Serie
Selección de Textos

S | T

Discusiones contemporáneas en filosofía de la mente

Voces Locales

Pablo López-Silva

(Editor)

Universidad de Valparaíso
CHILE
Facultad de Humanidades
Instituto de Filosofía

serie

Selección de Textos

S | T

Serie Selección de Textos

Juan Redmond (Director)
Rodrigo Lopez-Orellana & Jorge Budrovich S. (Editores)

Volumen 7

Discusiones contemporáneas en filosofía de la mente

Voces Locales

Pablo López-Silva

(Editor)

Universidad de Valparaíso
Facultad de Humanidades
Instituto de Filosofía

2019

Edición:
Instituto de Filosofía, Universidad de Valparaíso
Impreso en Valparaíso, Chile. Noviembre de 2019

Serrano 546, Valparaíso. Chile

ISBN 978-956-401-448-7

Discusiones contemporáneas en filosofía de la mente.
Voces Locales

Editor: Pablo López-Silva
Serie Selección de Textos, Volumen 7
Primera edición. Valparaíso, 2019

Comité Científico del Volumen

Índice

Prólogo 9

1. Personas y mentes
   José Tomás Alvarado Marambio 13

2. Una noción contextual de emergencia y un ejemplo en análisis de redes neuronales
   Esteban Céspedes y Rubén Herzog 39

3. ¿Hemos respondido la pregunta "puede pensar una máquina"?
   Rodrigo Alfonso González Fernández 71

4. En contra de la visión de la valencia afectiva como reforzadores internos
   José M. Araya 97

5. Del problema emoción-cognición a la integración de la fenomenología y la intencionalidad de los estados mentales
   Rodolfo Bachler 123

6. El camino no elegido: indiferencia ontológica en la filosofía de la mente
   Sebastián Sanhueza Rodríguez 151

7. La mente agencial: elementos para una teoría de las atribuciones de agencia mental
   Pablo López-Silva, Andrea Arancibia, Gabriel Cordero y Leonardo Henríquez 179

Prólogo

Cuando realizaba mi doctorado tuve la oportunidad de vivir con un amigo que se encontraba doctorando en astrofísica. Recuerdo muy bien que cuando celebrábamos la exitosa defensa de su tesis sostuve una conversación bastante profunda con algunos de los miembros de su laboratorio, todos astrofísicos de renombre mundial. De forma interesante, todos ellos estaban de acuerdo al sostener que el principal desafío para la ciencia actual no venía —por usar una metáfora— desde lo que 'estaba afuera'. En realidad, para ellos el universo de la astrofísica, si bien está siempre en expansión y parece ser infinito, está disponible 'afuera', públicamente, para ser medido, para ser explorado y para ser accedido mediante métodos que parecen ser bastante efectivos. No, el desafío para ellos no era lo que está 'afuera', sino que lo que está 'dentro', esto es, la mente humana. Ya desde las formas en que recolectamos datos sobre nuestra propia vida mental, el desafío de investigar la mente parece problemático. ¿Cómo puede acceder el otro a mis experiencias conscientes, las cuáles, parecen ser privadas? ¿Es acaso posible? ¿Y si no lo es, cómo sabe el otro que no soy un simple robot y que realmente tengo una mente al menos parecida a la suya? ¿Cómo podemos generar conocimiento objetivo sobre aquello que es por definición subjetivo? ¿qué pasa si las conductas del otro no son nada más que el producto de algoritmos introducidos por un programador, y que estas personas no son nada más que robots altamente complejos?

Desde hace varios años, la filosofía de la mente se ha preocupado de examinar las preguntas más fundamentales que surgen al observar el fenómeno de lo mental (sea lo que sea que es). A la luz de la naturaleza de su tarea, esta disciplina no solo encontró un nicho importante dentro de la filosofía de las ciencias, sino que se extendió más allá como un ejercicio interdisciplinario generando diálogos fructíferos con la biología, la neurofisiología, la psicología y la antropología, entre muchas otras disciplinas. Esto, podríamos

decir, se debe a que la filosofía de la mente no solo buscar lidiar con preguntas respecto de la ontología de los conceptos mentales, sino que también con asuntos relativos a su estructura subjetiva (fenomenológica), a la dimensión psicológica de la mente y a la relación entre tales estados y nuestra arquitectura cerebral. Ahora bien, el auge que la filosofía de la mente comenzó a experimentar desde los años 60' a la luz de las revoluciones cognitivas y los avances en neurociencia no ha pasado desapercibido en nuestra región. Ya desde hace varios años, diversos filósofos y psicólogos han comenzado a explorar distintas preguntas sobre lo mental a la luz de debates específicos dentro de las ciencias empíricas y sociales. Poco a poco, colegas en Colombia, Perú, Argentina y Brasil han comenzado a poblar el escenario local de la filosofía de la mente con ideas originales en el contexto de las discusiones acerca de la inteligencia artificial, la cognición, percepción, la conciencia, la agencia, la enfermedad mental, entre muchos otros.

La compilación 'Discusiones Contemporáneas en Filosofía de la Mente: Voces Locales' es un esfuerzo por visualizar el trabajo de filósofos chilenos en torno a diversas discusiones dentro de la disciplina. Este esfuerzo es profundamente motivado por la idea de la democratización del acceso al conocimiento y la idea de que los recursos digitales pueden ayudar a expandir de forma libre el conocimiento local. Teniendo en cuenta la complejidad de nuestro objeto de estudio, la compilación incluye contribuciones en diversas áreas de la filosofía de la mente; mientras algunas se avocan a aspectos puramente ontológicos, otras combinan los *insights* de las neurociencias, la psicopatología y la fenomenología en la comprensión de estados y procesos mentales específicos. Tal riqueza es, sin duda, muestra de la diversidad de contextos, trasfondos y voces que se unifican *en* y *para* el estudio de la mente humana. Así, la compilación inicia con un análisis ontológico respecto del concepto de persona y su relación con el cuerpo y los estados mentales, un problema que finalmente subyace a cualquier aproximación más funcional a la vida mental humana. Luego de esto, el segundo capítulo analiza un concepto fundamental dentro del acercamiento ontológico al problema-mente-cuerpo, esto es, la noción de emer-

gencia, una idea que cada vez es más popular a la hora de entender el origen de las características mentales de un sujeto dentro de la filosofía de la mente y neurociencias actuales. El tercer capítulo se adentra en una de las discusiones clásicas dentro de la disciplina, esto es, la pregunta respecto de si realmente sabemos si las máquinas pueden pensar. El sentido revisionista de este capítulo nos ayudará a mapear el estado de la cuestión y las reales posibilidades de respuesta a tal relevante debate. Por su parte, el cuarto y quinto capítulo se focalizan en discusiones que rodean el concepto de emoción y su relación con otros estados mentales. Ambos capítulos actuarán como un buen complemento para adentrarse a la filosofía de la emoción en su estado actual. Contrastando, el sexto capítulo no lidia con estados mentales específicos, sino que con algunas preguntas que surgen de la examinación del problema madre de la filosofía de la mente, esto es, el problema mente-cuerpo. Este ejercicio es siempre necesario para poder clarificar y dilucidar el proyecto de investigación que busca desarrollar la filosofía de la mente actual. Finalmente, la colección termina con una aproximación a uno de los aspectos más fundamentales del fenómeno de la agencia humana en el contexto de la actividad cognitiva. Tal como podemos ver, la compilación constituye un despliegue analítico de ideas clásicas y exploratorias, todo esto, con el fin de motivar la lectura profunda de nuestra audiencia.

Finalmente, junto con agradecer el apoyo del Instituto de Filosofía y la Escuela de Psicología de la Universidad de Valparaíso, quisiera agradecer profundamente a los miembros del comité científico de la compilación, quienes contribuyeron de forma importante con sus comentarios y recomendaciones; a los pares evaluadores, quienes sin duda mejoraron la calidad de todas las contribuciones; y a los directores de la Serie Selección de Textos de la Facultad de Humanidades de la Universidad de Valparaíso por su apoyo fundamental en distintas etapas del desarrollo de esta iniciativa. Además, quisiera agradecer el apoyo en el proceso editorial brindado por mis estudiantes de seminario de investigación Andrea Arancibia, Gabriel Cordero y Leonardo Henríquez, quienes comienzan a sumergirse en las aguas de la filosofía de la mente, y,

por supuesto, al 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 por proveer y promover espacios para pensar y re-pensar varios de los procesos administrativos y mentales que llevaron a la producción del presente volumen.

Dr. Pablo López-Silva

Profesor Adjunto
Escuela de Psicología
Universidad de Valparaíso
Chile

Capítulo 1

Personas y mentes

José Tomás Alvarado Marambio

Resumen

Sea lo que sea aquello a lo que refiere el pronombre de primera persona "yo" cuando es usado por cada uno de nosotros, eso es esencialmente una persona. Al menos desde Locke se ha supuesto además que lo que el pronombre de primera persona "yo" designa es también esencialmente una mente. No es extraño, entonces, que para esta tradición lockeana nuestras condiciones de identidad sean psicológicas. La concepción lockeana contrasta con la concepción tradicional de la persona, de acuerdo a la cual las personas son hypostaseis o supposita de una naturaleza racional. Las personas —entendidas de este modo— poseen estados mentales en algunos tiempos, pero no siempre, y podrían no haber sido mentes. En esta perspectiva, una persona humana es un animal racional, por lo que sus condiciones de identidad son las condiciones de identidad de un animal, tal como ha sido defendido por el 'animalismo' en las últimas décadas. Se argumenta aquí que la concepción tradicional de la persona, así como las posiciones 'animalistas', son por mucho más verosímiles y mejor apoyadas por nuestras intuiciones acerca de qué somos.

Palabras clave: personas, mentes, identidad personal, ontología de la persona, animalismo.

1. Introducción

Durante los últimos cien años, las discusiones acerca de la identidad de las personas en el tiempo y acerca de la ontología de la persona se han desarrollado dentro de un marco teórico cuya noción central es la de *mente* —y, correlativamente, la noción de *estado mental*—. La principal cuestión que se ha estado considerando es si las condiciones de identidad de una persona están fundadas en la continuidad de los estados mentales o no. De un modo análogo se ha discutido *ad nauseam* si los estados mentales deben o no reducirse a estados físicos, deben fundarse en estados físicos, o deben eliminarse por estados físicos. Ha interesado dilucidar si la mente es o no reducible, eliminable por el cerebro. Estas cuestiones son fundamentales porque —se supone— nosotros somos mentes. Hay varios presupuestos sustantivos en la forma en que se plantean estas cuestiones que han parecido obvias para buena parte de la comunidad filosófica. Estos presupuestos han hecho difícil comprender perspectivas diferentes tales como el llamado "animalismo", pero lo mismo puede decirse de las perspectivas más clásicas de nuestra tradición filosófica.

En efecto, se ha supuesto que el término "persona" designa aquello que cada uno de nosotros es en el nivel más fundamental (cf. Olson, 1997, 2007; Snowdon, 2014). Hay algo a lo que cada uno de nosotros hace referencia al usar el pronombre 'yo'. Se supone que lo que quiera que sea a lo que hace referencia ese pronombre cuando es usado por alguien es a una 'persona'. Está abierto a discusión si somos o no organismos biológicos, animales de cierto tipo, cerebros o un *ego* cartesiano, pero no está abierto a discusión que somos *personas*. Pues bien, al menos desde Locke en adelante ha parecido obvio que las condiciones de identidad en el tiempo de una persona deben estar fundadas en la naturaleza de sus estados psicológicos. Persona es algo que ha de tener estados mentales y conciencia de sus propios estados mentales. En un conocido pasaje señala Locke:

> [Una persona] es un ente pensante inteligente que posee razón y reflexión, y que puede considerarse a sí mismo como sí mismo, la misma cosa pensante (*thinking thing*) en diferentes tiempos y lugares; lo que hace sólo por aquella conciencia (*consciousness*) que es inseparable del pensamiento y —según me parece— es esencial a este: es imposible para alguien percibir sin percibir que percibe. (Locke, 1689, *Essay*, II, chap. 27, § 9).

Es persona, de acuerdo a Locke, un sujeto de estados mentales que han de ser transparentes para el mismo sujeto[1]. La 'mente' ha designado aquel ámbito 'interior' constituido por los pensamientos, creencias, preferencias, emociones, sentimientos, sensaciones y percepciones que cada uno de nosotros, desde la perspectiva de primera persona, posee y sabe que posee, así como el sujeto de tales estados. Desde la perspectiva propuesta por Locke, entonces, una persona es esencialmente una *mente*. No es extraño que se haya propuesto que las condiciones de identidad de tales personas estén fundadas en la cualidad intrínseca de los estados mentales de esas mentes. En efecto, como es bien conocido, en la llamada concepción "psicológica" de la identidad personal —la concepción heredera de Locke— dos personas en distintos tiempos son la misma persona si y sólo si son psicológicamente continuas. La continuidad psicológica está constituida por la continuidad de creencias, preferencias, intenciones, rasgos de carácter y, especialmente, la memoria de haber sido sujeto de estados mentales en el pasado. Esta memoria es lo que Locke denomina *consciousness*.

---

[1] Es obvio que hay aquí influencias cartesianas. Descartes es famoso por haber sostenido que somos una *res cogitans*, pero no tiene un tratamiento explícito sobre qué sea una persona. Cf. Descartes, *Meditatio secunda*, AT VII, 28: *"Sed quid igitur sum? Res cogitans. Quid est hoc? Nempe dubitans, intelligens, affirmans, negans, volens, nolens, imaginans quoque, & sentiens."* Por supuesto, no se puede proyectar esta concepción de la persona como una mente a todos los pensadores modernos. Se trata, sin embargo, de la idea que ha resultado dominante en las discusiones del siglo pasado y cuyo origen se encuentra en estos grandes referentes filosóficos del siglo XVII.

Una vez que se acepta este punto de vista lockeano, incluso las resistencias a la teoría psicológica de la identidad personal son motivadas por lo que se cree que es una mejor comprensión de qué sea una mente. Por ejemplo, si los estados mentales están fundados en estados físicos del cerebro —o son reducibles a estos estados—, entonces la identidad personal tendrá que estar fundada en los estados en que pueda estar un cerebro en distintos tiempos (cf. Olson, 2007). Una posición así, aún cuando aparece corrigiendo la teoría psicológica, sigue pensando que somos esencialmente mentes. Se trata de que la mente y los estados mentales poseen una naturaleza que impone restricciones a la continuidad psicológica. Otros han propuesto que la identidad en el tiempo no está fundada en la continuidad de los estados psicológicos, sino que debe suponerse como algo no reducible a otros hechos, pues la mente es un sujeto inmaterial simple (cf. Swinburne, 1984, 2012; Zimmerman, 2012; Madell, 2015). Tanto quienes proponen reducir o fundar la identidad personal a la identidad del cerebro, como quienes proponen no reducir la identidad personal a algo ontológicamente más básico, sin embargo, siguen asumiendo que una persona es una mente. Las coordenadas establecidas por Locke para plantear la cuestión sobre cuáles son las condiciones de identidad de una persona se mantienen.

Lo que se va a proponer aquí es que esta premisa lockeana es un error. Una persona no es una mente. No somos esencialmente mentes. Por lo menos, se intentará mostrar que la identificación de las personas con mentes es algo controversial, una tesis metafísica que requiere una justificación adicional que pocas veces se ha dado. En lo que sigue, se va a hacer primero un breve excurso acerca de los antecedentes de la noción de 'persona' en la tradición filosófica. Luego se explicarán algunas razones para preferir la perspectiva clásica más bien que la lockeana. Esto ayudará a entender por qué son también preferibles las teorías 'animalistas' contemporáneas. Tal como se va a mostrar, el 'animalismo' está en continuidad con esa tradición filosófica.

2. La hipóstasis de una naturaleza racional

El término "persona" comenzó a ser utilizado para designar lo que somos esencialmente desde fines de la Antigüedad clásica a propósito de las polémicas trinitarias y cristológicas. Una formulación de la noción, de acuerdo a las funciones teóricas que viene a cumplir, es la que hace Boecio: "una persona es la sustancia individual de una naturaleza racional" (Boecio, *De persona* III 1343 C13-D1: *persona est naturae rationalis individua substantia*). El término "persona" se propone como equivalente del griego *hypostasis*[2]. En las declaraciones dogmáticas de los concilios de Nicea y Constantinopla I (DH 125-126, 150-151) se había establecido la unidad de *ousía* entre el Padre, el Hijo y el Espíritu Santo —son *omousion*— y su diferencia como *hypostaseis*. En las declaraciones dogmáticas de los concilios cristológicos de Efeso y Calcedonia (DH 250-251, 300-303) se había establecido que Jesucristo es una *hypostasis* pero con dos *ousiai* o *physeis*: una divina y una humana. Una 'persona' es exactamente lo que en estas formulaciones se está designando como *hypostasis*. Se puede ver que los dogmas trinitario y cristológico obligan a introducir una distinción entre la 'naturaleza' y la 'persona'. Esta distinción se da entre la persona de Jesucristo y sus naturalezas humana y divina, y se da también en cada uno de nosotros entre nuestra naturaleza humana y la persona que somos. Con posterioridad, los latinos inventaron una expresión como equivalente del griego *hypostasis*, siguiendo estrechamente su etimología: *suppositum*[3]. Santo Tomás de Aquino

---

[2] "La subsistencia individual de una naturaleza racional la han llamado <los griegos> con el nombre hypostasis" (Boecio, *De persona* III 1344 A5-7: *naturae rationalis individua subsistentiam* hypostáseôs *nomine vocaverunt*).

[3] La introducción de este neologismo se hizo necesaria especialmente porque *substantia* es una expresión que adolece de cierta ambigüedad. En algunos casos se utiliza para designar lo mismo que designa *suppositum*, pero en otras ocasiones designa lo que designan *natura* o *essentia*. Cf. Santo Tomás de Aquino, *Summa* I, q. 29, a. 2, c.: "*substantia dicitur dupliciter. Uno modo dicitur substantia quiditas rei, quam significat definitio, secundum quod dicimus quod definitio significat substantiam rei: quam quidem substantiam Graeci usiam vocant, quod nos essentiam dicere possu-*

indica que una persona es una 'sustancia individual' en cuanto es "un sujeto o *suppositum* que subsiste en el género de la sustancia" (*Summa* I, q. 29, a. 2, c.: *subiectum vel suppositum quod subsistit in genere substantiae*). Hay varias características importantes que debe poseer una persona de acuerdo a una concepción en estas líneas.

En primer lugar, una persona ha de ser una 'sustancia particular'. Es característico de una entidad de este tipo ser algo que persiste en el tiempo siendo idéntica en distintos instantes, o ser algo temporalmente extendido, aunque simple. Una misma sustancia puede, por esto, poseer determinaciones incompatibles entre sí en diferentes tiempos. Puede creer que p en el tiempo t_1, pero luego no creer que p en el tiempo t_2. Esta característica de una sustancia particular puede parecer obvia, pero contrasta con algunas de las posiciones que se han llegado a defender en la discusión acerca de las condiciones de identidad en el tiempo. Precisamente, las dificultades que han enfrentado las teorías psicológicas han conducido a postular que las personas sean entendidas como fusiones de 'partes temporales' de duración instantánea (cf. Lewis, 1983). Cada una de tales partes temporales de persona sería una persona de vida infinitesimalmente corta. Lo que concebimos normalmente como una persona completa con un curso de vida temporalmente distendido sería lo que resulta de fusionar tales partes temporales de persona. Otras concepciones menos radicales han sostenido que no es realmente la identidad en el tiempo lo que interesa cuando se trata de la duración temporal de una persona, sino una relación más débil entre las 'etapas personales' de *supervivencia* que podría, eventualmente, darse entre una persona en un tiempo y varias personas diferentes y simultáneas en otro tiempo (cf. Parfit, 1971, 1986). Nada de esto sería siquiera inteligible en la perspectiva clásica. Cuando se trata de una persona, sencillamente no tiene

---

mus. Alio modo dicitur substantia subiectum vel suppositum quod subsistit in genere substantiae". El dogma trinitario es que Dios es uno en *substantia*, *natura* o *essentia* y trino en *personae* o *supposita*. Algo semejante sucede con la expresión griega *ousía* que puede designar el sujeto de los accidentes (*hypokeimenon*) o la esencia (*tò tí ên eînai*). Cf. Aristóteles, *Metafísica* VII, 3-6.

sentido hablar de sus 'partes temporales'. Sócrates no posee partes temporales. Si se quiere, hay algo así como partes de la *vida de Sócrates*, que es un evento conformado por todo lo que le sucede a Sócrates desde el principio hasta el fin de su vida, pero esto es una entidad numéricamente diferente de Sócrates y dependiente ontológicamente de él.

En segundo lugar, una sustancia ha de ser un sujeto ontológicamente independiente de las propiedades accidentales que pudiese estar instanciando. Hay varias concepciones en competencia acerca de qué sean las propiedades de un objeto (cf. Allen, 2016). Algunos sostienen que se trata de universales que poseen una naturaleza tal que pueden existir simultáneamente instanciados en diferentes objetos. La instanciación de un universal en un objeto particular se ha denominado un "estado de cosas", y es una entidad ontológicamente dependiente tanto del objeto como del universal que lo integran. Otros sostienen que las propiedades son entidades particulares, tan particulares como el objeto que están determinando. Estas propiedades particulares han recibido diferentes nombres en la historia de la filosofía. Actualmente se las denomina "tropos", pero en la tradición filosófica se las ha conocido como "accidentes" o "modos". Otros filósofos nominalistas han rechazado por completo la existencia de propiedades. Las funciones normalmente atribuidas a estas propiedades son satisfechas por clases de *semejanza* de objetos, o clases *naturales* de objetos, o alguna otra construcción teórica. Nótese que cualquiera sea la alternativa que se adopte, un 'estado mental' debe ser algo ontológicamente dependiente del sujeto de atribución. Ya sean estados de cosas dependientes de un universal, o tropos, o sea la semejanza del objeto de que se trate con otros con los que conforma una clase de semejanza, los estados mentales dependen del objeto que los posee. Y una persona es el sujeto o sustrato particular del que dependen sus propiedades. Se puede ver, entonces, que para la concepción clásica no tendría sentido pensar en las personas como 'cúmulos de estados de mentales' o 'secuencias de estados mentales conectados entre sí' (cf. Olson, 2007). Ninguna 'impresión sensible' podría existir de manera independiente de lo que quiera que sea su sujeto.

En tercer lugar, una hipóstasis posee una *naturaleza* o *esencia* que funda ontológicamente sus condiciones de identidad en el tiempo y entre diferentes mundos posibles. El desarrollo que pueda tener una sustancia particular en el tiempo queda restringido por el tipo de sustancia que es algo. Por ejemplo, algo que sea un 'gato', precisamente por serlo sólo puede persistir en el tiempo dentro de ciertos márgenes determinados. Un gato crece y se desarrolla de modos característicos. A un gato no pueden crecerle alas, ni podría reproducirse por mitosis. Una sustancia particular de un cierto tipo posee un 'perfil modal' acerca de lo que podría haberle sucedido aunque no le haya sucedido de hecho. Un gato, por ejemplo, no podría haber sido un cocodrilo o una bacteria. Cualquiera sea la forma en que pudiesen ser las cosas —lo que se ha denominado un "mundo posible"— una sustancia particular de un tipo de específico o especie no podría variar el ser una sustancia de tal tipo. Un ser humano no podría no ser un ser humano. Un caballo no podría no ser un caballo. Esto es lo que se ha denominado el "sortal" de un objeto (cf. Wiggins, 2001; cf. Lowe, 2009)[4]. Dado que el sortal tiene una función ontológicamente determinante de la evolución temporal de una sustancia y de lo que podría acaecerle a esa sustancia en mundos posibles diferentes del mundo actual, el sortal resulta temporalmente invariante para una sustancia y, además, invariante entre diferentes mundos posibles.

Es una cuestión abierta si *ser persona* es, por sí mismo, un sortal (cf. Kanzian, 2012). Se trataría, tal vez, de un 'esquema de sortal', pues una persona ha de poseer algún u otro sortal específico. Una persona debe ser un ser humano, o un ángel, o Dios. Esto vale de

[4] De acuerdo a Wiggins un concepto sortal puede ser caracterizado por la satisfacción de varias restricciones teóricas. Por ejemplo: "a = b si y sólo si existe un concepto sortal f tal que: (1) a y b caen bajo f; (2) decir que x cae bajo f o de que x es un f es decir que x existe (en el sentido que Aristóteles ha fijado); (3) a es el mismo f que b, esto es, coincide con b bajo f en el modo de coincidencia requerido para miembros de f (...)" (Wiggins, 2001, 56). El término "sortal" fue introducido por Locke (cf. *Essay*, III, chap. 3, § 15), pero corresponde a lo que Aristóteles denomina una *deutera ousía* (sustancia segunda; cf. *Categorías*, 5, 2a 14-18).

manera general para toda sustancia particular, que ha de poseer un sortal específico u otro que gobierne su persistencia en el tiempo y su identidad en diferentes mundos posibles. En todo caso, ninguna sustancia particular o hipóstasis podría dejar de serlo en algún tiempo. Ninguna sustancia podría pasar a ser un accidente, por ejemplo. Ser una sustancia es temporal y modalmente invariante para una sustancia. Lo mismo sucede para una persona, precisamente por ser una sustancia particular. En la discusión que se hará más abajo se va a tratar el 'ser persona' como un sortal por estas características modales y temporales. Lo mismo se hará respecto del carácter de 'ser una mente' para la misma discusión.

En cuarto lugar, lo que diferencia a una persona de otros *suppositа* es que una persona es la hipóstasis de una naturaleza racional. Que una sustancia particular posea una naturaleza racional no implica que tal sustancia posea en todo tiempo de su existencia —y en todos los mundos posibles en que exista— capacidades cognitivas desarrolladas para el pensamiento con contenido conceptual y decisiones intencionales libres. Mucho menos se requiere que tal sustancia tenga en todo tiempo actos de pensamiento y de voluntad consciente. Una naturaleza racional es un principio intrínseco que está dirigido al desarrollo de tales capacidades, si es que no es impedido por factores externos, o no se frustra por alguna enfermedad o la falta de los elementos del tipo requerido para que la sustancia pueda florecer. Así, una sustancia de naturaleza racional podría no llegar nunca a tener actos de pensamiento. Podría morir antes de estar en condiciones de ello. O puede suceder que una sustancia de naturaleza racional sufra una enfermedad grave que coarte sus capacidades cognitivas. El hecho de que tales capacidades no puedan desplegarse, sin embargo, no hace que la sustancia deje de tener su naturaleza propia. Por ejemplo, una araucaria podría ser cortada al tener un tamaño muy pequeño y dejar de crecer. Esto no haría, sin embargo, que la araucaria no sea una araucaria, tendida por un principio intrínseco a desarrollarse de la forma característica en que lo hacen tales árboles. Si no fuese impedida la tendencia de desarrollo, llegaría a crecer como árbol. Se puede apreciar aquí que una naturaleza racional no es una mente. Tampoco lo es, en-

fáticamente, una persona entendida en los términos que han sido expuestos —como la hipóstasis de una naturaleza racional—. Una mente está constituida sólo por estados mentales, *v. gr.*, estados de los que uno ha de estar inmediatamente consciente desde la perspectiva de primera persona. A lo más puede estar designando parte del curso de la vida de una persona humana. No hay mente si alguien está afectado por un coma profundo. No hay mente si alguien posee capacidades cognitivas demasiado incipientes, como sucede con un niño pequeño. No hay mente cuando alguien está durmiendo —al menos en varias de las fases del sueño.

En quinto lugar, una persona humana es la hipóstasis de una naturaleza animal. Esto no vale, naturalmente, para personas angélicas. En esta concepción, todo ser humano es esencialmente un animal y, por ello, un viviente. Por supuesto, no es cualquier animal, ni cualquier viviente, pues a las potencias características de un ser vivo o de un animal se añaden las potencias específicamente humanas ligadas a la racionalidad. Es la misma sustancia, sin embargo, la que es un ser vivo, un animal y un ser racional. Todos los actos de esa sustancia son actos que deben ser atribuidos a la misma hipóstasis. Los latinos decían *actiones sunt suppositorum*: las acciones son del *suppositum*. La actividad propia de la 'persona' entendida de este modo, por lo tanto, incluye cosas como la digestión, la circulación sanguínea, la síntesis de hormonas, el crecimiento del pelo y también, por supuesto, el pensamiento racional.

3. Por qué no somos mentes

Desde la última década del siglo pasado se ha estado proponiendo la idea de que las condiciones de identidad de una persona —o 'nuestras' condiciones de identidad— son las condiciones de identidad de un organismo biológico (cf. Snowdon, 1990, 2014; Olson, 1997, 2007, entre otros). Esta posición ha sido conocida como "animalismo" (*animalism*). Aunque de entrada el animalismo puede parecer una concepción muy opuesta a lo que hayan pensado Boecio o Santo Tomás de Aquino, se trata de una recuperación de la tradición filosófica en la que se inscriben tales autores.

En la tradición filosófica, tal como se ha explicado, una persona humana es una hipóstasis de una naturaleza animal y racional. No es, por esto, nada de extraño que las condiciones de identidad que rigen nuestra persistencia en el tiempo sean las condiciones de identidad de un animal. Los defensores del animalismo han puesto de relieve, por ejemplo, que cada uno de nosotros es idéntico a algo que en un tiempo pasado fue un feto (cf. Olson, 1997). No podríamos identificarnos con un feto en un tiempo pasado si no pudiésemos conectar el sujeto pensante que somos ahora con un organismo biológico que no lo es. Los animalistas han puesto de relieve, también, que admitir una persona pensante diferente de un animal humano implicaría postular *dos pensadores* co-localizados espacial y temporalmente. Para el defensor de la perspectiva psicológica las personas no son animales. El animal humano que coincide con nosotros en un instante dado en que uno tiene un pensamiento, es también capaz de realizar el mismo pensamiento. Así, cada pensamiento tendría dos sujetos: la persona y el animal humano, lo que resultaría absurdo (cf. Olson, 2003, 2007).

El animalismo ha resultado difícil de aceptar para muchos teóricos. Tal vez lo que explique en parte esto es que el animalista no está sosteniendo que las condiciones de identidad de una persona sean de tipo biológico porque, por ejemplo, la mente se reduzca a hechos biológicos o deba ser eliminada por hechos biológicos. El animalista no está ofreciendo una concepción diferente de la mente. Lo que el animalista está haciendo es rechazar que *seamos* mentes. Desde la perspectiva animalista cuál sea la naturaleza de la mente es irrelevante para la cuestión de qué es lo que seamos y, con ello, para la cuestión de qué sea una persona.

Me serviré ahora de lógica de primer orden con identidad y sortales, de acuerdo a la notación introducida por Jonathan Lowe (cf. Lowe, 2009). La expresión "*a*/*F* " se debe leer como "*a* es un *F* ", donde "*F* " designa un sortal. Una predicación ordinaria, en cambio, se expresará, como es usual, como "*Ga*". Aquí "*G*" puede o no estar designando un sortal. Aunque se ha explicado arriba que está abierto a discusión si el carácter de *ser persona* es o no un sortal, será útil utilizar la notación introducida por Lowe

para destacar el carácter ontológicamente fundamental, necesario y temporalmente invariante que tiene el *ser persona*. Lo mismo vale para el *ser una mente* de acuerdo a las corrientes herederas de Locke. El predicado "ser una persona" se designará por "*P*" y el predicado "ser una mente" se designará por "*M*". La forma usual en que se trata la noción de 'persona' es como designando qué es lo que somos en el nivel más fundamental, esto es:

(1) $\forall x\,((x = \text{yo}) \rightarrow x/P)$

Este enunciado (1) debe verse, en realidad, como un 'esquema' que expresa diferentes proposiciones en diferentes contextos de uso. En especial, el valor semántico de "yo" en un contexto en que un hablante *b* profiere (1) es *b*. (1) debe entenderse como enunciando que, si hay algo que soy yo o, si se quiere, si hay algo que es idéntico conmigo, esto es una persona[5]. Tal como se ha explicado arriba, es característico de un sortal que tiene una función regulativa para la persistencia en el tiempo y para la identidad en diferentes mundos posibles. Por esto, (1) implica —manteniendo fijo, naturalmente, el contexto de uso:

(2) $\Box\forall x\,((x = \text{yo}) \rightarrow Px)$

Aquí "□" es el operador modal de necesidad para hacer una atribución esencial a lo que quiera que sea que designa "yo" en el contexto de uso. Lo que (2) enuncia es que necesariamente, si hay algo que yo soy, eso es una persona. Nótese que en (2) la atribución '*Px*' es una atribución ordinaria, no sortal. De un modo semejante, (1) implicaría —también manteniendo fijo el contexto de uso:

(3) $\forall t\forall x\,((x =_t \text{yo}) \rightarrow Px\text{-en-}t)$

---

[5] Nótese que, aunque el enunciado (1) es una cuantificación universal que está indicando algo que debe valer para todos los objetos del dominio de cuantificación, a lo más uno de esos objetos puede satisfacer el antecedente, pues no más de un objeto del dominio de cuantificación puede ser idéntico conmigo. Yo no podría inteligiblemente ser idéntico con dos o más objetos diferentes entre sí.

En este enunciado (3) la variable '*t*' tiene como rango tiempos, la expresión "$=_t$" es la restricción de la identidad al tiempo *t* y la expresión "*Px*-en-*t*" es la restricción de '*Px*' al tiempo *t*. Lo que se enuncia en (3), entonces, es que, si hay algo que yo soy, eso es una persona en todo tiempo en que yo exista.

Es característico de la perspectiva lockeana acerca de la naturaleza de una persona y de la identidad personal que a (1) se agrega:

(4) $\forall x\,((x = \text{yo}) \rightarrow x/M)$

Este enunciado (4) nuevamente debe verse como un 'esquema' con valores semánticos variables de acuerdo a los contextos de proferencia. (4) sólo difiere de (1) en que se sustituye el 'ser persona' por 'ser una mente'. Pues bien, es característico de una posición animalista, así como de la tradición filosófica, el rechazo de (4). Considérese, en efecto, que de un modo análogo a lo que sucede con (1) respecto de (2) y (3), el enunciado (4) debería implicar —manteniendo fijo el contexto de uso:

(5) $\Box\forall x\,((x = \text{yo}) \rightarrow Mx)$

(6) $\forall t\forall x\,((x =_t \text{yo}) \rightarrow Mx\text{-en-}t))$

Esto es, deberíamos cada uno de nosotros ser esencialmente mentes y deberíamos ser mentes en todo tiempo de nuestra existencia. Es obvio, sin embargo, que tanto (5) como (6) son falsos. En este caso se produce una situación peculiar, pues aunque (4) parezca un enunciado que difícilmente podría ser falso al ser proferido por alguien, o al ser objeto de pensamiento por alguien, sus consecuencias sí parecen ser falsas. En efecto, cualquier sujeto que sea un hablante y esté en condiciones de efectuar aseveraciones debe ser también alguien con estados mentales de algún tipo. Parece, por lo tanto, obvio que cualquier sujeto que pueda proferir o juzgar sinceramente que (4) ha de ser una mente y ha de ser consciente de ser una mente. Esto debería ser tan obvio para cada uno como que uno existe al pensar algo. Sucede, sin embargo, que lo que se está enunciando en (4) es más fuerte que simplemente

que uno es una mente: lo que se está aseverando es que el *sortal* de cada uno de nosotros es ser una mente, lo que implica que el ser una mente es algo que determina ontológicamente nuestras condiciones de identidad entre diferentes tiempos y diferentes mundos posibles[6]. Por esto es que (4) implica (5) y (6), que tienen que ver precisamente con tales consecuencias modales y temporales.

El enunciado (5) parece falso, en primer lugar, pues claramente yo podría no haber sido una mente. Lo mismo puede constatar cada uno considerando, desde su respectiva perspectiva de primera persona, qué le hubiese o no podido acaecer. Si yo hubiese sufrido una enfermedad invalidante muy pequeño o si hubiese muerto muy pequeño jamás hubiese llegado a ser una mente. Para ser una mente, uno debe haber desarrollado ciertas capacidades cognitivas superiores. En una edad muy temprana yo no las tenía y podría no haberlas desarrollado nunca. El enunciado (6) también parece falso. Hay tiempos en los que yo existía y no poseía ninguna de las capacidades cognitivas requeridas para ser una mente. Tal vez existan tiempos en el futuro en los que seguiré existiendo, pero con mis capacidades cognitivas tan dañadas o disminuidas que ya no seré más una mente. No sé en qué sentido pueda decirse que soy una mente, además, si estoy en un periodo de sueño profundo o bajo el efecto de una anestesia total. Ciertamente en cualquiera de esas situaciones ni poseo estados mentales, ni tengo conciencia de tener estados mentales y de mi propia existencia.

Un defensor de la perspectiva lockeana objetará, naturalmente, que estas intuiciones acerca de la falsedad de (5) y (6) están presuponiendo que no somos esencialmente mentes, que es precisamente lo que está en discusión. Sólo si uno asume que uno no es esencialmente una mente puede uno contemplar escenarios en los que uno existe y no es una mente. Esto es efectivo. Es importante constatar, sin embargo, que nuestras intuiciones ordinarias acerca

---

[6] Lo que se está enunciando en (4) es que [x/M] no que [Mx], para un valor adecuado de la variable x. Recuérdense las prevenciones indicadas arriba para considerar las nociones de 'ser persona' y de 'ser una mente' como sortales.

de qué somos son desde ya intuiciones que van contra la suposición central de la tradición lockeana. Por supuesto y como sucede en otras áreas, estas no son razones inatacables, pero imponen una presunción a favor de la concepción clásica continuada ahora por el animalismo. Para cualquiera, en efecto, es obvio que uno ha existido en tiempos en que uno no era una mente, y podría ahora no ser una mente. Cualquiera que quiera sostener que estas intuiciones son inadecuadas y que, por lo tanto, tenemos que hacer una reforma drástica de nuestra manera de comprendernos a nosotros mismos, tiene la carga de la argumentación.

