancialista ha sido ampliamente criticado. De hecho, como se ha visto, algunas críticas procedían de los que se ha llamado individualistas (Hobbes, 1651/2005, pp. 34 y 35). No obstante, hay una crítica al sustancialismo —que aplica al individualismo— que corta precisamente en la dirección de este trabajo: la falacia filosófica de John Dewey. Esta falacia consiste en tomar abstracciones conceptuales —como «el agente» o «el yo»— y tratarlas como si fueran realidades sustanciales. En lugar de reconstruir las condiciones experienciales que dan lugar a estas nociones, se las toma como punto de partida ontológico. Este enfoque individualista ha sido objeto de una crítica sistemática por parte de Mustafa Emirbayer (1997), quien, influido por la pragmática de John Dewey, centra su enfoque en las relaciones mediante lo que llama *relational pragmatics*. Emirbayer identifica en el sustancialismo una tendencia a reificar al agente como una entidad estática y predefinida.

Emirbayer diferencia entre sustancialismo de las «auto acciones» y de interacciones. El primero concibe a los agentes como entidades completamente autosuficientes que actúan por sus propias motivaciones internas, independientemente del contexto. El segundo reconoce cierta interacción entre agentes, pero los sigue considerando como entidades ontológicas previas a tales interacciones (Emirbayer, 1997, p. 283). El enfoque de Emirbayer implica que sin la relación no existe el agente.

Emirbayer propone una concepción relacional del agente que se aleja radicalmente del enfoque sustancialista tradicional. En lugar de concebir al agente como una entidad fija, autónoma y dotada de atributos constantes, Emirbayer lo entiende como una figura dinámica, cuyas capacidades emergen en la interacción con contextos sociales, temporales y estructurales específicos. Esta perspectiva implica que la agencia no es una propiedad inherente al individuo, sino una práctica situada que se ejerce y se transforma continuamente. Al rechazar el esencialismo del agente, Emirbayer aporta una visión más

[6] Se encuentra en la 2ª nota al pie del Prólogo.

fluida y compleja de la acción humana, subrayando su carácter interdependiente y contextual. Junto a Ann Mische dice: «Conceptualizamos el yo no como una sustancia o entidad metafísica, como el «alma» o la «voluntad» (véase White, 1995), sino como una estructura dialógica, completamente relacional en sí misma. En otras palabras, nuestra perspectiva es relacional en todos los sentidos» (Emirbayer y Mische, 1998, p. 974)[7].

Emirbayer distingue tres elementos de la agencia: el iterativo (orientado al pasado), que refiere a la reactivación selectiva de patrones pasados de pensamiento y acción; el proyectivo (orientado al futuro), que implica la imaginación creativa de posibilidades futuras; y el evaluativo (orientado al presente), que concierne a la capacidad de realizar juicios entre posibles trayectorias de acción (Emirbayer y Mische, 1998). Esta concepción del agente, aunque trate de librarse de la tradición individualista que concibe al agente como una entidad metafísica, sigue considerando un ente unificado tanto en el actuar como en el tiempo. Por esto considero que sigue habiendo dualismo del ego o reificación sutil del agente.

Impermanentismo

Ante la posición de Emirbayer y Mische, que trata al agente como un ente extendido en el tiempo, Galen Strawson (2004, p. 430) distingue entre personas «diacrónicas», que experimentan su existencia como una narrativa extendida temporalmente, y personas «episódicas», que experimentan su «yo» presente como discontinuo respecto a sus encarnaciones pasadas y futuras. En *Selves* (2009) también traza la distinción, aunque ahora bajo los nombres de «endurantistas» e «impermanentistas» y, aunque también relaciona a los primeros con la narratividad y los segundos con la no-narratividad (pp. 14-5), destaca que hay gran variabilidad y que no hay más que una tendencia, pudiendo haber endurantistas no narrativos e impermanentistas narrativos (pp. 221-2). Respecto a enfoques que tratan a las personas como una fuente poco fiable que fabula constantemente —como el aquí expresado—, Strawson defiende que hay otras cuya memoria no es distorsionante.

> Tenemos aquí otra dimensión profunda de la diferencia psicológica humana. Algunas personas son fabuladoras hasta la médula. En otras, la memoria autobiográfica es fundamentalmente no distorsionante, independientemente de los procesos automáticos de remodelación y refundición que pueda implicar invariablemente (2004, p. 444).[8]

Strawson piensa que ser una cosa, que sea sujeto, que sea mental y que sea único son necesarias para sus «yoes». Pero no solo eso, sino también mínimamente suficientes. Siendo la agencia, la persistencia, la personalidad y la distinción características no esenciales de la experiencia del «yo» (e incluso posiblemente ausentes en la infancia temprana) (Strawson, 2009, p. 57).

[7] Las traducciones de las obras citadas en otros idiomas son mías salvo que se indique expresamente lo contrario. De las obras que uso traducción refiero a quién es el traductor en las referencias.
[8] Dice que es variable entre culturas, pero que el fenómeno básico es universal (2009, p. 35; 1999, p. 407)

> Existo y sé que existo; Descartes tenía razón en que se puede saber esto. La pregunta es: ¿qué es este yo que conozco?" (Descartes 1641: 1.18). Yo respondo: Soy un ser humano, un producto de la evolución por selección natural, un organismo vivo, un animal, un hombre, un objeto físico —un objeto totalmente físico. (…) Asumiré la verdad del materialismo o, más exactamente, la opinión de que todo lo que existe concretamente (que incluye todo lo mental) es físico (Strawson, 2009, p. 4).

Al inicio de su obra, Strawson advierte sus conclusiones con un condicional interesante: «Al final, mi escrito del yo me lleva a concluir que hay muchos yoes de vida breve o yoes pasajeros, si es que los hay» (Strawson, 2009, p. 9).[9] La visión transitoria o pasajera —también llamada de perlas (Strawson, 1999, p. 424) por la semejanza entre los yoes y las perlas enlazadas por una cuerda— constituye el producto de su trabajo metafísico (argumentación a partir de Descartes) y fenomenológico (como impermanentista y no narrativista).

En el capítulo «Phenomenology: the general question», Strawson introduce los *sesmets* como la estructura mínima necesaria para la autoexperiencia y argumenta su existencia dado que la experiencia misma es cierta. *Sesmets* sería un acrónimo de sujetos de experiencia que son cosas mentales únicas (*Subjects-of-experience-that-are-single-mental-things*; Strawson, 2009, p. 172). Para Strawson, cada *sesmet* es un «yo» experiencialmente unificado en un instante, pero no hay una entidad subyacente que unifique a todos los *sesmets* a lo largo del tiempo. De aquí su posición impermanentista. La sensación de continuidad del «yo» es un efecto de la narrativa personal (de ciertas personas, no de todas), pero no refleja una realidad metafísica subyacente.

Strawson dice que los *sesmets* son objetos, pero esto no significa que no sean procesos. De hecho, dice que todos los objetos son procesos (Strawson, 2009, p. 12). Cuando Strawson afirma que los *sesmets* son objetos procesos, no está postulando una sustancia estática. Más bien, entiende por objeto algo momentáneamente identificable en el flujo de la experiencia: una unidad mínima de autoexperiencia (*SELF-experience*) que, aunque procesual y efímera, cumple funcionalmente con el rol de un «yo» en el instante. Así, cada *sesmet* es un «yo» puntual, sin que exista un sujeto que los unifique más allá de ese momento.

Conclusión del apartado

Aunque el sustancialismo filosófico se ilustra con claridad en autores como Descartes, su eco puede encontrarse también en corrientes modernas que, sin sostener un dualismo, tienden a concebir la agencia como un atributo estable, individual e interno, como es el caso de ciertas lecturas de Hobbes, Locke o incluso Kant, y que se ha llamado individualistas. Este modelo, tal como señalan Emirbayer y Mische (1998, pp. 964-5),

[9] El condicional, así como la existencia del libro de Strawson, nos indica, entre otras cosas, que los «yoes», a pesar de ser intuitivos, deben ser argumentados y no dados por hecho. Creo que es necesario indicarlo en este trabajo, no solo para reconocer la honestidad de Strawson, sino para sentar las bases del debate acerca del «yo».

configura gran parte de la tradición occidental moderna, asociando agencia con voluntad deliberativa y cálculo racional. Se ha identificado la posición de estos autores con un modelo relacional de la agencia que, aunque trataba de acabar con el agente como entidad metafísica previa a la acción, seguía manteniendo una reificación sutil del agente. Por último, el modelo de Strawson habla del «yo» como «sujeto de experiencia que es una cosa mental única», entendiendo este *sesmet* como una cosa física instanciada en cada momento específico de experiencia y que sirve de estructura para que se pueda dar esta. A pesar de las diferencias entre dualismo, individualismo, relacionismo y la posición strawsoniana, todas estas concepciones presuponen algún tipo de entidad real detrás de la agencia, ya sea como causa, como estructura o como experiencia inmediata.

3. El eliminativismo y la agencia como ficción

Ante estos enfoques —que denominamos dualistas del ego— surge un enfoque que trata de explicar la agencia como una herramienta funcional, como una ficción útil. De este modo, el enfoque eliminativista prescinde del peso ontológico asociado a la noción de agente para reconocerlo como una mera herramienta del pensamiento. Esto es suficiente para explicar tanto la fenomenología de la agencia como las implicaciones evolutivas de la existencia de esta.

La teoría de la postura intencional (*intentional stance*) desarrollada por Daniel Dennett constituye uno de los pilares fundamentales para la concepción eliminativista de la agencia (o de la agencia como ficción útil sin peso ontológico). Dennett propone que atribuimos estados mentales intencionales (creencias, deseos, intenciones) a sistemas cuyo comportamiento queremos predecir, no porque estos estados existan ontológicamente como entidades discretas en su interior, sino como una estrategia predictiva pragmáticamente eficaz (Dennett, 1987/2018, p. 15).

Según Dennett, la atribución de agencia es una estrategia interpretativa que «consiste en tratar al objeto cuyo comportamiento se quiere predecir como un agente racional con creencias y deseos" (Dennett, 1987/2018, p. 14). Sin compromiso ontológico, ciertos sistemas exhiben patrones comportamentales que resultan inteligibles y predictibles cuando son interpretados *como si fueran* agentes intencionales.[10]

Para Dennett, no hay un lugar en el cerebro donde las decisiones se tomen (o «teatro cartesiano») (Dennett, 1991/1995), sino procesos distribuidos que forman la conducta que interpretamos como agencia. Esta posición es adoptada por, entre otros, Michael Gazzaniga. La expresa titulando dos capítulos de su libro *¿Quién manda aquí?* como «el cerebro paralelo y distribuido» y «el intérprete»[11]. Un breve extracto tratando la reacción a lo que podría haber sido una serpiente expresa el punto:

[10] La postura de Dennett (1991) implica que estos patrones son reales, definiendo jocosamente «reales» como que puedes ganar dinero apostando por ello. Aquí yo difiero de su posición.

[11] Como se verá más adelante, aunque pueda sonar a una palabra que deseamos evitar, Gazzaniga es totalmente consciente y explica a la perfección cómo se puede dar tal organización sin un controlador (o mandamás).

> (…) la realidad es que salté mucho antes (…) de percatarme de la presencia de la serpiente. No tomé la decisión consciente de saltar y luego la ejecuté conscientemente. Cuando respondí esa pregunta, estaba fabulando (…): aportando un relato ficticio de un acontecimiento pasado que creía cierto (Gazzaniga, 2012; p. 101).[12]

Siguiendo la línea eliminativista, Thomas Metzinger, —en su influyente obra *Being No One* (2003)— desarrolla una teoría representacionalista de la conciencia que niega la existencia de un «yo» sustancial mientras explica la potente fenomenología de ser un agente. Según Metzinger, «en el mundo no existen los yoes: Nadie ha sido ni ha tenido nunca un yo» (Metzinger, 2003, p. 1). Lo que existe, según Metzinger, es un automodelo fenoménico (*phenomenal self-model*), generado por el cerebro como parte de su modelo global de la realidad (Metzinger, 2003, p. 299; 2018, p. 3).

La característica definitoria de este modelo del «yo» es su transparencia, es decir, el hecho de que no lo experimentamos como un modelo o representación, sino como la realidad misma. «La transparencia de las estructuras de representación es el criterio decisivo para convertir un modelo en una *apariencia*, en una *realidad aparente*. No experimentamos la realidad que nos rodea como el contenido de un proceso de representación ni representamos sus componentes como marcadores de posición internos, actores causales, por así decirlo, de otro nivel externo de la realidad. Simplemente la experimentamos como el mundo en el que vivimos»[13] (Metzinger, 2003, p. 169). Esta incapacidad para detectar el carácter construido del modelo genera la ilusión de estar en contacto directo con un «yo» real y sustancial. En este sentido, dice Metzinger, «la transparencia es una forma de oscuridad».

En el contexto de la agencia, Metzinger habla en diferentes ocasiones del síndrome de la mano ajena.[14] En *El túnel del yo* (2018) dice que esta mano lleva un comportamiento que «*parece* guiado por una representación finalística explícita» (p. 161; la cursiva es mía). Sigue diciendo que «…toda teoría convincente de la identidad deberá poder explicar la disociación entre propiedad y agencia». En ambas obras menciona la esquizofrenia como ejemplo en que la persona que la padece deja de tener sensación de agencia (*sense of agency*) o sensación de propiedad (*sense of ownership*).[15] Metzinger (2018) también comenta la adaptabilidad evolutiva de la ilusión de identidad, de un «automodelo narrativo» (pp. 171-2).

---

[12] La concepción del intérprete del hemisferio izquierdo es mucho anterior a este libro. Ya se encuentra en (Gazzaniga, 1985).

[13] La cursiva es mía. Deseo destacar la diferencia entre apariencia y realidad. No quiero quitar la legitimidad al testimonio de la gente, pero debemos darnos cuenta de que, como decía Oliver Sacks, el problema de los testimonios es que dos personas pueden dar declaraciones contradictorias sobre el mismo hecho y ninguno estar mintiendo. En definitiva, trazo esta distinción para dejar claros estos dos conceptos y usarlo más adelante al atajar la argumentación de Strawson.

[14] El síndrome de la mano ajena o anárquica es un trastorno neurológico en el que la mano del paciente hace movimientos involuntarios e incontrolables debido a daño en diversas partes de la corteza o cuerpo calloso.

[15] Metzinger habla de agencia y propiedad, respectivamente.

William James rechaza tanto lo que llama la teoría del alma como la teoría ficcional que presenta al sujeto con el ser imaginario denotado por el pronombre «yo» (James, 1890). No obstante, la postura de James —que podríamos encajar como la asociación de la figura del pensador al propio pensamiento— se parece a la de David Hume, que entendía al «yo» no como una entidad, sino más bien como un «haz de sensaciones» (1739/2011, p. 222)[16]. Sus diferencias radicarían en que Hume se podría catalogar como conexionista[17] (o teórico de haces, *bundle theorist*; Parfit, 1984/2004) y, aunque a James se le podría incluir en el mismo grupo, diría que la posición de que los «yoes» no existan es extrema.

Susan Blackmore mantiene que «podemos rechazar cualquier entidad persistente que corresponda con nuestra sensación de ser un yo (*self*)» (2005; p. 76). Utilizando la distinción entre conexionistas y teóricos del «yo», cataloga a estos últimos como —lo que podríamos llamar— dualistas del ego (*ego dualists*), de la misma forma que Dennett hablaría de «materialismo cartesiano» (Dennett, 1991, p. 121).

> Hay que tener en cuenta que el dualismo es sólo una forma de teoría del yo, y que no es necesario ser dualista para creer en un yo continuo. De hecho, como veremos, muchas teorías científicas modernas que rechazan el dualismo siguen intentando encontrar los correlatos neuronales del yo o explicar el yo en términos de estructuras duraderas en el cerebro. Se trata, pues, de teorías del yo (Blackmore 2005, cap. 5, p. 64).

Una convergencia entre Blackmore, Dennett y Metzinger es la caracterización de la agencia como un fenómeno interpretativo más que como una propiedad ontológica intrínseca. Desde la postura intencional de Dennett hasta el modelo fenomenal del «yo» de Metzinger y la memética de Blackmore (1999; idea que toma de Dawkins, 1976/1985), estos enfoques conciben la agencia como una estrategia interpretativa útil, una característica de selección cultural o un modelo representacional transparente, no como una propiedad metafísica inherente a ciertos sistemas (Dennett, 1987/2018; Metzinger, 2005; Blackmore, 1999).