4. El problema de la 'identidad personal'

Tal como se ha indicado arriba, las disputas sobre ontología de la persona se han concentrado en la cuestión sobre las condiciones de identidad de las personas en el tiempo y, en especial, sobre la teoría psicológica de la identidad personal. La influencia de Locke ha sido responsable de esta concentración, pues gran parte de los trabajos se han orientado a mejorar y refinar, o a criticar y sustituir la teoría lockeana. Pues bien, el debate a favor y en contra de la teoría psicológica tiende a dejar implícitos sus supuestos. En efecto, cuáles sean las condiciones de identidad de un tipo de entes está fundado en cuál sea su naturaleza. Esto hace que la dilucidación de cuáles sean estas condiciones de identidad sea una cuestión filosófica habitualmente fructífera, pero también hace que se trate de una cuestión derivativa respecto de los problemas más centrales acerca de la naturaleza de un tipo de entes. Sucede también que la posibilidad de entregar condiciones de identidad informativas para los entes de un dominio presupone que esos entes sean ontológicamente dependientes de otros. Obviamente, no todo ente es dependiente de otros. Si se trata de entidades independientes, no hay condiciones de identidad informativas que entregar. La forma general de una especificación de condiciones de identidad es la siguiente:

$$(7) \qquad \forall x \forall y \, ((Fx \wedge Fy) \rightarrow ((x = y) \leftrightarrow Rxy))$$

Esto es, para entidades de tipo *F*, *x* e *y*, *x* es idéntico a *y* si y sólo si hay una relación *R* entre *x* e *y*. El bicondicional entre [$x = y$] y [Rxy] es, de por sí, neutral respecto del orden de dependencia entre lo que se describe en su lado derecho y lo que se describe en su lado izquierdo. Es trivial que si [$x = y$] entonces se sigue que hay infinitas sustituciones válidas de *R* que hacen (7) verdadera. Por ejemplo, si [$x = y$], entonces *x* e *y* son *el mismo F.* En cualquiera de estos casos, sin embargo, la relación entre *x* e *y* está fundada en la identidad [$x = y$]. Lo que hace filosóficamente interesante un principio con la forma de (7), sin embargo, es si hay prioridad ontológica de [Rxy] respecto de [$x = y$]. Esto es, se trata de especificar los hechos que *determinan* ontológicamente la identidad entre entes del tipo *F*. Un ejemplo clásico es el principio de extensionalidad para conjuntos:

$$(8) \qquad \forall x \forall y \, ((x = y) \leftrightarrow \forall z \, ((z \in x) \leftrightarrow (z \in y)))$$

Se supone en (8) que las variables '*x*' e '*y*' tienen como rango conjuntos. Lo que se enuncia aquí es que dos conjuntos son idénticos si y sólo si poseen exactamente los mismos elementos. Un conjunto depende ontológicamente de qué elementos posea, pues su esencia está conformada por cuáles sean aquellos[7]. La dependencia de los conjuntos en sus elementos permite especificar condiciones de identidad informativas tal como las enunciadas en (8).

---

[7] Como es habitual, se supone aquí que la 'dependencia ontológica' es una noción primitiva que no admite ser analizada en términos de otras nociones más familiares. La dependencia ontológica entre entidades *x* e *y* implica [$\Box((x$ existe) $\rightarrow$ (y existe))], esto es, que necesariamente la existencia de *x* (la entidad derivativa) debe ir acompañada por la existencia de *y* (la entidad ontológicamente prioritaria). Pueden darse casos en que exista una covariación modal entre dos entidades sin que esto tenga que ver con la dependencia de una de ellas respecto de la otra. La existencia de cualquier cosa está acompañada modalmente, por ejemplo, por la existencia del número 3, pero esto no hace que cualquier cosa sea dependiente del número 3. La dependencia ontológica es, por lo demás, un orden estricto irreflexivo, asimétrico y transitivo.

Cuando se trata de la identidad personal en el tiempo, la teoría psicológica ha pretendido hallar los hechos ontológicamente básicos que fundan la identidad personal en la continuidad psicológica. Esto es, para variables 'S_1' y 'S_2' que tienen como rango personas existentes en los tiempos t_1 y t_2 respectivamente:

(9) $\forall S_1 \forall S_2 ((S_1\text{-en-}t_1 = S_2\text{-en-}t_2) \leftrightarrow (S_2\text{-en-}t_2$ es psicológicamente continua con $S_1\text{-en-}t_1)$

Si este va a ser un principio aceptable para las condiciones de identidad en el tiempo se requiere que la persistencia en el tiempo de las personas dependa ontológicamente de la continuidad psicológica entre los estados mentales de las personas involucradas en esos tiempos. Esto genera problemas inmediatos que han sido advertidos muy pronto por los críticos de la teoría psicológica (cf. Butler, 1736; Reid, 1785). Una relación formalmente aceptable para ser una sustitución de *R* en (7) debe ser una relación de equivalencia, pues la identidad lo es, esto es, debe ser una relación reflexiva, simétrica y transitiva[8]. La continuidad psicológica, de por sí, no es ni simétrica, ni transitiva. Supóngase que Juan-en-2018 es psicológicamente continuo con Juan-en-2016, pues Juan-en-2018 recuerda haber pensado en 2016 que hay muchos gatos en el vecindario, junto con poseer una razonable continuidad en creencias, preferencias y rasgos de carácter. Juan-en-2016, naturalmente, no puede recordar lo que vaya o no a pensar Juan-en-2018. El recuerdo de los estados mentales que se han poseído sólo funciona retrospectivamente hacia el pasado, pero no hacia el futuro. Juan-en-2018 puede tener continuidad psicológica con Juan-en-2016 por los motivos indicados, y luego puede haber continuidad psicológica entre Juan-en-2016 y Juan-en-1980, pues Juan-en-2016 recuerda haber pensado en 1980 que hay pocos gatos en el ve-

[8] La identidad, en efecto, puede ser caracterizada como la relación de equivalencia más 'pequeña', pues implica toda otra relación de equivalencia. Si 'Q' es una relación de equivalencia cualquiera, vale que $[\forall x \forall y ((x = y) \rightarrow Qxy)]$. Si falla Q para objetos x e y cualesquiera, entonces se sigue que $[x \neq y]$.

cindario. Esto no garantiza, sin embargo, que Juan-en-2018 sea continuo psicológicamente con Juan-en-1980. Juan-en-2018 puede haber olvidado completamente qué estados mentales poseía en 1980. Puede suceder también que Juan-en-2018 sea psicológicamente continuo con Juan-en-1980, pero no con Juan-en-2016, aún cuando Juan-en-2016 sea también psicológicamente continuo con Juan-en-1980.

La continuidad psicológica es, entonces, una relación que *prima facie* es muy inapropiada para fundar la identidad de personas. Pero esto no ha arredrado a los defensores de la teoría psicológica. Hay formas de construir lógicamente una relación de equivalencia a partir de una relación diádica cualquiera, aún cuando no sea ni simétrica ni transitiva. La reflexividad no es problema en este caso, como es obvio, pues trivialmente cualquier sujeto de estados mentales en un tiempo es psicológicamente continuo consigo mismo en ese tiempo. Para una relación diádica Q que no sea simétrica se puede definir la relación disyuntiva $[Qxy]$ ó $[Qyx]$. Para una relación diádica no transitiva Q se puede definir el 'ancestral' de Q. La relación ancestral de Q se da entre x e y si y sólo si $[Qxy]$, o hay un z tal que $[Qxz]$ y $[Qzy]$, o hay z_1, z_2, tal que $[Qxz_1]$, $[Qz_1z_2]$ y $[Qz_2y]$, o hay z_1, z_2, …, z_n tal que $[Qxz_1]$, $[Qz_1z_2]$, …, $[Qz_ny]$, etc. Esto es, dos objetos x e y están conectados por el ancestral de Q si y sólo si hay una secuencia de relaciones Q que conectan a x e y. Por lo tanto, una teoría psicológica viable en términos de continuidad psicológica debería formularse haciendo apelación al ancestral de la suma lógica de la continuidad psicológica y su conversa. Sea Ψ la relación asimétrica y no transitiva de continuidad psicológica. La conversa de Ψ se designará como "Ψ^{-}". El ancestral de una relación R se designará como "R^{*}". Entonces, manteniendo las mismas variables que en (9), la teoría psicológica se debería formular como:

$$(10) \qquad \forall S_1 \forall S_2 \, [(S_1\text{-en-}t_1 = S_2\text{-en-}t_2) \leftrightarrow (\Psi(S_1\text{-en-}t_1, S_2\text{-en-}t_2) \vee \Psi^{-}(S_2\text{-en-}t_2, S_1\text{-en-}t_1))^{*}]$$

Se puede ver, entonces, que los defensores de la teoría psicológica han ido refinando el concepto de 'continuidad psicológica' para resolver estas dificultades (cf. Grice, 1941; Quinton, 1962) y otras bastante más serias. Por ejemplo, el recuerdo de que uno ha tenido de un estado mental en el pasado debe estar fundado en la identidad entre el sujeto que recuerda haber tenido un estado mental en el pasado y el sujeto que en el pasado ha tenido tal estado mental (cf. Butler, 1736; un famoso intento de respuesta en Shoemaker, 1970). ¿Cómo podrían esos estados de recuerdo ser lo que funda la identidad en el tiempo de una persona, si su corrección depende de esta misma identidad? Otra dificultad es que la identidad es una relación no-vaga que determinadamente se da o determinadamente no se da. La continuidad psicológica, en cambio, es vaga (cf. para una discusión, Lewis, 1983). El problema más célebre y más discutido, sin embargo, es el que tiene que ver las 'fisiones' y 'fusiones' de personas. La identidad es una relación uno-a-uno. Un objeto no puede ser idéntico a dos objetos diferentes entre sí. Si, por hipótesis, $b_1 = b_2$ y $b_1 = b_3$, entonces, por transitividad no podría ser que $b_2 \neq b_3$. Parece metafísicamente posible, sin embargo, que una misma persona sea psicológicamente continua con dos personas futuras diferentes entre sí. Si la identidad personal está fundada en la continuidad psicológica, entonces estos deberían ser casos en que una persona es idéntica a dos personas futuras diferentes entre sí, lo que es absurdo. Son estos problemas los que han motivado las drásticas reformas propuestas por Parfit (cf. 1971, 1986; cf. Lewis, 1983), quienes han sustituido la identidad en el tiempo por una relación mucho más débil de 'supervivencia' o por la relación mereológica de 'ser parte de'.

No es necesario hacer una discusión de estas dificultades. Lo que interesa destacar, en cambio, es que uno podría rechazar la teoría psicológica de la identidad personal sin necesidad de rechazar la tesis de acuerdo a la cual 'ser una mente' es nuestro sortal propio. Ha sido habitual que filósofos de inclinaciones dualistas —que siguen identificándonos con una mente— hayan rechazado que la continuidad psicológica sea un buen candidato para las condiciones de identidad de personas, pero para estos filósofos las di-

ficultades indicadas arriba para la teoría psicológica son una razón para adoptar una concepción dualista de la mente (cf. Swinburne, 1984). La forma en que se razona para sacar esta conclusión es bien indicativa. Hay dos premisas centrales para ello. En primer lugar, o bien nuestras condiciones de identidad son primitivas —esto es, no fundadas en otros hechos más básicos ontológicamente— o bien no lo son. Si no lo son, es porque nuestras condiciones de identidad deben estar fundadas en hechos psicológicos o en hechos físicos. Pero ni hechos de continuidad psicológica ni hechos físicos son adecuados para entregar condiciones de identidad personal en el tiempo[9]. Las dificultades indicadas arriba son una justificación suficiente para sostener tal cosa. Esto debería implicar que nuestras condiciones de identidad en el tiempo son 'simples' o no fundadas en otros hechos más básicos. ¿Por qué concluir de esto que *debemos* ser una sustancia puramente pensante diferente de la sustancia física, el cuerpo? Pues, porque somos mentes esencialmente. Y esta premisa no es objeto de discusión. Parece tan obvia que la falla en las teorías reductivas —y de la teoría psicológica, en especial— se ven como motivos para adherir de manera inmediata a una forma de dualismo. Nótese que estas mismas motivaciones que empujan a alguna forma de dualismo ante las dificultades de la teoría psicológica se convierten en motivos para aceptar la teoría psicológica, cueste lo que cueste. Si uno no está dispuesto a aceptar por ningún motivo el dualismo sobre la mente, entonces uno se verá motivado a insistir con la teoría psicológica, aún cuando sea completamente increíble como una teoría de la *identidad* de las personas en el tiempo. Si no hay más remedio, entonces se declara que 'no importa' la identidad, o que nada es idéntico en distintos tiempos. Esto es exactamente lo que han hecho Derek Parfit y David Lewis.

---

[9] La concepción de acuerdo a la cual nuestras condiciones de identidad son 'físicas' no ha sido objeto de discusión, pues tradicionalmente se la ha puesto siempre en un lugar subordinado respecto de la teoría psicológica. Parece obvio que nuestra persistencia en el tiempo no está fundada en la persistencia de algún objeto físico que nos constituya. Es propio de cualquier ser vivo estar modificando permanentemente los objetos físicos que lo constituyen en diferentes tiempos.

El 'animalismo', así como la concepción clásica de la persona no deben verse como teorías que están ofreciendo otros candidatos de condiciones de identidad para personas en el tiempo. Estas teorías están rechazando el gran presupuesto de toda la discusión sobre las condiciones de identidad: que somos esencialmente mentes. Postular que somos esencialmente animales —aunque dotados normalmente de capacidades racionales— no es todavía ninguna teoría específica acerca de cuáles sean nuestras condiciones de identidad. Tal vez existan condiciones de identidad sustantivas para animales y otros organismos biológicos. Tal vez la identidad en el tiempo de organismos biológicos sea un hecho primitivo, ni fundado en otros, ni reducible en otros. La concepción clásica —de acuerdo a la cual somos la hipóstasis de una naturaleza animal y racional—, así como el animalismo son posiciones neutrales respecto de cuáles sean tales condiciones de identidad.

5. Conclusiones

Se ha sostenido en este trabajo que la tradición que ha sido inaugurada por Locke acerca de la identidad personal y la ontología de la persona ha dependido de un supuesto central: que somos esencialmente mentes. Esto es, ser una mente sería nuestro sortal propio, de lo que se seguiría que nuestras condiciones de identidad en el tiempo y nuestras condiciones de identidad entre diferentes mundos posibles tendrían que estar regidas por nuestro carácter de ser mentes. Incluso muchas teorías que han pretendido corregir la teoría psicológica de la identidad personal —o sustituir por completo tal teoría— han dependido del mismo supuesto central.

Esta concepción típicamente moderna contrasta fuertemente con la concepción de una persona —esto es, de lo que somos en el nivel más fundamental— en la tradición filosófica anterior. La idea de que "persona" designa lo que cada uno es más propiamente es la idea de que una 'persona' es la hipóstasis o *suppositum* de una naturaleza racional. Tal hipóstasis llegará normalmente a poseer estados mentales en algunos tiempos, pero es obvio que no en todos los tiempos de existencia una persona tendrá tales estados, así como

no necesariamente una persona tendrá estados mentales. Cuando se trata de una persona humana, sus condiciones de identidad no son las de una mente conformada por estados internamente escrutables en primera persona, sino las de un *animal racional*, que es, en efecto, un *animal*. Es, por esto, interesante constatar cómo las corrientes animalistas defendidas en los últimos decenios son una continuación de nuestra venerable tradición filosófica.

No se ha hecho en este trabajo una discusión detallada de la multitud de argumentaciones que han sido propuestos a favor y en contra, tanto de la concepción lockeana psicológica de la identidad personal, como de las teorías animalistas. Pero sí se han presentado algunas razones intuitivas de un carácter muy general por las que una concepción en las líneas de la tradición filosófica resulta con mucho más verosímil que cualquiera de sus alternativas lockeanas. Parece obvio que hay tiempos en que no somos mentes y parece obvio que podríamos no haber sido mentes. Tales intuiciones son incompatibles con que seamos esencialmente mentes[10].

Referencias bibliográficas

Allen, S. (2016). *A Critical Introduction to Properties*. London: Bloomsbury.

Aristóteles (2009). Categorías. En J. Mittelmann (ed.), *Sobre la interpretación.* Buenos Aires: Losada.

Aristóteles (1998). *Metafísica.* En T. Calvo (ed.), *Traducción, introducción y notas.* Madrid: Gredos.

Santo Tomás de Aquino (1952). *Summa theologiae.* Cura et studio Petri Caramello con textu ex recensione leonina.

Boecio (1847). De Persona. Liber de persona et duabus naturis contra Eutychen et Nestorium ad Iohannem diaconum ecclesiae Romae. En J-P. Migne (ed.), *accurante, Patrologiae*

[10] Este trabajo ha sido redactado en ejecución del proyecto de investigación Fondecyt 1160001 (Conicyt, Chile).

cursus completus, Series Prima, Tomus LXIV, 1337-1354. Parisiis: venit apud editorem in via dicta d'Amboise.

Butler, J. (1736). The Analogy of Religion. En J. Perry (ed.), *Personal Identity*, pp. 99-105. Berkeley: University of California Press.

Descartes, R. (1979). Méditations métaphysiques. Objections et réponses suivies de quatre lettres. En J-M. Beyssade y M. Beyssade (eds.), *Chronologie, présentation et bibliographie*. Paris: Flammarion.

Gasser, G., Stefan, M. (eds.) (2012). *Personal Identity. Complex or Simple?* Cambridge: Cambridge University Press.

Grice, H.P. (1941). Personal Identity. En J. Perry (ed.), *Personal Identity*, pp. 73-95. Oxford: Oxford University Press.

Kanzian, C. (2012). Is 'Person' a Sortal Term? En G. Gasser y M. Stefan (eds.), *Personal Identity. Complex or Simple?* pp. 192-205. Cambridge: Cambridge University Press.

Lewis, D. (1983). Survival and Identity. En A. Oksenberg Rorty (ed.), *The Identities of Persons*, pp. 16-40. Berkeley: The University of California Press.

Locke, J. (1689). An Essay Concerning Human Understanding. En P.H. Nidditch (ed.), *Edited with an Introduction*. Oxford: Clarendon Press.

Lowe, E. (2009). *More Kinds of Being. A Further Study of Individuation, Identity, and the Logic of Sortal Terms*. Oxford: Wiley-Blackwell.

Madell, G. (2015). *The Essence of the Self. In Defense of the Simple View of Personal Identity.* London: Routledge.

Olson, E. (1997). *The Human Animal. Personal Identity Without Psychology.* New York: Oxford University Press.

Olson, E. (2003). An Argument for Animalism. En M. Raymond y J. Barresi (eds.), *Personal Identity*, pp. 318-334. Oxford: Blackwell.

Olson, E. (2007). *What are we? A study in Personal Ontology.* Oxford: Oxford University Press.

Parfit, D. (1971). Personal Identity. En J. Perry (ed.), *Personal Identity*, pp. 199-223. Berkeley: University of California Press.

Parfit, D. (1986). *Reasons and Persons.* Oxford: Oxford University Press.

Perry, J. (ed.) (2008). *Personal Identity*. Berkeley: University of California Press.

Quinton, A. (1962). The Soul. En J. Perry (ed.), *Personal Identity*, pp. 53-72. Berkeley: University of California Press.

Reid, T. (1785). Of Memory. En J. Perry (ed.), *Personal Identity*, pp. 113-118. Berkeley: University of California Press.

Shoemaker, S. (1970). Persons and their Pasts. En S. Shoemaker (ed.), *Identity, Cause, and Mind Philosophical Essays*, pp. 19-48. Oxford: Clarendon Press.

Snowdon, P. (1990). Persons, Animals, and Ourselves. En C. Gill (ed.), *The Person and the Human Mind: Issues in Ancient and Modern Philosophy*, pp. 83-107. Oxford: Clarendon Press.

Snowdon, P. (2014). *Persons, Animals, Ourselves.* Oxford: Oxford University Press.

Swinburne, R. (1984). Personal Identity: The Dualist Theory. En S. Shoemaker y R. Swinburne (eds.), *Personal Identity*, pp. 368-384. Oxford: Blackwell.

Swinburne, R. (2012). How to Determine Which is the True Theory of Personal Identity. En G. Gasser y M. Stefan (eds.), *Personal Identity. Complex or Simple?* pp. 105-122. Cambridge: Cambridge University Press.

Wiggins, D. (2001). *Sameness and Substance Renewed.* Cambridge: Cambridge University Press.

Zimmerman, D. (2012). Materialism, Dualism, and 'Simple' Theories of Personal Identity. En G. Gasser y M. Stefan (eds.), *Personal Identity. Complex or Simple?* pp. 206-235. Cambridge: Cambridge University Press.

Sobre el autor

José Tomás Alvarado Marambio es profesor asociado en el Instituto de Filosofía de la Pontificia Universidad Católica de Chile. El profesor Alvarado realiza investigación en Metafísica analítica; filosofía del lenguaje y epistemología. Contacto: jalvaram@uc.cl

Capítulo 2

Una noción contextual de emergencia y un ejemplo en análisis de redes neuronales

Esteban Céspedes y Rubén Herzog

Resumen

¿Cómo pueden los estados emergentes de un sistema estar determinados por los elementos más básicos del mismo, pero, al mismo tiempo, ser funcionalmente irreducibles a ellos? Para abordar esta pregunta, que ha estado presente desde los primeros estudios en torno al concepto de emergencia, proponemos que los estados emergentes son reducibles e irreducibles a la vez, dependiendo de una noción contextual de emergencia. Además, según nuestra propuesta, debe haber un contexto donde la observación de un estado emergente involucre novedad y esté correlacionada con un cambio abrupto en la complejidad asociada al sistema en cuestión. Ejemplificando nuestra propuesta con el caso del modelamiento matemático de redes neuronales, consideramos como estado emergente el conjunto de patrones generado por una red. Este conjunto, en ciertos contextos, no puede ser reducido a los estados de sus partes más básicas (neuronas), mientras que, en otros, sí es reducible a sus partes más básicas y a las interacciones de a pares entre neuronas. Así, ejemplificamos una noción contextual de emergencia que permite dar cuenta de los aspectos novedosos, reducibles e irreducibles del sistema en observación.

Palabras Clave: propiedades emergentes, complejidad, irreducibilidad, redes neuronales.

1. Introducción

Las propiedades emergentes, en un sentido amplio, son propiedades que se observan en un sistema compuesto de elementos que interactúan entre sí. Un aspecto crucial de las mismas es que, a pesar de poder tener conocimiento acabado de dichos elementos, no podemos deducir a partir de ellos descripciones que involucren propiedades emergentes. La propuesta que presentamos aquí requiere de i) la noción de complejidad, asociada a la forma no trivial en la que interactuarían los elementos del sistema y ii) la noción de irreductibilidad asociada a las propiedades o estados emergentes. Utilizando la noción de reducción funcional, según la cual un estado puede ser reducido a otro estado siempre y cuando ambos tengan el mismo rol causal, nos enfrentamos al llamado desafío de consistencia: ¿Cómo pueden los estados emergentes *E* estar determinados por los elementos más básicos *B* pero, al mismo tiempo, ser irreducibles funcionalmente a ellos? Para abordar este desafío que ha estado presente desde los primeros estudios en torno al concepto de emergencia, proponemos que los estados emergentes son reducibles e irreducibles a la vez, dependiendo de una noción contextual de emergencia. Introducimos la noción de contexto epistémico, que establece un orden de relevancia de un conjunto de expresiones, con la cual definimos emergencia contextual siempre que i) haya un contexto en que *E* se pueda reducir a *B*, ii) haya un contexto donde *E* no se puede reducir a *B* y iii) haya un contexto donde la observación de *E* incluye novedad y está correlacionada con un cambio abrupto en la complejidad asociada a *B*, pudiendo asociar *E* y *B* a una misma clase de referencia. Esta definición nos lleva a recaracterizar los fenómenos y propiedades emergentes, precisando también la noción de novedad y los tipos de reducción asociados a distintos contextos. Finalmente, ejemplificando nuestra propuesta con el caso del modelamiento matemático de redes neuronales, consideramos como estado emergente el conjunto de patrones generado por una red. Este conjunto, en ciertos contextos, no puede ser reducido a los estados de sus partes más básicas (neuronas), mientras que, en otros, sí es posible reducirlo a sus partes más básicas y a las interacciones entre pares

de neuronas. Además, presentamos un contexto en el que, si las redes son muy grandes, es necesario considerar interacciones de orden superior, lo que sugeriría un rol causal descendente de los estados de nivel superior hacia los estados de las partes más básicas. En ese mismo contexto introducimos una noción de complejidad sensu lato que nos permitirá mostrar cómo el estado *E* de la red neuronal se vuelve irreducible cuando hay un cambio abrupto en la complejidad y cómo, sin embargo, cuando la complejidad no varía abruptamente, el sistema puede ser reducido fácilmente a los estados de sus componentes básicos.

2. Propiedades emergentes y algunos problemas

Las propiedades emergentes pueden ser entendidas como características de un sistema que surgen desde las partes constituyentes del mismo pero que a la vez son irreducibles a tales partes. Uno de los conceptos clásicos de propiedad emergente fue desarrollado por el filósofo británico Charles Dunbar Broad (1925), cuyas caracterizaciones podrían ser concentradas y planteadas de la siguiente manera.

2.1 Propiedad emergente. Una propiedad *F* de un sistema *S* es una propiedad emergente siempre y cuando (cf. Beckermann, 1992, p. 17):

(2.1.1) los constituyentes de *S* interactúen entre sí de una forma determinada,

(2.1.2) exista una ley, según la cual todos los sistemas que poseen la misma organización o estructura que *S* exhiben *F* y,

(2.1.3) ninguna descripción de *F* pueda ser deducida de un conocimiento acabado de las propiedades de los componentes de *S*.

La primera condición menciona (quizás implícitamente) que las propiedades emergentes aparecen en sistemas con cierto nivel

de complejidad, es decir, en sistemas cuyas partes interactúan de manera relevante[1]. No estamos simplemente ante un montón de cosas inconexas (ver Figura 1).

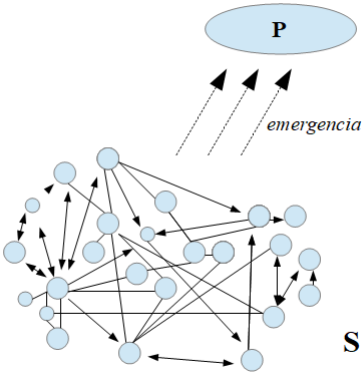

Fig. 1

Según la condición (2.1.2), existe cierta dependencia entre una propiedad emergente y las partes del sistema. Lo relevante de esta dependencia no se encuentra en las características de las partes consideradas individualmente, sino en cómo están organizadas.

Así, pensando en la tercera condición (2.1.3), decimos que no es posible deducir una descripción de la propiedad emergente en cuestión considerando únicamente las propiedades de los elementos particulares del sistema del que surge. En este sentido, ésta es una condición de irreducibilidad.

En lo que nos interesa ahora, no sólo podemos usar el término "propiedad emergente", sino también el de "emergencia", cuyo concepto correspondiente puede definirse (dependiendo de la definición de propiedad emergente) al menos de dos formas.

[1] Surge inmediatamente una pregunta sobre la noción de relevancia: ¿Con respecto a qué es relevante una relación de interacción? Como veremos más adelante, la noción de contexto epistémico será fundamental para responder esta pregunta. Por ejemplo, un tipo de interacción será considerada relevante, según un contexto, si es observable en ese contexto.

Podemos hablar de emergencia refiriéndonos al *proceso* en el cual aparece una propiedad emergente o bien podemos referirnos a la *relación* que existe entre las partes constituyentes de un sistema y sus propiedades emergentes. Por ejemplo, podemos decir que las propiedades de un tornado surgen como resultado de procesos que involucran moléculas de aire y de agua. En este caso, consideramos la emergencia como un proceso (o un conjunto de procesos). Alternativamente, podemos decir que las propiedades del tornado son emergentes con respecto a las propiedades de un conjunto de moléculas interactuando de formas determinadas. Aquí consideramos la emergencia como una relación.

Es necesario pensar en un problema fundamental asociado a la noción de emergencia. Aun cuando es ampliamente aceptado que tanto la irreducibilidad como la complejidad son nociones necesarias para una correcta definición del concepto de propiedad emergente, no parece haber claridad sobre cómo deberían estar relacionadas esas nociones en tal tipo de definición.

Según algunas propuestas recientes, las propiedades emergentes son propiedades que surgen en un sistema cuando su complejidad varía de forma abrupta (cf. El-Hani y Pereira, 2000). Pero, dependiendo de la idea de complejidad que adoptemos, esta condición no parece ser suficiente para entender la relación de emergencia. En principio, es posible describir un sistema considerando sólo niveles básicos y encontrar cambios abruptos en su complejidad sin mencionar las propiedades de niveles superiores que surgirían de la misma. En otras palabras, una propiedad emergente es algo que ha de ser posible observar, detectar o al menos singularizar lingüísticamente. Y en esto radica la condición de irreducibilidad.

Según una noción funcional de irreducibilidad, decimos que la eficacia causal de los estados emergentes va más allá de la eficacia causal de las partes del sistema de las cuales surgen. Esta noción puede ser definida como lo hacemos a continuación.

2.2 Reducción funcional. Un estado *A* es reducido a un estado *B* siempre y cuando el rol causal de *A* sea el mismo que el rol causal de *B* (cf. Kim, 1997).

De esta manera, la irreducibilidad de los estados emergentes puede sustentar la idea de que éstos poseen poderes causales que muestran cierta independencia de los poderes causales de las partes constituyentes del sistema[2]. Las crisis económicas, los estados cognitivos y las extinciones ecológicas poseen efectos que no son explicables simplemente en términos de los comportamientos microeconómicos, de los procesos neuronales y de las interacciones de los miembros de un ecosistema. Un estado emergente puede tener efectos en su mismo nivel, como cuando un estado mental es causa de otro. Podemos decir que este tipo de casos son casos de *causalidad horizontal*. Un estado emergente puede también tener efectos en niveles más bajos, como cuando una crisis económica afecta el comportamiento de los consumidores. Llamamos a esto *causalidad descendente*. Además, los estados constituyentes pueden influir en niveles superiores, como cuando una decisión política desata una crisis social. Podemos clasificar estos casos como casos de *causalidad ascendente*. Una teoría apropiada sobre estados emergentes debe dar buena cuenta de estos aspectos.

El caso de la causalidad descendente es especialmente problemático en los debates sobre emergencia, como nos muestra el muy conocido argumento por exclusión (cf. Kim, 2005). Supongamos que todo evento que posee una causa, posee una causa física y que las causas físicas suficientes para explicar un evento excluyen otro tipo de causas. Además, asumamos que no hay casos genuinos de sobredeterminación causal, es decir, que si un evento E tiene una causa física suficiente C, entonces no existe otro evento diferente de c que pueda ser considerado una causa suficiente de E. Ahora bien, si es posible considerar que un estado emergente E es causa de otro estado macro E^*, entonces E debería ser capaz de determinar causalmente también los estados que fundamentan E^* desde un nivel inferior. Pero esto choca con la primera de las asunciones mencionadas: Dichos estados inferiores estarían determinados

[2] Por ahora, nos centramos sólo en la idea de reducción funcional, ya que creemos que no está sujeta a tantos supuestos metafísicos como otras, que consideraremos brevemente más adelante.

tanto por *E* como por los estados que fundamentan *E*. Así, manteniendo los supuestos planteados, el estado emergente *E* parece no tener eficacia causal descendente.

Tenemos, considerando lo brevemente expuesto hasta aquí, dos problemas principales asociados a la caracterización de los estados emergentes (cf. Kim, 2006). Primero, es preciso aclarar en qué sentido los estados emergentes son irreducibles. Segundo, pensando que podemos optar por la irreducibilidad funcional, es preciso aclarar en qué sentido poseen eficacia causal. Estas cuestiones implican lo que podemos llamar el desafío de consistencia (cf. Sartenaer, 2016).

> *2.3 Desafío de consistencia*. Es difícil comprender cómo los estados emergentes están determinados por los elementos básicos del sistema en el que surgen y, al mismo tiempo, son irreducibles a ellos.

Creemos que éste es un desafío fundamental que ha estado presente desde los primeros estudios en torno a la noción de estado emergente y que, probablemente, expresa un intento de responder a los principales obstáculos que han impedido un desarrollo serio de esta noción. El desafío de consistencia nos mueve, junto con los problemas mencionados arriba, a proponer aquí una caracterización con la que podríamos intentar abordarlo satisfactoriamente. Según nuestra propuesta, los estados emergentes son reducibles e irreducibles a la vez. Por supuesto, esto es contradictorio en un primer momento. Deberemos reformular esta caracterización gruesa con el fin de poder incluirla en una teoría coherente.

3. Emergencia contextual

Para disolver la aparente contradicción señalada en 2.3, acudiremos a la noción de contexto. Así, podremos decir que un estado emergente es reducible e irreducible a la vez, pero bajo distintos respectos. Diremos que un contexto es un conjunto de expresiones en función de las cuales podemos evaluar enunciados acerca de estados emergentes. No queremos decir que en un contexto deben

estar descritas todas las condiciones objetivas que servirían para describir un estado emergente. En principio, sólo considerar las condiciones conocidas en un determinado momento de una investigación bastará para nuestros intereses. Así, nos enfocaremos en lo que podemos llamar *contextos epistémicos*. No queremos tomar en cuenta sólo el conjunto de expresiones epistémicas asociadas a un periodo dado de una investigación. Según la noción comprendida aquí, cada contexto lleva consigo, además de sus elementos epistémicos, un ordenamiento de relevancia funcionando sobre cada uno de ellos. En este sentido, un contexto es más que un conjunto de expresiones: Es un conjunto de expresiones ordenadas por relevancia.

Debemos notar que hablamos de expresiones en un sentido general. Así, un contexto no será entendido sólo como un conjunto de enunciados descriptivos ordenados, sino como un conjunto que puede incluir también enunciados y expresiones no proposicionales referidas a normas metodológicas, intereses e incluso experiencias subjetivas[3]. Considerando todo esto, podemos ahora caracterizar la noción de contexto epistémico de la siguiente forma.

---

[3] Podemos referirnos, mediante una expresión nominal, a una experiencia subjetiva sin describirla. Podemos asumir, por ejemplo, que φ simboliza la experiencia de una determinada persona al ver el mar por primera vez. Dada la riqueza cualitativa de las experiencias subjetivas, parece plausible pensar que ningún conjunto de descripciones podrá capturar lo que tal persona está viviendo en ese momento. Por supuesto, esto no quiere decir que no podamos describirla. Supongamos que para describir φ tenemos que usar expresiones como "Ella sentía que". Podemos interpretar esta expresión como un operador, es decir, como una expresión no-proposicional. Estas movidas son también aplicables al considerar normas metodológicas o intereses (p. ej. "Es necesario redefinir este concepto" o "Necesitamos una muestra representativa"). Así, el contexto epistémico de un equipo de investigación es también un buen ejemplo. En éste serían relevantes los intereses de los investigadores o sus experiencias de vida y cómo éstas los conducen hacia su investigación, además de las normas académicas, políticas y sociales en las cuales está embebida.

3.1 Contexto epistémico (K). Un contexto epistémico es una tupla <*R*, *Q*>, donde *Q* es un conjunto de expresiones (descriptivas, normativas o nominales) ordenado mediante un conjunto de funciones de relevancia *R*.

Podemos postular que las expresiones de *Q* para un contexto dado pueden estar referidas a otros contextos o a elementos de otros contextos. Con esto, *R* podrá también, para un contexto dado, determinar ordenamientos de contextos (o sea, expresiones sobre contextos).

Sea *K* el contexto epistémico desde el que consideramos si un estado de un sistema es emergente o no. Además, simbolicemos mediante K_1, K_2 y K_3 tres contextos a los cuales *K* puede referirse. Ahora podemos definir la noción de estado emergente.

3.2 Emergencia contextual. Según *K*, un conjunto de estados simbolizados por *E* es emergente a partir de otro conjunto de estados simbolizados por *B* siempre que se cumplan las siguientes condiciones.

(3.2.1) Hay un contexto K_1, según el cual *E* es reducido a *B*.

(3.2.2) Hay un contexto K_2, según el cual *E* es irreducible a *B*.

(3.2.3) Hay un contexto K_3, según el cual *E* involucra propiedades que son nuevas en contraste con *B* y cuya observación está correlacionada con un cambio abrupto en la complejidad asociada a *B*, de tal manera que es posible asociar *E* y *B* a una misma clase de referencia[4].

[4] Esta caracterización del concepto de estado emergente es compatible con otras nociones contextuales de emergencia, como las que ha desarrollado Bishop en colaboración con Atmanspacher (2006). Nuestra propuesta ha sido influenciada en medida importante por sus trabajos. Sin duda, sería importante considerar detalladamente las diferencias conceptuales entre estas perspectivas y la nuestra, pero es una tarea que preferimos no abordar acá.

Notemos que esto puede ser interpretado como la definición de una noción relacional de estado emergente o bien, alternativamente, como la relación de emergencia. Es decir, la definición nos permite no sólo decir bajo qué condiciones *E* puede ser considerado un estado emergente, sino también bajo qué condiciones podemos decir que *E* emerge a partir de *B*. Además, debemos poner énfasis en los aspectos conceptuales y lingüísticos involucrados en esta definición. En cierto punto, debe ser posible asignar a *E* y *B* un mismo referente. Nos concentraremos en esto en la sección 4.

Además del concepto de emergencia como relación, tradicionalmente hablamos usando las nociones de *fenómeno emergente* y de *propiedad emergente*. Éstas pueden ser entendidas como nociones derivativas de la definición 3.1. Podemos caracterizar así la de propiedad emergente.

> *3.3 Propiedad emergente*. Sea *E* un estado emergente en *K* y *K3*, un contexto considerado desde *K* en el cual es posible observar las nuevas propiedades de *E*. Éstas pueden ser llamadas propiedades emergentes, según *K*.

Continuando, podemos entender de la siguiente manera el concepto de fenómeno emergente.

> *3.4 Fenómeno emergente*. Tanto un fenómeno particular como un tipo de fenómeno correspondiente a la observación de una propiedad emergente puede ser llamado fenómeno emergente, según *K*.

Así, podemos caracterizar la noción de fenómeno emergente sobre la base de la definición de propiedad emergente, lo cual nos permite alcanzar cierta simpleza en el análisis. Por supuesto, podríamos haber definido el concepto de fenómeno emergente sin acudir a (3.3), basándonos únicamente en (3.2). O podríamos haberlo definido directamente, sin acudir ni a (3.3) ni a (3.2), pero en tal caso la formulación sería menos simple que (3.4).

Antes de aplicar estas definiciones a casos concretos, debemos decir algunas cosas más sobre las nociones de novedad y de reducción, partes fundamentales de la definición 3.2. A continuación, definiremos la primera.

> *3.5 Novedad*. Según un contexto *K*, una propiedad *P* es nueva siempre y cuando podamos considerar dos estados S_1 y S_2, tales que *P* sea observable sólo en uno de ambos estados. Si *P* es observable sólo en S_2, diremos que *P* es una propiedad nueva con respecto al estado S_1. Si sólo es observable en S_1, diremos que es una propiedad nueva con respecto a S_2.

Notemos que intentamos definir aquí la noción de novedad con respecto a estados y no con respecto a un orden temporal previamente determinado. Si nos basáramos en un orden temporal, podríamos decir que una propiedad es nueva si hay dos estados, uno previo al otro, tales que la propiedad es observada en el estado posterior, pero no en el anterior. Hay al menos dos razones de por qué preferimos no basarnos en esta caracterización. En primer lugar, nos llevaría a la difícil tarea de buscar un orden temporal fijo. Esta tarea es filosóficamente muy interesante, no queremos descartarla. Pero preferimos no abordarla en el presente trabajo[5]. En segundo lugar, caracterizar la novedad sin tener que aludir a un orden temporal fijo no nos impide aludir a él de todas formas cuando sea conveniente, con el fin de hacer la distinción entre novedad diacrónica y novedad sincrónica, una distinción especialmente relevante para el estudio de la noción de estado emergente (cf. Rueger, 2000). La primera requiere de un orden temporal determinado y puede ser definida como lo acabamos de hacer: Considerando dos estados separados temporalmente, una propiedad nueva es observada en el posterior y no en el anterior. La noción de novedad sincrónica expresa la diferencia entre un estado emergente y los estados básicos del sistema en el cual surge. Pensemos, por ejemplo, en los estados mentales. Un estado mental presenta cuali-

[5] Este orden temporal podría ser establecido en función de los parámetros de *K*, el contexto epistémico general desde el cual es evaluada la atribución de emergencia, siguiendo nuestra definición.

dades que no podemos observar en los estados de las redes neuronales que lo subyacen y que son realizados en el mismo periodo de tiempo. Ahondaremos en este tipo de casos en la sección 5. Estas distinciones nos permiten, además, hablar de novedad sincrónica ascendente y de novedad sincrónica descendente, asumiendo diferentes escalas de los estados observados. Si observáramos primero un estado mental y luego los estados correspondientes a las redes que lo constituyen, estaríamos en un caso de novedad descendente. Si el orden de las observaciones fuera el opuesto, sería un caso de novedad ascendente.

Consideremos ahora brevemente algunas distinciones relevantes que podemos hacer en torno al concepto de reducción. Es posible hacer una primera distinción importante entre las nociones de *reducción ontológica* y *reducción epistémica* (cf. Van Gulick, 2001).

Por un lado, la *reducción ontológica* es una relación que vincula objetos. Al hablar de objetos en este caso, no es necesario pensar que la relación en cuestión o que tales objetos, tal y como puedan ser postulados y caracterizados, son reales con absoluta independencia de cualquier contexto epistémico. Consideremos algunos tipos de reducción ontológica. Un caso es el de *reducción por identidad*. Si en una ontología determinada, *a* es idéntico a *b*, entonces *a* puede ser reducido a *b*. Otro caso es el de la *reducción por composición*. Si *a* está completamente compuesto de *b*, entonces *a* es reducible a *b*. Podríamos considerar, por ejemplo, cómo según ciertas ontologías postuladas en química orgánica los compuestos son reducibles a sus componentes y a cierta estructura que los une. Por otro lado, la *reducción epistémica* es una relación entre representaciones (cf. Van Gulick, 2001). Un caso es el de la *deducción*. Si una proposición *A* es deducible de *B*, entonces *A* es reducible a *B*. Otro es el de la *equivalencia expresiva*. Si *A* representa el mismo conjunto de objetos (la misma extensión) que *B*, entonces *A* es reducible a *B*. Además, la reducción funcional podría ser considerada como un tipo de reducción epistémica si asumimos una noción epistémica de causalidad.

4. Causalidad descendente

George Ellis (2016) propone una noción de estado emergente que no sólo permite la causalidad descendente, sino que también es caracterizable en concordancia con las teorías físicas. En esta sección queremos presentar brevemente la perspectiva de Ellis, mostrando su plausibilidad y de qué manera es coherente con nuestro concepto de emergencia contextual. Una de las bases del argumento central de Ellis (2016) consiste en pensar cómo surgen los estados constituyentes de un sistema y cómo surgen de estos, a su vez, otros estados emergentes.

4.1 Origen de estados. Las estructuras y estados de niveles bajos que generan comportamientos complejos de niveles altos no pueden surgir de una forma puramente ascendente (*bottom-up*). La explicación de esto es que dependen de una conjunción precisa de estructuras de niveles altos.

Algo crucial de estas estructuras de niveles altos es que involucran características funcionales. Éstas pueden estar basadas en objetivos o propósitos y, en este sentido, son llamadas funciones objetivo. Los objetivos pueden ser impuestos extrínsecamente o bien pueden surgir en comportamientos adaptativos. De cualquier forma, las características funcionales permiten organizaciones determinadas de un sistema (es decir, organizan sus niveles inferiores). Un ejemplo de funciones objetivo es el de los acuerdos sociales (Ellis, 2016).

4.2 Acuerdo social. El poder legislativo de un país puede ser organizado de una determinada manera mediante acuerdos sociales. Esta organización es establecida siguiendo ciertos objetivos de una sociedad, postulados de diversas maneras (que dependen del régimen político)[6]. Si bien en un estado democrático

[6] Por supuesto, las fuerzas sociales que interactúan en este tipo de procesos son diversas. Podríamos extender el ejemplo y mencionar movimientos comunales, la corrupción u otros poderes estatales, entre otras. Cada una de estas fuerzas establece funciones objetivo que influyen en el orden legislativo.

> la sociedad puede participar mediante elecciones, el comportamiento del poder legislativo no emerge directamente de las interacciones entre todos los miembros de la sociedad. Pensemos ahora en un grupo de países organizados de esta manera y que interactúan entre sí, generando un sistema de relaciones económicas que puede mostrar un nivel alto de complejidad. El punto es éste: Las partes constituyentes de dicho sistema—entre ellas, los estados y los mercados—no surgen de manera puramente ascendente desde los estados físicos (ni desde el conjunto total de ciudadanos), sino por medio de propósitos de niveles altos, como ha sido señalado recién.

Este caso involucraría, en cierto sentido, funciones objetivo impuestas (es decir, no-adaptativas). Las normas formuladas por el poder legislativo no son otra cosa. Considerando una de las distinciones introducidas anteriormente, podemos decir que este caso muestra cómo la emergencia sincrónica depende de la emergencia diacrónica (cf. Ellis, 2016). Si bien la forma en la que está organizada una sociedad en un período dado puede ser descrita en términos de lo que hacen los ciudadanos en ese periodo, tal organización no puede haber surgido sólo mediante procesos ascendentes. Así, hay un sentido en el que esta organización social, entendida como un estado general emergente, es reducible a sus partes constituyentes (mediante reducción por composición, si queremos) y, a la vez, un sentido en el que es irreducible (pues es resultado de procesos emergentes diacrónicos).

El otro tipo de características funcionales que señala Ellis (2016) es el de las funciones adaptativas. Un buen ejemplo es el siguiente (p. 108).

> *4.3 Información biológica*. La información genética determina, en parte, las formas en las que un organismo puede comportarse y así interactuar con su entorno[7]. Pero al parecer no es

[7] Para evitar malentendidos, es preciso mencionar que la secuencia genética no contiene toda la información sobre toda clase de comportamiento de niveles más altos. En este sentido, quizás sería más apropiado hablar sólo de secuencia genética y no de información genética. Así, usamos

> posible predecir la secuencia de nucleótidos en el ADN a partir de la física o de la microbiología. Tal información es producto de interacciones con el ambiente durante periodos históricos muy largos. El punto central aquí es el mismo que el señalado antes. Las estructuras genéticas pueden ser consideradas como partes constituyentes de un sistema del que emergen comportamientos complejos de niveles altos. Pero tales partes no se constituyen de una manera puramente ascendente, sino que dependen de interacciones con el entorno que involucran causalidad descendente.

Así, en estos casos, si pudiéramos modificar suficientemente la historia de las interacciones entre una especie y su entorno, podríamos modificar también su estructura genética. También, pensando en escalas temporales menores, podemos modificar, por supuesto, la conducta y los hábitos de un individuo realizando modificaciones en su entorno. Estos son ejemplos de influencia descendente desde estados emergentes hacia partes constituyentes del sistema en el que surgen dichos estados. Y, nuevamente, son casos en los que es posible hablar de reducción, en algún contexto epistémico relevante, y de irreducibilidad, según otro contexto epistémico.

Debemos señalar que Ellis da fundamento a estas ideas desde la física y, considerando esto, un contexto podría ser entendido como un conjunto de expresiones sobre circunstancias materiales en medio de las cuales ocurre un proceso emergente, entendido también en términos físicos. Cuando modificamos un entorno ecológico, modificamos materialmente el entorno que habitan determinados individuos; podemos cambiar, por ejemplo, condiciones climáticas, condiciones químicas o condiciones biológicas de otros organismos. Ahora bien, esto no quiere decir que debamos ver esta noción de contexto como completamente desligada de la noción epistémica de contexto introducida en la sección anterior.

---

el término "información biológica" admitiendo sus limitaciones. Cabe mencionar aquí el área de la epigénetica, cuyo objeto es estudiar cómo variables ambientales y conductuales modulan y regulan la expresión de varios genes, sin modificar la secuencia de ADN.

De hecho, una de las caracterizaciones que ofrece Ellis (2016) de la noción de contexto es claramente epistémica (p. 261). Podemos formularla de la siguiente manera.

> *4.4 Contexto de variables*. En una investigación sobre estados emergentes, un contexto puede ser entendido como un conjunto de variables realizadas de nivel alto (macro) que pueden influir en un conjunto de variables realizadas de nivel bajo (micro).

Una variable realizada es una variable junto a la asignación de uno de sus valores posibles. Por ejemplo, podemos considerar la variable *temperatura* y uno de sus valores posibles, 30°C. Si asignamos dicho valor a esta variable podemos tratarla como variable realizada y a su valor como la realización de la variable.

Ahora bien, la noción de contexto sobre la que formulamos nuestra propuesta (3.1) es quizás más general que ésta. Por supuesto, un contexto de variables es un contexto epistémico o puede ser considerado como parte de un contexto epistémico. Pero, en principio, todo contexto epistémico puede ser extendido y convertirse en algo aún más general, si pensamos, por ejemplo, en los criterios de relevancia sobre la base de los cuales queremos seleccionar un conjunto particular de variables macro. Claro, estos mismos criterios podrían ser tratados a su vez en una estructura de variables, las que deberían ser seleccionadas mediante otros criterios aún más generales y así sucesivamente. En algún punto, debido a razones pragmáticas, el contexto puede ser demarcado con cierta claridad, pero difícilmente con absoluta especificidad.

5. Reducción como identidad

La noción de emergencia que proponemos aquí expresa, como hemos visto, no sólo que un estado emergente es un estado que no es reducible y que surge en un sistema que alcanza cierto cambio de complejidad, sino que también es un estado reducible bajo ciertas condiciones. Queremos exponer ahora brevemente la teoría desarrollada por Herbert Feigl (1967), que muestra en qué sentido

es plausible identificar estados mentales con estados neuronales. Esto nos permitirá luego caracterizar un tipo de reducción basada en la noción de identidad que podría formar parte de una interpretación para la definición 3.2. Basándonos en el argumento de Feigl, podemos caracterizar la identidad entre esos tipos de estados de la siguiente manera.

> *5.1 Identidad referencial de lo mental y lo físico*. Sea *M* una expresión correspondiente a un lenguaje de introspección (un lenguaje fenomenológico, si queremos) y *N* un conjunto de expresiones correspondientes a un lenguaje neurocientífico, ambas pueden ser consideradas idénticas si poseen el mismo referente (cf. Feigl, 1967).

Esta idea de identidad nos permite entender la relación de reducción así:

> *5.2 Reducción por identidad*. Es posible reducir una expresión *M* a otra expresión *N* cuando *M* es referencialmente idéntica a *N*.

Según Feigl, cada persona tiene acceso epistémico privilegiado a las propiedades mentales subjetivas (*qualia*), tales como el dolor o los colores. Los términos "dolor" y "rojo" son parte de lo que podemos llamar lenguaje introspectivo o fenomenológico, por medio del cual podemos referirnos a estas cualidades. Además, podemos referirnos a las mismas cualidades mediante expresiones correspondientes a un lenguaje neurocientífico. Este tipo de lenguaje es intersubjetivo y es físico en este sentido muy general, es decir, en la medida en que permite construir conocimiento descriptivo. Con esto, podemos identificar expresiones referidas a lo mental con expresiones referidas a lo físico[8]. Así, ésta es una forma en la que lo mental puede ser reducido a lo físico.

---

[8] Es preciso notar que la identidad de referencia no está sostenida (únicamente) sobre bases empíricas, sino sobre bases lingüísticas, epistémicas y metodológicas. Así, la correferencialidad es postulable desde la correlación empírica, pero esto no significa que la identidad de referencia sea una relación empírica.

Es preciso poner énfasis en que las diferencias entre lo mental y lo físico relevantes para Feigl son diferencias entre lenguajes o sistemas conceptuales y no diferencias ontológicas. Pensando en la terminología asociada a nuestra propuesta sobre la noción de estados emergentes, podemos decir que la diferencia entre lo mental y lo físico es una diferencia entre contextos epistémicos, una diferencia epistémica y contextual.

Considerando estas ideas, podemos caracterizar también una forma de identificar cualquier estado emergente en general con conjuntos de estados micro. Recordemos la definición de emergencia contextual que ofrecimos anteriormente (3.2). Cuando un estado es emergente, consideramos un conjunto de expresiones *B* que describe un sistema y sus partes; una expresión *E*, correspondiente a un contexto epistémico en el que es posible observar nuevas propiedades del sistema; y un contexto según el cual *E* y *B* deben tener el mismo referente. Éste es un sentido en que un estado emergente es reducible a los estados micro de los cuales surge.

6. El caso de las redes neuronales

Sobre la base de las discusiones y definiciones propuestas anteriormente, analizaremos algunos resultados conocidos sobre redes neuronales e intentaremos, de ser posible, proponer contextos epistémicos en los que se pueda hablar de reducción y otros en los que no. Para esto, introduciremos brevemente el marco teórico de las redes neuronales y ciertas formas de representarlas matemáticamente.

Las neuronas son células especializadas que conforman el sistema nervioso en todas las especies del reino animal, formando redes a través de conexiones llamadas sinapsis. Éstas son conexiones físicas en donde las neuronas pueden estimularse eléctricamente y compartir diversos tipos de moléculas[9]. Dentro de los muchos

[9] Existe una corriente importante en neurociencias, según la cual los estados mentales de distintos tipos surgen a partir de estados neuronales.

tipos neuronales que existen, están las neuronas que emiten potenciales de acción, lo que corresponde a un aumento abrupto y transitorio (del orden de 1 milisegundo) en el voltaje intracelular. Estos potenciales de acción pueden ser registrados eléctricamente (desde el espacio juxtacelular) de forma masiva y en paralelo, consiguiendo registros de redes neuronales compuestas por cientos de neuronas, cada una disparando hasta 300 potenciales de acción por segundo. Esta tasa de descarga varía tanto entre neuronas como entre especies.

Con ese tipo de registros, podemos observar un conjunto de potenciales de acción ordenados en el espacio (conjuntos de neuronas que disparan juntas) y en el tiempo (secuencia de disparos). Aquí estableceremos el primer contexto epistémico, llamado *K*. En este contexto, las neuronas son consideradas variables binarias, es decir, que en un instante de tiempo disparan un potencial de acción o están en silencio y la combinación total de patrones espaciales (neuronas disparando juntas) posibles en una red neuronal de N neuronas es igual a 2^N. Este conjunto de patrones es llamado "ráster" en la comunidad neurocientífica. Este ráster puede ser caracterizado por un conjunto de patrones espaciales σ (de dimensión N, una dimensión por neurona) y sus probabilidades asociadas. El conjunto de probabilidades de todos los patrones es llamado distribución de probabilidad $P(\sigma)$ y será considerado como el estado emergente *E*. Sus elementos básicos, *B*, serán las neuronas. Entonces, el contexto de variables incluye, al menos, las neuronas (variables de más bajo nivel) y patrones conformados por ellas (variables de alto nivel). De ser posible alguna reducción, por simplicidad, será sincrónica (pero existen formalismos para reducción diacrónica también).

Si bien no nos enfocaremos aquí en las relaciones de emergencia que pudieran ser establecidas entre tales estados, indicaremos caracterizaciones de emergencia entre estados de las redes neuronales mismas. Estas caracterizaciones sentarían algunas bases para comprender los procesos emergentes —sin duda, mucho más complejos— que ocurren entre estados neurales y estados mentales.

Para considerar $P(\sigma)$ como un estado emergente de la red neuronal, debemos, al menos, establecer una noción de incertidumbre y novedad entre los estados del sistema. Así, podremos reconocer la novedad entre estados microscópicos (las neuronas) y estados macroscópicos ($P(\sigma)$). En el contexto de la teoría de la información y estadística, la entropía estadística cuantifica el nivel de incertidumbre sobre la observación de una variable aleatoria dada. Así, si observamos varias realizaciones de una misma variable aleatoria discreta[10] x, obtendremos una distribución de probabilidad (los valores posibles de x y sus probabilidades asociadas). La información que podemos extraer sobre x, mirando sólo su estadística (es decir, su distribución de probabilidad), es cuantificada por la función de entropía $S(x)$:

$$S(x) = -\sum p(x)\log(p(x)) \qquad (1)$$

Aquí, $p(x)$ es la probabilidad de cierto valor de x y *log* es la función logaritmo. La sumatoria recorre toda la distribución de probabilidad. Esta función es máxima cuando tenemos la mayor incertidumbre sobre x y será cero cuando tengamos total certeza sobre su valor. Entonces, en K, representamos con $P(\sigma_i)$ la distribución de probabilidad de la neurona i y, de existir algo nuevo en el estado emergente $P(\sigma)$ respecto a sus componentes básicos, encontraríamos la siguiente relación:

$$S(\sigma) < \sum_{i}^{N} S(\sigma_i) \qquad (2)$$

Es decir, al observar la estadística del estado emergente $P(\sigma)$ aprendemos más de la red que si observamos sólo la estadística de sus partes básicas: Hay novedad en el estado emergente respecto a observar a las partes por sí solas. Si reemplazamos la desigualdad mostrada en (2) por una igualdad, diremos que sólo mirando la

[10] Asumimos, por simplicidad, que la variable es discreta, pero el concepto también es aplicable a variables continuas.

estadística de las partes ya sabemos la estadística del todo; en este caso, el estado $P(\sigma)$ no es novedoso respecto a observar sus partes básicas y $P(\sigma)$ puede ser reducido a sus partes.

Entonces, de observar novedad en $P(\sigma)$ respecto a sus partes básicas, nos gustaría ser capaces de identificar ciertos roles causales de las variables sobre la base de los cuales sea posible generar la distribución $P(\sigma)$ observada. Este intento puede ser abordado desde los modelos de máxima entropía, en los cuales uno busca generar una distribución $P(\sigma)$ lo más entrópica posible, estadísticamente, pero restringida a reproducir ciertos valores observados empíricamente en el ráster. Estos modelos toman la siguiente forma:

$$P(\sigma) = \frac{\exp(\beta H(\sigma))}{Z} \qquad (3)$$

La función *exp* es la función exponencial, β es la llamada temperatura inversa[11] y Z es la llamada *función de partición* o *constante normalizadora*, que asegura que la suma de todas las probabilidades sea igual a 1. $H(\sigma)$ es la llamada *función de energía* y expresa las restricciones del modelo, es decir, nuestros supuestos sobre los roles causales de las variables que generan $P(\sigma)$. La expresamos como:

$$H(\sigma)=\sum_{l}^{L} h_l m_l \qquad (4)$$

donde h_l son parámetros libres que ajustaremos para reproducir el valor empírico del promedio de la variable m_l. L corresponde al número de parámetros o variables del modelo. En suma, tenemos un parámetro por variable y la elección de las variables quedan a criterio del investigador. Por ejemplo, uno podría construir una función de energía que sólo reprodujera las tasas de disparo (número de potenciales de acción por unidad de tiempo) de las neuronas observadas, quedando de la siguiente forma:

$$H(\sigma)=\sum_{i}^{N} h_i \sigma_i \qquad (5)$$

[11] Se suele tomar que β=1. Más adelante veremos su rol.

En este caso, σ_i corresponde al estado de la neurona i en el patrón σ. Recordemos que, en el contexto K, la neurona puede estar activa o inactiva solamente.

6.1 Contexto de irreducibilidad

Supongamos ahora un contexto K_1, según el cual el rol causal de las neuronas es puramente ascendente, de tal manera que podemos generar $P(\sigma)$ sólo considerando las actividades de las neuronas independientes. Este supuesto podría sostenerse sobre un criterio de eficiencia y economía, pensando que cada neurona hace algo diferente (son independientes), de tal manera que se optimizan los recursos metabólicos del organismo. En ese contexto, una función de energía como la mostrada en (5) sería la elección. Como muchos resultados muestran (Gardella, Marre, y Mora, 2019), un modelo como éste falla dramáticamente al reproducir $P(\sigma)$ para distintos tipos de redes neuronales, lo que sugiere, al menos, que las neuronas no son independientes en su actividad y que $P(\sigma)$ no puede ser construida de forma puramente ascendente. Así, en el contexto K_1, el estado emergente E no puede ser reducido sólo a la actividad de los elementos básicos B.

6.2 Contexto de reducción

Por otro lado, asumamos que en un contexto K_2 los investigadores piensan que, dado que una red neuronal es también un tejido biológico y parte de un animal, las neuronas no pueden ser totalmente independientes, considerando sus condiciones comunes. Es decir, podría haber causalidad descendente desde las propiedades del animal (por ejemplo, ritmos metabólicos, historia de desarrollo y evolución) hacia las propiedades de las neuronas. En este contexto, los supuestos de los investigadores incluyen tanto descripciones sobre la actividad de cada neurona individual como también sobre las interacciones de pares de neuronas. La función de energía queda expresada de la siguiente forma:

$$H(\sigma)=\sum_{i}^{N} h_i\sigma_i + \sum_{i<j}^{N} J_{ij}\sigma_i\sigma_j \qquad (6)$$

donde h_i es el mismo de la ecuación (5); J_{ij} es un parámetro libre que se ajusta para reproducir el promedio del disparo en conjunto entre la neurona i y la neurona j; y $\sigma_i\sigma_j$ corresponde a que la neurona i y la neurona j estén activas juntas en el patrón σ. En este caso, las interacciones son simétricas, es decir, $\sigma_i\sigma_j = \sigma_j\sigma_i$, y por esto la sumatoria está expresada asumiendo que $i<j$. Así no se repite ninguna interacción. En la literatura de mecánica estadística esta función de energía es una generalización del modelo de Ising[12]. Para redes en las que $N = 40$ neuronas, por ejemplo, estos modelos son capaces de reproducir $P(\sigma)$ con mayor precisión (respecto al modelo (5), observada en varios tipos de redes neuronales. En este contexto, podríamos decir que $P(\sigma)$ es reducida por composición a las neuronas y a sus interacciones[13]. Aquí incluimos un nuevo conjunto de variables de segundo orden: las interacciones de a pares.

6.3 Contexto de correlación

Ahora, en un contexto K_3, donde es posible registrar más neuronas en paralelo, asumiendo, por ejemplo, que $N = 100$, un modelo como (6) comienza a fallar. En específico, conociendo $P(\sigma)$, es fácil derivar $P(k)$, la probabilidad de k neuronas disparando juntas en el mismo patrón. $P(k)$ reflejaría un rol causal descendente, donde interacciones de alto orden influyen sobre elementos de bajo orden. Es decir, la actividad de las neuronas ahora está causada, en parte, por su tendencia a disparar en conjunto. Al comparar el

---

[12] El modelo de Ising representa el comportamiento colectivo de un material ferromagnético, donde el estado de cada partícula ferromagnética (llamado espín) depende tanto de fluctuaciones térmicas como del estado de sus vecinos más cercanos espacialmente. En el caso de las neuronas, se consideran todas las interacciones, no sólo las interacciones entre las vecinas más cercanas.

[13] Dada la naturaleza cuantitativa del estudio, la reducción se define en torno a un umbral de precisión.

modelo con los datos, se observa que un modelo como (6) subestima, en especial, la probabilidad (observada en el ráster) de muchas neuronas disparando juntas, es decir, $P(k)$ cuando k se acerca a N. Esto ha sido considerado como un fallo del modelo al reproducir las interacciones de alto orden (interacciones entre más de dos neuronas).

Por lo mismo, los investigadores pueden considerar una nueva forma de la función de energía, que contiene explícitamente un término asociado a $P(k)$:

$$H(\sigma)=\sum_{i}^{N} h_i\sigma_i + \sum_{i<j}^{N} J_{ij}\sigma_i\sigma_j + V(k) \qquad (7)$$

donde los dos primeros términos de la ecuación corresponden con los de (4) y $V(k)$ es una función que se ajusta para reproducir $P(k)$. De esta forma, el modelo incluye i) la actividad de cada neurona, que les permitirá incluir la diversidad individual, ii) las interacciones entre pares de ellas, que les permitirá inferir la topología funcional de la red y iii) $P(k)$, que les permitirá incluir interacciones de orden mayor. Utilizando (5), los investigadores son capaces de reproducir $P(\sigma)$ y $P(k)$ con mayor precisión que usando (4), para casos de redes más grandes ($N > 100$). Así, en K_3 es posible reducir (según cierto umbral de precisión) el estado emergente E (aquí representado por $P(\sigma)$ a las partes básicas del sistema, a sus interacciones de a pares y a la tendencia de la red de tener k neuronas activas.

Volviendo a la ecuación (3), ahondaremos un poco en el rol de β, que simboliza lo que llamamos temperatura inversa ($\beta = 1/(T\,\mathrm{kB})$, donde T es la temperatura del sistema y kB la constante de Boltzmann). Durante el proceso de ajustar los parámetros para reproducir los valores observados en el ráster, β es considerado igual a 1, ya que no se tiene una noción clara de cuál sería la temperatura en este sistema. Sin embargo, una vez que hemos obtenido los parámetros del proceso de ajuste, podremos realizar simulaciones computacionales donde es posible generar muchas $P(\sigma)\beta$ mediante el uso de los parámetros inferidos, pero multiplicados por diferentes

valores de β. Así, simulamos la actividad de redes neuronales que contienen la misma estructura interna (sus parámetros), pero donde todos los parámetros son escalados por este factor β.

Así, recordando la condición (3.2.3) de la definición (3.2), que exige la correlación entre la aparición del estado E y un cambio abrupto en la complejidad del sistema, introduciremos la noción de variabilidad de $P(\sigma)$. Llamaremos *Var log*$(P(\sigma))$ a la variabilidad de los valores de *log*$(P(\sigma))$, la cual será nuestra aproximación *sensu lato* a la complejidad del sistema. Esta relación se puede establecer sobre la base de que si *Var log*$(P(\sigma))$ es muy pequeña, quiere decir que todos los patrones tienden a tener la misma probabilidad, indicando muy poca complejidad del sistema y que el sistema está en máxima entropía. Por otro lado, si *Var log*$(P(\sigma))$ es muy grande, quiere decir que el sistema explora un amplio rango de valores de $P(\sigma)$, exhibiendo patrones con muy alta probabilidad como otros con muy baja. Finalmente, si el sistema visita con alta probabilidad muy pocos patrones, y visita al resto de los patrones con muy poca probabilidad (pero similar entre ellos), *Var log*$(P(\sigma))$ sigue siendo pequeña, ya que la mayoría de los patrones tienen la misma probabilidad (aunque baja). Entonces, *Var log*$(P(\sigma))$ es baja cuando el sistema visita con la misma probabilidad todos los estados, como también cuando visita muy pocos estados preferentemente, mientras que se hace alta cuando $P(\sigma)$ toma un amplio rango de valores, es decir, $P(\sigma)$ presenta una estructura estadística no trivial. En suma, considerando la laxa relación entre la función *Var log*$(P(\sigma))$ y la idea de complejidad, podemos variar β para modificar $P(\sigma)$ y así poder explorar cómo varía la complejidad respecto a un parámetro de control (ver Figura 2). Esto nos permitiría acercarnos a la noción de cambio abrupto en la complejidad del sistema.

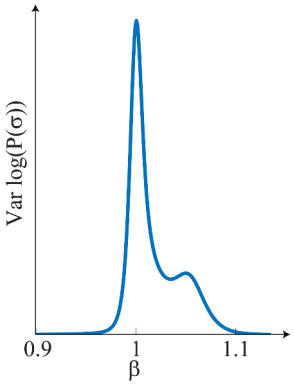

Fig. 2

Por ejemplo, para redes de 100 neuronas, se observa un máximo de complejidad cuando β es cercano a 1 (el valor con el cual se obtuvieron los parámetros desde los datos), mientras que la complejidad decae a medida que nos alejamos de β=1 (ya sea subiendo o bajando dicho valor).

Por un lado, tener valores muy altos de β (es decir, una temperatura simulada muy baja) hace que todos los patrones con una o más neuronas activas tengan muy baja probabilidad (y requieran de mucha energía), haciendo que el sistema tienda a explorar sólo el patrón en el que todas las neuronas están inactivas: el silencio. El resto de los patrones presentan una probabilidad semejante entre ellos, pero muy baja respecto al silencio. En este estado el sistema tiene la mínima entropía (sólo visita un patrón) Podríamos pensar que está "congelado". Aquí, $P(\sigma)$ podría ser reducida por identidad al silencio, ya que todas las neuronas están inactivas.

Por otro lado, si bajamos mucho β (aumentando así la temperatura), todos los patrones comienzan a ser explorados por el sistema con la misma probabilidad, dejando de importar las interacciones entre neuronas y la diversidad neuronal. Aquí el sistema muestra la máxima entropía y podríamos considerar al sistema como "hirviendo", donde no muestra preferencias por un patrón u otro y no somos capaces de identificar estructuras en $P(\sigma)$. En este caso, $P(\sigma)$

es reducible por deducción a la distribución uniforme (todos los patrones tienen la misma probabilidad), donde la probabilidad de cada patrón es $1/(2N)$, por lo que $P(\sigma)$ puede ser deducida a partir de N, el número de neuronas.

Finalmente, como fue anticipado, el sistema muestra su mayor complejidad cuando β es cercana a 1, indicando que si utilizamos una reducción como la realizada para un valor de β muy alto o muy bajo (identidad y deducción, respectivamente)[14], concluiríamos que cuando β=1, $P(\sigma)$ no puede ser reducido ni al silencio ni deducido a partir del número de neuronas. Respecto a los otros dos estados, cuando β=1, hay novedad y requerimos incluir otros roles causales para generar $P(\sigma)$. En específico, una función de energía como la mostrada en (7) parece ser suficientemente amplia respecto a los roles causales que incluye, tal que, en un cierto estado, nos permite reducir $P(\sigma)$ a un solo elemento (el silencio), mientras que, en otro estado, es posible deducirla desde el número de elementos del sistema. Incluso, podemos generar $P(\sigma)$ en el estado de máxima complejidad, pero requerimos de la inclusión de variables de órdenes mayores, tal como $P(k)$.

Así, hemos mostrado que en un contexto general K, donde el sistema S es llamado ráster, es posible definir la distribución de probabilidad $P(\sigma)$ como un estado emergente, teniendo como condición, al menos, que observar el estado macroscópico $P(\sigma)$ reduce la incertidumbre, es decir entrega novedad, respecto a observar sus estados microscópicos (las neuronas). Esto fue mostrado en el contexto K_1, donde no es posible generar $P(\sigma)$ sólo a partir de la actividad de las neuronas. Por otro lado, en un contexto como K_2, para redes en las que N<50, es posible reducir, por composición, el sistema a sus partes más básicas y a sus interacciones, pero no así cuando se consideran redes más grandes. Finalmente, según un contexto como K_3, donde consideramos

[14] En el caso del sistema congelado, se reduce la propiedad emergente a todas las neuronas inactivas, mientras que, en el caso del sistema de alta temperatura, la propiedad emergente se deduce del tamaño de la red neuronal.

redes más grandes (N > 100), explorar las variaciones de la complejidad del sistema nos permite producir estados emergentes que son fácilmente reducibles a un solo elemento (por ejemplo, al silencio o al número de neuronas), como también otros estados emergentes donde necesitamos diversos roles causales para reproducir $P(\sigma)$. Estos roles causales son i) *ascendentes*, en el sentido que la actividad de cada neurona contribuye a generar $P(\sigma)$ y ii) *descendentes*, en el sentido que las interacciones de alto orden dan forma a la actividad de cada neurona. Como nota final, mencionamos que una ecuación como (7) no reduce el sistema a sus partes básicas, sino que además requiere de la inclusión de variables de órdenes más altos, como la interacción entre pares de neuronas y la tendencia de más de dos neuronas a disparar juntas.

7. Conclusión

Hemos considerado aquí algunos problemas clásicos relacionados con la caracterización de la noción de emergencia, basándonos principalmente en el denominado desafío de inconsistencia. Básicamente, éste es el desafío de aclarar en qué sentido los estados emergentes de un sistema dependen, por una parte, de los estados constituyentes del mismo y son, por otra parte, irreducibles a los mismos. Para abordarlo, ofrecemos una propuesta basada en la idea de contexto epistémico. Según nuestra caracterización, debe haber un contexto epistémico según el cual un estado emergente es reducible, pero también debe haber contextos epistémicos que nos permitan determinar tanto su irreducibilidad como los cambios de complejidad asociados a su aparición. Esta propuesta nos permite también dar cuenta de propiedades asociadas tradicionalmente a la noción de estado emergente, como la causalidad descendente y la novedad, entre otras. Por supuesto, la caracterización que ofrecemos es aplicable a fenómenos de diversos ámbitos, como las ciencias sociales y la biología. Con el fin de centrar la discusión en cuestiones asociadas a la filosofía de la mente, hemos aplicado, en la última sección, este concepto contextual de emergencia a cuestiones correspondientes al análisis matemático de redes neuro-

nales. Hemos mostrado cómo es posible satisfacer las condiciones de la definición que proponemos para estados de redes neuronales, en particular, para distribuciones de probabilidad obtenidas de las mismas. Es más que plausible, por supuesto, pensar que las relaciones de emergencia que podamos determinar sobre la base del análisis de redes neuronales sirven para establecer relaciones de emergencia entre estados mentales y estados neuronales. No nos hemos detenido en los detalles de esta tesis, pero creemos que lo expuesto aquí establece un suelo teórico riguroso para desarrollarla.

Agradecimientos

E.C. Agradece a Peter Baumann, George Ellis y a Miguel Fuentes por discusiones y aclaraciones sobre algunos de los puntos centrales del presente trabajo. También quisiera agradecer el apoyo de CONICYT (FONDECYT N° 3160180, N° 1181414 y N° 11180624).

R.H. Agradece a Rodrigo Cofré por las discusiones sobre modelos de máxima entropía y criticalidad. También quisiera agradecer a Fernando Rosas por discusiones sobre complejidad y fenómenos emergentes en el contexto de la estadística y la computación. Finalmente, agradece a la Beca de Doctorado Nacional de CONICYT.

Referencias

Beckermann, A. (1992). Introduction: Reductive and Nonreductive Physicalism. En A. Beckermann, H. Flohr, J. Kim (eds.), *Emergence or reduction? Essays on the prospects of nonreductive physicalism*, vol. 17. Berlin: Walter de Gruyter.

Bishop, R., Atmanspacher, H. (2006). Contextual Emergence in the Description of Properties. *Foundations of Physics*, 36(12): 1753-1777.

Broad, C. D. (1925). *The mind and its place in nature.* Cambridge: Routledge.

El-Hani, C. N., Pereira, A. M. (2000). Higher-level descriptions: why should we preserve them. En E. Andersen, C. Emmeche, N.O. Finnemann y P. Christiansen (eds.), *Downward Causation.* Aarhus: University of Aarhus Press.

Ellis, G. (2016). *How can Physics Underlie the Mind?* New York, Berlin, Heidelberg: Springer-Verlag.

Feigl, H. (1967). *The mental and the physical: The essay and a postscript.* Minnesota: University of Minnesota Press.

Gardella, C., Marre, O., Mora, T. (2018). Modeling the correlated activity of neural populations: A review. *Neural Computation*, 31(2): 233-269.