Parfit (1984/2004; apéndice J) dice que el Buda es el primer conexionista, y tanto Blackmore como Metzinger coinciden (Blackmore, 2005, p. 64; Metzinger, 2018, p. 189). Se destaca la frase atribuida a Buda: «Existen las acciones y existen sus consecuencias (el mérito y el demérito), pero la persona que actúa no existe. No hay nadie para descartar esto o asumir lo otro. No existe el individuo, es solo un nombre convencional para describir un conjunto de elementos» (atribuido a Vasunbandhu y citado en Metzinger (2018, p. 189)).

La posición de Blackmore, Metzinger y Dennett es parecida, aunque tienen sus diferencias. Y algunas de ellas residen en el concepto de existencia. Dennett niega la existencia de los «yoes», sin embargo, Blackmore dice que existen, pero utiliza fielmente

[16] «Bundle or collection of different perceptions» (Libro 1, parte 4, sección 6, Of personal identity).

[17] Utilizo aquí el término *conexionista* no en su sentido contemporáneo de modelos computacionales de redes neuronales, sino en el sentido filosófico clásico asociado a Hume: una teoría del «yo» como conjunto de percepciones conectadas (*bundle theory*), sin una sustancia subyacente que las unifique.

el diccionario para decir que son ilusorios, alegando a la definición de ilusorio como «no ser lo que parecen» (Blackmore, 2005, p. 48). Metzinger afirma que no existen, pero otorga relevancia ontológica a las representaciones que Dennett rechaza. Blackmore y Metzinger coinciden en las razones de la producción de la sensación de identidad: «La ilusión de continuidad se produce porque cada «yo» temporal viene acompañado de recuerdos que dan una impresión de continuidad» (Blackmore, 2005, p. 66) y «En cuanto a la identidad transtemporal subjetivamente experimentada del «yo», obviamente, dos factores son de principal importancia: continuidad transtemporal y la invariabilidad de los contenidos de retorno de la autoconciencia (por ejemplo, la experiencia de la agencia), y la aparición de la memoria autobiográfica» (Metzinger, 2003, p. 315). En esta línea, la comprensión de la agencia como ficción útil se sigue de la teoría de los memes, diciendo que somos «máquinas meméticas» cuyas decisiones y acciones están sustancialmente influenciadas por unidades culturales autorreplicantes (memes) que compiten por expresarse a través de nosotros —sus vehículos— (Dawkins, 1976/1985; Blackmore, 1999).

4. La neurociencia de la agencia

Patrick Haggard (2002) desarrolló el concepto de sensación de agencia (*sense of agency*) mediante un diseño experimental que permitía observar diferencias entre los actos voluntarios e involuntarios. Utilizando técnicas como la estimulación magnética transcraneal (TMS) y tareas de sincronización temporal, Haggard demostró que la sensación de agencia puede ser manipulada artificialmente.

Sensación de agencia y unión intencional

La metodología es parecida a la de otros experimentos tipo Libet (1983)[18]. Sin embargo, los participantes juzgaban los tiempos de percepción sensorial y de acción, en lugar de juzgar el surgimiento de su impulso o intención de moverse. En el primer grupo, las personas escuchaban un tono y movían su dedo índice derecho voluntariamente. En el segundo, el movimiento del dedo era producido por TMS. En el grupo de control, la TMS se orientaba a otra zona parietal, efectuando el mismo «*click*» pero sin producir contracción del dedo. Haggard observó que, al realizar una acción voluntaria justo después de un evento sensorial (como un tono en este caso), los participantes percibían una atracción temporal entre la acción y el evento. Sin embargo, cuando la acción era generada exógenamente ―provocada de manera involuntaria mediante TMS―, esta sincronización se perdía, viéndose una repulsión temporal entre los eventos. Este fenómeno de atracción temporal sugiere que el cerebro construye activamente la experiencia de agencia mediante mecanismos neurales que integran acción e intención

[18] Los experimentos de tipo Libet tienen una estructura similar en el sentido de que, en estos, los participantes miraban un reloj y daban los tiempos en los que eran conscientes de haber tenido el impulso (*urge*) o la decisión de actuar. Los experimentos de Libet eran diferentes porque medían la actividad eléctrica de las cortezas motoras, suplementarias y pre-suplementarias. En cambio, Haggard (2002) estimula con TMS.

(Haggard et al., 2002, p. 383). A este fenómeno se le llama «unión intencional» (*intentional binding*)[19].

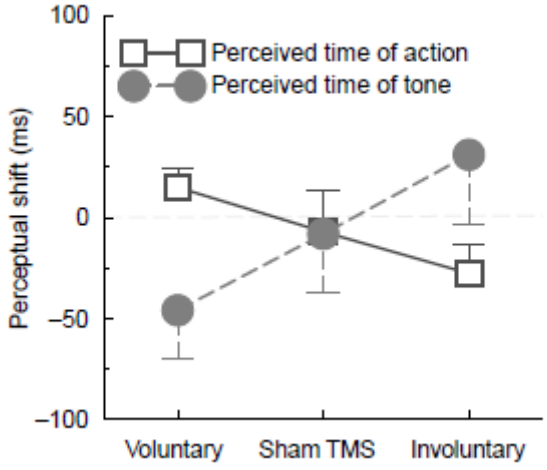

Figura 1: Atracción y repulsión temporal del efecto de la unión intencional (Haggard, 2002, p. 383).

La sensación de agencia se experimenta de forma retrospectiva, una vez que se recibe la retroalimentación del resultado de la acción. Este juicio puede ser fácilmente sesgado, incluso por estímulos inconscientes relacionados con los efectos de la acción (Haggard, 2017). Un ejemplo de esto —dice Haggard— son estudios de unión intencional como el que acabamos de ver:

> …los estudios sobre la unión intencional han distinguido entre estos dos componentes [los componentes prospectivos y retrospectivos de la acción] variando la probabilidad de que una acción voluntaria produzca un tono. Por ejemplo, un estudio demostró que, cuando la probabilidad de que un tono siga a una acción es del 50%, la vinculación de la acción hacia el tono es mayor en los ensayos en los que se presentan tonos que en los ensayos sin tonos. Esto sugiere que el tono alteró retrospectivamente el momento percibido de la acción que lo provocó (Haggard, 2017, p. 6).

Según Haggard y Chambon (2012, p. 390), la sensación de agencia comprende «la experiencia de controlar las propias acciones y, a través de ellas, los eventos en el mundo externo». A pesar de ello, Haggard diferencia, como Metzinger, la sensación de agencia (*sense of agency*) de la sensación de propiedad (*sense of ownership*).[20] De hecho, dice que la sensación de agencia implica la sensación de control y el control consciente y voluntario, también conocido como la experiencia de ser la fuente de la acción (Haggard, 2017, p. 2).

[19] Para una revisión: Moore y Obhi (2012).
[20] Como ya decía Metzinger (2018), esta diferencia es importante, sobre todo al explicar disociaciones entre ambas sensaciones, como se ve en patologías como la esquizofrenia.

Sesgo de autoservicio

El carácter de «informe interno» de la sensación de agencia es problemático porque está sujeto a sesgos que impregnan toda la experiencia humana. Un claro ejemplo es el sesgo de autoservicio (*self-serving bias*) (Haggard, 2017). Este sesgo consiste en la atribución del crédito por las acciones que son exitosas, pero no en los fracasos. Junto a este sesgo, existen otros factores que aumentan la atribución de autoría. Tener varias opciones entre las que elegir fortalece la unión intencional en el momento de la elección, aumentando la sensación de agencia. Asimismo, esta unión intencional es más fuerte cuando las elecciones predicen los resultados de la acción a cuando no lo hacen; cuando se decide por uno mismo que cuando otros deciden por uno; cuando los resultados son positivos antes que negativos (solo cuando es una consecuencia predecible de la acción); cuando los estímulos láser producidos por la acción voluntaria son menos dañinos que cuando producen mayor dolor; y cuando los resultados de acciones financieras son positivos que cuando son negativos (Haggard, 2019). Es decir, cuando los resultados de nuestras acciones son mejores, tenemos una mayor sensación de que hemos sido los autores de esas acciones. Y, continuando con la construcción de narrativa esperada, no se encontraron diferencias entre resultados neutros frente a negativos cuando estos los experimentaba un tercero. Esto quiere decir que si los resultados de la acción los sufre otra persona, que sea malo o neutro no cambia nuestra sensación de agencia.[21] Estos datos sugieren la existencia de una relación entre la permisibilidad moral de un acto y la autoría atribuida a las acciones que afectan a terceros. Esta relación surge de la sensación de agencia suprimida o deteriorada en virtud del resultado de la acción, de si la sufre —o goza— un tercero o uno mismo, etc.