Kim J. (1997). The Mind-Body Problem: Taking Stock after Forty Years. En J. Tomberlin (ed.), *Mind, Causation, and World*, pp. 185-207. Oxford: Blackwell.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Kim, J. (2006). Emergence: Core ideas and issues. *Synthese*, 151(3): 547-559.

Rueger, A. (2000). Physical emergence, diachronic and synchronic. *Synthese*, 124(3): 297-322.

Sartenaer, O. (2016). Sixteen Years Later: Making Sense of Emergence (Again). *Journal for General Philosophy of Science*, 47(1): 79-103.

Van Gulick, R. (2001). Reduction, emergence and other recent options on the mind/body problem: A philosophic overview. *Journal of Consciousness Studies*, 8(9-10): 1-34.

Sobre los autores

Esteban Céspedes recibió el grado de licenciado en Filosofía de la Pontificia Universidad Católica de Valparaíso y el de doctor en Filosofía de la Universidad Goethe de Fráncfort del Meno. Actualmente trabaja, como investigador asociado a la Universidad de Valparaíso, enfocado en cuestiones relacionadas con las nociones de contexto, emergencia, representación y de sistema abierto. Contacto: estebancespedes@aol.com

Rubén Herzog Amunátegui, biólogo de formación, cursa su segundo año del Doctorado en Biofísica y Biología Computacional en la Universidad de Valparaíso. Sus intereses de investigación principales son las redes neuronales, el efecto de las drogas psicodélicas en el cerebro humano y la relación cerebro-consciencia. Contacto: rubenherzog@ug.uchile.cl

Capítulo 3

¿Hemos respondido la pregunta "puede pensar una máquina"?

Rodrigo Alfonso González Fernández

Resumen

Este trabajo examina si la pregunta "¿puede pensar una máquina?" ha sido respondida de manera satisfactoria. La primera sección, justamente, examina el *dictum* cartesiano según el cual una máquina no puede pensar en principio. La segunda trata sobre una rebelión en contra de Descartes, encabezada por Babbage. A su vez, la tercera describe una segunda rebelión encabezada por Turing. En ambas se examina, primero el lenguaje mentalista/instrumentalista para describir a una máquina programada y segundo, el reemplazo de la pregunta por el Juego de la Imitación. En la cuarta sección sostengo que la evidencia aportada por dicho juego nos devuelve a la pregunta, pese a Turing. Por último, la quinta sección versa sobre cómo la Habitación China de Searle, paradójicamente, apoya el *dictum* cartesiano, y lo hace porque tal experimento mental se basa en una capacidad mental falible: la introspección. Se concluye que la pregunta acerca de si puede pensar una máquina es derivada de otra pregunta filosófica compleja: la de la naturaleza de lo mental y el problema mente-cuerpo.

Palabras clave: Descartes, dictum, máquina, inteligencia, mente.

Considero que el problema mente-cuerpo es un problema totalmente abierto y sumamente confuso.
Saul Kripke, *El Nombrar y la Necesidad*

1. Introducción

La filosofía es una disciplina que nos enfrenta a preguntas difíciles de responder. Preguntas acerca de la naturaleza de la verdad, del conocimiento, del bien, de la belleza, de lo correcto, de qué es la mente, entre muchas otras, son filosóficamente abiertas. Es decir, el preguntar sobre tales tópicos *no* suscita consenso, sino más bien un debate amplio. Sin embargo, esto no es impedimento para no enfrentar los tópicos a los que refieren dichas preguntas. Es importante notar que los investigadores tratan tales problemas, incluso para decir que no tiene sentido hacerlo. Tal como Kripke (1980) sostiene, no hay consenso filosófico con relación a que es la mente, lo cual muestra que estamos ante una pregunta filosófica abierta, e incluso confusa. Desde que Descartes propuso qué era el pensamiento, esencia de la conciencia y de nuestra vida mental, ha habido múltiples reflexiones, desde aquellas que intentan diluir el problema mente-cuerpo, hasta las posturas que sugieren que la mente simplemente no existe. En consecuencia, la pregunta por la naturaleza de lo mental no solo suscita disenso, sino que además se relaciona con otras preguntas filosóficas difíciles y abiertas, como la que interroga por la posibilidad de vida mental en máquinas programadas o computadores.

En este trabajo muestro que la pregunta "¿puede pensar una máquina?" no ha sido respondida, y que sigue, al igual que el problema mente-cuerpo, siendo un problema filosófico abierto. Ciertamente, las dificultades para caracterizar qué es la mente contaminan la mencionada pregunta, al punto de que hacen difícil una respuesta definitiva. Justamente, en la primera sección examino el *dictum* cartesiano según el cual las máquinas no pueden pensar en principio. Las secciones dos y tres se concentran en las rebeliones en contra del *dictum*, con las propuestas de Babbage y de Turing, respectivamente. Aquél sostiene que el lenguaje mentalista para

describir máquinas es un instrumento, mientras que este parece augurar que la mencionada pregunta no suscitará nunca consenso, por lo que debe ser reemplazada por el Juego de la Imitación. La quinta sección analiza de qué forma dicho juego aporta evidencia inductiva no demostrativa, y ello, como se argumenta, nos devuelve a la pregunta que Turing quiere eliminar. Finalmente, la sexta sección se enfoca en la Habitación China y en cómo, pese a Searle, esta no refuta la IA fuerte, sino que da razones para dudar de que el funcionalismo puede ser una aproximación adecuada a la mente. Finalizo este trabajo con una conclusión en la que muestro que las preguntas filosóficas son en esencia abiertas, y ese es justamente el caso de "¿Tiene X mente?" y "¿puede pensar una máquina?", ambas estrechamente relacionadas.

2. El *dictum* cartesiano: la imposibilidad en principio de que máquinas piensen

La mente es un fenómeno complejo, y lo es por su carácter subjetivo e interno. En efecto, a diferencia de otros fenómenos, especialmente los físicos, la mente es accesible a sí misma, por ejemplo, cuando sentimos un dolor. Los dolores duelen para quien los experimenta, así conocemos que tenemos dolor, un fenómeno de suyo complejo para los demás. Pero, no solo la mente es accesible en un sentido epistémico, de cómo conocemos el fenómeno mismo de manera interna. Su *modo de existencia* subjetivo hace que la mente sea un fenómeno único en el mundo natural. Es decir, existe en tanto la experimentamos conscientemente, y ello la hace uno de los fenómenos más complejos, pero a la vez más fascinantes de estudiar. Es claramente un desafío para la filosofía, disciplina que aspira a la verdad y objetividad, incluso en problemas difíciles como el llamado mente-cuerpo.

El examen de la mente fue propuesto por René Descartes en el siglo XVII, cuando, en su intento de refutar a escépticos y ateos, propuso la existencia del *cogito* (González, 2017). Este es ontológicamente diferente de las cosas materiales por no tener extensión, ni ser divisible y limitado, lo cual traerá importantes consecuencias

para la imposibilidad de que las máquinas piensen, según Descartes. Si bien el francés es juzgado como responsable de habernos legado el problema mente-cuerpo, precisamente por la distinción dualista radical entre el cogito y las cosas materiales, no cabe duda de que funda las bases de la filosofía de la mente, al menos en relación con el examen de la naturaleza de lo mental. Nunca un filósofo había examinado qué era la mente, como un fenómeno que parece distinto de lo material. Este pasaje caracteriza qué es el dualismo cartesiano mediante el argumento de la intuición modal:

> En primer lugar, puesto que ya sé que todas las cosas que concibo clara y distintamente pueden ser producidas por Dios tal y como las concibo, basta con poder concebir clara y distintamente una cosa sin otra, para estar seguro de que la una es distinta de la otra, ya que, al menos en virtud de la omnipotencia de Dios, pueden darse separadamente […] Por lo tanto, como sé de cierto que existo y, sin embargo, no advierto que convenga necesariamente a mi naturaleza o esencia otra cosa que ser cosa pensante, concluyo rectamente que mi esencia consiste en ser solo una cosa pensante, o una substancia cuya esencia o naturaleza toda consiste solo en pensar. Y aunque acaso (o mejor, con toda seguridad, como diré enseguida) tengo un cuerpo al que estoy estrechamente unido, con todo, puesto que, por una parte, tengo una idea clara y distinta de mí mismo, en cuanto que soy solo una cosa que piensa –y no extensa–, y, por otra parte tengo una idea distinta del cuerpo, en cuanto que él es solo una cosa no extensa –y no pensante–, es cierto entonces que ese yo (es decir, mi alma, por la cual soy lo que soy), es enteramente distinto de mi cuerpo, y que *puede existir sin él* (Descartes. 1977, pp. 65-66, AT 78 y 79, énfasis mío).

Pero ese no es el final de la historia dualista, la cual sienta las bases del problema mente-cuerpo, o de cómo dos cosas metafísicamente diferentes se relacionan. Luego, en la misma 6ª Meditación Metafísica agrega:

> Hay una gran diferencia entre el espíritu y el cuerpo; pues el cuerpo es siempre divisible por naturaleza, y el espíritu es enteramente indivisible. En efecto: cuando considero mi espíritu, o sea, a mí mismo en cuanto que soy solo una cosa pensante, no puedo dis-

> tinguir en mí partes, sino que me entiendo como una sola cosa, sola y enteriza. Y aunque el espíritu todo parece estar unido al cuerpo todo, sin embargo, cuando se separa de mi cuerpo un pie un brazo o alguna parte, sé que ello no le quita algo a mi espíritu (p. 71, AT 86).

De este modo, Descartes zanja que espíritu y cuerpo son diferentes, separables, y distintos metafísicamente. Tal dualismo, esencia del problema mente-cuerpo, es fundamental para comprender el *dictum* cartesiano en contra de que las máquinas piensen.

En efecto, en su obra *Discurso del* Método (Frondizi, 1994) Descartes ya había hecho un anticipo de tales argumentos, tanto del descubrimiento del cogito como de su naturaleza metafísica. Un poco antes de proponer el argumento del *cogito ergo sum*, y del de la intuición modal arriba expuesto, introdujo el *dictum* acerca de la imposibilidad de que una máquina piense en principio. Su argumento es el siguiente:

> [Una máquina] no podría ordenar las palabras de formas diferentes para responder al significado que se dice en su presencia, como incluso el menos inteligente de los humanos puede hacer […] Incluso, aunque tales máquinas podrían hacer algunas cosas tal como nosotros les hacemos, o de mejor forma, inevitablemente fallarían en otras, lo cual revelaría que no están actuando de acuerdo con su entendimiento, sino por la pura disposición de sus órganos (p. 113).

En concreto, las máquinas son objetos físicos, y si lo son, tienen partes divisibles y limitadas: engranajes, palancas y mecanismos que hacen que su respuesta al ambiente sea también limitada y automática. Una máquina, entonces, solo responde a los estímulos ambientales por su mera disposición material. Esta opera causalmente, y por conllevar automatismo, esto es, por funcionar en términos de causa y efecto, hace que una máquina sea predecible. A mi juicio, algo similar sucede con los signos naturales (e.g. el humo, los anillos concéntricos de un árbol, etc.), los cuales son efectos de causas. Los animales, si son máquinas, responderían a la misma lógica, cuestión que hace a Descartes presa de críticas

por doquier. Ciertamente, si los animales son máquinas, entonces tienen reacciones al ambiente que son limitadas y predecibles, en función de sus órganos, y por tanto solo emiten signos naturales en virtud de la disposición de dichos órganos.

Por el contrario, el ser humano no puede ser solo una máquina, y esto, estimo, lo capacita para tener respuestas al ambiente que son flexibles, en términos de otros signos, los convencionales. Esto vale algunas aclaraciones previas. En primer lugar, el ser humano, por no ser solo máquina, tiene la capacidad de reaccionar de manera flexible al ambiente. Si esto es así, entonces los seres humanos pueden manejar signos convencionales lingüísticos de modo de significar lo mismo de distintas maneras. Por ejemplo:

O1 Juan ama a María
O2 María es amada por Juan
O3 Juan está enamorado de María

En todos estos casos, un ser humano puede enunciar mediante distintos signos convencionales, ordenándolos de distinta manera. Pero, lo más importante es que puede significar lo mismo incluso utilizando distintos signos lingüísticos.

En segundo lugar, es importante notar que el dominio de un lenguaje, dada la cuestión mencionada de los signos convencionales lingüísticos, indica racionalidad. La razón emplea el lenguaje para comunicar ideas y pensamientos, de modo que los mencionados signos son vehículos de estos últimos. El uso de lenguaje, según el francés, es mediado por el uso de la razón, del cogito. En consecuencia, cualquier organismo que use lenguaje es racional e inteligente.

Finalmente, la razón y el entendimiento son causa de las acciones. En el ser humano sucede todo lo contrario a lo que acontece en los animales-máquina. Estos reaccionan automáticamente, por la mera disposición de sus órganos. En cambio, el ser humano usa su entendimiento para actuar, y así su conducta es siempre guiada por *razones*. Entonces, el entendimiento es visto por Descartes como un instrumento universal, que otorga flexibilidad, tanto en

el uso del lenguaje como en la acción. La conducta humana es guiada por razones, mientras que la conducta animal es explicada por causas y efectos, a partir de la mencionada noción de signo natural.

En síntesis, el *dictum* cartesiano es una consecuencia lógica del dualismo del francés, es decir, de dividir el mundo en cosas mentales y cosas físicas, esencia del problema mente-cuerpo. Como las cosas físicas son limitadas, los *outputs* de las máquinas también lo serán, cuestión crucial con relación a la vida mental. Las máquinas, y los animales, piensa el Descartes metafísico, no tienen mente, y dos signos de ello es que no pueden usar lenguaje y sus acciones son automáticas e inflexibles. De esto se sigue que, en principio, no podría haber una máquina que usara el lenguaje de manera genuinamente humana, y que por ello la Inteligencia Artificial parece completamente condenada al fracaso: si bien es posible crear máquinas que empleen signos convencionales lingüísticos, dicho uso es solo una remembranza de cómo opera la razón. El dualismo, entonces, trae consecuencias importantes para el desarrollo de la Inteligencia Artificial futura. Ello porque, si el francés tiene la razón, la respuesta a "¿pueden pensar las máquinas?" sería negativa en principio y, más aún, definitiva, lo que provocará dos rebeliones materialistas.

3. La primera rebelión de la IA contra Descartes: Babbage

El materialismo, una filosofía monista, es radicalmente diferente al dualismo cartesiano. Básicamente, sostiene que hay una sola clase de substancias en el mundo: las cosas materiales. Todo estaría condicionado por la física, y la mente, pese a ser un fenómeno complejo, sería un fenómeno material más. Ello acontecería pese a la experiencia consciente, la subjetividad y su carácter interno. Ninguno de estos aspectos de la mente impediría que sea un fenómeno físico, porque todo es finalmente material. En consecuencia, el monismo materialista postula que la mente es un fenómeno que, dado el mundo material, sería material también, y ello pese a la subjetividad de la experiencia consciente ligada al *cogito* cartesiano.

Babbage (en Swade, 2000) adscribe a la filosofía materialista, en clave decimonónica. Es decir, en el contexto de la época industrial. En efecto, en tal época se creyó que todo podía ser resuelto con máquinas y que, incluso, la mente era un fenómeno material más. Descartes, por tanto, es el enemigo natural de los materialistas y lo es de Babbage también, pese a que sus ideas no refieren directamente al francés. Con todo, los materialistas buscan refutar a Descartes, argumentando, de una forma u otra, que el hecho de que la mente sea más fácil de conocer que el cuerpo no implica que la mente sea un fenómeno diferente. Más aún, para algunos materialistas como Babbage incluso Dios es una substancia material, tesis que se conoce como panteísmo materialista. Así, para Babbage (en Swade, 2000), *todo*, incluso Dios, es material.

Cabe destacar que sus ideas materialistas responden a una necesidad práctica del siglo XIX, época en que se intenta reducir el mundo a números y máquinas. Hay, ciertamente, una cuestión práctica que aquejó a Babbage, y que lo inspiró para crear y diseñar sus máquinas: la de las Diferencias y la Analítica. A diferencia de lo que ocurre actualmente, antaño no había calculadoras y las máquinas eran diseñadas sin medidas estándar. Ello implicaba una serie de problemas técnicos y prácticos, pero el más importante sin duda era cómo calcular y matematizar el mundo *con certeza*, la finalidad misma de las matemáticas. En el siglo XIX solo existían unas tablas de cálculo y mediante estas se hacían los cálculos complejos de navegación, ingeniería, finanzas, entre otras actividades humanas. Los errores eran frecuentes, cuestión que provocaba, además de pérdidas materiales y humanas, *incertidumbre*.

Las tablas de cálculo estaban plagadas de errores humanos. Teniendo presente este problema, Babbage se propuso crear máquinas que erradicaran el error humano de una vez y para siempre. Es decir, su sueño consistió en que el pensamiento matemático fuera mecanizado, una idea en concordancia con el espíritu de la Revolución Industrial. Su plan consistió en mecanizar las cuatro etapas de la confección de las tablas de cálculo: fórmulas, cálculo, lectura de prueba e impresión. De hecho, Babbage con la Máquina de las Diferencias pretendió abarcar dichas etapas, transformando

la fuerza de palancas y engranajes en pensamiento matemático. Su cometido, no obstante, tuvo una serie de problemas que impidieron la construcción de la máquina en su totalidad. Entre ellos destacó la falta de estandarización de herramientas, problemas financieros y desavenencias entre Babbage y el gobierno británico, producto de su carácter difícil.

Pese a tales problemas, introdujo una serie de mejoras en una segunda Máquina de las Diferencias, lo cual sentó las bases para el diseño de la Máquina Analítica. Esta, gracias a su diseño con unidad de procesamiento, control, memoria e inputs, le valió ser considerada el primer computador de la historia. Y ello sucedió a pesar de que el propio Babbage no imaginó el potencial multipropósito de dicha máquina. En efecto, fue Agusta Ada, Lady Lovelace, la que concibió que los engranajes no solo podrían representar números y operaciones, sino toda clase de entidades: notas musicales, posiciones en juegos de ajedrez y damas, y similares. Es por esto que Ada ha sido considerada por la historia como la descubridora del potencial de la Máquina Analítica, y no Babbage.

Sin embargo, el matemático decimonónico sí descubrió algo digno de notar. Gracias a las mejoras de la segunda Máquina de las Diferencias, y el posterior diseño de la Máquina Analítica, introdujo vocabulario mentalista para describir su funcionamiento. En particular, comenzó a utilizar expresiones como "la máquina recuerda", "memoriza", "piensa", etc. (Swade, 2000) Dicha introducción de vocabulario mentalista fue crucial para el nacimiento de una nueva disciplina filosófica: la Filosofía de la Inteligencia Artificial. Ello porque la pregunta principal que atraviesa a esta es si las máquinas realmente piensan, contra Descartes, o si el uso del vocabulario mentalista es puramente instrumental. Si lo es, entonces no es posible considerar que las máquinas, como los computadores, realmente tienen estados mentales. La atribución de estos solo sería una movida para comprender mejor su funcionamiento. Pero, ¿qué postura tuvo Babbage al respecto?

Su tratamiento del problema filosófico-cartesiano es el siguiente, según Swade (2000): Babbage habla de que al motor [Analíti-

co] "se le puede enseñar a prever". En otro lugar habla de que el motor "conoce". Estaba claro de que usar esta manera de referirse a él era apropiada, y evidentemente sintió que *antropomorfizar* los mecanismos requería justificación o excusa: "La analogía entre estos actos y las operaciones de la mente casi me forzó al uso *figurativo* de estos términos. Fueron adecuados y expresivos y prefiero usarlos en vez de sustituirlos por largos circunloquios" (Babbage en Swade, 2000, pp. 103-104, énfasis mío).

De esta forma, la primera rebelión anti cartesiana busca, mediante el reemplazo del pensamiento matemático por engranajes y mecanismos, sustituir la pregunta "¿puede pensar una máquina?" por un lenguaje que, aunque mentalista, refleje una postura instrumentalista. Tal postura no desecha la pregunta del todo, sino que concentra esfuerzos en contestarla, algo muy similar a lo que busca Alan Turing con su famoso y controvertido Juego de la Imitación.

4. La segunda rebelión de la IA contra Descartes: Alan Turing

Luego de los desarrollos de Turing a propósito del *Entscheidungsproblem* de David Hilbert, y de la pregunta por el potencial de los algoritmos (Turing, 1936), éste concibió un juego que pudiera *reemplazar* la pregunta "¿puede pensar una máquina?" En concreto, con esa iniciativa Turing buscó evitar definir conceptos como "máquina" e "inteligencia", toda vez que no solo llevan a una discusión filosófica sobre el uso común de estos, sino que además al estudio de su significado con base en encuestas tipo *Gallup*. En particular, Turing sostiene lo siguiente en 1950:

> Si la exploración del significado de términos como "máquina" y "piensa" se debe efectuar a partir del análisis de cómo estos se usan regularmente, es difícil evitar la conclusión de que el significado y la respuesta a la pregunta, ¿pueden pensar las máquinas?, debe encontrarse a través de una investigación estadística similar a una encuesta Gallup (p. 40).

Mediante dichas encuestas, piensa Turing, se podría establecer el uso más frecuente de los mencionados términos. Sin embargo, esto sería claramente insatisfactorio para dar fundamento a la Inteligencia Artificial. En efecto, determinar qué piensa la gente comúnmente acerca del significado del término "pensar", por ejemplo, no lleva a un cese de la discusión. Esto es, no lleva a consenso alguno, como sí lo haría la evidencia empírica recabada por un test. Por tanto, Turing descarta la discusión del significado de términos como "pensar", que parece demasiado cercana a filósofos como Descartes y el problema mente-cuerpo.

Como una manera de evitar tal discusión y polémica, plantea un test basado en un juego, el famoso y controvertido Juego de la Imitación. Turing, de hecho, ya había planteado un juego de similares características antes (1948), y luego lo hizo teniendo un presente una filosofía particular (Turing, 1950). En términos generales dicho juego tiene una inspiración funcionalista, al adscribir al principio de realizabilidad múltiple: en relación con propiedades funcionales, por ejemplo, asociadas a estados mentales, no importa el material en que estas se instancian. Tal como un carburador puede ser de cobre, bronce o aluminio, y puede desempeñar la función de mezclar el oxígeno y el combustible, los estados mentales pueden instanciarse en diferentes materiales: en un sistema basado en carbono o silicio, por ejemplo. Más aún, una tesis fundamental del funcionalismo y del principio mencionado es que las funciones son *separables* de los materiales en que se instancian, de un modo similar a cómo la mente es separable del cuerpo, lo que le ha valido a Turing ser acusado de ser dualista encubiertamente y, por tanto, de sostener una postura que no es auténticamente materialista y anti cartesiana (González, 2011).

Independiente de esas acusaciones, es claro, a pesar de lo que sostiene la tradición, que el Juego de la Imitación es funcionalista. Lo es porque la primera versión consiste en que hay un hombre en una pieza, una mujer en otra pieza, y jueces que hacen rondas de preguntas para determinar quién es el hombre y quién la mujer. El rol del hombre es hacerse pasar por una mujer, respondiendo como si fuera una, mientras que el rol de la mujer es ayudar a los jueces a

descubrir que ella es de sexo femenino. Las preguntas son simples, con rondas de no más de 5 minutos de duración. Por ejemplo, una pregunta típica es: ¿Tiene Ud. el cabello largo? La simplicidad de las preguntas es fundamental porque el computador digital no puede estar en una desventaja obvia frente a los jueces.

Por otra parte, un aspecto fundamental del juego, y que habla de su funcionalismo, es que la determinación del sexo de los participantes no es un tópico casual (González, 2015). Según Turing, en la medida que el hombre se hace pasar por una mujer, puede emular la inteligencia femenina, sin que se requiera de cerebro femenino. Lo mismo, por lo demás, acontecería si la mujer se hiciese pasar por un hombre. Para emular la inteligencia de este, no se requeriría que la mujer tuviera cerebro masculino. En ambos casos, los cerebros masculinos y femeninos están constituidos por los mismos materiales, pero se asocian con distintos tipos de inteligencia. Nótese que los materiales, por ser los mismos, son irrelevantes para la existencia de los dos tipos de inteligencia. Así, Turing sostiene la tesis de que esta es una propiedad funcional independiente de los materiales, y es separable de estos últimos, lo que hace que la postura funcionalista de Turing resulte profundamente anti-biológica. Lo es porque no importa la instanciación de la inteligencia en el cerebro; ella podría ocurrir en cualquier sistema que fuera *funcionalmente* equivalente al cerebro, según el funcionalismo. Más aún, es suficiente pero no necesaria la instanciación de la inteligencia en el cerebro.

Ahora bien, el reemplazo de la pregunta "¿puede pensar una máquina?" ocurre en la segunda versión del juego. Turing pregunta al lector: ¿Qué sucedería si un computador reemplazase al hombre en la primera habitación? Con ello, afirma taxativamente, se lograría el reemplazo de aquella pregunta. Nótese que, de manera similar a Babbage, planteó el reemplazo de la inteligencia por la imitación de esta, porque imitar tal inteligencia es suficiente para tenerla. Es decir, el que una máquina sea capaz de imitar la inteligencia hace que esta sea inteligente también. Como una manera de

remarcar que el Juego de la Imitación no involucra una definición de inteligencia, Turing (en Copeland, 2000) afirma lo siguiente en una entrevista concedida a BBC en 1951:

> No quiero dar una definición de qué es pensar, pero si tuviese que darla, probablemente sería incapaz de expresar nada más acerca de ésta que decir que fue un tipo de zumbido mental [*buzzing*] en mi cabeza. Pero no veo que tengamos que estar de acuerdo en una definición en modo alguno. Lo relevante es tratar de distinguir entre las propiedades de un cerebro o las de un hombre, que queremos discutir, y aquellas que no queremos. Para ponernos en un caso extremo, no estamos interesados en el hecho de que el cerebro tenga la consistencia de la papilla. No queremos decir "esta máquina es muy compleja, luego no es un cerebro y no puede pensar". Me gustaría sugerir *una clase particular de test* que uno pudiese aplicarle a una máquina. Ud. podría querer llamar a este un test para ver si la máquina piensa, *pero sería mucho mejor no formularlo así y caer en la petición de principio*, diciendo que las máquinas que pasen el test serán, por decirlo de algún modo, máquinas grado A. *La idea del test es que la máquina tiene que simular ser un humano a través de responder a las preguntas que se le hacen, y sólo pasará este test si la simulación es suficientemente convincente* (p. 6, énfasis mío).

Esta propuesta es, incluso, más radical de lo que parece en primera instancia, al menos desde el punto de vista de la tesis según la cual la inteligencia no requiere de cerebro. En efecto, dado el funcionalismo de Turing, este piensa que una Máquina de Turing puede imitar el comportamiento de cualquier máquina cuya conducta sea en principio calculable. Si esto es así, y el cerebro es una máquina, un computador digital que imite su comportamiento, también tendrá el *output* de un cerebro y, en consecuencia, se podrá sostener que piensa (vuelvo sobre esto más abajo). Con tales ideas, Turing asume que el Juego de la Imitación es prueba suficiente de que las máquinas programadas piensan, o al menos, de que no hay razones para asumir que no lo hacen. Por esto, y como una manera de resumir lo anterior, la tradición ha formulado el Juego de la Imitación estándar, en que se deja fuera el sexo de los participantes.

La versión simplificada del Test de Turing busca aportar evidencia empírica para evitar negar que las máquinas programadas piensan, o de considerar que esto es un sinsentido. En particular, el juego ahora consiste en un computador programado en una pieza, una persona en otra, y rondas de jueces fuera de ambas. La dinámica es similar a las versiones 1 y 2: los jueces formulan preguntas, el computador responde como si fuera humano, y la persona responde de manera sincera, tal como se puede apreciar en la siguiente figura:

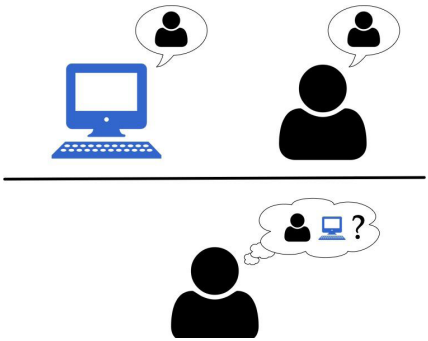

Fig. 1. La versión estándar del Test de Turing, tal como la tradición la ha planteado

Como se puede apreciar, Turing prosigue adhiriendo al funcionalismo, incluso en la versión simplificada estándar del test. No es necesario que una máquina programada o computador tenga el cerebro de un humano para ser inteligente; por el contrario, es suficiente que imite la inteligencia de este para ser inteligente.

En relación con este punto, Turing (1950), incluso, formuló una interesante predicción, con base en el Juego de la Imitación, y a raíz de si puede pensar una máquina:

> Creo que en alrededor de cincuenta años será posible programar un computador, con una capacidad de memoria de 109, para que participe en el Juego de la Imitación tan eficientemente que un interrogador lego no tendrá más de un 70% de probabilidad de hacer la identificación correcta después de cinco minutos de interrogatorio. Creo que la pregunta original, ¿pueden las maquinas

> pensar?, es demasiado absurda para seguir analizándola. No obstante, pienso que a finales de este siglo las ideas de la gente ilustrada y el uso de las palabras habrán cambiado de manera tal que uno será capaz de hablar de máquinas que piensan sin incurrir en contradicción alguna. (p. 49, énfasis mío)

El punto descrito aquí es crucial: el Test de Turing describe un método para la obtención de evidencia *inductiva* que apoye la hipótesis según la cual computadores y máquinas programadas poseen vida mental. Una propuesta, nuevamente, funcionalista y profundamente antibiológica. Sin embargo, el carácter inductivo del test es tan revolucionario como polémico. Y lo es porque, tal como se analiza en la siguiente sección, la evidencia empírica inductiva aportada, que es no demostrativa, no es concluyente en relación con si las máquinas pueden pensar.

5. La evidencia del Test de Turing: ¿fundamenta el reemplazo de la pregunta "puede pensar una máquina"?

Tal como se analizó en la sección previa, Turing reemplaza la pregunta "¿puede pensar una máquina?" por el Juego de la Imitación. Tal reemplazo no considera una definición de inteligencia, sino el aporte de evidencia empírica inductiva que muestre que no tiene sentido negar estados mentales a los computadores digitales o máquinas programadas (Moor, 1976, 1987; Copeland 2000). En este sentido, Turing tampoco pretende que el test sea una condición necesaria de atribución de inteligencia, ni que sea siquiera una condición suficiente. Lo primero, porque no pasar el test claramente no muestra que un sistema no sea inteligente, e.g. los animales. Mientras que lo segundo, la condición suficiente, no se establece claramente en la medida que no hay una definición de inteligencia involucrada (tal como la cita de arriba muestra).

Pero una cuestión que conviene analizar es si el tipo de evidencia proporcionada por el Juego de la Imitación es concluyente con respecto a si una máquina piensa. Parece el caso que Turing confunde dos cosas cuando afirma que el test puede reemplazar

la pregunta de si las máquinas programadas pueden pensar. En efecto, tal parece que confunde un problema del ámbito de la epistemología, esto es, si una máquina podría tener la capacidad de convencer a jueces de que existe inteligencia, con un problema ontológico, a saber, cuáles son las condiciones que son necesarias para la existencia de inteligencia (González, 2007). Más aún, una cosa es *saber* acerca de la existencia de otras mentes, lo cual refiere al problema de las otras mentes, y otra cosa diferente es si dichas mentes existen. Turing parece confundir ambas cosas al plantear que la evidencia empírica inductiva es *suficiente* para el reemplazo de la pregunta "¿pueden pensar las máquinas?".

Puesto de otra manera, confunde el saber acerca de propiedades mentales con la existencia de dichas propiedades. Esto es claro especialmente en lo que respecta a la imitación: nadie diría que imitar la inteligencia *es* necesariamente ser inteligente. Piénsese en una especie de mimo cognitivo que imitase la conducta inteligente de otros humanos, pero al hacerlo no tuviese un criterio de decisión al respecto. En estas circunstancias, ante un error humano, el mimo cognitivo imitaría la conducta, y podría convencer a jueces del caso, incluso si no es inteligente. Por lo tanto, la imitación, base de la persuasión de los jueces, no es garantía suficiente para que *exista* auténtica inteligencia. Ciertamente, una cosa es saber acerca de esta propiedad y otra muy distinta es la existencia de la misma, tal como una cosa es la existencia de otras mentes, y otra el saber acerca de ellas.

El Test de Turing reduce la pregunta acerca de si las máquinas piensan a un método de *verificación* que recabe evidencia empírica de que hay inteligencia. Pero ello no es correcto. El verificacionismo tiene problemas a la hora de responder preguntas ontológicas, al abstenerse de hacerlo producto de su fijación en la conducta observable, científicamente testeable. Tal verificación es adecuada a la hora de corroborar, por ejemplo, si la hipótesis acerca de una partícula sub-atómica tiene sentido. Pero, no sucede lo mismo con la mente y los estados mentales. Tal como Putnam (1965) destaca, con el experimento mental de los super super espartanos, es

perfectamente posible que existan estados mentales sin que exista conducta asociada, o viceversa. Por tanto, la conducta inteligente es contingente en relación con la existencia de estados mentales.

Es importante insistir en que la conducta, que es observable de manera pública y manifiesta, no es útil para responder preguntas de carácter metafísico en relación con la mente. Esto es particularmente relevante en el caso de la metafísica de la mente y de los estados mentales, es decir, de cómo estos existen, un problema subsidiario del problema mente-cuerpo. Justamente, muchos detractores del Juego de la Imitación, en vista de la evidencia que recaba, que es puramente observacional, lo han calificado de conductista, criticándolo como un método *passé* para determinar si los computadores programados tienen efectivamente estados mentales (Block, 1990). Dichos críticos han observado que la conducta lingüística no es suficiente para determinar con certeza que los computadores programados efectivamente poseen mente, estados mentales e inteligencia. En consecuencia, la conducta lingüística, que es observable, no es suficiente para responder de manera tajante preguntas metafísicas acerca de la mente; por ejemplo, si las máquinas programadas tienen estados mentales como los nuestros.

Imitar la existencia de estados mentales no es suficiente para aseverar que dichos estados han sido *replicados*. Claramente no es lo mismo *imitar* la propiedad F que replicar esta: piénsese en la imitación de las condiciones de un huracán y los efectos del huracán mismo. Si alguien afirmara que la imitación de las condiciones es suficiente para tener un huracán, estaría confundiendo el proceso mediante el cual la simulación de este convence a observadores acerca de sus propiedades con la existencia de estas, esto es, con un proceso causal, objetivo e independiente de los observadores. La cuestión de la imitación *versus* la replicación ha sido debatida en filosofía de la mente, especialmente en relación con el experimento mental de la Habitación China de John Searle (como se examinará en la siguiente sección). Turing (en Copeland, 2000), cae, justamente, en este error en una entrevista concedida a BBC en 1951:

> Para lograr que nuestro computador imite a una máquina sólo es necesario programarlo para que calcule lo que la máquina en cuestión haría bajo ciertas circunstancias […] Ahora bien, si una máquina en particular puede describirse como un cerebro, tenemos que solamente programar nuestro computador digital para imitarlo *y también será un cerebro*. Si se acepta que los cerebros reales, descubiertos en animales, y en especial en el hombre, son una clase de máquina, se seguirá entonces que nuestro computador digital, debidamente programado, *se comportará como un cerebro*. Este argumento presupone una idea que puede ser razonablemente cuestionada […] que esta máquina debiera ser de una naturaleza cuya conducta sea en principio predecible mediante cálculo […] Nuestro problema es, entonces, cómo programar una máquina para *imitar al cerebro*, o si lo pudiésemos expresar de una manera más breve y menos rigurosa, *para que piense* (p. 11, énfasis mío).

Pero, hay un elemento que es incluso más importante de discutir, y es el de la evidencia estadística recabada por el Test de Turing. Esta no es suficiente para reemplazar la pregunta de si las máquinas programadas piensan, porque no brinda ninguna clase de respuesta definitiva. Paradójicamente, el propio Turing desecha las encuestas como un método fiable en la determinación del significado de términos como "piensa" y "máquina", y luego propone un juego cuya esencia consiste en brindar evidencia estadística de que los computadores piensan. O al menos, de que no tiene sentido alguno negarles la existencia de estados mentales. Esta paradoja no es trivial, porque conduce de lleno a la pregunta de si las máquinas piensan *de nuevo*. En consecuencia, el método de Turing no es capaz de *reemplazar* definitivamente la pregunta, producto de la evidencia que aporta, puramente observacional, y que ciertamente no es demostrativa en relación con la existencia de estados mentales en máquinas. En efecto, la evidencia observacional no es nunca definitiva en relación con las preguntas que emanan de la metafísica de la mente.

Una segunda crítica que se le hace al test se liga con lo anterior. John Searle (1980) plantea un experimento mental que muestra que, incluso si el test convenciera al 100% de los jueces, ello no

implicaría que una máquina piensa, al menos en los términos como Turing concibe su famoso y controvertido Juego de la Imitación. De este modo, tal juego no solo no ayuda a reemplazar la pregunta, sino que incentiva a intentar una respuesta, desde la arena de la filosofía de la mente. El intento de fundamentar la Inteligencia Artificial por parte de Turing, si bien tiene como objetivo reemplazar la pregunta ¿puede pensar una máquina?, nos devuelve a esta, y lo hace desde una consideración filosóficamente más profunda: la aceptación de que las preguntas sobre la mente no se resuelven con ayuda de la pura evidencia observacional.