Modelos de agencia

Se han elaborado muchos modelos para dar cuenta de la sensación de agencia. Uno de ellos es el modelo comparador, que sostiene que el cerebro genera continuamente predicciones sobre las consecuencias sensoriales de las acciones y compara estas predicciones con la retroalimentación sensorial real (Frith et al., 2000). Este modelo, basado originalmente en teorías de control motor, propone que el cerebro utiliza copias eferentes de comandos motores para predecir los efectos sensoriales de los movimientos antes de que ocurran. Cuando estas predicciones coinciden con los resultados reales, el movimiento se experimenta como generado endógenamente. En contraste, cuando hay discrepancias entre predicciones y resultados, el sentido de agencia se debilita (Frith et al., 2000).

[21] Una consideración por tener en cuenta con estos resultados es si el tercero es una persona que no conocemos más allá del contacto en el entorno experimental o si pertenece a nuestro círculo cercano. Para los propósitos del argumento no es relevante indagar más.

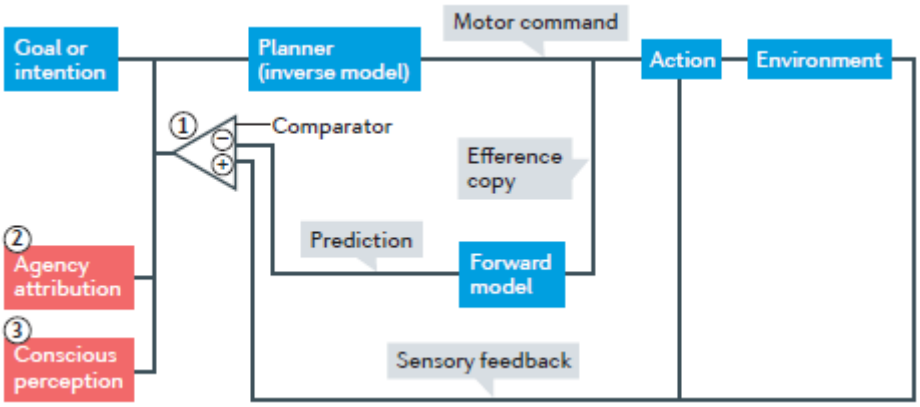

Figura 2: Modelo comparador de la agencia (Haggard, 2017)

Entre las críticas al modelo comparador encontramos que el modelo: 1) se enfoca en procesos sensorimotores, pero deja fuera influencias cognitivas de orden superior, 2) falla al explicar atribuciones erróneas (ya que hay personas que pueden sentir agencia incluso cuando no hay coincidencia sensorimotora, o no sentirla aunque la haya) y 3) no explica la experiencia consciente de agencia. Es decir, el modelo comparador puede explicar una «prerracionalización» de agencia o FoA (*feeling of agency*), pero no su integración en la experiencia consciente y narrativa o JoA (*judgment of agency*) (Synofzik et al., 2008).[22]

En segundo lugar, el modelo multifactorial explica no solo la predicción motora y su contribución a la sensación de agencia, sino también cómo esta se ve influida por señales *post hoc* de agencia como el *feedback* visual o la valencia del resultado de la acción, especialmente relevantes en pacientes que sufren esquizofrenia, por ejemplo (Synofzik et al., 2013).

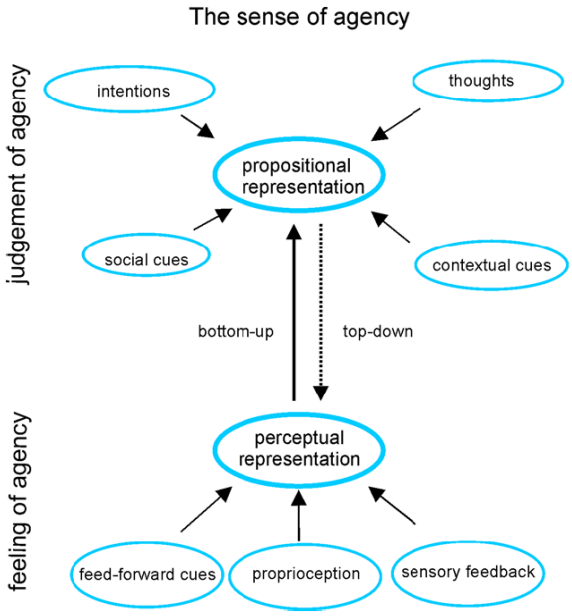

Figura 3: Modelo multifactorial (Synofzik et al., 2008).

[22] En el sentido de esta distinción, Strawson *juzga* y *siente* que él no fue el que «estuvo ahí ayer» (Strawson, 2009, pp. 225-226).

Haggard ha investigado experimentalmente cómo la coincidencia entre intención y resultado motor puede modular la experiencia de control, pero también ha señalado que la sensación de agencia emerge de un sistema distribuido que incluye tanto mecanismos motores como procesos cognitivos de alto nivel (Haggard, 2017). Estos dos niveles hacen alusión al FoA y JoA, respectivamente, pero Haggard habla de sensación de agencia como término paraguas y distingue entre los procesos cognitivos de alto nivel y de nivel sensorial.

En tercer lugar, tenemos al modelo bayesiano (Hohwy, 2013), que dice que el cerebro es como un sistema que predice constantemente sus propios estados. El sentido de agencia surge cuando la predicción reduce el error entre expectativas y resultado sensorial. El modelo bayesiano tendría al modelo comparador como una situación particular donde la forma de minimizar error de predicción no implica factores más allá de los motores con el *feedback* sensorial. Esto está muy ligado a la teoría del cerebro predictivo. James Moore, por su parte, ha cuestionado la suficiencia del modelo comparador para explicar casos de agencia en ausencia de acción, defendiendo una aproximación de integración de señales que se basa en principios bayesianos: el juicio de agencia sería el resultado de la ponderación de múltiples pistas internas y externas según su fiabilidad contextual (Moore & Fletcher, 2012). Este tipo de enfoque permite explicar por qué a veces se experimenta agencia incluso cuando no hay una coincidencia clara entre la copia eferente y el *feedback* sensoriomotor.

Un último modelo que me gustaría comentar es el de la «causación mental aparente» (Wegner, 2002). Los estudios de Wegner y sus colaboradores donde los participantes y un cómplice movían conjuntamente un puntero sobre un tablero con diversas imágenes muestran la maleabilidad de la sensación de agencia. Cuando los participantes escuchaban previamente el nombre de una imagen y el puntero se detenía sobre ella poco después, tendían a experimentar un sentido de agencia sobre este movimiento, aunque en realidad era el cómplice quien controlaba el puntero (Wegner y Wheatley, 1999). La teoría de la «causación mental aparente» de Wegner (2002) postula que nuestro sentido de agencia es fundamentalmente una atribución retrospectiva basada en tres factores: 1) Prioridad, es decir, un pensamiento relevante debe preceder a la acción, 2) Consistencia, el pensamiento debe ser congruente con la acción, y 3) Exclusividad, por lo que no deben existir causas alternativas salientes (Wegner y Wheatley, 1999).

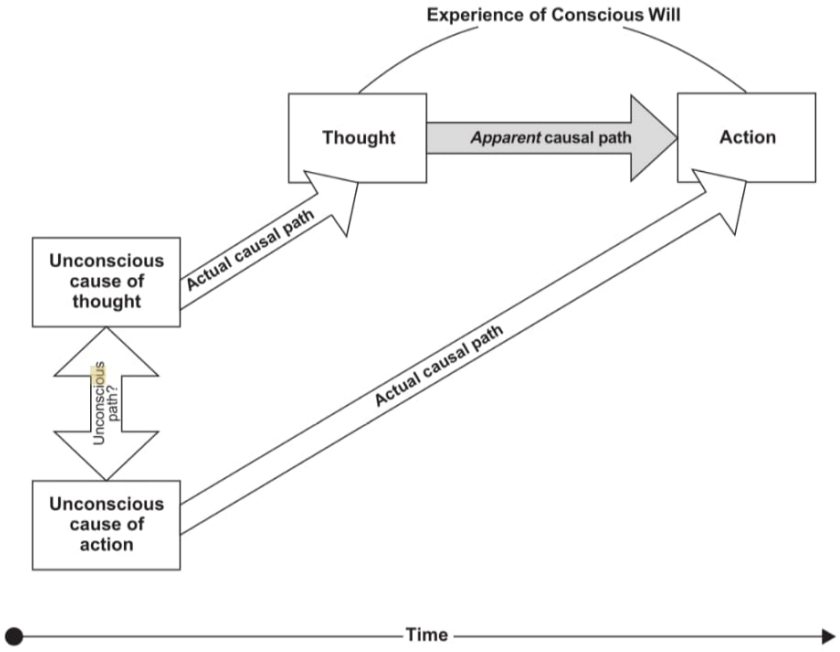

Figura 4: Modelo de la causación aparente (Wegner, 2002, p. 63)

Según declara el propio Wegner (2002): «Esta teoría de la causalidad mental aparente depende de la idea de que la consciencia no sabe cómo funcionan los procesos mentales conscientes». Estas palabras hacen que rescatemos este modelo, aunque sea más antiguo. Este modelo destaca el componente retrospectivo y fabulatorio de la sensación de agencia y señala a la esfera de las razones —al igual que este trabajo— como pertenecientes al reino de las apariencias. «Las personas experimentan una voluntad consciente cuando *interpretan* su propio pensamiento como la causa de su acción» (La cursiva es mía; Wegner y Wheatley, 1999). Este se corresponde con el JoA del modelo multifactorial (Synofzik et al., 2008), aunque también se podría formular la oración señalando a que las personas interpretan que son los causantes de sus acciones cuando sienten el *feedback* sensoriomotor adecuado, atajando así el FoA.

Además de la maleabilidad y fabulación de la sensación de agencia que he comentado según diferentes condiciones (Haggard, 2017), las decisiones morales también revelan cómo fabulamos agencia, dado que los juicios morales y las atribuciones de responsabilidad surgen *a posteriori*, bajo condiciones de narración, expectativas sociales y sesgos cognitivos. El efecto de los jueces hambrientos es un claro ejemplo de esto (Danziger et al., 2011). En este estudio, se vio que el factor que cobraba mayor importancia en la toma de decisiones del juez era el tiempo desde su última ingesta. En términos más rigurosos: el nivel de glucosa en sangre. La última sentencia del artículo es reveladora: «De hecho, la caricatura de que la justicia es lo que el juez desayunó podría ser una caricatura apropiada para la toma de decisiones humanas en general» (p. 6892). Cualquier juez apoya sus juicios morales en sólidas razones y evidencias, pero eso no impide que dichas razones sean fabuladas *a posteriori*.[23] Otro ejemplo más cercano a la

[23] En este trabajo no se tratará la racionalidad de los juicios morales ni, por tanto, el argumento de Alvin Plantinga (2011) contra el naturalismo evolucionista.

neurociencia cognitiva se encuentra en Jeurissen et al. (2014), donde TMS en la unión temporoparietal afectaba directamente al juicio de la persona sobre la permisibilidad moral de un acto. Robert Sapolsky tiene una sentencia que se ha vuelto famosa recientemente: «La libertad es biología que desconocemos». Aunque el debate del libre albedrío no sea el objetivo de este trabajo, aquí podríamos decir: «La agencia es biología que desconocemos».

Patologías relevantes

Además de la ya mencionada esquizofrenia, donde se observaba una pérdida de la sensación de propiedad sobre las voces en la cabeza del paciente (Metzinger, 2018; esto explicaba la necesidad del JoA de Synofzik et al., 2008), existen otras patologías que nos pueden servir para comprender mejor la sensación de agencia. Como dice Metzinger (2018, p. 173), personas con mutismo acinético[24], a pesar de experimentar una sensación de su cuerpo y estar despiertos, «no son agentes: no actúan de ningún modo». También Metzinger (2018, p. 161) habla del síndrome de la mano ajena[25] —o anárquica—, y dice que esta mano lleva un comportamiento que «parece guiado por una representación finalística explícita». En este síndrome, el paciente carece de sensación de agencia (no tiene control ni se siente la fuente del acto), pero conserva la sensación de propiedad. La depresión puede entenderse como una alteración global de la experiencia del mundo, de uno mismo y de la propia capacidad de actuar (Ratcliffe, 2015).

La agencia, así como el concepto de «yo», son ilusiones, en el sentido de que «no son lo que parecen» (Blackmore, 2005) y en el sentido de que se les atribuye un estatus ontológico que no tienen. Esta concepción se vuelve especialmente relevante cuando se analiza la vivencia depresiva como una erosión sistemática de la sensación de poder actuar, decidir y querer (Ratcliffe, 2013).[26] Esta característica marcada de cuadros depresivos (sobre todo los graves) puede tener alcance explicativo. Diversos estudios sostienen que «los individuos deprimidos experimentan menos control sobre los resultados que los individuos sanos» debido a «un sesgo hiperagéntico de los individuos sanos», es decir, «las personas deprimidas tienen una valoración realista de lo limitada que es realmente su capacidad de acción» (Haggard, 2017, citando a Alloy et al., 1979). En resumen, estas patologías no solo ilustran fallos en los mecanismos de generación de sensación de agencia, sino que revelan su naturaleza esencialmente construida y vulnerable. Estas patologías cuestionan la solidez de la sensación de agencia (y la de

[24] El mutismo acinético es un trastorno neurológico en el que el paciente no tiene capacidad de moverse o hablar, aunque se encuentre despierto y con sus funciones sensoriales preservadas y se encuentre en estado de vigilia.

[25] En este síndrome, el paciente se sorprende por acciones de su propia mano que escapan a su control y que, a menudo, parece que llevan propósitos antagónicos a los del paciente. Hay ejemplos más cómicos de esto, como que la mano desabroche la camisa que el paciente intenta abrochar, y otros más dramáticos, como intentos de estrangulamiento.

[26] El deterioro de la agencia que se da en la depresión se refleja tanto en actos deliberativos como en acciones automáticas (Ratcliffe, 2015, p. 156), mostrando que hay un mecanismo subyacente a toda toma de decisiones y producción de la sensación de agencia que se encuentra deteriorado. Este dato es relevante para estudios futuros, donde trataré la toma de decisiones en relación con la sensación de libertad y donde, por tanto, la distinción entre actos deliberativos y aquellos de intención más «urgente» (*urges*) cobra importancia.

libertad) tal como la experimentamos la población no clínica de manera rutinaria (Ratcliffe, 2013; Metzinger, 2003; Gallagher, 2012).

5. Discusión filosófica

Tal como se ha visto, la sensación de agencia viene influida por la unión intencional, y esta se da mediante: 1) coherencia entre la copia eferente y el *feedback* sensoriomotor (ver figura 2), 2) por la miríada de factores que influyen en nuestro sesgo de agencia (*self-serving bias*) y, en definitiva, 3) por la racionalización *post-hoc* elaborada a partir de otros pensamientos, intenciones y señales sociales (JoA; ver figura 3) que Gazzaniga (1985) denominó «intérprete». Lo que fenomenológicamente parece un teatro cartesiano (Dennett, 1991/1995), no es sino un compendio de factores que, a varias escalas temporales y mediante un procesamiento distribuido y paralelo, sirve un propósito evolutivo de engaño unitario extremadamente adaptativo.