6. La segunda ola de críticas al test: la Habitación China

Dado el *dictum* cartesiano, una serie de investigadores en la IA concentraron sus esfuerzos en crear *chatbots* capaces de usar lenguaje natural. Por ejemplo, han creado programas como ELIZA, PARRY, SHRDLU, entre otros. Con excepción de Weizenbaum (1984), creador de ELIZA, los investigadores han sostenido que el manejo de lenguaje natural en conversaciones es conducente a pasar el Test de Turing, signo de inteligencia. Paradójicamente, asumieron el *dictum* cartesiano con el objetivo de refutarlo. En vista de este problema, relacionado con el entendimiento lingüístico, John Searle (1980) distingue entre dos tipos de Inteligencia Artificial, la fuerte y la débil. En particular, propone lo siguiente:

> Encuentro útil distinguir entre la IA "fuerte" y la "débil" (o cautelosa). De acuerdo con la IA débil, el principal valor del computador en el estudio de la mente es que nos brinda una poderosa herramienta. Por ejemplo, nos permite formular hipótesis de modo más riguroso. Pero de acuerdo con la IA fuerte, el computador no es meramente una herramienta en el estudio de la mente; *el computador adecuadamente programado es una mente,* en el sentido de que los computadores con *los programas adecuados pueden decirse que literalmente entienden y tienen otros estados cognitivos,* los programas no son meras herramientas que nos permiten testear explicaciones psicológicas; más bien, los programas son en sí dichas explicaciones (p. 417, énfasis mío).

Tal comentario se fundamenta en un programa específico: SAM, *Script Applier Mechanism,* de Schank y Abelson (1977). Este simula el entendimiento lingüístico de historias, mediante una base de datos que contiene libretos, los cuales son concatenaciones de eventos de manera causal. Por ejemplo, hay un libreto para *Restaurante*, otro para *Trabajo*, y así sucesivamente. Ahora la cuestión interesante es que la aplicación de un libreto a situaciones por parte del computador permitiría a este inferir información que no está explícita en una historia, y ello sería signo de entendimiento lingüístico. Por ejemplo, si un mozo atiende mal mi mesa, a partir de *Restaurante* el computador infiere que no dejaré propina, o que esta será magra. De esta manera, Schank y Abelson simulan el entendimiento lingüístico de historias, lo cual incita a los investigadores de la IA fuerte a pensar que el computador es una mente porque esta entiende historias a la manera de SAM.

Justamente, contra la IA fuerte, y como una manera de *falsar* la creencia de los computadores programados que manejan lenguaje son inteligentes, Searle propone un experimento mental: la Habitación China. Dicho experimento contempla la existencia de una habitación en que hay alguien, hablante de inglés que no habla nada de chino, a quien se le envían ideogramas en este idioma. Gracias a un banco de ideogramas chinos, más un libro de reglas para correlacionar estos, Searle manda hacia afuera de la habitación ideogramas chinos. Sin saberlo, la primera tanda de signos es una historia, la segunda refiere a preguntas y la tercera son respuestas a dichas preguntas. Sin saberlo también, Searle responde de manera perfecta, tal como lo haría un hablante nativo de chino, solo con base en la manipulación de símbolos. Es claro que, bajo estas condiciones, la Habitación China pasaría el Test de Turing, y lo haría convenciendo al 100 % de los jueces de que hay un hablante nativo de chino en la habitación. Con ello, entonces, Searle "falsaría" la IA fuerte (vuelvo sobre esto abajo), en la medida que hace trabajar la mente como esta describe el funcionamiento de la cognición, aunque no habría entendimiento lingüístico genuino.

Hay un fárrago de réplicas a la Habitación China. Teniendo presente el objetivo de este trabajo, no vale la pena ahondar en

todas ellas. Solo me concentraré en dos: la réplica del sistema y la réplica de las otras mentes, pues están directamente relacionadas con la pregunta "¿puede pensar una máquina?" En efecto, la réplica del sistema ataca un punto débil del argumento de Searle: este, con el libro de reglas que es un símil de un programa, es solo una parte del sistema. Luego, incluso si Searle no entiende chino, que es una parte del sistema, no se puede sostener que éste no entiende. La respuesta de Searle es que, si se internalizan en su mente todos los elementos constituyentes del sistema, sigue sin entender. Independiente de la corrección de la respuesta del filósofo norteamericano, es claro que hay un solo elemento que no puede internalizar: él mismo (González, 2012). Este elemento muestra que su experimento mental, al aludir a la introspección, base del análisis de todo experimento mental, tiene una cercanía con el cartesianismo, muy a pesar de lo que Searle piensa. En efecto, la operación llevada a cabo, i.e., la manipulación de símbolos, es evaluada desde la introspección, y ello muestra que solo mediante esta podemos responder la pregunta de si una máquina, en este caso un programa, puede pensar. Como la introspección no es un método totalmente fiable, en tanto puede arrojar resultados experimentales *a priori* falibles no puede concluirse que la Habitación China refuta la IA fuerte de manera definitiva, sino que *solo sienta las bases para dudar del carácter adecuado de tal aproximación* a la mente. Es decir, la Habitación China, un experimento mental *a priori* basado en la introspección, podría, pese a Searle, no refutar la IA fuerte, sino más bien sentar una duda razonable con relación a su verdad.

Hay, además, otro bache en el argumento de Searle: las otras mentes. Dado que la Habitación China se basa en la introspección, no resulta muy claro cómo se puede salvar la objeción de que no es posible saber de la existencia de otras mentes con certeza, cuestión que afectaría a la Habitación China. Si bien Searle responde que lo esencial no es la epistemología acerca de otras mentes, sino la ontología, o las condiciones para la existencia de ellas, queda la duda sembrada, nuevamente, del real alcance de la Habitación China en relación con la refutación de la IA fuerte. Es decir, es claro que sabemos con certeza acerca de nuestras propias mentes, pero no

acerca de la existencia de otras, y ello representa un escollo fundamental para responder adecuadamente si la Habitación China entiende y tiene estados mentales. En consecuencia, no es claro, nuevamente, que dicho experimento *refute* de manera tajante la IA fuerte. Nuevamente, solo puede sostenerse que se *siembra una duda razonable* acerca de la adecuación de esta aproximación a la mente y cognición humana.

La Habitación China no es capaz de brindar certeza con relación a la respuesta a la pregunta "¿puede pensar una máquina?" Ciertamente, tiene la limitación de basarse en la introspección, y aunque se repliquen las condiciones de la IA fuerte, o de otras aproximaciones a la mente a propósito de la mencionada pregunta, no es claro que pueda sostenerse de modo tajante una respuesta definitiva. No hay posibilidad de consenso, entonces, y ello muestra que la Habitación China de Searle es un experimento que no otorga una respuesta definitiva. Por tanto, la pregunta sigue abierta, al igual que lo que ocurre con el famoso y controvertido Test de Turing, todo lo cual es subsidiario del difícil y confuso problema mente-cuerpo.

7. Conclusión

En este trabajo se ha examinado la cuestión de si hemos respondido a la pregunta "¿puede pensar una máquina?" El análisis se efectuó especialmente a la luz del *dictum* cartesiano según el cual es imposible en principio que una máquina piense. Ello ocurriría porque las máquinas son cosas físicas constituidas de partes y engranajes, y por eso son limitadas en relación con los *outputs*, por ejemplo, los lingüísticos. Sin embargo, el *dictum*, subsidiario del problema mente-cuerpo, ha recibido críticas filosóficas importantes. Se examinaron dos: la de Babbage y la de Turing. Mientras que el primero sostiene que la pregunta puede responderse en vista del lenguaje instrumental que se emplea para describir el funcionamiento de una máquina, el segundo sostiene que se puede evitar la pregunta mediante el Juego de la Imitación. Este aportaría evidencia inductiva respecto de que no tiene sentido negar vida mental

a las máquinas programadas o computadores. Luego se analizaron dos respuestas a las aproximaciones de Babbage y Turing: el problema de la evidencia del Juego de la Imitación y el experimento mental de la Habitación China de Searle. Ambas, si bien no son definitivas en cuanto al *dictum*, sí sientan bases para dudar de que hayamos respondido adecuadamente la pregunta "¿puede pensar una máquina?"

Dicha pregunta sigue siendo abierta, y es *filosóficamente* fértil, porque nos lleva a cuestionar la naturaleza de lo mental. Pese a los argumentos de Turing, han pasado más de 50 años a partir de su predicción, y aún no hay consenso en relación con si las máquinas pueden pensar. No existe, así, una respuesta definitiva, lo que indica que la pregunta sigue siendo un incentivo para la investigación en filosofía de la mente. La situación entre partidarios y detractores del *dictum* queda en tablas, y estas, insisto, muestran que la pregunta "¿puede pensar una máquina?" sigue siendo una cuestión filosófica a debatir, un problema que muestra, en todo su esplendor, por qué la filosofía no es una disciplina con respuestas definitivas. La falta de consenso sugiere, además, que la búsqueda filosófica continúa tal como acontece con respecto al problema mente-cuerpo. Hay, pese a Turing y Searle, una relación estrecha entre la naturaleza de lo mental y nuestra posibilidad de responder la pregunta "¿puede pensar una máquina?" Ciertamente, dicha pregunta abre un horizonte de posibilidades y como siempre en filosofía, nos invita a seguir debatiendo y discutiendo en las movedizas y siempre inestables arenas del problema mente-cuerpo.

Referencias bibliográficas

Block, N. (1990). The computer model of the mind. En D.N. Osherson y E.E. Smith (eds.), *Thinking: An Invitation to Cognitive Science*, pp. 247-289. Cambridge, Mass.: MIT Press.

Copeland, B.J. (2000). The Turing Test. En J.H. Moor (ed.), *The Turing Test: The Elusive Standard of Artificial Intelligence*, pp. 1-21. Dordrecht: Kluwer Academic Publishers.

Descartes, R. (1977). *Meditaciones Metafísicas*. Traducción. y notas de Vidal Peña. Madrid: Alfaguara.

Descartes, R. (1994). *Discurso del Método.* Traducción, estudio preliminar y notas de Risieri Frondizi. Madrid: Alianza Editorial.

González, R. (2007). El Test de Turing: dos mitos, un dogma. *Revista de Filosofía Universidad de Chile*, 63: 37-53.

González, R. (2011). Descartes, las Intuiciones Modales y la IA. *Revista Alpha*, 32: 181-198.

González, R. (2012). La pieza china: un experimento mental con sesgo cartesiano. *Revista Chilena de Neuropsicología*, 7(1): 1-6.

González, R. (2015). ¿Importa la determinación del sexo en el Test de Turing? *Revista de Filosofía Aurora*, 27(40): 277-295.

González, R. (2017). La refutación cartesiana del escéptico y del ateo: Tres hitos de su significado y alcance. *Revista Anales del Seminario de Historia de la Filosofía*, 34(1): 85-103.

Kripke, S. (1980). *El nombrar y la necesidad.* México, D.F.: UNAM.

Moor, J.H. (1976). An Analysis of the Turing Test. En S. Shieber (ed.), *The Turing Test: Verbal Behaviour as the Hallmark of Intelligence*, pp. 297-306. Cambridge, Mass.: MIT Press.

Moor, J. H. (1987). Turing Test. En S.C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence*, pp. 1126-1130. New York, Wiley.

Putnam, H. (1965). Brains and behaviour. En J. Heil (ed.), *Philosophy of Mind: A Guide and Anthology*, pp. 96-104. Oxford: OUP.

Schank, R.C., Abelson, R.P. (1977). *Scripts, Plans, Goals, and Understanding.* Hillsdale, N.J.: Erlbaum.

Searle, J. (1980). Minds, Brains and Programs. *The Behavioral and Brain* Sciences, 3(3): 417-424.

Swade, D. (2000). *The Difference Engine: Charles Babbage and the Quest to build the First Computer*. London: Penguin.

Turing, A.M. (1936). On computable numbers, with an application to the Entscheidungsproblem. En M. David (ed.), *The Undecidable.* New York: Raven Press.

Turing, A.M. (1948). Intelligent Machinery. En B. Meltzer y D. Michie (eds.) *Machine Intelligence*, pp. 3-23. Edinburgh: Edinburgh University Press.

Turing, A.M. (1950). Computing intelligence and machinery. En M.A. Boden (ed.), *The Philosophy of Artificial Intelligence*, pp. 40-66. Oxford: OUP.

Weizenbaum, J. (1984). *Computer Power and Human Reason: From Judgement to Calculation.* Harmondsworth: Pelican.

Sobre el autor

Rodrigo Alfonso González Fernández es PhD in Philosophy, por la Katholieke Universiteit Leuven. Actualmente es profesor asistente del Departamento de Filosofía, y Director del Centro de Estudios Cognitivos, de la Facultad de Filosofía y Humanidades, Universidad de Chile. Sus áreas de investigación son la Filosofía de la Mente y de la Inteligencia Artificial y la Ontología Social.

Capítulo 4

En contra de la visión de la valencia afectiva como reforzadores internos

José M. Araya

Resumen

La valencia afectiva es un constructo clave en las ciencias afectivas. Lo que se puede denominar la *teoría de la valencia como señal no-sensorial* (VNS) sostiene que la valencia afectiva consiste en una señal interna—que no es ella misma un estado perceptual o conceptual de ningún tipo—que marca a las representaciones sensoriales como buenas o malas (Carruthers, 2011, 2017; Prinz, 2004, 2010). En la versión de Prinz de VNS, la valencia consiste en señales internas de reforzamiento. Estas últimas son señales de aprendizaje que otorgan a las emociones su característica fuerza motivacional. En este artículo argumento que la versión de Prinz de VNS, la visión de las señales internas de reforzamiento, es problemática. Por un lado, esta visión predice una disociación entre el empuje (*oomph*) motivacional y la intercepción que no está bien justificada, y sus argumentos que intentan mostrar que la valencia no está fundada en el sistema interoceptivo pueden enfrentarse mediante los recursos teóricos provistos por el marco del procesamiento predictivo. Por otro lado, la visión de Prinz implica una base neural de la valencia afectiva que resulta implausible.

Palabras clave: afecto, valencia, activación, interocepción, procesamiento predictivo.

1. Introducción

La valencia es un constructo clave en filosofía de las emociones y ciencias afectivas (Barrett, 2006; Barrett y Russell, 1999; Carruthers, 2017). Las emociones se clasifican como emociones positivas o emociones negativas en virtud del carácter de su valencia. No solo las emociones tienen valencia como componente. Los estados de ánimo (*moods*), las motivaciones homeostáticas (hambre, sed, dolor, picazón, etc.), entre otros estados afectivos, también exhiben dicho carácter positivo o negativo. En efecto, la valencia—junto con la activación (*arousal*)—es una dimensión que define a los estados afectivos en general (Barrett, 2006; Barrett y Russell, 1999). Como enfatiza Carruthers (2017), determinar la naturaleza de la valencia es entonces clave para la comprensión de la naturaleza del afecto.

Prinz (2004, 2010) ha ofrecido una atractiva visión acerca de la naturaleza de la valencia afectiva. Él defiende una versión de lo que se puede denominar la *teoría de la valencia como señal no-sensorial* (VNS). Según esta clase de teoría, la valencia consiste en una señal interna—que no es ella misma un estado perceptual o conceptual de ningún tipo—que marca a las representaciones sensoriales como buenas o malas (deseadas o indeseadas). En línea con una cierta tradición en ciencias afectivas, en estas visiones, la valencia se considera entonces como algo que se "pega" a las representaciones sensoriales/perceptuales, siendo así la valencia algo distinto de las representaciones sensoriales mismas. En la visión de Prinz, la valencia consiste en señales internas de reforzamiento. Estas últimas son señales de aprendizaje, no fundadas (*grounded*) en ninguna modalidad sensorial, que otorgan a las emociones su característica fuerza motivacional. La visión de Prinz, la visión de las señales internas de reforzamiento (VSIR), exhibe varias ventajas explicativas, lo que la vuelve una más que prometedora visión, en comparación a los enfoques tradicionales sobre la valencia.

En este artículo, argumento que VSIR es problemática, independientemente de la plausibilidad de las teorías en competencia sobre la valencia. Por un lado, esta visión predice una disociación

entre el empuje (*oomph*) motivacional y la intercepción que no está bien justificada, y sus argumentos que intentan mostrar que la valencia no está fundada en el sistema interoceptivo pueden enfrentarse mediante los recursos teóricos provistos por el marco del procesamiento predictivo. Por otro lado, la visión de Prinz implica una base neural de la valencia afectiva que resulta implausible.

Comienzo caracterizando brevemente la noción de valencia (sección 2.), enfatizando su rol evaluativo y motivacional (sección 2.1.). En la sección 3., introduzco algunos desiderata para las teorías sobre la valencia. Luego, en la sección 4., presento brevemente algunas de las principales familias de teorías en competencia sobre la valencia. En la sección 5., presento la versión de Prinz de VNS, la visión de las señales internas de reforzamiento (VSIR), y sus ventajas explicativas (sección 5.1.). Como muestro en la sección 6., VSIR enfrenta problemas decisivos, lo que la vuelve una candidata insatisfactoria para una teoría plenamente plausible acerca de la naturaleza de la valencia.

2. Caracterizando el componente de valencia de las emociones

Nos afanamos en tener ciertos tipos de emociones, y nos afanamos en evitar tener otros tipos de emociones. Ciertas emociones se sienten bien, mientras que otras emociones se sienten mal. Esto es, hay emociones positivas y emociones negativas. Por ejemplo, la alegría, el orgullo, y el amor son típicamente emociones positivas; mientras que el enojo, el miedo, la culpa, y el desdén son típicamente emociones negativas. Las emociones se clasifican de este modo en virtud del carácter de su valencia. Las emociones positivas tienen como componente valencia positiva; mientras que las emociones negativas tienen como componente valencia negativa[1] [1](e.g., Barrett, 2006; Prinz, 2004, 2010).

[1] Noten que este modo de caracterizar la valencia afectiva deja abierta la posibilidad de que un cierto tipo de emoción *E* puede tener distinto valor de valencia en diferentes ocasiones.

La valencia no es tan solo parte de nuestra comprensión de psicología-de-sentido-común respecto de la naturaleza de las emociones. La valencia es un constructo que también juega un rol fundamental en el estudio científico de las emociones (ver, e.g., Barrett, 2006; Russell, 2003; Berridge y Kringelbach, 2015), al punto que, para algunos teóricos, la valencia es uno de los principales pilares de todos los estados afectivos (Barrett, 2006; Carruthers, 2017; Russell, 2003). Noten entonces que la noción de valencia en la que estoy interesado en este artículo es una noción no-normativa que juega un rol explicativo en psicología. Así, contrariamente a una posible lectura de lo que algunos investigadores han señalado (e.g., Charland, 2005; Picard, 1997; Solomon, 2003), cuando se dice, en las ciencias afectivas, que una emoción es *positiva* o *negativa* (i.e., que tiene valencia positiva o negativa) no se está diciendo que tal emoción es positiva o negativa en el sentido de ser *buena* o *mala* normativamente, en algún sentido ético o prudencial. En el sentido en el que estoy interesado en este artículo, la valencia no es un constructo ético ni prudencial; la valencia es un constructo psicológico descriptivo. Simplemente asumo entonces que la valencia es una clase natural que juega un rol explicativo en ciencias afectivas (ver Carruthers, 2017; Prinz, 2010).

2.1 El rol evaluativo y motivacional de la valencia

¿Cuál es el rol funcional de la valencia? Una respuesta no controvertida es que el rol de la valencia es hacer que las cosas le importen al agente positiva o negativamente así facilitando e impulsando la conducta relativa a los ahora relevantes y valorados aspectos del ambiente. La valencia hace que las cosas nos importen y consecuentemente empuja a la acción. Esto es, la valencia juega un rol evaluativo y motivacional.

La afirmación de que la valencia es principalmente invocada como un constructo que juega un rol evaluativo y motivacional queda evidenciada por el modo en el cual la valencia ha sido caracterizada lo largo de su historia en psicología (ver Colombetti, 2005). Por ejemplo, Tolman (1949) entendía la valencia como las

'fuerzas atractivas o repulsivas' que los objetos tienen para los organismos. Lewin (1935) entendía la valencia como 'hechos ambientales imperativos' que guían la conducta y dan lugar a un mundo de significancia para el organismo. Schneirla (1959) entendía la valencia en términos de la dirección de la conducta dirigida a objetivos. Más recientemente, Fridja (1986) sostenía que los objetos con valencia resultan atractivos o aversivos para los organismos, definiendo así la significancia de una situación. Davidson (1993) sostenía que la valencia conlleva conductas de aproximación y evitación, las que, según él, son intrínsecamente placenteras y displacenteras, respectivamente. Las visiones que entienden la valencia como tipos de evaluación (Ben-Ze'ev, 2000; Lazarus y Lazarus, 1991; Ortony, Clore y Collins, 1988) también pueden tomarse como postulando la valencia como un constructo motivacional, dado que los contenidos evaluativos, contrariamente a los contenidos meramente indicativos, recomiendan cursos de acción. La valencia también se ha caracterizado como una representación de objetivos (Dyer, 1987; Izard, 1991; Panksepp, 2005; Rozin, 2003), la que puede mantenerse que son inherentemente motivacionales. Carruthers (2011, 2017) sostiene que la valencia consiste en una señal no-sensorial que confiere valor (bueno o malo) a los estímulos atendidos, y que motiva su persecución o evitación. También Prinz (2004, 2010) sostiene que la valencia hace que los estados afectivos nos importen, y funciona como una señal motivacional.

Puede que resulte útil clarificar brevemente lo que quiero decir aquí con 'evaluación' y 'motivación'. Con 'evaluación' simplemente me refiero a un ítem mental que, cuando se asocia o combina con un cierto estado mental, hace de este último algo positivo o negativo para el agente. Ahora bien, ya que las teorías sobre la valencia son teorías acerca de qué hace a los estados afectivos positivos (negativos) tener un carácter *positivo* (*negativo*), la naturaleza de tal positividad (negatividad) va a depender de la teoría de la valencia que resulte verdadera. Con 'estados motivacionales' me refiero a la presteza para actuar que se puede sentir antes de seleccionar una acción específica en el ambiente externo. Esto es, el *empuje* (*oomph*) *motivacional* que nos impulsa a regular estados afectivos

internos, impeliéndonos a modificarlos (desde luego, esto típicamente nos lleva a buscar cómo actuar en el ambiente externo). Las motivaciones homeostáticas como el hambre y las ansias de fumar son ejemplos típicos de estados motivacionales en el sentido que intento dar a entender. Así, con 'motivación' no quiero referir a razones para tomar cursos de acción específicos en el ambiente externo, que resultan del razonamiento práctico (i.e., motivos).

Ahora bien, el rol evaluativo y motivacional de la valencia es heredado por las emociones, en la medida en que estas últimas contienen un marcador de valencia. De esta manera, el contenido representado por cierta emoción (i.e., su *core relational theme*, e.g., algo perdido en el caso de la tristeza) se vuelve algo que le importa al agente, de modo que éste es impelido por su estado emocional a actuar de un modo relevante (Prinz, 2004). La afirmación de que las emociones heredan su aspecto evaluativo/motivacional de su componente de valencia es comúnmente mantenido en la investigación sobre las emociones (ver Frijda, 2008).

3. Desiderata para las teorías sobre la valencia

Hay ciertas propiedades generales que puede mantenerse que la valencia debe tener. Estas propiedades son truismos fundamentales que constituyen metas explicativas, o desiderata, para las teorías sobre la valencia.

Una teoría sobre la valencia debe satisfacer los siguientes desiderata[2]. En primer lugar, una teoría sobre la valencia debe tener suficiente alcance para aplicarse a todos los casos claros de emoción (*desiderátum del alcance*). Esto es, todos los casos claros de emoción deben exhibir la propiedad o mecanismo que se propone que da cuenta de la valencia. En segundo lugar, una teoría sobre la valencia debe acomodar nuestras taxonomías pre-teoréticas relati-

[2] No creo de ningún modo que esta sea una lista exhaustiva. Es una lista tentativa. Aún así, estos desiderata capturan criterios compartidos que se piensa que una teoría sobre la valencia debe satisfacer para contar como satisfactoria (ver Prinz, 2010).

vas a las emociones que típicamente cuentan como positivas y las emociones que típicamente cuentan como negativas (*desiderátum de la taxonomía pre-teorética*). En tercer lugar, una teoría sobre la valencia debe explicar el truismo de que las emociones positivas se sienten bien y que las emociones negativas se sienten mal (*desiderátum del sentimiento* (*feeling*)). Finalmente, una teoría sobre la valencia debe dar cuenta del rol evaluativo y motivacional de la valencia. Esto es, por qué la valencia nos vuelve las cosas algo positivo y negativo (*desiderátum de la evaluación*); y cómo la propiedad o mecanismo que se propone que da cuenta de la valencia logra dar cuenta de su rol motivacional característico (*desiderátum de la motivación*) (ver Prinz, 2010).

Intuitivamente, si una teoría sobre la valencia *A* satisface más de estos desiderata que una teoría *B*, entonces *A* ha de ser preferida por sobre *B*.

4. Teorías en competencia sobre la valencia

Parcialmente siguiendo a Prinz (2004, 2010), distingo cuatro familias principales de teorías sobre la valencia. (a) *Teorías de la aproximación/evitación*. Estas teorías identifican la valencia positiva y negativa con conductas (o tendencias de acción) de aproximación y evitación, respectivamente (e.g., Maclean, 1993). (b) *Teorías evaluativas*. Según una versión de esta visión, la valencia positiva y negativa se identifica con la evaluación de una situación como congruente o incongruente con los objetivos del agente, respectivamente (Lazarus y Lazarus, 1991). Según otra versión de esta visión, la valencia positiva y negativa se identifica con el juicio de que el objeto intencional de la emoción relevante es bueno o malo, respectivamente (e.g., Ben-Ze'ev, 2000). (c) *Teorías hedónicas*. Según estas teorías, la valencia positiva y negativa se identifica con el placer y el displacer, respectivamente (e.g., Barrett, 2006; Damasio, 1994). (d) Lo que se puede denominar la *teoría de la valencia como señal no-sensorial* (VNS) (e.g., Carruthers, 2011, 2017; Prinz, 2004, 2010). Esta visión sostiene que la valencia afectiva consiste en una señal interna—que no es ella misma un estado

perceptual o conceptual de ningún tipo—que marca a las representaciones sensoriales como buenas o malas (Carruthers, 2011, 2017; Prinz, 2004, 2010).

En este artículo, me centro en discutir críticamente la versión de Prinz de VNS por las siguientes razones. En primer lugar, la versión de Prinz de VNS es una de las pocas propuestas filosóficas recientes acerca de la naturaleza de la valencia afectiva. En segundo lugar, como veremos más abajo, la visión de Prinz, VSIR, tiene varias ventajas explicativas, comparándose favorablemente en relación con otras clases de enfoques sobre la valencia. Es relevante entonces mostrar que la versión de Prinz de VNS exhibe supuestos problemáticos.

5. Teoría de la valencia como señal no-sensorial: la visión de las señales internas de reforzamiento

Lo que se puede denominar la *teoría de la valencia como señal no-sensorial* (VNS) (e.g., Carruthers, 2011, 2017; Prinz, 2004, 2010) identifica la valencia con una señal interna—que no es ella misma un estado perceptual o conceptual de ningún tipo—que marca a las representaciones sensoriales como buenas o malas (deseadas o indeseadas). Estas visiones se alinean con una cierta tradición en ciencias afectivas, que consiste en considerar el aspecto afectivo (de valencia) de las experiencias sensoriales/perceptuales como algo que se "pega" a las representaciones sensoriales/perceptuales. En otras palabras, el aspecto de valencia de una experiencia sensorial se considera como algo 'extra' a las representaciones sensoriales mismas. Por ejemplo, según la tradición en cuestión, comer un pastel dulce se siente bien porque un ítem mental afectivo (valencia) se "pegó" a la representación sensorial de dulzura, siendo esta última un ítem mental distinto que el primero. Esto es, solo cuando las representaciones sensoriales/perceptuales (e.g., dulzura, un paisaje, la música, etc.) tienen un "barniz hedónico" añadido por el afecto es que estas representaciones se vuelven algo que se siente bien (o mal). Tal "barniz hedónico" se considera que es un ítem no-sensorial en el mobiliario de la mente, distinto de cual-

quier tipo de representación sensorial/perceptual o trozo de conocimiento de alto-nivel (ver, e.g., Berridge y Kringelbach, 2010, 9). Al argumentar que la versión de Prinz de VNS no es plenamente convincente tal como se la ha propuesto, estaré entonces también sugiriendo indirectamente que la tradición en cuestión debe ser revisada.

Pongámonos manos a la obra. Prinz (2004) identifica a las emociones con percepciones de cambios corporales. En el enfoque de Prinz, las emociones, entendidas de este modo, incluyen una señal para su propio cese o continuación. La valencia es tal señal. Más precisamente, según Prinz, la valencia equivale a lo que él denomina *reforzadores internos*. Los reforzadores internos son dispositivos internos que señalan comandos estructurados de modo no-proposicional. Estos comandos especifican si acaso la emoción considerada (i.e., la percepción de estado corporal relevante) debe ser mantenida y tenida más frecuentemente en ocasiones futuras (valencia positiva), o si acaso la emoción considerada debe ser terminada y no tenida en ocasiones futuras (valencia negativa).

Prinz ilustra el funcionamiento de los reforzadores internos con imperativos que, en relación con cierta emoción, dicen algo así como "¡más de esto!" (valencia positiva) o "¡menos de esto!" (valencia negativa). Esto es, los reforzadores internos son señales que median la conducta relativa a la mantención o cese de las percepciones corporales relevantes (emoción). Así, en la teoría de Prinz sobre la valencia, las emociones positivas son aquellas que impulsan su propia mantención, y las emociones negativas son aquellas que impulsan su propia terminación. De esta forma, los reforzadores internos hacen del estado mental que tienen por "blanco" algo positivo o negativo: ellos hacen de las emociones algo que le *importa* al agente (Prinz, 2004).

En cuanto señales con contenido tipo-imperativo[33], los reforzadores internos son inherentemente motivadores: "Las emocio-

[3] La afirmación de Prinz de que la valencia equivale a una señal interna que tiene el rol funcional de mandar la continuación (o cese) de esta-

nes ejercen fuerza motivacional por medio de los marcadores de valencia" (Prinz, 2004, 242). Los reforzadores internos impulsan a los agentes a hacer algo, a saber, cambiar un estado interno llevándolos a tener que determinar cómo actuar en el ambiente externo, i.e., seleccionar cursos de acción que puedan llevarlos a mantener o terminar una cierta emoción positiva o negativa, respectivamente. Noten que los reforzadores internos no motivan cursos de acción específicos en el ambiente externo, diseñados para alcanzar la meta de mantener o terminar cierta emoción, como, por ejemplo, decidir ir a una fiesta para terminar un estado de tristeza. En otras palabras, los reforzadores internos no comandan estrategias específicas de regulación emocional. Los reforzadores internos tan solo nos empujan a modificar estados internos. Ellos comandan "¡menos de este estado interno! ¡haz lo que sea para lograr eso!", lo que puede llevar luego a decidir alguna acción regulatoria en específico en el ambiente externo (e.g., ir a una fiesta).

Importantemente, los reforzadores internos son señales de recompensa y castigo. En cuanto tales, ellos juegan un rol clave en el aprendizaje por refuerzo y en el condicionamiento, además de su rol motivacional comentado más arriba. Los reforzadores internos les permiten a los agentes aprender qué estímulos externos cuentan como recompensas y qué estímulos externos cuentan como castigo. Los estímulos externos encontrados en el pasado, y que gatillaron una cierta emoción, serán buscados en el futuro, puesto que se le "pegó" un marcador de valencia positiva que comanda la mantención de esa emoción. Esta clase de asociaciones se almacenan en la memoria. Así, los estímulos asociados en la memoria con las emociones que contienen un marcador de valencia positivo aumentaran la probabilidad de una respuesta apetitiva—i.e., esos estímulos cuentan como recompensa para el agente. Lo mismo

---

dos corporales no debe tomarse como queriendo decir que el carácter fenoménico de los estados con valencia está dado por tal contenido imperativo (en donde la noción de contenido relevante demanda que el sujeto tenga acceso a nivel-personal a tal contenido). Para una defensa es este último tipo de visión en relación al estado afectivo del dolor, ver, e.g., Klein (2007) y Martínez (2011).

puede decirse de los marcadores de valencia negativos, mutatis mutandis. En este sentido, los reforzadores internos son también señales de aprendizaje.

Crucialmente, los reforzadores internos y las percepciones de cambios corporales son componentes disociables de la emoción (Prinz, 2004). Así, los reforzadores internos pueden, en principio, "pegarse" a estados mentales no-emocionales. Ahora bien, considerando que, en la teoría de Prinz, las representaciones constituidas por las modalidades sensoriales corporales y los reforzadores internos son entidades separadas, distintas dentro del mobiliario de la mente, y los reforzadores internos no están fundados en ninguna modalidad, la valencia es una señal *no-sensorial*. Consecuentemente, la valencia no es algo que se pueda sentir (*feel*) (más de esto más abajo). Sin embargo, los reforzadores internos y los cambios corporales van típicamente juntos, haciendo que los últimos le parezcan buenos o malos al agente[44].

6. Ventajas explicativas

VSIR parece acomodar los desiderata presentados más arriba (sección 3). En primer lugar, considerando que los reforzadores internos son señales motivadoras, VSIR no debiera tener problemas acomodando el *desiderátum de la motivación*. En segundo lugar, no parece implausible pensar que todas las emociones incluyen un componente motivacional que nos impulsa a buscar maneras de mantenerlas o eliminarlas. El solo hecho de que existan fenómenos de regulación y desregulación emocional habla a favor de este supuesto. Así, VSIR satisface el *desiderátum del alcance*. VSIR también satisface el *desiderátum de la taxonomía pre-teorética*. Casos claros de emociones negativas, como el enojo, la culpa, y el miedo

[4] Prinz no presenta evidencia empírica para esta visión. Prinz respalda VSIR argumentando que ésta tiene más ventajas explicativas que las teorías en competencia, y argumentando que estas últimas son problemáticas por sí mismas (ver Prinz, 2010).

son emociones respecto de las cuales nos sentimos motivados a deshacernos de ellas; y casos claros de emociones positivas, como la alegría, la euforia, y el amor son emociones respecto de las cuales nos sentimos motivados a mantener. Finalmente, VSIR parece también satisfacer los desiderata del *sentimiento* y de la *evaluación*. En la medida en que estamos impulsados a eliminar las emociones negativas y a mantener las emociones positivas, la valencia hace de las emociones algo que nos importa. En este sentido, consideramos a las emociones negativas y positivas como malos y buenos sentimientos, respectivamente (*desiderátum del sentimiento*). Finalmente, las emociones contienen reforzadores internos. Como recién mencioné, los reforzadores internos hacen de las emociones algo que nos importa. Considerando que las emociones pueden asociarse con representaciones de situaciones externas, éstas pueden volverse situaciones positivas o negativas para nosotros (*desiderátum de la evaluación*) (ver Prinz, 2010).

Las otras teorías en competencia fallan en acomodar todos los desiderata. Consideren, por ejemplo, la *teoría de la aproximación/evitación*, que identifica la valencia positiva y negativa con conductas (o tendencias de acción) de aproximación y evitación, respectivamente. Esta visión falla en satisfacer los desiderata de la *evaluación* y del *sentimiento*, puesto que no es en absoluto claro cuál es el vínculo, si es que lo hay, entre aproximarse (o alejarse de) a algo y tenerlo como bueno (malo) y sentirse bien (mal). Además, la teoría de la aproximación/evitación también falla en satisfacer el desiderátum del alcance, puesto que no todas las emociones conllevan tipos distintivos de conducta o tendencias de acción. Esta teoría también enfrenta problemas relativos al desiderátum de la taxonomía pre-teorética, ya que casos claros de emociones negativas (e.g., enojo) típicamente conllevan conductas de aproximación (ver Prinz, 2004, 2010).

O consideren la *teoría evaluativa*. Según una versión de esta visión, la valencia positiva y negativa se identifica con la evaluación de una situación como congruente o incongruente con los objetivos del agente, respectivamente (Lazarus, 1991). Según otra versión de esta visión, la valencia positiva y negativa se identifica

con el juicio de que el objeto intencional de la emoción relevante es bueno o malo, respectivamente (e.g., Ben-Ze'ev, 2000; Ortony et al., 1988). La teoría evaluativa no acomoda el *desiderátum del alcance*, ya que es ampliamente aceptado que los animales no-humanos tienen estados de valencia. Sin embargo, es improbable que, digamos, los tordos tengan el aparato conceptual requerido para formar evaluaciones y juicios propiamente tales. Esta visión también falla en satisfacer el *desiderátum del sentimiento*, puesto que no es claro cómo las evaluaciones y juicios se pueden sentir como algo.

Interesantemente, VSIR puede explicar la plausibilidad intuitiva de las teorías en competencia, al tiempo que evita sus problemas más típicos (Prinz, 2010). Para tomar tan solo un ejemplo, VSIR puede explicar fácilmente por qué tendemos a evitar el objeto de las emociones negativas. Tendemos a comportarnos de esa manera porque, frecuentemente, un modo efectivo de obedecer el imperativo "¡menos de esta emoción!" es alejarse del evento o situación que está causando la emoción relevante. Al mismo tiempo, y contrariamente a la *teoría de la aproximación/evitación*, VSIR puede acomodar el caso del enojo. Tendemos a aproximarnos al objeto del enojo porque poner término a la situación que causa el enojo interviniendo en ella es un modo efectivo de deshacerse de lo que está ocasionando la emoción en cuestión. Finalmente, dado que los reforzadores internos son meras etiquetas que funcionan como señales de comando o indicadores de valor (i.e., ellos son señales no-conceptuales), VSIR evita el problema de convertir la valencia en algo cognitivamente demasiado demandante como para ser poseída por animales no-humanos y niños pequeños, como sí lo hace la teoría evaluativa.

VSIR puede acomodar más desiderata que las teorías en competencia, y al mismo tiempo puede explicar lo intuitivo de éstas. Así, VSIR se compara favorablemente a las visiones en competencia sobre la naturaleza de la valencia afectiva.