No obstante, cuando hablamos de agencia, parece que asumimos al agente y lo dotamos de dignidad ontológica. Sin embargo, se puede hablar con un lenguaje que dote de unidad, sin necesidad de asumir estas conclusiones. Cuando hablamos de una enfermedad, no mencionamos todos los síntomas conocidos, sino que le damos un nombre que, en ocasiones, es el de la persona que la descubrió. En este sentido, el nombre hace las veces de resumen de los síntomas, pero no asume que de lo que se habla sea una cosa inamovible, estanca y unificada. De la misma manera que hablamos teniendo esta concepción no reificada de la enfermedad o de cualquier situación fisiológica, podemos preservar este lenguaje útil acerca de las acciones y los agentes siendo conscientes de que, a pesar de que lo parezca, no hay un agente unificado que actúe en el mundo, sino una miríada de factores —o pequeños robots, como dice Dennett— orquestados evolutivamente (sin director de orquesta).

Parecer y ser

Teniendo esto en cuenta, todo lo que podemos decir acerca del «yo» y del agente es que parecen, sin duda alguna, ser algo. No obstante, ¿es legítimo pasar del *parecer* al *ser*? En multitud de ocasiones se ha argumentado en contra del marco eliminativista que, cuando se habla de conciencia, si admitimos que parece algo, tenemos que aceptar que la conciencia es algo, ya que el concepto de conciencia solo ambiciona la apariencia.[27] En este mismo sentido, cuando hablamos de un triángulo, no exigimos ver en la naturaleza un triángulo para decir que es, sino que solo exigimos que sea coherente. De tal forma, podemos decir que un triángulo específico no existe cuando se predica de él algún atributo contradictorio, como que sea rectangular o circular. ¿Es el concepto de «yo» o agente ambicioso?

[27] Justo este argumento se podría poner en contra de la tesis de Strawson que aquí se critica. Podría decirse que el mismo argumento que se emplea para decir que la conciencia debe de ser algo, aunque ese algo sea solo apariencia, es el que condena al sujeto delgado de Strawson a quedar confinado al reino de la apariencia. De esta manera, podríamos seguir toda su argumentación y acabar alegando únicamente que lo que ha defendido no son «yoes» reales, sino aparentes.

Galen Strawson afirma que, si los *sesmets* son considerados «yoes» (*selves*), entonces los «yoes» existen. Para ello, primero define los *sesmets*, aclara y defiende cada uno de los conceptos que engloba el término y ataja la cuestión de la existencia. Centraré mi crítica en la concepción del «yo» como objeto real en Strawson, tanto desde la aparente unidad como desde la fenomenología.

Crítica al sujeto delgado como objeto

Debo preludiar esta crítica exponiendo que Strawson reconoce que casi todos los que quieren que haya un «yo» quieren que haya un «yo» persistente. De hecho, en este punto, Strawson es un aliado de la tesis que en este trabajo se mantiene: «La propuesta de Dennett (…) no se toma en serio el escrito en este sentido: niega que existan realmente cosas como los yoes (y creo que es correcta, en la medida en que se considera que los yoes son cosas que persisten durante largos periodos de tiempo)» (Strawson, 1999, p. 100). Strawson (2009), dice en la introducción que el texto se puede leer en clave eliminativista sin problema (p. 5), lo que nos indica que su concepto es más sutil e inmune a la crítica eliminativista del «yo» extendido en el tiempo.

Su visión de los «yoes» como perlas tiene sentido dentro de su marco, ya que esto validaría llamar objetos a los «yoes», al entender «objeto» como algo que tiene una «unidad fuerte» (Strawson, 2009, p. 298). Para contestar a la propuesta de Strawson, por tanto, no es suficiente con sugerir la invalidez o inestabilidad de un concepto de «yo» distendido en el tiempo, pues, aunque ese sea el concepto de «yo» que, como bien señala Strawson, la mayoría acoge[28], no es el que Strawson argumenta.

Strawson habla del «sujeto delgado» (*thin subject*), es decir, el fenómeno de la experiencia del «yo» en el «momento vivido de la experiencia» (*living moment of experience*) (Strawson, 2009, p. 256) como constituyente de los *sesmets*. Esto muestra que Strawson parte de la fenomenología para luego construir lo que, a su forma de ver, tiene que constituir un concepto de «yo» mínimo. Este tiene que ser, como se ha dicho, un *sesmet*, es decir, un sujeto de experiencia que sea una cosa mental única. Dice Strawson: «Los sujetos delgados considerados en el momento vivo de la experiencia ciertamente califican como unidades fuertes objetivas si alguna cosa lo hace» (Strawson, 2009, p. 368).

La disolución de las condiciones de transtemporalidad y memoria autobiográfica en contextos depresivos desestabiliza la experiencia unificada de la identidad y la capacidad de actuar con coherencia, favoreciendo una discontinuidad narrativa (Ratcliffe, 2015). Este fenómeno podría relacionarse con la visión «de perlas» de Strawson, ya que muestra cómo, en ciertas condiciones patológicas, los sujetos experimentan el «yo» de una manera que parece coincidir con la descripción de Strawson del «sujeto delgado». Sin embargo, aunque estas experiencias patológicas podrían parecer una manifestación de su teoría, mi crítica radica en que incluso en estos casos, donde la continuidad del «yo» se ve

[28] Por buenas razones, como las apuntadas por Metzinger (2003, p. 315) y Blackmore (2005, p. 66), así como por la fenomenología que Strawson solo achaca a los endurantistas. Aunque vaya a criticar la visión impermanentista, el endurantismo suscita críticas parecidas, así como la paradoja del barco de Teseo.

comprometida, la concepción de Strawson sigue reificando al sujeto. La experiencia fragmentada de la identidad en contextos patológicos revela que el «yo» descrito por Strawson, aunque aparentemente efímero y sin sustancia, sigue siendo entendido como una entidad, aunque se trate de una entidad fragmentada. Así, la crítica debe centrarse no solo en la disolución de la identidad en contextos clínicos, sino también en cómo la teoría de Strawson, al enfocarse en la fenomenología de la experiencia, sigue manteniendo una noción de «yo» que, al final, sigue siendo reificada.

Desde la fenomenología

Como he indicado, incluso ahí donde él dice que hay una unidad fuerte y objetiva, existe maleabilidad de la sensación de agencia y fabulación de esta en favor de la narrativa. Por tanto, incluso si se reconoce que la unidad aparente del «yo» es más fuerte seccionando la experiencia en tramos arbitrarios de tiempo donde nos encontramos con un flujo de conciencia libre de hiatos (Strawson, 1997, pp. 413-4), se debe tener en cuenta que la evidencia mostrada —y mucha evidencia no mostrada— apunta a un sujeto de experiencia totalmente construido, cuya fenomenología se mueve en una esfera de razones fabuladas debido a la opacidad de los procesos fisiológicos que verdaderamente producen lo que llamamos «yo». No solo las experiencias en contextos patológicos —como las estudiadas por Ratcliffe en el caso de la depresión—, sino toda experiencia en población no clínica donde actuamos muestra que todo nuestro entendimiento de nosotros mismos radica en la fabulación de razones que dan sentido a la narrativa y, por tanto, es inevitablemente falsa por su incapacidad de acceder a la esfera de las causas reales. No deseo realizar una crítica a la arbitraria variabilidad temporal del «yo» strawsoniano, ni discuto que gozamos de una aparente unidad fuerte. Sin embargo, es notable que estas perlas (o canicas), que representan la unidad en la experiencia, pueden ser —y, de hecho, lo son— menos sólidas de lo que parecen, y no solo en contextos clínicos. En favor de Strawson diré que este distingue entre la experiencia de tener un «yo» y la realidad ontológica de este, pero su construcción argumentativa posterior se basa en la fenomenología de los sujetos delgados para proponer los *sesmets*, que luego pasarán a ser tan o más reales que cualquier objeto. Asimismo, tanto el modelo orwelliano de la conciencia presentado por Dennett (1991) como el relato presentado por Gazzaniga para ilustrar la narrativa creada por un «intérprete» (2012) que no puede acceder a la información subpersonal son rechazadas por Strawson como otra magnitud variable de la condición humana, diciendo que la fabulación se restringe a un espectro variable, y no es una característica universal.

Reificación y ontología

Strawson (1997, p. 408) dice que la agencia es una propiedad no esencial al «yo» mediante el «argumento de la reducción», es decir, expone que la agencia no es necesaria para la experiencia mínima. Encontramos evidencia en favor de esto en patologías como el comentado mutismo acinético y, a su vez, encontramos evidencia en contra en trabajos que unen acción y percepción como un único proceso indisoluble mediante, por ejemplo, el estudio de los movimientos sacádicos oculares necesarios para la percepción visual. No deseo argumentar en estas direcciones en este trabajo, sino atacar la noción de agente

como ente asumido al observar una acción. Pero este trabajo no se restringe al «yo» que actúa, sino a toda noción reificada del «yo» encarnada en la postura de los llamados «dualistas del ego». Para este propósito, por tanto, podemos incluso asumir la existencia hipotética de «observadores puros», es decir, criaturas inmóviles, cognitivamente bien equipadas, altamente receptivas, autoconscientes y que están bien informadas sobre su entorno (Strawson, 2009, p. 187).

¿Está reificando Strawson al sujeto? Ciertamente es lo que parece: su propuesta se basa en defender los «yoes» como objetos realmente existentes. Strawson habla de que los «yoes» son estados cerebrales complejos o sinergias (Strawson, 2009, p. 273). Sin embargo, dice: «Al decir que un yo es una cosa ontológicamente distinta, quiero decir al menos que no es lo mismo que cualquier otra cosa identificada de forma ordinaria o natural como una cosa. Pero no quiero decir que sea "una entidad independiente o que exista por separado" (Parfit, 1995, p. 18)» (Strawson, 1997, p. 425).

Sobre el estatus ontológico de los objetos como sillas o gatos, dice, es dudoso cuando la metafísica se pone seria (Strawson, 2009, p. 221). A pesar de esto, dice que los «yoes» son «tan reales como conejos o partículas Z», «tan objetos o cosa como un grano de sal» y «tan cosa física como un vaso sanguíneo o una vaca» (Strawson, 1997, p. 425). Asimismo, compara la posición de Van Inwagen acerca de los átomos que están «atrapados en la vida de un organismo» con aquellos «atrapados en la vida de un yo mental» (Strawson, 1997, p. 425). No obstante, esta formulación suscita una tensión: ¿cómo puede un elemento quedar «atrapado» en la vida de un «yo» mental si ese elemento constituye —ontológicamente— ese mismo «yo»? ¿No implica esto asumir una distinción entre la entidad que atrapa y lo atrapado, es decir, una forma sutil de dualismo en la concepción del «yo»? Reconozco que Strawson evita el dualismo clásico, pero me veo obligado a señalar que su lenguaje deja abierta la posibilidad de una lectura dualista o sustancialista implícita. Podría tomar los escritos de Strawson en clave eliminativista e integrarlos a la visión del «yo» como una abstracción o ficción útil, pero encuentro en ellos un claro ejemplo de posesión del meme del dualismo del ego a un materialista convencido.

Como se ha visto, el cerebro es una máquina que se encuentra constantemente lanzando predicciones (ver figura 2) y, según la coherencia entre estas predicciones, nuestras acciones y otra serie de estímulos (ver figura 3), fabulamos retrospectivamente la agencia y sentimos más o menos responsabilidad por los resultados de las acciones. Esto no solo atañe a la agencia —que, en el marco de Strawson, queda relegada en un segundo plano— sino que trata directamente de la fenomenología del «yo». Incluso en un escenario donde concibamos al «yo» como instanciado en un periodo arbitrariamente extenso de tiempo donde la conciencia se encuentre libre de hiatos, la evidencia proveniente de la experimentación —y especialmente, de la neurociencia cognitiva— sugiere que este «yo» es construido y no dado previa o subyacentemente. Y que, por tanto, la causalidad que nos atribuimos retrospectivamente pertenece al reino de la apariencia (Wegner, 2002).

Una contradicción

A pesar de la coherencia del pensamiento de Strawson y de que mi crítica no abarca la coherencia de su enfoque, sino más bien su falsedad, encuentro una contradicción. Strawson dice: «Muchos, entre los que me incluyo, piensan que basta con decir que un sujeto de experiencia es una entidad que es de tal clase que hay, para ella, "algo que es como" serlo» (Strawson, 2009, p. 63). Sin embargo, Strawson, que distinguía entre la experiencia de tener un «yo» y la realidad ontológica de este, aquí parece no reflejar esa distinción. Esta afirmación constituye, en mi opinión —así como en la de Blackmore, Parfit, Dennett y probablemente Metzinger—, una petición de principio en la que se da por hecho el sujeto al que se quiere llegar. La argumentación para el sujeto de experiencia es que existe experiencia (aunque siempre se encarga de destacar que podría haberla sin sujeto). Más allá de esto, no tengo ningún problema con que se diga que para que se dé la experiencia debe haber un elemento estructurador de la experiencia (yo diría que son muchos elementos los que estructuran la experiencia de manera distribuida y paralela). Pero decir que para que haya experiencia de cierto tipo (*SELF-experience*) se necesita un «yo» estructurador de la experiencia como sujeto que sea una cosa mental, única y que sea agente, persistente y distinta del humano al completo y que sea un objeto físico (Strawson, 2009, p. 266) es una afirmación algo más pesada.[29]

6. Conclusiones

Este trabajo ha defendido una tesis fuerte: el «yo» es ilusorio. Esta tesis ha sido defendida desde dos perspectivas. Principalmente desde el punto de vista del agente que parece estar implícito en las acciones (o inferirse a partir de ellas), calificándolo como una construcción narrativa retrospectiva. Y, de manera más secundaria —y centrada en la postura de Strawson— del «yo» como entidad mínima asumida por la experiencia, siendo relegada a un segundo plano la agencia como propiedad no necesaria para el «yo» mínimo. Por tanto, y como se ha visto, la agencia, lejos de ser una propiedad ontológicamente real del sujeto, es una sensación construida, maleable, falible y fabulada, que se sitúa en el reino de la apariencia y no en el de la realidad. A partir de una metodología interdisciplinar y un enfoque de corte eliminativista, se ha mostrado que tanto la filosofía como la neurociencia contemporánea ofrecen fundamentos sólidos para rechazar la existencia de un «agente» o «yo» como entidad sustancialmente unificada.

A lo largo del trabajo se ha argumentado que muchas teorías clásicas —desde el sustancialismo (Descartes) hasta formas modernas de individualismo (Hobbes, Locke, Kant) y relacionalismo (Emirbayer y Mische)— conservan un núcleo problemático: lo que se ha llamado el dualismo del ego, siguiendo a Blackmore y Parfit. Se ha encontrado en el relacionismo una reificación sutil del agente en su relación al actuar y se ha presentado someramente la visión «de perlas» (Strawson). Tras esto, para fundamentar mi crítica, me he apoyado en varias perspectivas que he englobado en el concepto de

[29] Como dice Carl Sagan (1979): «*Extraordinary claims require extraordinary evidence*».

eliminativismo (Dennett, Metzinger, Blackmore, Parfit, Hume y, en cierta medida, James).

Asimismo, se ha mostrado cómo la neurociencia cognitiva —a través de modelos predictivos, atribucionales y comparadores— respalda una lectura eliminativista de la agencia: lo que sentimos como voluntad o autoría no prueba la existencia de un agente, sino que resulta de procesos reconstructivos e interpretativos. La agencia, en este marco, opera más como una coordenada narrativa de referencia —al estilo de lo que Dennett describe como un «centro de gravedad narrativa»— que como una propiedad sustancial o funcionalmente autónoma que precise de un agente.

Por último, he tratado con algo más de profundidad la posición de Strawson, que lejos de ser una posición que haga una reificación fuerte del agente como unidad extendida en el tiempo, se ha identificado con el eliminativismo (en este sentido). Su tesis «de perlas» del «yo» permite aplicar las evidencias presentadas y el marco formulado a un «yo» puntual en el tiempo. Si mi crítica es correcta, abarcaría no solo la figura del «yo extendido», sino también formas supuestamente mínimas como el *sesmet* propuesto por Strawson. Como se ha argumentado, incluso el «yo» puntual —estructurador de la experiencia sin hiatos— resulta ser una construcción fabulada, dependiente de inferencias narrativas, expectativas internas y ceguera causal. El marco eliminativista propuesto en este trabajo, por tanto, aplica tanto al «yo» duradero como al «yo» momentáneo.