7. VSIR: algunos problemas

VSIR es una visión atrayente. Sin embargo, enfrenta algunos desafíos. En primer lugar, VSIR predice una disociación entre el empuje (*oomph*) motivacional y la intercepción que no está bien justificada. En segundo lugar, VSIR implica una base neural de la valencia afectiva que resulta implausible.

8. ¿Son la percepción corporal y el empuje (*oomph*) motivacional separables?

Según la teoría de Prinz sobre las emociones (Prinz, 2004), estas están constituidas por dos componentes, a saber, percepciones de cambios corporales (percepción interoceptiva) y marcadores de valencia (reforzadores internos). Estos componentes distintos sirven funciones diferentes. Por un lado, las percepciones de cambios corporales se supone que sirven dos funciones: (1) representar relaciones organismo-ambiente (i.e., *core relational themes, e.g.*, *pérdida* en el caso de la tristeza), y (2) asignar recursos fisiológicos con el fin de facilitar la acción adaptativa, independientemente de cualquier tipo de motivación. Por otro lado, los reforzadores internos sirven la función de señalar la urgencia de actuar para cambiar estados corporales internos, lo que hace que tales percepciones de estados corporales le importen al agente. Esto es, los reforzadores internos juegan un rol motivacional; mientras que los cambios corporales facilitan la acción, y sirven una función semántica.

Crucialmente, como mencioné más arriba, los reforzadores internos y las percepciones de cambios corporales son componentes disociables de la emoción. Esto es, durante la emoción, VSIR asume una disociación entre la clase relevante de motivación (i.e., el empuje motivacional característico de los episodios emocionales) y la interocepción (i.e., la percepción de la condición fisiológica del cuerpo entero). Este supuesto es problemático. Déjenme explicar.

La motivación, entendida como un sentido de urgencia, como la urgencia de cambiar un estado interno, es un estado mental que se puede sentir. Ahora bien, solo las representaciones sensoriales/

perceptuales pueden sentirse[55]. Si tal empuje motivacional es algo que usualmente sentimos, y solo las representaciones perceptuales pueden sentirse, entonces esta clase de motivación es algo que tiene lugar en alguna modalidad sensorial. Obviamente, el sentimiento de motivación en cuestión no es algo que tenemos debido al olfato, ni la visión, ni ninguna modalidad exteroceptiva o propioceptiva—y no hay razones para afirmar que alguna combinación de aquellas puede hacer el trabajo. El mejor candidato parece ser entonces el sistema interoceptivo. Esto es, en este respecto, la intuición es que el empuje (*oomph*) motivacional que exhiben los estados de valencia consiste en percibir que nuestros cuerpos están fisiológicamente preparados (o no preparados) para entrar en acción. Si el componente motivacional de la emoción está fundado en la interocepción, y la valencia juega tal rol de impulsarnos a actuar de modo de cambiar un estado interno durante los episodios emocionales, entonces hay razones para pensar que no hay disociación entre interocepción y el empuje motivacional característico de la valencia, como VSIR asume.

Esta intuición es respaldada por ciertas líneas de evidencia que sugieren que las regiones interoceptivas son críticas para los estados conscientes de valencia. La corteza insular (particularmente la ínsula anterior), una región clave para la percepción interoceptiva consciente, es crítica para la experiencia de los impulsos (*urges*) (ver Berman, Horovitz, y Hallet, 2013; Craig, 2015). Por ejemplo, tras el daño de la ínsula, los adictos al tabaco se vuelven anhedónicos respecto del tabaco, y pierden su motivación de fumar (Navqi y Bechara, 2010). Considerando que la ínsula es una región critica para la percepción interoceptiva, este hecho apoya la idea de que el empuje motivacional característico de la valencia no es un componente independiente de la interocepción, en el modo en que VSIR requiere.

---

[5] En este artículo, y siguiendo a Prinz (2012), simplemente asumo que la conciencia fenoménica está poblada por nada más que perceptos, i.e., que no hay cualidades fenoménicas más allá de las representaciones sensoriales/perceptuales. Este supuesto simplemente niega que puede haber vehículos representacionales no-sensoriales con carácter cualitativo.

Es más, es importante notar que la valencia es una parte componente de todas las clases de estados afectivos (Barrett 2006; Carruthers, 2017), incluyendo las motivaciones homeostáticas como el hambre. La ausencia de la disociación predicha en cuestión gana apoyo intuitivo de este tipo de estados afectivos. Piensen en el hambre. El hambre es un impulso (*urge*) con valencia que se siente. En concordancia con la afirmación de que probablemente no hay disociación entre la interocepción y el empuje motivacional en cuestión, evidencia anatómica y funcional muestra que la experiencia de hambre (experiencia interoceptiva) está exhaustivamente fundada en estructuras interoceptivas del cerebro (ver Craig, 2015).

Desde luego, todavía se podría argumentar que el empuje motivacional característico de los estados de valencia surge en el sistema interoceptivo, pero solo en caso de que las representaciones interoceptivas se vuelvan el objetivo de la acción modulatoria de los reforzadores internos. En otras palabras, tal empuje motivacional surge solo en caso de que un marcador de valencia se "pegue" a una representación corporal interna, en dónde ésta es la única clase de representación que puede ser el objetivo (*target*) propio de los marcadores de valencia. Así, sin representaciones interoceptivas, el sentimiento de motivación en cuestión no puede tener lugar. Sin embargo, sigue el argumento, las representaciones interoceptivas pueden tener lugar sin marcadores de valencia. Esto podría explicar la intuición de arriba, y las líneas de evidencia de más arriba que vinculan la interocepción con el empuje motivacional característico de los estados afectivos. En este respecto, VSIR predice entonces que puede haber percepciones de estados corporales internos asociados con un cierto estado afectivo que simplemente carecen de valencia, y que esta anomalía no es dependiente de anomalías debidas a representaciones interoceptivas mal formadas (así la asimbolia del dolor y casos similares parecen estar excluidos, en consideración de que la asimbolia del dolor conlleva el malfuncionamiento de regiones interoceptivas en el cerebro, tales como la ínsula posterior y anterior (ver Craig, 2015)); sino que depen-

de más bien de una anomalía en otro componente con un perfil funcional en línea con los reforzadores internos. No obstante, no parece haber casos de este tipo en específico.

Es más, es importante notar que la réplica del párrafo de arriba parece ser una petición de principio. Como comenté más arriba, la fenomenología del empuje (*oomph*) motivacional característico de los estados de valencia—más el supuesto de que la conciencia se agota en las representaciones sensoriales/perceptuales (ver pie de página 5)—indica que el sentimiento de motivación en cuestión se agota en las representaciones interoceptivas. Insistir que se necesita postular otro componente además de estas últimas representaciones parece entonces redundante, sin consideraciones fenomenológicas o evidencia empírica que justifique especular acerca de la existencia de este otro componente—los reforzadores internos—que pueda modular las representaciones interoceptivas en el modo requerido. Esto es, sin un caso para la afirmación de que las representaciones interoceptivas por ellas mismas son motivacionalmente inertes, y sin evidencia positiva de disociación doble entre el empuje motivacional y las representaciones interoceptivas, la objeción de arriba es una petición de principio. Más aún, parece no haber casos de sentimientos fundados en la interocepción—tales como las motivaciones homeostáticas—que se disocien de su valencia, como VSIR predice. Por ejemplo, parece no haber tal cosa como hambre o sed sin valencia, y los sentimientos de temperatura y picazón siempre tienen valencia, aún cuando muchas veces su valor de valencia puede ser muy sutil en el caso de que no sean atendidos.

No obstante, Prinz (2004) argumenta que la emoción de la sorpresa es precisamente un caso en el cual la percepción corporal se disocia de la valencia. La sorpresa a veces se siente como una emoción positiva, y otras veces la sorpresa se siente como una emoción negativa. La sorpresa puede ser positiva o negativa. Sin embargo, en ambos casos, especula Prinz, la sorpresa está constituida por el mismo patrón de cambios corporales. El mismo patrón de cambios corporales puede entonces exhibir diferente valor de valencia.

Por lo tanto, el valor de valencia tiene que estar determinado por un componente separado de la percepción corporal: la percepción de cambios corporales y la valencia se disocian.

Esta línea de razonamiento no me convence. Asumiendo, en aras del argumento, que la sorpresa sí conlleva un patrón distintivo de cambios corporales, y que hay sorpresas positivas y negativas, todavía hay una interpretación distinta de este tipo de casos. Como discutiré abajo, la sorpresa positiva y negativa pueden verse como conllevando experiencias corporales *diferentes*, ya que el *mismo* patrón de cambios corporales puede dar lugar a perceptos corporales diferentes. Así, si esta explicación está bien encaminada, el patrón de cambios corporales característico de la sorpresa *no* es percibido como el mismo tipo de estado corporal en el caso de las sorpresas positivas y negativas. Los argumentos de Prinz fallan entonces en establecer su conclusión, simplemente porque la emoción de la sorpresa (positiva y negativa) *no* cuenta como un caso en el cual el mismo percepto corporal conlleva diferente valor de valencia. Esa premisa no se sostiene. Así, el hecho de que las sorpresas positiva y negativa se experimenten de modo diferente se puede explicar simplemente apelando a la diferencia en las experiencias corporales que cada una de ellas conlleva, en lugar de mediante una diferencia en marcador de valencia que se "pega" a la misma percepción corporal. Esta explicación alternativa puede entonces explicar este tipo de caso sin postular un componente distinto de la percepción corporal. En lugar de eso, esta explicación puede acomodar este tipo de caso mostrando que el mismo proceso mediante el cual se forman los perceptos de modo estándar puede operar de tal modo que diferentes tipos de perceptos interoceptivos pueden surgir del mismo tipo de estímulo corporal. Puesto que esta hipótesis alternativa es más parsimoniosa, se debería al menos considerar antes de saltar a la conclusión de que el valor de valencia está dado por un componente separado de la percepción corporal en casos como el de la emoción de la sorpresa.

El aspecto del proceso estándar de formación de perceptos que tengo en mente es la modulación atencional. La modulación atencional juega un rol clave durante la formación de perceptos. Muy

toscamente, según algunos influyentes enfoques, como el enfoque del procesamiento predictivo (Clark, 2016; Hohwy, 2013), la modulación atencional (toscamente, en este enfoque, 'precisiones', es decir, fiabilidad estimada de la señal sensorial) determina a qué aspectos de la señal sensorial entrante se le da más peso, y qué aspectos de la señal sensorial entrante son ignorados, dependiendo de cuán fiable se infiera que es la señal sensorial. Esto ocurre a través de todos los niveles de la jerarquía perceptual. Varios factores contextuales influencian tal asignación de pesos, como, por ejemplo, expectativas sensoriales descendientes basadas en conocimiento de alto-nivel (ver Clark, 2016; Hohwy, 2013). Este proceso de regímenes de asignación de pesos determina el conjunto particular de rasgos que constituirán un percepto en una cierta ocasión, y qué conjunto de rasgos constituirán un percepto en otra ocasión, dado un cierto estímulo fijo. Esto es, en ocasiones y contextos diferentes, dadas diferentes asignaciones de pesos de 'precisión' (i.e., diferentes estimaciones de fiabilidad de la señal sensorial) y el mismo estímulo, el percepto que resulta de este proceso estará compuesto de diferentes rasgos ligados (*bounded*), cambiando así la configuración del percepto que eventualmente se experimentará. Para tomar un ejemplo típico, dependiendo de cuáles aspectos de la señal sensorial se atienden en diferentes niveles de la jerarquía perceptual, a partir del mismo estímulo, como el ruido blanco, alguien puede formar o bien el percepto auditivo de una canción familiar, o el percepto de una conversación, o tan solo tonos al azar. No hay razones para negar esto para el caso de los perceptos de la condición interna del cuerpo. Entonces, dadas diferentes clases de expectativas sensoriales sensibles al contexto, pesando de modo diferente distintos aspectos del *mismo* patrón de cambios corporales mediante mecanismos atencionales (el patrón de cambios corporales que supuestamente individúa la sorpresa), el tipo de percepto interoceptivo del cuerpo que se forma en el caso de la sorpresa positiva puede diferir del percepto del cuerpo que se forma en el caso de la sorpresa negativa. No se requiere entonces ningún ítem mental adicional para explicar diferencias en valor de valencia, dada la misma clase de estímulo del cuerpo. Las expectativas sensitivas al contexto pueden hacer el trabajo. De hecho,

diferentes regímenes de asignaciones de peso ('precisiones', i.e., estimaciones sobre la fiabilidad de la señal sensorial) probablemente son responsables de las diferencias en valor de valencia observadas en casos como la analgesia del placebo vs no-analgesia (ver Büchel, Geuter ,Sprenger y Eippert, 2014). No se necesita entonces considerar la sorpresa y casos similares como casos en los cuales la misma percepción corporal exhibe distinto valor de valencia. En consideración de que tales casos se pueden tomar como casos en los cuales perceptos corporales *diferentes* conllevan un valor de valencia *diferente*, la hipótesis de que en casos como la sorpresa y similares el valor de valencia está determinado por la sola percepción corporal debe excluirse antes de saltar a la conclusión de que algo ajeno la maquinaria sensorial misma se requiere para dar cuenta de estos casos. Esto es, el hecho de que la sorpresa positiva y negativa se experimenten de modo diferente se puede explicar simplemente apelando a la diferencia en las experiencias corporales que cada una conlleva, en lugar de apelar a una diferencia en el marcador de valencia que se le "pega" a la misma percepción corporal.

9. La base cerebral de los reforzadores internos

VSIR enfrenta otro desafío. Déjenme considerar tan solo el caso de la valencia positiva para facilitar la exposición. Los reforzadores positivos internos son dispositivos internos que no solo motivan, sino que también juegan un rol en el aprendizaje por refuerzo. Esto es, VSIR asume que, en este respecto, el mismo dispositivo interno que motiva durante las emociones también equivale a una señal de aprendizaje por recompensa. Ahora bien, las estructuras neurales que son las mejores candidatas para realizar los marcadores de recompensa son las regiones ricas en dopamina del mesencéfalo. Más precisamente, las señales de aprendizaje por recompensa probablemente están realizadas en área tegmental ventral (ATV) y en la sustancia negra parte compacta (SNpc). No solo otros investigadores han señalado que estas estructuras dopaminérgicas son buenas candidatas para realizar los reforzadores internos positivos de Prinz (Corns, 2014), sino que ellas tienen

algunas propiedades que encajan de buena manera el constructo de Prinz: además de ser responsables de las señales de aprendizaje por recompensa (Schultz, Dayan y Montague, 1997), ellas no realizan estados hedónicos por sí mismas (Schroeder, 2004), y (más controversialmente) ellas también juegan un rol en la motivación basada en la recompensa (Berridge, 2007). No es entonces arbitrario considerar a ATV y SNpc como realizadores de los reforzadores positivos internos[6].

Ahora bien, recuerden que las emociones tienen valencia como componente constituyente. Así, en consideración de lo de arriba, VSIR hace la siguiente predicción: si la valencia positiva equivale a una señal interna de recompensa, entonces ATV y SNpc deben activarse consistentemente durante las emociones positivas. Ciertamente, los estudios de neuroimagen actuales sobre las emociones son limitados. No obstante, estos estudios inequívocamente muestran que estas estructuras no están significativa y consistentemente involucradas durante las emociones positivas (e.g., Murphy, Nimmo-Smith y Lawrence, 2003; Phan, Wager, Taylor, y Liberzon, 2002; Vytal y Hamann, 2010). Así, no parece probable que la valencia positiva equivalga a una señal internas de recompensa, como VSIR sostiene.

Interesantemente, otras regiones que se han asociado con la valencia, como la ínsula anterior, la corteza cingulada anterior, y las cortezas orbitofrontales (Carruthers, 2017), forman parte de una red responsable de la percepción interoceptiva (i.e., *salience network*) (e.g., Seth, 2013). Estas regiones sí se encuentran activas consistentemente durante las emociones (e.g., Murphy et al., 2003; Phan et al., 2002; Vytal y Hamann, 2010). Esto apoya aún más la idea de que la disociación entre la valencia y la percepción

[6] Prinz (2004, 161-162) menciona algunas regiones que se observan activas durante algunas tareas que gatillan estados afectivos. Él hace esto con el fin de respaldar la afirmación de que de hecho hay tal cosa como la valencia. La valencia es un fenómeno real. Sin embargo, Prinz no está explícitamente comprometido con ningún realizador neural de los marcadores de valencia.

corporal implicada por VSIR, y por lo que denominé la teoría de la valencia como señal no-sensorial, (Carruthers, 2011, 2017; Prinz, 2004, 2010), debe ser revisada.

10. Conclusión

Elucidar la naturaleza de la valencia es mandatorio no solo porque la valencia es un constructo clave en filosofía de las emociones y ciencias afectivas, sino también porque la valencia cumple un rol fundamental en varias funciones de alto-nivel (Carruthers, 2017).

En este artículo, discutí la versión de Prinz de la teoría de la valencia como señal no-sensorial (e.g., Carruthers, 2011, 2017; Prinz, 2004, 2010), a saber, la visión de los reforzadores internos. Argumenté que esta versión de la teoría de la valencia como señal no-sensorial es problemática. Esto es el caso ya que la visión en cuestión predice una disociación entre el empuje motivacional y la intercepción que no está bien justificada. Por otro lado, la visión de Prinz también implica una base neural de la valencia afectiva que resulta implausible.

Ahora bien, como mencioné más arriba (sección 5), las teorías de la valencia como señal no-sensorial se alinean con una cierta tradición en ciencias afectivas, que consiste en considerar el aspecto afectivo (de valencia) de las experiencias sensoriales como algo que se "pega" a éstas. La valencia es vista entonces como un "barniz hedónico" que tiñe a las representaciones, siendo de este modo algo distinto de cualquier tipo de representación sensorial o trozo de conocimiento de alto-nivel. Al argumentar que la versión de Prinz de esta clase de teorías es problemática, estoy también sugiriendo entonces que tal tradición en ciencias afectivas debe ser revisada.

Referencias

Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1): 20-46.

Barrett, L. F., Russell, J. A. (1999). Structure of current affect. *Current Directions in Psychological Science*, 8(1): 10-14.

Ben-Ze'ev, A. (2010). The thing called emotion. En, P. Goldie (ed). *The Oxford handbook of philosophy of emotion*, pp. 41-62. Oxford: Oxford University Press.

Berman, B. D., Horovitz, S. G., Hallett, M. (2013). Modulation of functionally localized right insular cortex activity using real-time fMRI-based neurofeedback. *Frontiers in human neuroscience*, 7(638): 1-11.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, 191(3): 391-431.

Berridge, K. C., Kringelbach, M. L. (2010). In *Pleasures of the brain*. New York: Oxford University Press.

Berridge, K. C., Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron*, 86(3): 646-664.

Büchel, C., Geuter, S., Sprenger, C., Eippert, F. (2014). Placebo analgesia: a predictive coding perspective. *Neuron*, 81(6): 1223-1239.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.

Carruthers, P. (2017). Valence and value. *Philosophy and Phenomenological Research*, 97(3): 658-680. doi:10.1111/phpr.12395

Charland, L. C. (2005). The heat of emotion: Valence and the demarcation problem. *Journal of consciousness studies*, 12(8-9): 82-102.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.

Colombetti, G. (2005). Appraising valence. *Journal of consciousness studies*, 12(8-10): 103-126.

Corns, J. (2014). Unpleasantness, motivational oomph, and painfulness. *Mind & Language*, 29(2): 238-254.

Craig, A.D. (2015). *How do you feel? an interoceptive moment with your neurobiological self*. New Jersey: Princeton University Press.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason and the human brain*. New York: Putnam.

Davidson, R. J. (1993). Cerebral asymmetry and emotion: Conceptual and methodological conundrums. *Cognition & Emotion*, 7(1): 115-138.

Dyer, M. G. (1987). Emotions and their computations: Three computer models. *Cognition and emotion*, 1(3): 323-347.

Frijda, N. H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

Frijda, N. H. (2008). The psychologists' point of view. En M. Lewis, J.M, Haviland-Jones, y L.F. Barret (eds.), *Handbook of emotions*, 3rd. ed., pp. 68-87. New York: Guilford Press.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Izard, C.E. (1991). *The Psychology of Emotions*. New York: Plenum.

Klein, C. (2007). An imperative theory of pains. *Journal of Philosophy*, 104(10): 517-532.

Lazarus, R. S., Lazarus, R. S. (1991). *Emotion and adaptation*. En L.A. Pervin (ed.), Handbook of Personality: Theory and Research, pp. 609-637. New York: Guilford.

Lewin, K. (1935). *A Dynamic Theory of Personality: Selected Papers*. Trans. D.K. Adams y K.E. Zener. New York: McGraw-Hill.

MacLean, P. D. (1993). Cerebral evolution of emotion. En M. Lewis, y J.M. Haviland (eds.), *Handbook of emotions*, pp. 67-83. New York: Guilford Press.

Martínez, M. (2011). Imperative content and the painfulness of pain. *Phenomenology and the Cognitive Sciences*, 10(1): 67-90.

Murphy, F.C., Nimmo-Smith, I., Lawrence, A.D. (2003). Functional neuroanatomy of emotions: a meta-analysis. *Cognitive and Affective Behavioral Neuroscience*, 3(3): 207-233

Naqvi, N.H., Bechara, A. (2010). The insula and drug addiction: an interoceptive view of pleasure, urges, and decision-making. *Brain Structure and Function*, 214(0): 435-450.

Ortony, A., Clore, G.L., Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Panksepp, J. (2005). On the embodied neural nature of core emotional affects. *Journal of Consciousness Studies*, 12(8-10): 158-184.

Phan, K.L., Wager, T., Taylor, S.F., Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage*, 16(2): 331-348.

Picard, R.W. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

Prinz, J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford University Press

Prinz, J. (2010). For valence. *Emotion Review*, 2(1): 5-13.

Prinz, J. (2012). *The conscious brain*. Oxford: Oxford University Press.

Rozin, P. (2003). Introduction: Evolutionary and cultural perspectives on affect. En R.J. Davidson, K.R. Scherer y H.H Goldsmith (eds.), *Handbook of Affective Sciences*, pp. 839-852. Oxford: Oxford University Press.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1): 145-172.

Schneirla, T.C. (1959). An evolutionary and developmental theory of bi-phasic processes underlying approach and withdrawal. En M.R. Jones (ed.), *Nebraska Symposium on Motivation*, pp. 1-42. Lincoln: University of Nebraska Press.

Schroeder, T. (2004). *Three faces of desire*. Oxford: Oxford University Press.

Schultz, W., Dayan, P., Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306): 1593-1599.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, 17(11): 565-573.

Solomon, R.C. (2003). Against valence. En C. Solomon (ed.), Not passion's slave, pp. 135-147. New York: Oxford University Press.

Tolman, E.C. (1949). *Purposive Behavior in Animals and Man*. Berkeley: University of California Press.

Vytal, K., Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of cognitive neuroscience*, 22(12): 2864-2885.

Sobre el autor

José M. Araya es Doctor en Filosofía por la Universidad de Edimburgo (Escocia), Magíster en Estudios Cognitivos y Licenciado en Filosofía por la Universidad de Chile. Su trabajo se centra en la integración de resultados empíricos provenientes de las ciencias afectivas y ciencias cognitivas con el fin de abordar problemas en filosofía de las emociones y filosofía de la mente. Actualmente trabaja como investigador postdoctoral en el Instituto de Filosofía y Ciencias de la Complejidad (IFICC, Chile), y como profesor en la Universidad de Talca (Chile).

Capítulo 5

Del problema emoción-cognición a la integración de la fenomenología y la intencionalidad de los estados mentales

Rodolfo Bachler

Resumen

Se analiza el problema emoción/cognición con el objetivo de mostrar que este es subsidiario de otro dilema previo y fundamental: el de las relaciones existentes entre las propiedades fenoménicas e intencionales de la mente. A partir de esa premisa, se argumenta que, en lugar de priorizar el análisis de las relaciones emoción/cognición, una estrategia más eficaz para avanzar en la comprensión de la mente corresponde al examen de los diferentes modos de integración que existen entre la fenomenología y la intencionalidad de los estados mentales. Siguiendo este enfoque, se examinan dos modos de combinación entre ambos "ingredientes" de lo mental, uno de los cuales daría origen a la construcción de las emociones y el otro, a una variedad específica de pensamiento. Se concluye reforzando la inconveniencia de la dicotomía emoción/cognición y reflexionando, además, sobre las razones que dificultan apreciar la integración entre los dos tipos de propiedades fundamentales de la mente señalados.

Palabras Clave: emociones, cognición, fenomenología, intencionalidad, *qualia*.

1. Cognición y emoción: una dicotomía que se diluye

Según una clásica definición de cognición, ésta corresponde a "todos los procesos por los que la información de los sentidos se transforma, reduce, elabora, guarda, recupera y utiliza" (Neisser, 1967). Además, en el corazón del "mundo cognitivo", se asume que el objeto de los procesamientos descritos por Neisser son las representaciones o, dicho de forma más general, que es el carácter intencional de los estados mentales aquello que permite definirlos como cognitivos (Rabosi, 1995). En este contexto, tradicionalmente, se considera que pensar, quizá el proceso cognitivo más prototípico de todos, consiste en manipular símbolos representacionales en base a reglas (Searle, 1996). De esta forma, lo que ocurriría en nuestra mente cuando pensamos, es que producimos y manipulamos objetos intencionales (Brentano, 1935). Por ejemplo, si pienso en mi perro, puedo tener en mi mente una o varias imágenes de éste, veo su hocico, sus largas orejas y su cola peluda moviéndose. En este caso, las imágenes que se realizan en mi mente re-presentan, es decir, vuelven a presentarme, esta vez de forma mental, a mi perro físico real permitiéndome así pensar en él. Por otro lado, también pudiera ocurrir que piense en mi perro a través de una o varias palabras, en cuyo caso, estas re-presentarían también (aunque de un modo distinto que no analizaremos aquí) a mi mascota. En este último caso, puedo pensar realizando mentalmente términos como "Late" (el nombre de mi perro), "compañero fiel", "perro bueno", etc., cadenas de símbolos aprendidos que también representan a mi can. Pensar, es entonces, y desde este punto de vista, re-presentarse el mundo a través de un código icónico y/o lingüístico, y manipular esas representaciones de distintas formas.

Por otra parte, experimentar una emoción particular es el estado mental que de mejor forma ilustra aquello que en la literatura se conoce como "procesos afectivos". Se trata de un tipo de estado que, a diferencia de los denominados "cognitivos", se caracteriza principalmente porque consiste en experimentar en la mente algún tipo de propiedad cualitativa, también llamada "*quale*". Así, quien está en un estado emo-

cional "x", experimenta un tipo particular de fenomenología diferente, por ejemplo, de otra que pudiera asociarse a otro tipo de procesos psicológicos, tales como la percepción o la memoria o incluso a otro tipo de emociones específicas. Tener tristeza es, en términos nucleares, sentirse de un particular modo, mientras que alegrarse, es experimentar otro tipo de propiedad cualitativa específica y, en este caso, aquello que caracteriza a la emoción no es una conexión representacional con el mundo, como en el caso del pensamiento, sino que el carácter fenomenológico o experiencial que presenta.

Hasta aquí todo bien, ya que mirados de esta forma los procesos cognitivos y emocionales, éstos parecen ser dos tipos de estados claramente diferenciables. Los primeros serían procesos de carácter representacional o intencional, mientras que los segundos corresponderían a estados de tipo fenomenológico. No obstante, el problema se complejiza puesto que, desde hace algunos años, filósofos y neuro-científicos vienen debatiendo sobre el carácter borroso de los límites existentes entre ambos tipos de procesos. En este contexto, como analizaré a continuación, existen enfoques desde los cuales las emociones pueden concebirse como estados cognitivos, mientras que los procesos cognitivos por su parte, pueden ser considerados como fenómenos emocionales.

1.1 El carácter cognitivo de las emociones

Desde el ámbito de estudio de la filosofía de las emociones, diferentes autores han comenzado a referirse al carácter cognitivo de dichos estados (De Sousa, 1987; Nussbaum, 2008) afirmando con ello que, en un cierto sentido, las emociones no serían tan distintas del pensamiento. Dos argumentos son los que priman para sostener esta afirmación entre los autores. Por un lado, existe un enfoque que entiende la emoción como información implícita, un tipo de "conocimiento" que, a diferencia de aquel que se deriva de los procesos tradicionalmente concebidos como cognitivos, no es explícito o declarativo. Como afirma Solomon (2007), desde esta perspectiva, tener ira es tener distinguir con claridad. ¿Qué ocu-

rrirá si examinamos ahora este problema desde la perspectiva de la constitución de los llamados procesos cognitivos? una apreciación implícita respecto de que estoy siendo tratado injustamente. Del mismo modo, experimentar nostalgia, equivale a valorar la existencia de tiempos pretéritos por sobre un presente considerado como menos afortunado. Como puede apreciarse, la emoción así entendida es cognitiva puesto que nos provee de información sobre el mundo y sobre nosotros mismos.

Por otra parte, existe otro tipo de argumento que puede utilizarse para sostener la idea de las emociones como hechos cognitivos. Se trata de la constatación de que dichos estados no son meramente fenomenológicos, ya que suelen mantener conexiones con objetos representacionales. Como la experiencia cotidiana nos demuestra, tener miedo es "tener miedo de", al igual que enojarse, es "enojarse por". Desde este punto de vista, tener una emoción no es "tan sólo" experimentar algo, o dicho de otra forma, estar en un estado emocional "X" supone no sólo experimentar un *quale* particular, sino también, un contenido intencional. La anterior es la perspectiva de Barrett (Barret y Russell, 2015), quién considera que el término "emoción" debe restringirse sólo para aquellos tipos de afectividad que se encuentran ligados a contenidos representacionales o intencionales, reservando el concepto de "afecto nuclear" para aquellos estados fenomenológicos que no presentan intencionalidad. Examinaré esta perspectiva con mayor detalle más adelante.

Resumiendo lo examinado en este apartado existen al menos dos perspectivas diferentes que permiten afirmar que las emociones son estados cognitivos. La primera parte de la base de que las emociones son conocimiento implícito puesto que proveen de información no declarativa al organismo. La segunda, denominada construccionismo, considera que dichos estados son cognitivos en virtud de su vínculo con objetos representacionales. El análisis de ambas perspectivas nos sugiere que oponer las emociones a los procesos cognitivos no parece ser una buena estrategia para avanzar en la comprensión del funcionamiento de la mente, puesto que no se trataría, en verdad, de dos tipos de procesos que se puedan.

1.2 El carácter emocional de la cognición

En la vereda del frente, si consideramos las principales características de los procesos cognitivos, encontramos una serie de datos empíricos que dan cuenta de la estrecha integración entre emociones y procesos cognitivos que existe en el cerebro (Pessoa, 2013). Estudios experimentales muestran que la activación de áreas cerebrales como la ínsula anterior o la corteza somato-sensorial, se produce indistintamente, tanto para aquellos procesos que llamamos cognitivos, así como, para otros que solemos denominar como afectivos (Gu, Liu, Van Dam, Patrick, y Fan, 2012). Otros estudios muestran que la amígdala, una estructura del cerebro que tradicionalmente ha sido identificada con el funcionamiento afectivo (LeDoux, 2000), contiene proyecciones hacia un área frecuentemente asociada a los procesos cognitivos, tal cual es el caso de la corteza prefrontal (Kandel, Jessell, y Schwartz, 1999). Esto podría sugerir que la activación (emocional) de la amígdala implica activación (cognitiva) de la corteza prefrontal y viceversa. Hallazgos como los anteriores han llevado a algunos neurocientíficos a plantear que todo proceso cognitivo se encuentra constituido por propiedades fenoménicas, aquellas que solemos identificar con las emociones:

> Dado que, todos los objetos y los hechos tienen consecuencias somato-viscerales, las experiencias cognitivas y sensoriales están necesariamente impregnadas en algún grado por la afectividad. No hay tal cosa como un "pensamiento no afectivo". La afectividad juega un rol en la percepción y la cognición, incluso cuando la gente no puede sentir su influencia (Duncan y Barrett, 2007, p. 1185).

El tipo de antecedentes presentados anteriormente puede ser considerado como una invitación a pensar que todo proceso cognitivo contendría en su interior una dimensión afectivo cualitativa, derivada del funcionamiento al unísono de estructuras cerebrales de tipo afectivo. Incluso en aquellos casos de procesamiento cognitivo de alto nivel de abstracción, como pueden ser la lógica o las matemáticas, nos encontramos con conceptos que se deri-

van de metáforas que muchas veces tienen un sustento último en nuestra estructura corporal (Lakoff y Núñez, 2000) y son en un cierto sentido, fenoménicas o experienciales. Esto último, se deriva del hecho de que se encuentran sustentadas en mecanismos cognitivos tales como las metáforas conceptuales que "nos permiten proyectar la estructura inferencial surgida de un dominio fuente generalmente anclado en la *experiencia* concreta del mundo real (como la experiencia térmica, por ejemplo), hacia otro dominio generalmente más abstracto" (Núñez, 2018, pág. 274, énfasis del autor).

Finalmente, también es factible rastrear las bases de esta, al parecer ineludible integración cognitivo emocional, desde un punto de vista evolutivo. En este contexto, Jean Piaget, probablemente el más importante investigador sobre desarrollo cognitivo de todos los tiempos, elaboró su teoría pensando que las capacidades representacionales aparecían en el ser humano alrededor de los dieciocho meses de vida. Antes de ese momento, según dicho autor, el pensamiento de los bebés era exclusivamente sensorio-motriz, es decir, se configuraba a partir de las sensaciones experimentadas como consecuencia de la percepción del mundo y el movimiento del cuerpo únicamente (Piaget, 2001). Sin embargo, lo que parte de la investigación más reciente sugiere, es que las capacidades representacionales se encuentran presentes prácticamente desde el nacimiento (Papalia, Olds, y Feldman, 2009) e incluso, muchas de éstas son, en un inicio, representaciones plenamente encarnadas (Meltzoff y Moore, 1994). Dado este último punto, puede plantearse que la integración emocional cognitiva o dicho de mejor forma, fenomenológico representacional, es una característica, si no innata, al menos muy temprana en el desarrollo humano.

Hallazgos como los analizados en los apartados anteriores, permiten pensar que es probable que "en la cabeza" no haya nunca símbolos o representaciones carentes de *qualia*, por lo que el procesamiento cognitivo no tendría, en ningún caso, un carácter puramente sintáctico, al decir de Searle (1996). Las representaciones mentales que emergen cuando pensamos tendrían siempre una dimensión afectiva. Se trata de una característica de la mente que,

desde otro punto de vista, resulta evidente. ¿De qué forma pudiéramos percatarnos de estar pensando, si no fuese porque experimentamos nuestros procesos cognitivos en un sentido fenomenológico o afectivo? Los seres humanos no "vemos" nuestro pensamiento bajo la forma de un desfile de representaciones sobre un escenario. Los símbolos que se realizan en nuestra mente, sean palabras o imágenes, no pasan delante nuestro como en un "teatro de la conciencia", por utilizar la metáfora de (Baars, 1997). Lo que verdaderamente ocurre, es que sentimos lo que pensamos, y es esta *sentiencia* , aquello que permite que nos percatemos de que estamos teniendo una idea en la mente. Otras perspectivas sobre este problema corren el riesgo de hacernos caer en una posición dualista al exigir que asumamos que, por un lado, "marchan" las representaciones en nuestra mente, mientras que, por otro, estoy yo observando cómo los símbolos se despliegan. No es éste el funcionamiento de la mente y, actualmente, contamos con bastante evidencia acumulada que cuestiona la existencia de homúnculos de este tipo. El pensamiento fluye de otra forma, los seres humanos experimentamos o sentimos nuestro pensamiento en vez de "verlo". No obstante, con mucha frecuencia y de modo no autoconsciente, hacemos uso de herramientas culturales como el lenguaje para explicitar nuestras ideas, un mecanismo que tal vez alimente la ilusión de ver el pensamiento como símbolos desfilando en el espacio mental. Este mecanismo es un punto clave de la hipótesis que presentaré más adelante y que denomino "*qualia* como centro de los procesos cognitivos" la cual vendrá a afirmar que el pensamiento es en su origen, un estado fenomenológico que es re-formateado como símbolos o representaciones mentales. Cuando esto ocurre, reconfiguramos nuestro pensamiento, dejando en segundo plano a sus características fenomenológicas, una situación que tal vez contribuya a generar la errada sensación de cognición como proceso frío.

3. Del análisis de la relación emoción-cognición al estudio del vínculo existente entre las propiedades fenomenológicas y representacionales de la mente

En base a los antecedentes examinados, tanto respecto del carácter cognitivo de las emociones, así como sobre las propiedades afectivas de la cognición podría decirse que el problema emoción-cognición es un falso problema. Dicho de otra forma, el dualismo cognitivo emocional es una perspectiva que no se condice adecuadamente con algunos estudios empíricos actuales sobre el funcionamiento de la mente y el cerebro, puesto que, emociones y cognición no parecen ser dos tipos de estados mentales total y completamente diferentes como se nos ha planteado históricamente desde la filosofía (Casado y Colomo, 2006) y la psicología (Eich, Kihlstrom, Bower, y Forgas, 2003).