Esta visión abre nuevas líneas de investigación. En primer lugar, deseo destacar la estrecha relación entre la sensación de agencia y las sensaciones de responsabilidad y libertad, entre otras. Como ha señalado Haggard (2017, p. 10): «Responsabilidad individual depende en gran medida de los mecanismos cerebrales que subyacen a la sensación de agencia». Siendo así la sensación de agencia la principal responsable de la resultante sensación de responsabilidad y las consecuentes sensaciones de culpa u orgullo. Este camino nos llevaría a explorar la debatida cuestión del libre albedrío y la responsabilidad moral que, dado el enfoque aquí acogido, auguro llevaría a la negación de ambos, independientemente de que el escenario propuesto sea determinismo o indeterminismo. También se abre una posible aplicación en la ética clínica: concebir al agente como una ficción podría evitar formas de culpabilización innecesarias o perjudiciales en pacientes con condiciones que afectan el juicio, la iniciativa o la autopercepción. En este sentido, se han comentado algunas patologías que ayudan a entender un poco mejor la sensación de agencia y la de propiedad. Este camino también debe seguir explorándose, aunque esto no suponga que otros enfoques lesionales como el uso de la TMS deban ser suprimidos.

7. Referencias

Alloy, L. B. y Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: sadder but wiser? *J. Exp. Psychol. Gen*. 108, 441–485.

Baddeley, A. D., y Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press.

Blackmore, S. J. (1999). *The Meme Machine*. Oxford: Oxford University Press.

Blackmore, S. (2005). *Consciousness: A very short introduction*. Oxford University Press.

Danziger, S., Levav, J. y Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions, *Proc. Natl. Acad. Sci.* U.S.A. 108 (17) 6889-6892, https://doi.org/10.1073/pnas.1018033108

Dawkins, R. (1985). *El gen egoísta: Las bases biológicas de nuestra conducta* (J. Robles Suárez & J. Tola Alonso, Trad.). Barcelona: Salvat. (Original: *The Selfish Gene*, 1976, Oxford: Oxford University Press).

Dennett, D. C. (1995). *La conciencia explicada: Una teoría interdisciplinar* (S. Balari Ravera, Trad.). Barcelona: Paidós. (Original: *Consciousness Explained*, 1991, Boston: Little, Brown & Co.).

Dennett, D. C. (2018). *La actitud intencional* (C. Clemente, Trad.). Barcelona: Gedisa. (Original: *The Intentional Stance*, 1987, Cambridge, MA: MIT Press).

Descartes, R. (1999). *Meditaciones metafísicas* (M. García Morente, Ed. y Trad.). Madrid: Espasa Calpe. (Original: *Meditationes de prima philosophia: in qua Dei existentia et animæ immortalitas demonstrantur*, 1641, Paris: Clarendon Press).

Emirbayer, M. (1997). Manifesto for a relational sociology. *American Journal of Sociology*, 103(2), 281-317.

Emirbayer, M., y Mische, A. (1998). What is agency? *American Journal of Sociology*, 103(4), 962–1023.

Ferrero, L. (2022). An introduction to the philosophy of agency. En L. Ferrero (Ed.), *The Routledge handbook of philosophy of agency* (pp. 1–13). London: Routledge.

Frith, C. D., Blakemore, S.-J. y Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. Phil*osophical Transactions of the Royal Society B: Biological Sciences*, 355(1404), 1771–1788. https://doi.org/10.1098/rstb.2000.0734

Gallagher, S. y Zahavi, D. (2012). *La mente fenomenológica*. Barcelona: Alianza Editorial. (Original: *The Phenomenological Mind*, 2008, Londres: Routledge).

Gazzaniga, M. S. (1985). *The social brain: Discovering the networks of the mind*. New York: Basic Books.

Gazzaniga, M. S. (2012). *¿Quién manda aquí?: El libre albedrío y la ciencia del cerebro*. Barcelona: Paidós. (Original: *Who's in Charge? Free Will and the Science of the Brain*, 2011, New York: Ecco).

Haggard, P., Clark, S. y Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, 5(4), 382-385.

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, 18(4), 196–207. https://doi.org/10.1038/nrn.2017.14

Haggard, P. (2019). The Neurocognitive Bases of Human Volition. *Annual Review of Psychology*, 70(Volume 70, 2019), 9-28. https://doi.org/10.1146/annurev-psych-010418-103348

Hobbes, T. (2011). *Leviatán*. Buenos Aires: Fondo de Cultura Económica. (Original: *Leviathan*, 1651, London).

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hume, D. (2011). *The essential philosophical works*. Ware: Wordsworth Editions. (Original: *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*, 1739-40, London: Clarendon Press).

James, W. (1980). *Principios de psicología* (E. O'Gorman, Trad.; vols. 1–2). Madrid: Ediciones Morata. (Original: *The Principles of Psychology*, 1890, New York: Henry Holt & Co.).

Jeurissen, D., Sack, A. T., Roebroeck, A., Russ, B. E., & Pascual-Leone, A. (2014). TMS affects moral judgment, showing the role of DLPFC and TPJ in cognitive and emotional processing. *Frontiers in Neuroscience*, 8. https://doi.org/10.3389/fnins.2014.00018

Kant, I. (2003). *Crítica de la razón práctica* (J. Rovira Armengol, Trad.). Madrid: Editorial Tecnos. (Original: *Critique of Practical Reason*, 1788, Königsberg).

Libet, B., Gleason, C. A., Wright, E. W. y Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. *Brain*, 106(3), 623-642.

Locke, J. (1956). *Ensayo sobre el entendimiento humano* (E. O'Gorman, Trad.). México: Fondo de Cultura Económica. (Original: *Essay Concerning Human Understanding*, 1690, London).

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.

Metzinger, T. (2018). *El túnel del yo: Ciencia de la mente y mito del sujeto*. Barcelona: Basic Books. (Original: *The Ego Tunnel*, 2009, New York: Basic Books).

Moore, J. W. y Fletcher, P. C. (2012). Sense of agency in health and disease: A review of cue integration approaches. *Consciousness and Cognition*, 21(1), 59–68. https://doi.org/10.1016/j.concog.2011.08.010

Moore, J. W. y Obhi, S. S. (2012). Intentional binding and the sense of agency: a review. *Consciousness and cognition*, *21*(1), 546–561. https://doi.org/10.1016/j.concog.2011.12.002

Nietzsche, F. (2006). *Nihilismo: Escritos póstumos* (A. Valle, Trad.). Madrid: Alianza Editorial. (Original: *Nachgelassene Schriften*, 1887, Leipzig: C. G. Naumann).

Norman, D. A. y Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory* (Vol. 4, pp. 1–18). Plenum Press.

Parfit, D. (2004). *Razones y personas* (M. Rodríguez González, Trad.). Madrid: A. Machado Libros. (Original: *Reasons and Persons*, 1984, Oxford: Clarendon Press).

Plantinga, A. (2011). *Where the Conflict Really Lies: Science, Religion, and Naturalism.* Oxford University Press.

Ratcliffe, M. (2013). *Depression and the Phenomenology of Free Will*. En Fulford, K. W. M., Davies, M., Gipps, R., Graham, G., Sadler, J., Stanghellini, G., & Thornton, T. (Eds.), *The Oxford Handbook of Philosophy and Psychiatry* (pp. 574–591). Oxford University Press.

Ratcliffe, M. (2015). *Experiences of Depression: A Study in Phenomenology*. Oxford University Press.

Ryle, G. (1949). *The concept of mind*. London: Hutchinson's University Library.

Sagan, C. (1979). *Broca's brain: Reflections on the romance of science*. New York: Random House.

Strawson, G. (1997). The self. In S. Gallagher & J. Shear (Eds.), Models of the self (pp. 1–24). Exeter: Imprint Academic.

Strawson, G. (1999). The self and the SESMET. Journal of Consciousness Studies, 6(4), 99-135.

Strawson, G. (2004). *Against Narrativity*. Ratio, 17(4), 428–452.

Strawson, G. (2009). *Selves: An Essay in Revisionary Metaphysics*. Oxford: Oxford University Press.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, 17(1), 219–239. https://doi.org/10.1016/j.concog.2007.03.010

Synofzik, M., Vosgerau, G., & Voss, M. (2013). The experience of agency: An interplay between prediction and postdiction. *Frontiers in Psychology*, 4. https://doi.org/10.3389/fpsyg.2013.00127

Wegner, D. M. (2002). The illusion of conscious will. Cambridge, MA: MIT Press.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54(7), 480-492.