Lo anterior no equivale a afirmar que no existan las emociones como la alegría o la esperanza, ni los procesos cognitivos tales como el pensamiento o la reflexión. Evidentemente, ambos tipos de estados forman parte del espectro de lo mental, no obstante, lo que ocurre es que no existen como entidades naturales en el sentido biológico del término, ni tampoco, como unidades discretas de la mente. Respecto de lo primero, sabemos ya desde hace algunos años, que las emociones no se asocian con circuitos neuronales específicos y diferenciados, ni tampoco con un lenguaje facial determinado. Al decir de (Barrett, 2018, p 484), "la investigación no ha revelado una huella dactilar corporal consistente ni siquiera para una sola emoción". Por otro lado, y en parte como consecuencia de lo anterior, tampoco cabe referirse a las emociones como entidades discretas como ha venido haciendo la psicología durante las últimas décadas bajo el concepto de "emociones básicas" (Ekman, 2007; Izard, 1991). No existen el miedo, la rabia, la alegría, etc., como compartimentos estancos y discretos de la mente. Lo que hay detrás de aquello que llamamos "emociones" es un permanente fluir de estados cualitativos que, desde el construccionismo, ha sido llamado "afecto nuclear". Se trata de un "barómetro emocional" que varía en diferentes grados de intensidad entre el placer y desagrado, en sintonía con las condiciones de nuestro medio

ambiente interno y externo. Cambia el mundo, cambia nuestra experiencia. Se presentan obstáculos "allá afuera", nos sentimos incómodos "aquí dentro". Por otro lado, cuando las cosas se dan de buena forma en el mundo, nos sentimos internamente a gusto. Podemos nombrar a estos estados como rabia o alegría respectivamente, pero aquello corresponde a una construcción realizada por nuestra mente, no a entidades naturales y discretas para las cuales venimos biológicamente programados. Además, la gran parte del tiempo, este fluir permanece únicamente como un campo de experiencia difuso y no consciente, de una forma parecida a lo que algunos autores han llamado "Trasfondo" (Searle, 1996). Erróneamente, decimos que nos damos cuenta de nuestras emociones, cuando, en verdad, lo que hacemos es construirlas puesto que no existían como tales antes de ser nombradas.

Algo similar es lo que ocurre con el pensamiento. Tampoco hay ideas como entidades discretas en nuestra psique, el pensamiento es otra forma de flujo mental. Podemos explicitar e incluso formalizar una idea, utilizando, por ejemplo, un soporte externo como un cuaderno, un computador o una conversación con otra persona. Luego leeremos lo escrito y pensaremos que es así como se encontraba en nuestra mente desde un inicio. Pero no es correcto. En la mente nunca hubo una idea en esos términos. Como afirma la escritora estadounidense Joan Didion: "yo no sé lo que pienso hasta que no lo he puesto por escrito". Hablamos de nuestra experiencia como "emociones" al igual que nos referimos al pensamiento en términos de ideas, pero en ninguno de los dos casos, los términos refieren a entidades naturales y discretas. Del mismo modo que las emociones no son estados autónomos y biológicamente programados, el pensamiento no es una colección de estados representacionales organizados e independientes. Ambos son el flujo resultante de la combinación de ingredientes primarios de la mente humana.

El problema emoción cognición se encuentra encastrado entonces en un dilema anterior y fundante, cual es, el de las relaciones existentes entre las propiedades fenomenológicas y representacionales de la mente. En este contexto, si queremos avanzar

en su comprensión, deberemos abocarnos al estudio de los modos de integración de estos "ingredientes" más básicos de los estados mentales, los cuales dan lugar a las configuraciones complejas que llamamos emociones, pensamiento, creencias, actitudes, etc. Una de estas modalidades, es aquella que ha estudiado Barret y su equipo (Barret y Russell, 2015), la cual da origen a las emociones según he descrito en apartados anteriores, y que profundizare en la sección siguiente. No obstante, esta no la única forma de combinación posible. Para progresar en el camino de comprensión de la mente, resulta necesario examinar otros modos de integración. En particular, y dado su carácter prototípico ya señalado, me parece necesario analizar de qué forma se combinan la fenomenología y la intencionalidad de la mente para hacer emerger el proceso cognitivo que llamamos "pensamiento".

En lo que sigue, intentaré avanzar en este sentido, mostrando, primero, el modo bajo el cual se generan las emociones según el construccionismo, dando pie a continuación, a una propuesta de combinación entre fenomenología y representaciones que produciría un particular tipo de pensamiento. La propuesta que se presenta puede ser considerada como una especulación razonada que requerirá, además de la necesaria discusión filosófica, de un respaldo de investigación empírica que permita su rechazo o continuidad como hipótesis explicativa.

4. La construcción mental de las emociones

El siguiente ejemplo, tomado de Bächler y Pozo (2016) puede ayudarnos a entender a qué se refiere esta modalidad de integración entre las propiedades fenoménicas e intencionales de la mente:

> Juan vino a casa después de una larga jornada de trabajo. El día fue complejo, comenzando con una conversación con su jefe nada más llegar a la oficina. Durante este encuentro, Juan fue interpelado duramente por un problema que, en último término, no dependía de él. Sin embargo, Juan, que no es un tipo dado a las confrontaciones, asumió la situación sin reclamos y salió de la ofi-

> cina de su jefe en silencio. Durante el resto del día Juan se dedicó a trabajar sin pausa, como es su costumbre, y aunque a ratos su mente se teñía de un trasfondo emocional de incomodidad, no colocó su atención en ello. Por el contrario, ese día, Juan trabajó sin descanso, de forma más intensa aún que lo habitual, distrayendo su mente en las tareas laborales. Sin embargo, al llegar a casa por la noche, Juan se encontró con su pareja quién nada más verle le señaló efusivamente: ¡Qué mala cara traes! ¿Tuviste algún problema en la oficina? En ese preciso instante se agolparon en la mente de Juan las imágenes de la conversación mantenida con su jefe por la mañana y resonaron (ahora agresivamente) algunas de las palabras expresadas por su superior. Juan se percató entonces de una intensa sensación de rabia, emoción que le llevó a reflexionar que de forma permanente es maltratado por su jefe, y que, con su silencio, él avala esta relación.

¿De qué forma puede describirse teóricamente el proceso realizado por Juan? Lo primero que cabe decir al respecto, es que si consideramos la secuencia mental descrita habría que distinguir a lo menos dos grandes momentos que la componen. El primero, corresponde a la experiencia de Juan antes de llegar a casa. Lo que durante esta etapa ocurre, es que en la mente de nuestro protagonista emerge un *quale* específico que, "desde fuera", pudiéramos denominar como "nerviosismo" o "inquietud". Se trata de un estado consciente, es decir, que Juan experimenta como parte de su vida mental, junto con muchos otros que surgen momento a momento. Para describir adecuadamente este estado en particular, es necesario decir que se trata de una fenomenología "difusa" o "desorganizada" puesto que no ha sido conceptualizada, o, dicho de otro modo, se encuentra todavía sin categorizar. En palabras de Barrett (2011) corresponde a "afecto nuclear", es decir, cambios fisiológicos en el organismo que son experimentados con una valencia (desagradable en este caso) y un cierto grado de activación o intensidad. Este estado no es, sin embargo, auto-consciente, puesto que Juan no se percata de su realización.

El segundo momento, corresponde a lo que ocurre en la mente de Juan al llegar a casa. A partir de una breve conversación con su pareja, Juan realiza la operación de "examinar" su estado afecti-

vo, conectándolo o ligándolo con entidades representacionales, las cuales son, por un lado, imágenes de la conversación mantenida con su jefe y, por otra parte, juicios evaluativos respecto del trato que recibe de parte de su superior. Producto de la operación anterior, en la mente de Juan emerge un nuevo estado, el cual tiene algunas características distintivas respecto de aquel que se produjo en la etapa anterior. Podríamos decir que la mente de Juan ha mutado, en lo que a este aspecto específico se refiere, desde un estado afectivo consciente, desorganizado, y carente de contenido, hacia uno auto-consciente, categorizado o conceptualizado y con contenido intencional. La figura 1 ilustra este "tránsito de lo mental".

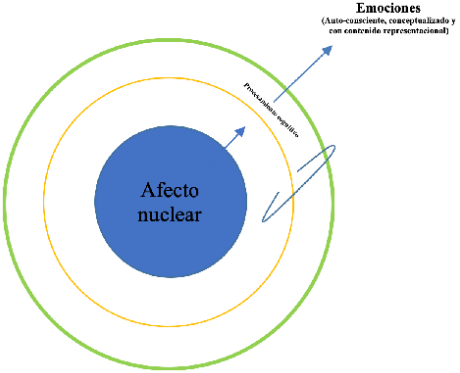

Figura 1. El afecto nuclear y su procesamiento cognitivo.

Barret (2006) denominaría al segundo estado de Juan bajo el término "emoción", un concepto que remite a un estado complejo compuesto de dos elementos al menos. Por un lado, una propiedad fenomenológica, que anteriormente describimos como nerviosismo o inquietud, y, por otra parte, una secuencia de representaciones mentales explicitadas, las cuales se "ligan" a la experiencia interna permitiendo su categorización. Es decir, junto a la sensación interna o subjetivamente experimentada, hay ahora cadenas de representaciones mentales que resultan indisociables del aspecto fenomenológico de las mismas. El tránsito desde el afecto nuclear hasta la emoción implica un procesamiento valorativo y

categorial de la experiencia afectiva: "ambos procesos, la valorización y la categorización modifican el estado de la persona creando un producto emergente que es a la vez afectivo y conceptual" (Barrett, 2006, pág. 36, traducción del autor). El corchete incluido en la Figura 1 muestra que se trata de un proceso que no es lineal, sino que se produce a través de mutuos constreñimientos entre el afecto nuclear y las distinciones representacionales que posibilitan la conceptualización del estado afectivo nuclear. Así, a través de distinciones cognitivas, preferentemente lingüísticas, nuestra experiencia emocional básica es ligada a determinados conceptos que permiten su organización otorgándole un carácter intencional que la dirige sobre algo distinta de ella misma. Podríamos decir que, a través de este proceso de categorización, la fenomenología rompe su encapsulamiento original y se expande o conecta con determinados objetos del mundo interno o externo de quien la experimenta. La característica principal de este nuevo estado, tal cual es experimentado por Juan, sigue siendo, no obstante, su *quale* o fenomenología intrínseca, propiedades que lo distinguen de otro tipo de estados complejos como aquellos que denominamos pensamiento, según analizaré a continuación.

5. ¿De dónde vienen los pensamientos?: el reformateo representacional de los afectos

En un trabajo anterior examiné parte del problema consistente en entender cómo se producen los pensamientos, proponiendo una distinción entre diferentes tipos de *qualia*, y, analizando, además, sus particulares roles en el funcionamiento de los procesos cognitivos (Bächler, 2018). Teniendo como base dicho artículo, en esta oportunidad me centraré en una exploración más detallada de un tipo de integración fenomenológico representacional que allí denominé "*qualia* como centro de los procesos cognitivos". Ésta era una propuesta explicativa acerca del papel que juega la dimensión fenomenológica de la mente en la generación del pensamiento.

El argumento que en esta oportunidad expondré, parte de la base de que antes de que el pensamiento tenga un carácter representacional o simbólico, existe mentalmente, bajo un formato cualitativo o fenomenológico. Así, pensar no sería un acto del todo o nada. Cuando "aparecen" pensamientos en la mente (en un sentido representacional), estos surgen gracias a un procesamiento realizado sobre nuestros estados fenomenológicos previos, por lo que la mente no se encontraría nunca vacía. Se trata de un procesamiento que opera fuera del foco de la auto-conciencia y que consiste en re-formatear los *qualia* originales. Como explicaré más adelante, cuando digo re-formatear, estoy señalando con ello, que, mediante un proceso que requiere del apoyo de prótesis culturales (principalmente el lenguaje verbal), lo que la mente hace, es traducir estados fenomenológicos hacia un formato representacional que deja en segundo plano su fenomenología. El resultado de esta operación deviene en pensamiento o ideas, tal cual las identificamos en nuestra experiencia privada y cotidiana, es decir, como cadenas de símbolos representacionales que manipulamos internamente para realizar inferencias y conocer el mundo. No obstante, éste es sólo el resultado final de un proceso que comienza con la emergencia de estados que originalmente eran fenomenológicos.

El producto que resulta de la operación de re-formateo descripto, no es uno de características simbólicas exclusivamente. Por el contrario, el pensamiento que emerge del procesamiento de nuestros *qualia* lleva en sí el "germen" fenomenológico original. Esta doble dimensionalidad (fenomenológico representacional) del pensamiento es una característica constitutiva de dichos estados, aun cuando en la experiencia privada de los seres humanos, concebimos el pensamiento como un estado representacional exclusivamente. Probablemente, lo anterior ocurre debido a que el estado cualitativo originario, permanece como trasfondo de unos símbolos que surgen como la principal figura del estado mental que llamamos pensar. En este contexto, entender de esta forma el pensamiento implica asumir también, que la distinción entre este proceso y las emociones no es tanto una cuestión ontológica, sino que, más bien, correspondería a un problema epistemológico, es

decir, referido al cómo conocemos o vivenciamos ambos tipos de estados. Experimentamos las emociones como estados de tipo fenomenológico, al igual que tenemos la sensación de que los pensamientos son procesos de carácter representacional exclusivamente. No obstante, ambos estados contienen los dos tipos de propiedades básicas en su interior.

La propuesta que presento a continuación, consistente en un análisis del proceso de re-formatear nuestros *qualia*, surge motivada en gran medida por los resultados de la investigación realizada sobre los procesos de toma de decisiones por el neurocientífico portugués Antonio Damasio y puede considerarse como una reinterpretación filosófica de dicha teoría. No obstante, esta idea bebe también de otras fuentes de conocimiento, tanto de corte científico, como de vertientes alternativas que explicitaré a medida que avance en su desarrollo. Por último, terminando con los prolegómenos, debo decir que, al igual como hice respecto de las emociones, para analizar la construcción del pensamiento, me serviré primero de un ejemplo, en este caso de una vivencia personal, el cual me permitirá ilustrar algunas características de este proceso.

6. Anticipando terremotos: sensaciones que piensan

La mañana del 25 de febrero de 2010 desperté algo más tarde de lo usual para aquella hermosa época de pañales en mi familia. Al ponerme de pie y dirigirme a la salida del cuarto, me encontré con una ruma de libros apilada al lado de la puerta, como impidiendo que la habitación se cerrase. Medio dormido aún, fui en dirección del dormitorio de mis hijos y, para mi sorpresa, otro montón de libros se hallaba junto a la puerta de esta segunda habitación. Recuerdo haber pensado: "mi mujer (a quien pondremos de nombre Javiera) está haciendo orden, seguro que los textos están aquí, en espera de su nuevo lugar en el hogar". A continuación, me dirigí al baño, y al mirar hacia la tina, me encontré con varios baldes y ollas en su interior, todas repletas de agua, como previniendo un inminente y desastroso corte de suministro. Recuerdo haberme incomodado por lo que en ese momento consideré como demasiadas

sorpresas para una tranquila mañana de vacaciones. Fui al salón, decidido a resolver el dilema, y al interrogar a Javiera por la razón de tantos extraños movimientos, recibí como respuesta, de forma seria y decidida, la siguiente expresión: ¿Cómo que qué pasa? ¿No has visto cómo está el ambiente?: ¡Va a haber un terremoto! Recuerdo haber pasado en cuestión de segundos de la sorpresa y el desconcierto, a una franca molestia. Tengo que decir que, por esa época, me encontraba estudiando ciencia cognitiva, y cualquier idea que tuviese un tufillo a esoterismo activaba inmediatamente mis defensas de científico en ciernes. Nada mejor entonces que mi pareja para dar una buena charla acerca de la imposibilidad científica de predecir terremotos. ¡Qué torpe fui! Exactamente dos días después, se produjo en mi país, el cuarto terremoto más grande de la historia del mundo, con una intensidad de 8,8 $M_{\mathbb{W}}$ y cientos de muertos debido al tsunami que arrasó las costas chilenas

Posteriormente, además de la conmoción que para todos los chilenos significó el movimiento tectónico, yo estuve bastante "revuelto" también, debido a la anticipación que de esta catástrofe había hecho mi cónyuge. No podía ser casualidad. Javiera, de nacionalidad española y con poco tiempo en Chile, no había estado jamás en un terremoto, por lo que no me "calzaban" las cosas. Por esta razón, me dediqué durante varias semanas a investigar qué había ocurrido con mi pareja, y como parte de mi indagación, fui insistente en pedirle qué me explicase qué había pasado por su "cabeza", qué era lo que le había sugerido que la tierra temblaría. En un principio, se limitó a decirme que no tenía ni idea, que tan sólo recordaba haber tenido "sensación de terremoto". ¡Pero eso no existe, decía yo! ¡Hay dolores, emociones, pensamientos, actitudes proposicionales, ¡pero en ninguna parte nadie ha hablado jamás de sensaciones de terremoto!

Tanta fue mi insistencia en el tema, que después de unos días, Javiera comenzó a percatarse de algunas cosas que le habían ocurrido antes del terremoto. Recordó, por ejemplo, que cerca de diez días previos a su "premonición" había sentido ganas de abrazar los árboles puesto que se sentía insegura e imaginar que se fundía con las raíces de éstos le tranquilizaba. Además, se dio cuenta de que

antes del seísmo, había llamado su atención el hecho de no ver ni escuchar pájaros por la mañana y que notaba que no había hormigas desplazándose por los troncos de los árboles como es frecuente en el patio de nuestra vivienda. Me habló de estas sensaciones y percepciones y la conclusión a la cual llegamos fue que ella había experimentado, implícitamente, un proceso que puede describirse más o menos de la siguiente forma:

1. Notó cambios en el ambiente, básicamente, el hecho de que no había hormigas en los árboles, ni pájaros cantando por la mañana en el patio de la casa.
2. Experimentó intranquilidad producto de no saber las razones del cambio, pero se dio cuenta de que imaginar una mayor conexión con la tierra a través del contacto con los árboles le tranquilizaba.
3. Supuso que si las hormigas y los pájaros no estaban donde siempre era porque se refugiaban de algo y se preguntó de qué pueden protegerse los animales en un país como Chile.
4. Contando con la información de que Chile es un país sísmico concluyó que era probable que se produjese un terremoto.

Por supuesto, todas estas inferencias son hipótesis explicativas, elaboradas a posteriori, respecto del proceso mental experimentado por mi cónyuge. Lo que Javiera tuvo en mente dos días antes del seísmo fue, tan sólo, una "sensación de terremoto". Es lo que en Chile y Bolivia se denomina una *tincada*, es decir, tener la intuición de que algo ocurrirá. Se trata de un término de raíz quechua (*t'inkay*), una palabra asociada a una ceremonia de agradecimiento a la tierra (*pachamama*) por las cosechas obtenidas. Probablemente el uso de este término guarde alguna relación con un tipo de conocimiento más cercano a lo terrenal y corporal, o como define la Real Academia Española de la Lengua la palabra intuición: "facultad de comprender las cosas instantáneamente, sin necesidad de razonamiento". Al menos, eso fue lo que ocurrió con mi pareja. Ella tuvo una comprensión implícita de que existía la probabilidad

de que se produjese un terremoto. Esta comprensión, se manifestó bajo la forma de una sensación, un *quale*, una experiencia subjetiva o una *tincada* de que aquello ocurriría. ¿De qué forma puede explicarse este fenómeno?

Si examinamos el ejemplo anterior a la luz del trabajo del neurocientífico portugués Antonio Damasio puede que encontremos algunas resonancias que nos permitan comprender de mejor forma el caso descrito. Damasio (2001) señala que "una estima secreta no consciente, precede a cualquier proceso cognitivo sobre el tema" (p.207). En este caso, cuando Damasio dice "estima secreta" está haciendo referencia a "marcadores somáticos", es decir reacciones fisiológicas que se anticipan a la cognición, entendida esta última como proceso representacional en el sentido tradicional del término, según examinamos al comienzo de este trabajo. Su investigación con el juego de cartas de Iowa (Bechara, Damasio, Damasio, y Anderson, 1994), mostró empíricamente, que antes de que las personas comprendan algo de forma explícita, el cuerpo, a través de reacciones fisiológicas de carácter afectivo, anuncia cursos de acción de los eventos de una forma que no es explícita o declarativa. Por otra parte, si bien su estudio presenta claras resonancias con el argumento que estoy desarrollando en este trabajo, debo decir que no es la única investigación que menciona la posibilidad de "adelantamiento experiencial" al surgimiento del pensamiento en un sentido representacional. En el campo de la psicoterapia por ejemplo, la hipótesis que denomino "los *qualia* como centro de la cognición" resulta coherente con los postulados del enfoque psicoterapéutico conocido como *focusing*. La anterior es una teoría que ha demostrado a través de la práctica clínica llevada a cabo por más de cincuenta años que muchos contenidos mentales autoconscientes se encuentran antecedidos por elementos fenomenológicos que, desde esta perspectiva, se denominan "sensaciones sentidas" o *felt meaning* (Gendlin, 1997). Por otro lado, también en el ámbito de las neurociencias, además de la investigación llevada a cabo por Antonio Damasio ya referida, existen estudios que dan cuenta de una anticipación por parte del cerebro a la toma de decisiones consciente (Bode, Hanxi Sel, Soon, Chung, Trampel y Turner,

2011), una situación que permite suponer la posibilidad de que con anterioridad al surgimiento de las representaciones mentales, pudiera haber estados de un orden distinto que constituyan el material sobre el cual se elabore lo que nosotros reconocemos como pensamiento. Esta es al menos, la idea de Russell Hurlburt, investigador de la Universidad de Nevada, quien afirma, a partir de los resultados de su investigación, que no todo pensamiento es simbólico representacional ya que existe una variedad caracterizada por la presencia de imágenes y bits sensoriales, entre otros elementos (Hurlburt y Akhter, 2008).

Los antecedentes anteriores pueden ser interpretados como evidencia a favor de la hipótesis de que, en la mente, antes de que se configuren entidades representacionales, existe una fenomenología no intencional. Sin embargo: ¿De qué forma pasamos de esta "fenomenología pura" al tipo de proceso representacional que popularmente denominamos pensamiento?

7. El reformateo representacional de nuestros afectos

Pozo (2006), en una revisión del modelo de Karmiloff-Smith (1992), plantea que el paso desde lo implícito a lo explícito durante la construcción mental de conocimiento consistiría en un proceso que ocurre en cuatro etapas que finalizan en una re-descripción representacional de lo implícito a través de las herramientas que nos provee la cultura. Según el autor, en directa alusión a la teoría de las emociones de Damasio, aquello que existe en la mente con anterioridad a su explicitación, corresponde a reacciones fisiólogicas que se producen en el cuerpo como consecuencia de las modificaciones que se perciben en el ambiente. Esta fenomenología[1] que tiene el carácter de información implícita, es re-descrita utilizando para ello el lenguaje que nos provee la cultura en la cual estamos insertos. Se trata de un proceso que permite pasar de

[1] Pozo (2008) denomina representaciones implícitas a este tipo de información.

información que es implícita, procedimental y fenomenológica, a un pensamiento de características explícitas, declarativas y representacionales.

Volviendo a los ejemplos presentados anteriormente, en el caso de Javiera, a diferencia de Juan, lo que ella hace no es categorizar su estado afectivo, sino que traducirlo a un formato simbólico o representacional. Específicamente, la sensación de intranquilidad o inseguridad, junto al deseo de conexión con la tierra, son estados que se re-formatean a unos términos simbólico representacionales que pueden sintetizarse bajo la forma sintáctica "tengo la sensación de que va a producirse un terremoto". Esta cadena de símbolos que se producen en la mente de Javiera emerge sinérgicamente como el resultado de la interacción entre sus estados cualitativos y la utilización de un código lingüístico aprehendido culturalmente, el cual incluye, además, un conjunto de conocimientos sobre Chile como país sísmico. Por lo mismo, el producto resultante en este caso, es uno diferente del que emerge en el caso de Juan. Lo que surge en la mente de Javiera es experimentado por ella como un pensamiento, es decir, un estado de características principalmente simbólicas o representacionales. Además, podemos decir que este nuevo estado, a diferencia de la fenomenología que lo origina, tiene como tal, un carácter fugaz o momentáneo, ya que todo lo pensado, una vez pensado desaparece en lo que a su formato simbólico se refiere, quedando sólo la fenomenología como "traza" de lo mental. Desde esta perspectiva, los símbolos son un resultado emergente del proceso de reconstrucción de nuestra fenomenología o experiencia a través del lenguaje, sin embargo, estos no existen como entidades mentales constitutivas de nuestra psique del mismo modo que las sensaciones y otros estados cualitativos. Dicho de otra forma, cada vez que pensamos (en el sentido simbólico representacional) hacemos uso de la herramienta culturalmente dada del lenguaje, la cual nos permite reformatear momentáneamente nuestros estados fenomenológicos. No obstante, una vez realizado este proceso, los símbolos se diluyen y lo pensado vuelve a su estado fenomenológico original. Se trata de un proceso que sucede de forma automática y sin auto-consciencia, en virtud de aprendizajes implícitos

ocurridos tempranamente en nuestro desarrollo y que producto de su constante uso, nos generan la ilusión de tener símbolos en la mente.

Por otra parte, esta concepción del pensamiento como tránsito desde lo implícito-cualitativo hacia lo explícito-representacional, plantea la necesidad de definir el carácter que adquieren las representaciones mentales una vez que son explicitadas. De acuerdo con la perspectiva que estoy defendiendo, el contenido mental resultante de la explicitación, sería una representación que si bien presenta nuevas propiedades (de tipo simbólico), no pierde, sin embargo, su carácter cualitativo original. Los estados implícitos iniciales resultarían momentáneamente modificados durante su explicitación, pero la representación resultante no sería, sin embargo, un estado puramente simbólico, como han expresado algunos cognitivistas clásicos como Fodor (1986). Por el contrario, las nuevas representaciones mentales que se producen mediante la explicitación, más allá de su carácter simbólico momentáneo, llevarían en su interior una dimensión cualitativa original. Además, dado que se trata de un proceso que requiere del uso del lenguaje, podríamos decir que los símbolos o representaciones lingüísticas no estarían verdaderamente en la mente de quien piensa, sino que serían una propiedad emergente y extendida, la cual surge "en el medio" de la interacción entre la fenomenología de un organismo y el lenguaje que "toma prestado" de su contexto cultural. Las características del proceso descrito anteriormente permiten comprender, además, fenómenos como aquel que denominamos "tener algo en la punta de la lengua". Se trataría en este caso, de saber que sabemos algo en virtud de tener un particular tipo de fenomenología realizado en la mente, sin contar, en un momento determinado, con las herramientas lingüístico culturales que permitan su momentánea traducción o reformateo representacional. Por supuesto, varios son los temas que quedan pendientes de examinar respecto de este proceso.

8. Conclusiones, proyecciones y asuntos pendientes

Al comenzar este trabajo he presentado una revisión de las características más distintivas de los procesos emocionales y cognitivos con el objetivo de mostrar que no se trata de entidades completamente distintas. De acuerdo con el argumento esgrimido, ambos tipos de procesos son construcciones mentales realizadas a partir de la combinación de al menos dos ingredientes básicos de todo estado mental: la fenomenología y la intencionalidad de los mismos. El argumento planteado me ha conducido a la necesidad de analizar las diferentes formas de combinación de ambos bloques constructivos de la mente, con el objetivo de comprender de qué forma se producen las emociones y cuáles serían las diferencias más importantes que presenta este proceso respecto de la construcción del pensamiento. Para ello he examinado primero la teoría de Barrett (2011) sobre la construcción de las emociones, y a continuación he presentado mi propia propuesta acerca del modo de producción de un particular tipo de pensamiento de características intuitivas, la cual he denominado "*qualia* como centro de los procesos cognitivos". La propuesta desarrollada permite sacar algunas conclusiones, no obstante, varias nuevas preguntas quedan abiertas para futuros trabajos en esta línea.

Respecto a las conclusiones quisiera destacar en primer lugar, el hecho de que una vez constatada la imposibilidad de disociar los aspectos cualitativos e intencionales de los estados mentales, no resultaría por tanto pertinente, la distinción entre procesos cognitivos y afectivos como tradicionalmente se ha estructurado el estudio de la mente (por ejemplo, Morris y Maisto, 2005). Antes bien, he defendido la tesis de que cada uno de estos procesos psicológicos contiene en su interior, tanto propiedades cualitativas específicas, así como también, una dimensión intencional que la organiza.

En segundo lugar, el estudio de esta modalidad de funcionamiento de la mente nos permite comprender que los seres humanos no conocemos nunca de una forma "fría", representándonos el mundo mediante símbolos "asépticos" carentes de *qualia*. Lo

que en realidad sucede, es que, en el proceso de pensar y conocer, participarían tanto fenómenos cualitativos como las sensaciones, el afecto nuclear y otros similares; así como estados intencionales o representacionales, tradicionalmente denominados como cognitivos. Probablemente, la característica anterior forme parte de aquello que lleva a algunos autores que adhieren al enfoque de la mente encarnada, a sostener que la cognición, no consiste en re-presentar un mundo pre dado, sino más bien, en plantear las cuestiones relevantes que van surgiendo en cada momento de nuestra vida. Habría que agregar, sin embargo, que "lo relevante" en este caso, es un hecho que se define por el carácter cualitativo de la experiencia, en tanto fenómeno subjetivo, sensitivo y corporal.

En tercer lugar, resulta indudable que más allá de la profunda integración existente entre las propiedades fenomenológicas e intencionales de la mente que he examinado, existen factores de diverso orden, que dificultan apreciar esta característica de funcionamiento de la mente. En primer lugar, en Occidente vivimos inmersos en una cultura dualista que al separar cuerpo y mente (Berman, 1987), nos incita, de forma indirecta, a dicotomizar también la razón de la emoción. Esta separación se produce bajo el entendido implícito de que las emociones residen en el cuerpo mientras que los procesos cognitivos morarían en el nivel de lo mental. Este modo de concebir el problema se ha visto reforzado, además, por la importancia que adquirió en el siglo recién pasado la metáfora computacional como modelo explicativo de las relaciones mente cerebro (Searle, 1996).

Asimismo, existen causas de otro orden detrás del mantenimiento de esta perspectiva. Específicamente, es probable que el mismo modo de funcionamiento de la mente descrito en este trabajo, sea también, un factor que se encuentre en la base de nuestra dificultad para apreciar la integración fenomenológico representacional de la psique. Lo anterior ocurriría puesto que las formas a través de las cuales nos percatamos de la realización de propiedades fenomenológicas o representacionales en la mente, es justamente un mecanismo que nos dificulta apreciar la integración existente entre ambas dimensiones. Así, cuando reformateamos nuestros

estados cualitativos, estos adquieren un carácter representacional más destacado, mientras que cuando conceptualizamos nuestra emoción, nos hacemos mayormente conscientes de sus propiedades cualitativas. De esta forma, en uno y otro caso, "perdemos de vista" la integración de ambas propiedades.

En último lugar, respecto de las preguntas que se abren con este trabajo, hay dos que me parecen las más importantes de plantear. En primer lugar, que resulta necesario clarificar con mayor detalle cuáles serían exactamente las operaciones que la mente realiza cuando categoriza o reformatea representacionalmente la fenomenología. Respecto de este punto, si bien aquí no he explicado en un sentido estricto de qué forma pasamos de un estado a otro, y que, en ese contexto, el principal aporte de este trabajo se encontraría más bien en distinguir la existencia de dos procesos o mecanismos para la producción de diferentes tipos de estados mentales.

Por otra parte, más allá de las explicaciones que quedan pendientes, queda abierta también un interrogante referido a la identificación del principio en virtud del cual cierto tipo de fenomenología se categoriza mientras que otra es reformateada. Mi impresión respecto de este punto, es que la diferencia se encontraría en las características del tipo de fenomenología que operaría como materia prima en uno y otro caso. En el caso de las emociones, se trata de *qualia* no perceptuales respecto de los cuales, somos conscientes de su subjetividad intrínseca. En el segundo caso en cambio, la fenomenología que daría origen al pensamiento correspondería principalmente a *qualia* perceptuales, es decir, sensaciones respecto de los cuales no somos completamente conscientes de su subjetividad (Bächler, 2018). Futuros trabajos deberán abocarse a la tarea de analizar estos asuntos pendientes.

De cualquier forma, pueden concebirse ambos mecanismos de producción de estados mentales como procesos complementarios que se superponen, ya que es probable que justamente en el mismo momento en el cual estamos conceptualizando una emoción, comenzamos también a reformatearla dando origen a pensamiento en el sentido representacional del término. Por tanto, los análisis

realizados en este trabajo serían sólo desagregaciones de mecanismos que operan de forma holista e integrada en un flujo continuo de experiencia mental que no distingue ningún tipo de separación entre estados y propiedades.

Referencias

Baars, B. (1997). *In the theater of consciousness: The workspace of the mind.* New York: Oxford University Press.

Bächler, R. (2018). Desagregando los qualia: un análisis de su función en los procesos cognitivos. *Universitas philosophica*, 35(70): 15-41.

Bächler, R., Pozo, J.-I. (2016). ¿Siento, luego enseño? Concepciones docentes sobre las relaciones entre las emociones y los procesos de enseñanza/aprendizaje. *Infancia y Aprendizaje*, 39(2): 1-38.

Barret, L.F. (2018). *La vida secreta del cerebro.* Barcelona: Paidos.

Barret, L. F., Russell, J. (2015). *The Psychological Contruction of Emotion.* London: Guilford Press.

Barrett, L. F. (2011). Constructing emotion. *Psychological Topics*, 20(3): 359-380.

Barrett, L. F., (2006). Solving the Emotion Paradox: Categorization and the Experience of Emotion. *Personality and Social Psychology Review*, 10(1): 20-46.

Bechara, A., Damasio, A. R., Damasio, H., Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50(1): 7-15.

Berman, M. (1987). *El Reencantamiento del Mundo.* Santiago de Chile: Cuatro Vientos.

Bode, S., Hanxi Hel, A., Soon, C., Chung, Trampel, R., Turner, R. (2011). Tracking the Unconscious Generation of Free Decisions Using UItra-High Field Fmri. *PLoS ONE*, 6(6): 1-13.

Brentano, F. (1935). *Psicología desde el punto de vista empírico.* Madrid: Revista de Occidente.

Casado, C., Colomo, R. (2006). Un breve recorrido por la concepción de las emociones en la filosofía occidental. *A Parte Rei*, 47: 1-10. Recuperado de http://serbal.pntic.mec.es/AParteRei/

Damasio, A. R. (2001). *El error de Descartes.* Barcelona: Crítica.

De Sousa, R. (1987). *The Rationality of Emotion.* Cambridge, MA: MIT Press.

Duncan, S., Barrett, L. F. (2007). Affect is a form of cognition: a neurobiological analysis. *Cognition and Emotion*, 21(6): 1184-1211.

Eich, E., Kihlstrom, J., Bower, G., Forgas, N. P. (2003). *Cognición y Emoción.* Bilbao: Desclée de Brouwer.

Ekman, P. (2007). *Emotions revealed: recognizing faces and feelings to improve comunication and emotional life.* New York: Henry Holt and Co

Fodor, J. A. (1986). *La modularidad de la mente: un ensayo sobre la psicología de las facultades*. Madrid: Ediciones Morata.

Gendlin, E. (1997). *Experiencing and the Creation of Meaning. A Philosophical and Psychological Approach to the Subjetive.* Illinois: Northwestern University Press.

Gu, X., Liu, X., Van Dam, N. T., Hof, P. R., Fan, J. (2012). Cognition–emotion integration in the anterior insular cortex. *Cerebral cortex*, 23(1): 20-27.

Horgan, T., Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. En D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: Oxford University Press.

Hurlburt, R. T., Akhter, S. A. (2008). Unsymbolized thinking. *Consciousness and Cognition*, 17(4): 1364-1374.

Izard, C. E. (1991). *The psychology of emotions.* New York: Plenum Press.

Kandel, E., Jessell, T., Schwartz, J. (1999). *Neurociencia y conducta.* Madrid: Prentice hall.

Karmiloff-Smith, A. (1992). *Beyond modularity*. Cambridge: Cambridge.

Lakoff, G., Núñez, R. (2000). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. New York: Basic Books.

LeDoux, J. (2000). *El cerebro emocional.* Barcelona: Planeta.

Meltzoff, A. N., Moore, M. K. (1994). Imitation, memory, and the representation of persons. *Infant behavior and development*, 17(1): 83-99.

Morris, C. G., Maisto, A. (2005). *Psicología.* México DF: Pearson Educación.

Neisser, U. (1967). *Cognitive Psychology.* Englewood Cliffs: Prentice-Hall.

Núñez, R. (2018). Praxis matemática: reflexiones sobre la cognición que la hace posible. *THEORIA*, 33(2): 271-283.

Nussbaum, M. (2008). *Paisajes del pensamiento. La Inteligencia de las emociones.* Barcelona: Paidos.

Papalia, D. E., Olds, S. W., Feldman, R. (2009). *Psicología del desarrollo: de la infancia a la adolescencia.* New York: Mc Graw Hill.

Pessoa, L. (2013). *The cognitive-emotional brain: from interactions to integration* (First ed.). Cambridge: MIT Press.

Piaget, J. (2001). *Inteligencia y afectividad.* Buenos Aires: Aique

Pozo, J. I. (2006). *Adquisición de conocimiento.* Madrid: Morata.

Pozo, J. I. (2008). *Aprendices y Maestros: La Psicología Cognitiva del Aprendizaje.* Madrid: Alianza Editorial.

Rabosi, E. (1995). *Filosofía de la mente y ciencia cognitiva.* Barcelona: Paidos.

Searle, J. (1996). *El redescubrimiento de la mente.* Barcelona: Grijalbo Mondadori.

Solomon, R. (2007). *Etica emocional: una teoría de los sentimientos.* Barcelona: Paidos.

Sobre el autor

Rodolfo Bachler es psicólogo y realizó un Magíster en Estudios Cognitivos en la Facultad de Filosofía de la Universidad de Chile, y posteriormente, un Doctorado en Psicología en la Universidad Autónoma de Madrid. Su línea de investigación es sobre las relaciones emoción-cognición. Actualmente se desempeña como académico del núcleo de psicología de la Universidad Mayor, Chile. Contacto: rodolfo.bachler@gmail.com

Capítulo 6

El camino no elegido: indiferencia ontológica en la filosofía de la mente

Sebastián Sanhueza Rodríguez

Resumen

El problema mente-cuerpo —esto es, la pregunta sobre cómo la mente existe en la naturaleza espacio-temporal— ha dominado buena parte de la filosofía de la mente del siglo XX. El objetivo de esta contribución es sugerir que esta tradición filosófica ha sufrido de cierta indiferencia hacia importantes preguntas sobre la estructura ontológica de nuestros fenómenos mentales. Para concluir, esbozaré cómo un renovado interés en la ontología de la mente está replanteando el hasta ahora estancado problema mente-cuerpo.

Palabras Claves: problema mente-cuerpo, materialismo, ciencias de la mente, procesos, estados.

1. Introducción

Después de definir buena parte de la discusión filosófica sobre la naturaleza de la mente humana a mediados de siglo XX, el problema mente-cuerpo —esto es, la pregunta sobre cómo la mente existe en la naturaleza espacio-temporal— fue rápidamente desplazado por investigaciones sobre aspectos más específicos de nuestra vida mental, e.g. su intencionalidad o su carácter fenomenológico. Aunque este drástico giro podría inicialmente atribuirse al constante devenir de modas filosóficas, aquí asumiré, primero, que la relevancia relativa del problema mente-cuerpo en la escena

filosófica contemporánea responde a razones profundas —esto es, implícitas y fundamentales— y, segundo, que una comprensión más clara de tales razones podría ayudarnos a re-evaluar la historia reciente y algunas perspectivas futuras de la discusión filosófica sobre la mente humana. Sobre ambos presupuestos, este ensayo sugerirá que una resistencia implícita pero poderosa a reflexionar sobre la estructura ontológica de nuestra vida mental —de aquí en adelante, una *indiferencia ontológica*— no sólo ha definido los términos en los cuales el problema mente-cuerpo fue inicialmente formulado, sino también la tendencia posterior a relegar dicho problema a un segundo plano a favor de cuestiones más específicas sobre la intencionalidad y la fenomenología de lo mental. Para finalizar, concluiré introduciendo una línea de investigación naciente en torno a la ontología de la mente para ilustrar cómo es posible reevaluar el problema mente-cuerpo de manera ontológicamente informada: asumiendo que la estructura temporal de un determinado tipo de fenómeno determina en buena medida nuestra comprensión ontológica del mismo —después de todo, el tiempo o la temporalidad es una categoría metafísica fundamental que abarca tanto el dominio de lo físico como el de lo mental— presentaré brevemente un debate filosófico que examina cómo una investigación sobre la estructura temporal de nuestros fenómenos mentales afecta nuestra comprensión del problema mente-cuerpo.

Esta tarea se dividirá así en tres partes. Primero, esbozaré la historia reciente del problema mente-cuerpo en términos muy crudos. Luego, me detendré en algunos aspectos más específicos de dicha historia, para explicar por qué una indiferencia o apatía ontológica ha parecido caracterizar el debate filosófico en cuestión. Finalmente, describiré brevemente cómo el problema mente-cuerpo puede ser informado por un nuevo debate ontológico sobre la estructura temporal de nuestros fenómenos mentales.

2. El problema mente-cuerpo

En términos generales, el problema mente-cuerpo puede entenderse como la pregunta sobre cómo lo mental se relaciona con

lo físico. El componente relacional de esta formulación debe ser entendido de manera extremadamente general. Esto es, la formulación del problema no debería —aunque, históricamente, lo ha hecho a menudo— presuponer en principio cuál es la naturaleza específica de la relación entre lo físico y lo mental: el problema debería ser neutral a si acaso tal relación es causal, reductiva o constitutiva, ya que su solución precisamente busca esclarecer tal cuestión. El problema mente-cuerpo sólo debería asumir que existe una relación, de algún tipo todavía no determinado, entre ítems de carácter mental o psicológico y al menos algunos ítems de carácter físico. En cuanto a las nociones de lo físico y lo mental, pueden pensarse como conceptos generales que pueden ser especificados de distintas formas: mundos físico y mental; eventos físicos y mentales; procesos o estados físicos y mentales; entre otros. Típicamente, el dominio físico abarca todos los objetos, las cualidades, los eventos y estados espacio-temporales que podemos identificar en la realidad cotidiana en la que vivimos (e.g. árboles, colores, explosiones, etc.) y en ciencias naturales como la física, la química y la biología (e.g. átomos, ondas, entre otros). Por ítems mentales, en cambio, se entienden todos aquellos ítems que pueden ser reconocidos por una psicología del sentido común y una científica (e.g. experiencias perceptivas, pensamientos, sensaciones, emociones, acciones, una gran gama de fenómenos y mecanismos subpersonales, entre otros). En esta discusión, adoptaré una formulación relativamente intuitiva del problema: ¿cómo los elementos de nuestras vidas mentales existen en la realidad espacio-temporal revelada tanto por nuestro conocimiento pre-científico como por las ciencias naturales básicas?

A mediados de siglo XX, la discusión de la pregunta anterior se articuló como una reacción sistemática en contra de una doctrina cartesiana sobre la interacción entre mente y cuerpo (cf. Descartes 1642, 1977, meditaciones II y VI; y 1999, 25-31). Aunque el mismo Descartes probablemente no contempló el problema mente-cuerpo entre sus preocupaciones filosóficas, sí defendió, en el contexto de un ambicioso proyecto sobre la justificación del conocimiento humano, un número de tesis que inspiran una respuesta

determinada al problema mente-cuerpo. En términos muy generales, la posición filosófica en cuestión consta de dos tesis principios que pueden expresarse del siguiente modo:

(1) Mente y cuerpo son substancias de distinta naturaleza.

(2) Mente y cuerpo interactúan entre sí mediante relaciones causales.

(1) es una forma concisa de expresar lo siguiente: nuestras vidas mentales son substancias u objetos de una naturaleza completamente distinta a aquella de los sustratos físicos —es decir, nuestros cuerpos o cerebros— gracias a los cuales interactuamos con la realidad espacio-temporal pre-científica y científica. Esta idea captura lo que tradicionalmente se conoce como el componente *dualista* de la respuesta cartesiana al problema mente-cuerpo. (2), mientras tanto, captura un elemento menos idiosincrático, pero igualmente vital, de la teoría en cuestión, a saber, la idea según la cual existe una relación de producción causal entre nuestras vidas mentales y el mundo espacio-temporal. Esta tesis captura el componente i*nteraccionista causal* de la doctrina cartesiana. De tal forma, el dualismo cartesiano se ha configurado en la discusión contemporánea como una teoría según la cual, a pesar de concebir mente y cuerpo como substancias de distintas naturalezas, son capaces de interactuar causalmente entre sí.

Las principales teorías materialistas proliferadas desde mediados del siglo XX en torno al problema mente-cuerpo —a saber, el conductismo, la teoría de identidad mente-cerebro, y el funcionalismo— tienden a ser definidas por dos condiciones: por una parte, el rechazo del dualismo cartesiano; y, por otra, un intento de resolver los problemas asociados a las teorías materialistas que cada una de ellas históricamente intenta reemplazar. Este rechazo casi transversal del dualismo cartesiano se deriva principalmente de una oposición en contra de (1), no de (2). El segundo elemento de la doctrina cartesiana sorprendentemente logra evitar cualquier tipo de escrutinio crítico sistemático. Como veremos en la próxima sección, esta curiosa inmunidad que ha gozado (2) a lo largo de

las últimas décadas es, en mi opinión, un síntoma importante de una tendencia contemporánea a ignorar la dimensión ontológica del problema mente-cuerpo.

El conductismo abarca una familia de teorías que comparten como denominador común la tesis según la cual la mente es simplemente la manifestación de nuestra conducta o la posesión de disposiciones a comportarse de ciertas maneras en respuesta a los estímulos del medioambiente, así como de nuestros procesos y estados internos (cf. Farrell, 1950; Skinner 1951, 2014). Por ejemplo, la experiencia visual que padezco cuando observo las plantas de mi jardín no es, según el conductista, un estado o proceso interno ontológicamente independiente de los procesos físicos que causan dicha experiencia o que se siguen de la misma: más bien, tal experiencia visual sería el conjunto de respuestas conductuales gatilladas por la llegada de información desde las plantas a mi cortex visual primario —donde el concepto de conducta pertinente en este contexto incluye no sólo respuestas conductuales observables (e.g. movimientos corporales a través del espacio), sino también respuestas conductuales no-observables (e.g. reacciones musculares, actos de discriminación o monitoreo visual del medioambiente, entre otras)—. Asimismo, el dolor que siento al tocar una cafetera italiana caliente no sería sólo una sensación introspectivamente accesible a mí y sólo a mí: más bien, sería el conjunto de respuestas conductuales que se siguen o que podrían seguirse del evento de tocar una cafetera caliente. De tal manera, la teoría conductista sostiene que nuestras vidas mentales no son el resultado de la actividad de substancias de una naturaleza distinta a los objetos espacio-temporales con los cuales estamos familiarizados: nuestras vidas mentales deberían ser más bien analizadas o entendidas en términos de las reacciones conductuales de objetos espacio-temporales perfectamente parroquiales, e.g. cerebros o cuerpos.

La posición conductista sin duda es favorecida por una motivación de corte metodológico: al reformular nuestros procesos y estados mentales en términos de respuestas conductuales actuales o posibles, el conductismo hace de la mente un fenómeno plenamente observable y, como tal, científicamente accesible. Al mismo

tiempo, la posición conductista no sólo se distingue del dualismo cartesiano por evitar una concepción de la mente como una substancia inmaterial: de hecho, el conductismo evita concebir la mente humana como una substancia en lo absoluto, inclinándose más bien hacia una concepción disposicional de la mente.

Dicho eso, la ventaja inicial de la posición conductista también tiene un costo crucial. Si ella es correcta, entonces pareciera que (2), la idea según la cual mente y cuerpo interactuan causalmente entre sí, no puede ser verdadera. El conductista concede que el mundo físico impacta causalmente nuestras vidas mentales: nuestras respuestas conductuales o nuestras disposiciones a comportarnos de ciertas maneras serían el resultado de la acción causal del mundo sobre nuestros cuerpos o sistemas nerviosos. Sin embargo, no podría acomodar con la misma elegancia el impacto causal de nuestra vida mental sobre el mundo físico: después de todo, los fenómenos de nuestra vida mental no serían eventos que causan ciertas respuestas conductuales, las cuales a su vez modifican nuestro cuerpo o nuestro entorno físico inmediato; más bien, ¡lo que concebimos como tales fenómenos serían idénticos a nuestras respuestas conductuales (o, al menos, a los respectivos estados disposiciones)! De tal manera, si bien preserva el impacto causal de lo físico sobre lo mental, el conductismo aparentemente menoscaba el impacto causal de lo mental sobre lo físico. En la medida en que esta implicancia ha sido considerada demasiado contraintuitiva, el materialismo posterior ha buscado formas alternativas de articular una respuesta no-cartesiana al problema mente-cuerpo.

Las propuestas materialistas no-conductistas posteriores se han movido en dos direcciones distintas: por una parte, reforzando la relación conceptual entre mente y cerebro (o tal vez los sistemas nerviosos central y periférico); y, por otra, distanciando la esencia de la mente de la base orgánica que la realiza. La teoría de identidad mente-cerebro —también conocida como la teoría materialista de estados centrales o *materialismo australiano* por sus principales defensores, J.J.C. Smart y U.T. Place— es una teoría del primer tipo y la respuesta histórica inmediata a la solución conductista (cf. Place, 1956; Smart, 1959; Feigl 1967). La teoría de identidad

sostiene que el conductismo es correcto en la medida en que evita la ontología mental distintiva del dualismo cartesiano y subraya la conexión íntima entre nuestra vida mental y su manifestación conductual. Sin embargo, también lo considera incompleto en tanto en cuanto no es capaz de reconocer que los fenómenos mentales son, en algún sentido relevante, *internos* a sus respectivos sujetos —una concepción de la mente que a su vez es en buena parte motivada por la idea de que nuestros procesos y estados mentales tienen un impacto causal en nuestra conducta y el mundo externo—. En respuesta a esta dificultad, la teoría de identidad sostiene que nuestros procesos y estados mentales son simplemente procesos y estados cerebrales: estos últimos son causados por la estimulación informacional del mundo externo, y también son capaces de producir efectos conductuales. De tal manera, aunque nuestra vida mental sería interna, no lo sería en un sentido que sugiera las misteriosas entidades del dualismo cartesiano: los fenómenos que la constituyen simplemente son intra-cutáneos o intra-craneales. Si bien tales procesos y estados no son evidentes a la simple observación, pueden ser objetivamente estudiados a través de disciplinas como la psicología cognitiva y la neuro-biología.

Ejemplificando la segunda dirección de desarrollo materialista, una teoría funcionalista evita atar la esencia de la mente a su base neuro-biológica. Según el funcionalista, la existencia de procesos y estados mentales tiene que ser, en algún sentido, distinta a la de la base neural que la realiza en mamíferos como nosotros (cf. Putnam, 1967, 1981; Lewis 1980). Aunque un pulpo, por ejemplo, posee una constitución biológica muy distinta a la nuestra, en principio esto no excluye que pueda tener experiencias perceptivas, sensaciones, entre otros estados y procesos mentales. Algo similar podría decirse sobre entidades mucho más exóticas, e.g. robots pensantes o ángeles, si no es incoherente concebirlas. Asimismo, nuestro conocimiento pre-científico y especializado de nuestras propias mentes y las de otras personas precede por muchos cientos de años a la historia relativamente modesta de las investigaciones neuro-biológicas. ¿Cómo es esto posible si no existe alguna diferencia substantiva entre mente y cerebro? En virtud de estas y

otras consideraciones, el funcionalista modela nuestros procesos y estados mentales en términos del rol causal que juegan dentro de nuestras vidas psicológicas. Aunque nuestras experiencias (e.g. percepciones visuales y auditivas) y sensaciones (e.g. dolores, cosquillas, etc.) sin duda existen en un determinado sistema nervioso, también se caracterizan por tener ciertas causas y efectos. Según el funcionalista, la identidad de tales procesos y estados es definida no tanto por los sistemas nerviosos específicos en los cuales son de tal manera realizados, sino más bien por los rasgos causales antes mencionados. Un determinado tipo de dolor, por ejemplo, sería la clase de estado mental que es en virtud del rango de estímulos que pueden causarlo, así como en virtud del rango de efectos que puede tener en sus respectivos sujetos —en pocas palabras, por la función causal que cumplen en nuestras vidas mentales—. Por cierto, esta posición funcionalista no implicaría que un fenómeno mental podría existir fuera de un determinado sistema orgánico: en este sentido, la teoría es fiel al espíritu materialista que comparte con el conductismo y la teoría de identidad. El punto es más bien que nuestra vida mental podría realizarse en uno u otro sustrato material, ya que este último no es lo que la define como tal. Sin embargo, las funciones causales que definen a nuestros fenómenos mentales deben de todas formas realizarse en algún tipo de base material.

De tal manera, el segundo tercio del siglo XX ha visto el desarrollo paulatino de respuestas materialistas cada vez más sofisticadas al problema mente-cuerpo. Mientras que la teoría de identidad mente-cerebro intentó refinar una posición conductista popular pero aparentemente problemática, el funcionalismo también emerge como una solución a problemas planteados por la teoría de identidad. Ahora bien, aunque la posición funcionalista es actualmente considerada la mejor respuesta disponible al problema mente-cuerpo, no está exenta de problemas (cf. Block, 1978; Putnam 1988, 1990; Kim 1988, 2002). Por ejemplo, parece problemática en la medida en que es posible atribuir estados funcionales a ítems que intuitivamente no clasifican como mentes o fenómenos mentales. El famoso experimento mental de la "nación china" sugiere

la posibilidad de crear una nación de individuos capaces de emular las interacciones eléctricas entre las millones de neuronas que componen al cerebro humano: sin embargo, aun si cada individuo de tal nación llega a imitar a cada neurona de mi cerebro, no es plausible sostener que dicha nación constituye una súper-mente. Asimismo, no parece del todo obvio que una descripción de nuestra vida mental en términos funcionales capture todo su carácter subjetivo. De tal forma, se ha desarrollado una intensa discusión en torno al experimento mental de los "zombies filosóficos", esto es, organismos que, a pesar de ser material y funcionalmente idénticos a nosotros, carecen del conocimiento introspectivo, esto es, aquel a través del cual cada persona típicamente tiene acceso a su propia mente —de forma muy simplista, un zombie filosófico es alguien que luce y se comporta como nosotros, pero que internamente no tiene experiencias o pensamientos—. Ambas dificultades, entre otras, todavía parecen estar vigentes.

En esta sección, he presentado una versión muy cruda del debate reciente sobre el problema mente-cuerpo. El punto crucial no es que esta visión panorámica sea completamente específica, sino que sea posible reconocer en ella el tono general de la discusión filosófica en cuestión. Sobre esta base, defenderé en la próxima sección que la filosofía de la mente contemporánea en general ha prestado poca atención a la estructura ontológica de nuestra vida mental —una tendencia que a su vez no sólo ha determinado el debate mismo sobre el problema mente-cuerpo, sino también las discusiones filosóficas que le han seguido—.

3. Filosofía de la mente sin ontología

En una comprensión muy general pero relativamente familiar de la palabra, una *ontología* puede ser entendida como una investigación filosófica de cualquier aspecto de la realidad, en sus términos más generales. De tal manera, una ontología de la mente puede ser ampliamente entendida como una investigación filosófica sobre nuestra realidad mental, en sus términos más generales posibles. Dicho esto, parecen haber al menos tres momentos cru-

ciales en los cuales los filósofos de la mente contemporáneos no han prestado suficiente atención a la ontología de la mente: primero, orientándose primariamente a ofrecer modelos explicativos de la relación entre mente y cuerpo en lugar de describir nuestra vida mental en sus términos más generales; segundo, asumiendo casi de forma automática cierta comprensión de la interacción causal entre mente y cuerpo; y, tercero, eventualmente dejando de lado el problema anterior para enfocarse en discusiones sobre propiedades más específicas sobre nuestra vida mental, como su carácter intencional y fenomenológico. A continuación, me referiré a cada momento en un poco más de detalle.

Como he enfatizado en la sección anterior, el debate en torno al problema mente-cuerpo crucialmente busca entender la naturaleza específica de las relaciones psico-físicas, dejando de lado un estudio pormenorizado de sus respectivos *relata* —esto es, los elementos así relacionados—. En otras palabras, el punto del debate no ha sido tanto entender lo físico o lo mental como entender la forma en la cual ambos dominios interactúan entre sí. En mi opinión, esta sutil elección de enfoque ya enfrenta una importante dificultad metodológica: pues ¿cómo es posible dilucidar la relación entre lo físico y lo mental sin una comprensión previa de los elementos así relacionados? Intuitivamente, la estructura ontológica de cierta clase de ítems determina el conjunto de relaciones posibles que ítems de ese tipo podrían mantener con ítems de otra clase. De tal manera, por ejemplo, comprender qué es un color, un sonido y un objeto —aun si esta comprensión es extremadamente general— nos ayuda a entender qué clase de relaciones pueden existir entre colores, sabores y objetos. Tal vez de forma más decisiva, tal comprensión nos permite entender qué clase de relaciones no pueden existir entre los ítems que caen bajo las categorías anteriores. Por ejemplo, una noción solamente genérica de qué es un color, un sabor y un objeto, nos permite apreciar que un color no puede tener un sabor (o viceversa) de la misma manera en que un objeto puede tener cierto color o sabor. De manera semejante,

un estudio de la relación entre mente y cuerpo se vería profundamente limitado sin una comprensión previa de los ítems mentales y físicos así estudiados.

En respuesta a la preocupación anterior, tal vez es natural pensar que las nociones pertinentes de lo físico y lo mental son suficientemente claras. Después de todo, mientras que la naturaleza general del mundo físico es claramente iluminada por la observación cotidiana y la investigación científica, la realidad de los fenómenos mentales parece ser uno de los hechos introspectivos más íntimos y familiares de la condición humana. Sin embargo, aun concediendo una familiaridad relativamente robusta con la estructura ontológica del mundo físico y la confirmación introspectiva de la realidad mental, no es obvio que una comprensión mínima de nuestras vidas mentales en sus términos más generales se derive de ellas. Aunque la mente humana es sin duda un elemento constitutivo de nuestra existencia, también es un fenómeno profundamente misterioso. Por ejemplo, aunque es razonablemente obvio que soy un sujeto de experiencias visuales —esto es, de eventos psicológicos gracias a los cuales soy consciente de los ítems del mundo— es improbable que mi conciencia introspectiva de tales experiencias sea semejante al tipo de observación sensorial que tengo del computador y el escritorio que tengo frente a mí en ese momento. Simplemente no puedo "observar" mis experiencias visuales de la misma forma en que puedo ver un computador o un escritorio. Aun si nuestros fenómenos mentales son objetos directos de nuestras capacidades sensoriales o cognitivas —algo que ya de por sí es materia de intenso debate filosófico— simplemente no parece plausible que la conciencia de nuestra geografía mental sea tan manifiesta como nuestra conciencia del mundo de objetos públicos

El diagnóstico anterior tiene precedente histórico. Liderados por filósofos como Wittgenstein y Ryle, los proyectos de topografía mental de la primera mitad del siglo XX precisamente parecen haberse concentrado en delinear la estructura conceptual u ontológica de nuestros fenómenos mentales. Si mi reconstrucción histórica expresa algún tipo de verdad, entonces resulta lamentable

que tales proyectos filosóficos hayan sido abruptamente interrumpidos por la llegada del problema mente-cuerpo, tal como ha sido descrito aquí.

La indiferencia ontológica en la discusión contemporánea del problema mente-cuerpo se manifiesta no sólo en un sutil pero decisivo énfasis en las relaciones psico-físicas por sobre sus respectivos relatos, sino también en las motivaciones que mueven a dicha motivación. Es innegable que, en una u otra medida, Descartes jugó un juego ontológico: aun si sus intereses primarios fueron epistemológicos, una determinada ontología mental era el costo necesario de su fundamentación del conocimiento humano. Sin embargo, el rechazo por parte de los materialistas australianos de las mentes cartesianas dentro de su visión de mundo no tiene tanto que ver con que tales ítems sean entidades de alguna manera ocultas o misteriosas: el que parecieran entidades misteriosas para conductistas tan prominentes como Skinner durante la primera mitad del siglo XX, por ejemplo, no fue obstáculo para que estados y procesos cerebrales fuesen, sólo algunas décadas después, integrados dentro la perspectiva materialista. Más bien, dos motivaciones prominentes detrás de dicho rechazo se relacionan: primero, a una concepción de la filosofía de la mente según la cual su rol principal consiste en articular coherentemente el marco conceptual que subyace a las ciencias de la mente —en particular, la psicología cognitiva y la neuro-biología de entonces— así como desactivar las objeciones filosóficas disponibles en contra del mismo; y, segundo, a la imposibilidad de incorporar los sujetos psicológicos postulados por el dualismo cartesiano al mundo de procesos y estados físicos a través de leyes psico-físicas. En breve, las posiciones materialistas dominantes en el problema mente-cuerpo no parecen rechazar al dualismo cartesiano tanto por razones ontológicas, sino más bien por consideraciones derivadas de la relación metodológica entre ciencia y filosofía, por una parte, y, por otra, la posibilidad de formular leyes naturales que relacionen a los sujetos psicológicos cartesianos con el mundo físico conocido y amado por todos[1].

[1] Esta segunda motivación se hace, con todas sus virtudes y contradiccio-

El segundo momento de indiferencia ontológica al cual me referiré dice relación con la suposición automática y prácticamente indiscutida de una importante tesis sobre la interacción entre mente y cuerpo. En forma concisa esta tesis podría ser expresada de la siguiente manera: los fenómenos de nuestra vida mental son la clase de ítems que pueden producir o causar fenómenos físicos u otros fenómenos mentales. En la medida en que la teoría de la causalidad más comúnmente aceptada entiende los relata de tales relaciones productivas como eventos, la tesis anterior se traduce en una concepción de las relaciones psico-físicas según la cual lo físico y lo mental está constituido de eventos físicos y mentales, respectivamente, que son a su vez capaces de producir o causar otros eventos físicos y mentales. Siguiendo el trabajo de Helen Steward, quien sobresalientemente identifica y describe este supuesto en el debate contemporáneo sobre el problema mente-cuerpo, la idea anterior se podría caracterizar como un *supuesto de redes causales*: esto es, un supuesto que hace alusión a una realidad exhaustivamente interconectada por redes causales; y, como tal, una donde el estatus ontológico de cada ítem depende crucialmente de que ocupe algún lugar dentro de esa compleja red (cf. Steward, 1997).

El supuesto de redes causales tal vez se sustenta en la tesis filosófica según la cual la identidad de un ítem —ya sea objeto, propiedad, evento, etc.— depende parcial o completamente de que tal ítem tenga algún tipo de potencialidad causal en el mundo (cf. Shoemaker, 2003). Pero además de este punto, dicho anterior tal vez es suficientemente vindicado por nuestras intuiciones cotidianas al respecto. Entre los datos más evidentes en una reflexión no-filosófica sobre nuestras mentes, es posible reconocer que el mundo físico tiene efectos causales en nuestras vidas mentales, y viceversa. Al enfrentar un escritorio y en condiciones perceptivas normales, probablemente se seguirá una experiencia visual de un escritorio. Cuando mi dedo haga contacto con una cafetera caliente, experimentaré un dolor. Y, a la inversa, mi decisión de comer

---

nes, particularmente expresa en el *monismo anómalo* de Donald Davidson (1970).

una fruta en lugar del helado me llevará a tomar una naranja en la fila del casino. Estas consideraciones podrían entonces sugerir que el supuesto de redes causales no es controversial: es un dato con el cual comienza nuestra reflexión filosófica sobre la mente y no un elemento de esta última.

Sin embargo, las cosas no son tan simples. Aun si concebimos el mundo en términos de mosaicos causales increíblemente complejos, no se sigue de ello que todos sus elementos estén conectados entre sí a través de relaciones productivas. En principio, existen diferentes formas en las cuales tales elementos podrían estar causalmente relacionados, sin que las relaciones en cuestión deban ser productivas. Para ilustrar este razonamiento, basta considerar un sencillo ejemplo relacionado al acto de encender un fósforo. En primer lugar, naturalmente existe una relación de productividad causal entre el acto E_1, encender un fósforo, y E_2, el proceso de combustión del fósforo. En virtud de una comprensión clásica de la causalidad, E_1 y E_2 deben ser entendidos en términos de eventos. Sin embargo, se pueden distinguir otros elementos en este simple ejemplo causal que trascienden a una relación productiva y sus respectivos relata. Por cada evento E_1 y E_2 se puede identificar un hecho o estado de cosas, a saber, que E_1 es el caso —llamémoslo H_1— y que E_2 es el caso —esto es, H_2—. Y aunque naturalmente pueden existir relaciones nomológicas o contra-fácticas entre H_1 y H_2–por ejemplo, que si yo no hubiese encendido el fósforo, éste último no se habría quemado —no deben ser entendidas como relaciones de productividad—. Asimismo, el fósforo debe gozar de un número de condiciones para poder ser encendido: estar seco, contar con suficiente oxígeno para generar el proceso de combustión, etc. Estas condiciones son sin duda causales, pero no producen el proceso de combustión. El punto de este ejemplo es mostrar cómo el mundo —o una parte muy pequeña de él— puede ser entendido como un mosaico causal, sin necesariamente reducir todos sus componentes a relata de relaciones productivas. Si esta línea de razonamiento es persuasiva, entonces no sería necesario que los elementos de nuestra vida mental deban ser relata de relaciones productivas.

Sobre esa base, el punto de fondo es que ni nuestras intuiciones cotidianas ni tesis más controversiales sobre individuación ontológica parecen definir qué rol exacto tienen los elementos de nuestras vidas mentales en el complejo mosaico causal del mundo natural. Tal vez ellas revelan que nuestro mundo es un mosaico causal y que nuestras mentes son parte del mismo. Pero nada de esto especifica algo que es propiamente parte de una ontología de la mente, a saber, si nuestros fenómenos mentales deben ser entendidos como eventos, condiciones, hechos, etc. Este tipo de pregunta nos lleva a un estudio de la realidad mental en sus términos más generales posibles: a su vez, solamente habiendo resuelto dicha pregunta podemos determinar qué rol causal desempeñan tales fenómenos. Este tipo de pregunta ontológica es, nuevamente, el tipo de preocupación que motivan a las filosofías de Oxford y Cambridge en la primera mitad del siglo XX: en contra de una práctica interpretativa hasta cierto momento habitual, no parece apropiado llamar "conductismo analítico" las teorías de la mente desarrolladas por Ryle y Wittgenstein porque, en lugar de buscar resolver el problema de cómo fenómenos físicos y mentales se relacionan productivamente entre sí, tales filósofos más bien se concentran en la tarea previa de describir nuestros fenómenos mentales, distinguiendo nítidamente sus aspectos episódicos, procesivos, disposicionales, etc. Con la llegada del problema mente-cuerpo a la escena filosófica, sin embargo, investigaciones de ese tipo prácticamente desaparecen del paisaje filosófico popular.

Por último, una apatía ontológica generalizada también parece reflejarse en el curso histórico que eventualmente tomó el debate en torno al problema mente-cuerpo. Entendiendo los fenómenos de nuestras vidas mentales como eventos capaces de producir o ser producidos por otros eventos físicos y mentales, la discusión sobre la relación mente-cuerpo naturalmente colapsó en un debate sobre el rol productivo de nuestra vida mental. Los resultados de este último tipo de investigación ya han sido suficientemente estudiados: nuestros fenómenos mentales parecen sobre-determinar causalmente las cadenas productivas entre eventos físicos; parecen ser efectos de causas físicas, pero ellos mismos causalmente inertes;

o, finalmente, simplemente son eliminados de una visión materialista del mundo en la medida en que no pueden encontrar una posición natural dentro del gran mosaico causal-productivo del mundo. Es relativamente claro que todos estos resultados posibles son insatisfactorios. Ahora bien, en virtud del escaso progreso en alcanzar una solución satisfactoria, el problema mente-cuerpo ha sido progresivamente desplazado del centro de atención filosófica durante las últimas dos décadas del siglo XX. En su lugar, problemas vinculados a las nociones de intencionalidad y subjetividad han ganado protagonismo (cf. Dennett, 1969, 2010, 1991, 2013, 1995; Searle 1983; Chalmers 1996, 1999). Los fenómenos de nuestra vida mental son intencionales en la medida en que generalmente refieren a aspectos del mundo que van más allá de los fenómenos mismos: por ejemplo, una descripción de una determinada experiencia visual no es remotamente completa a menos que hagamos referencia a los ítems visibles sobre los cuales ella nos hace conscientes; asimismo, la descripción de una determinada actitud proposicional —e.g. una creencia, un deseo, etc.— implica la alusión a hechos reales o posibles del mundo que constituyen el contenido de la actitud proposicional en cuestión. El carácter subjetivo de fenómenos sensoriales y cognitivos, mientras tanto, se refiere a cualidades relacionadas al modo en el cual accedemos a ciertos aspectos del mundo, las cuales a su vez están íntimamente relacionadas al hecho de que, como sujetos confrontados al mundo, tenemos sólo una perspectiva finita sobre el mismo. Usando un ejemplo clásico, un murciélago percibe el mundo de manera muy distinta a aquella en la cual yo lo hago, y esta diferencia por lo menos se relaciona a las diferentes perspectivas que yo y un murciélago tenemos del mundo. No parece exagerado sostener que gran parte de la literatura contemporánea en la filosofía de la mente se avoca a explorar la naturaleza específica de la intencionalidad y el carácter subjetivo de nuestros fenómenos mentales, así como sus relaciones mutuas.

Aunque las discusiones sobre los aspectos anteriores son sumamente ingeniosas, también descansan sobre una indiferencia generalizada en relación a la estructura ontológica de la mente.

Asumiendo que las nociones de intencionalidad y subjetividad refieren a aspectos de nuestra vida mental, me parece metodológicamente natural asumir que, para entender qué clase de propiedades tiene un fenómeno tan misterioso como la mente, es necesario tener, por más general que sea, una comprensión de qué clase de ítem es o podría ser la mente. Sin embargo, los debates contemporáneos acerca de estas propiedades mentales se ha desarrollado en general sin atender a cómo debemos entender ontológicamente los fenómenos mentales que instancian tales cualidades. De hecho, el supuesto parece ser una subordinación inversa: para iluminar la estructura ontológica de la mente consciente, es necesario específicar sus propiedades más específicas. Daniel Dennett —cuyo trabajo ha influido decisivamente la agenda de investigación en la filosofía de la mente desde las últimas décadas del siglo XX— identifica dos conceptos guía en la filosofía de la mente contemporánea, a saber, conciencia e intencionalidad: sobre esta base, pone el peso de la discusión filosófica sobre el misterio de la intencionalidad mental en la medida que, de lograr resolver este último misterio, el problema de la conciencia de alguna manera se hará cargo de sí mismo (cf. Guttenplan, 1994). Esta posición no sólo se deriva de una visión de la filosofía como una disciplina conceptual subordinada a la investigación científica del momento —visión que Dennett parece compartir con la investigación de los defensores tempranos de la teoría de identidad mente-cuerpo— sino también de un explícito rechazo de una reflexión metafísica u ontológica sobre la mente (cf. Dennett, 1991, 2013; van Fraassen, 2002).

A través de las viñetas anteriores, he hecho un esfuerzo muy modesto para demostrar cómo una apatía ontológica ha dado forma a la discusión contemporánea del problema mente-cuerpo y en alguna medida determinado el debate posterior en la filosofía de la mente contemporánea. Naturalmente, estas observaciones críticas no constituyen un argumento propiamente tal o una línea ordenada de razonamiento: sólo esperan constituir evidencia sugerente o, para usar una frase elegante de Alan Turing, recitaciones destinadas a generar una creencia. Durante los últimos diez o quince años, esta apatía ha ido perdiendo la influencia que solía tener

entre los filósofos de la mente —un fenómeno resultante en parte de la rehabilitación de la metafísica como una disciplina respetable en la filosofía de inspiración anglo-americana, y en parte del reconocimiento intra-disciplinar de que la filosofía de la mente no puede progresar genuínamente sin reflexionar rigurosamente sobre las bases ontológicas del fenómeno que quiere dilucidar—. Para ilustrar esta nueva tendencia pro-ontológica, esbozaré en la próxima sección algunos aspectos de una nueva línea de debate filosófico explícitamente arraigado en la ontología de la mente.

4. La ontología de la mente

Según la visión panorámica previa, la filosofía y las ciencias de la mente de la segunda mitad del siglo XX se han abocado a estudiar sistemáticamente la pregunta sobre cómo nuestros procesos y estados mentales (e.g. experiencias perceptivas, emociones, sensaciones, creencias, etc.) se relacionan a las operaciones de los sistemas nerviosos central y periférico. La discusión filosófica y científica de esta interrogante tendenciosamente da por sentado la distinción entre lo físico y lo mental, para así concentrar su atención en especificar la relación exacta entre ambos dominios. Después de todo, la psicología cognitiva y la neuro-biología estudian rigurosamente la base física de la mente, mientras que la dimensión subjetiva o cualitativa de nuestra propia vida mental es accesible introspectivamente. Como tal, el principal desafío para toda teoría de la mente parece ser entender cómo aquellos fenómenos introspectivamente accesibles se relacionan a elementos físicos como el cerebro humano y el complejo sistema nervioso que lo acompaña. El debate contemporáneo en torno al problema mente-cuerpo de tal manera se ha enfocado principalmente en delinear la relación entre lo físico y lo mental, y normalmente se asume que la relación en cuestión es aquella de identidad (cf. Kripke, 1981, 1995, lección 3). Esto es, las principales respuestas a este problema han oscilado entre los extremos de un reduccionismo materialista que identifica nuestros procesos y estados mentales con sus contrapartes físicas, por un parte, y, por otra, alguna forma de anti-re-

duccionismo que, al no identificar ambos dominios, enfrenta el desafío de aclarar cómo nuestras mentes tienen un impacto causal en la realidad física, y viceversa.

Ahora bien, aunque la reflexión filosófica sobre la mente humana se ha principalmente enfocado, a lo largo de las últimas siete décadas, en el problema mente-cuerpo tal como se ha descrito, no se agota en dicha interrogante. Como ya he mencionado, en lugar de discutir cómo nuestras mentes se relacionan al mundo físico, pensadores de la primera mitad del siglo XX (e.g. Gilbert Ryle, Ludwig Wittgenstein, C.D. Broad, etc.) se concentraron en dilucidar el funcionamiento de nuestro complejo vocabulario de términos psicológicos (cf. Broad, 1937; Ryle, 1949, 1954, 1956; Wittgenstein, 1953). Esta empresa tal vez fue inicialmente concebida como un proyecto de orden lingüístico, pero rápidamente pasó a tomar la forma de una investigación sobre la inteligibilidad y la realidad de los fenómenos a los cuales nuestros conceptos mentales refieren. El espíritu de esta línea de investigación puede, de tal manera, expresarse en términos de las siguientes preguntas: ¿Qué clase de ítems son los fenómenos de nuestras vidas mentales (e.g. experiencias perceptivas, creencias, emociones, y sensaciones)? ¿De qué hablamos cuando hablamos de experiencias perceptivas, de creencias, de emociones o sensaciones? Desafortunadamente, la popularidad del problema mente-cuerpo y el ocaso de la filosofía lingüística de Oxford y Cambridge a fines de los años sesenta eclipsaron el interés filosófico en dichas preguntas. La tarea de categorizar ontológicamente los elementos de nuestra vida mental tendría que esperar a finales del siglo pasado, cuando la persistente insolubilidad del problema mente-cuerpo sugeriría la necesidad de re-examinar los presupuestos sobre los cuales dicho problema fue planteado. Puesta de manera muy cruda, la sugerencia es la siguiente: tal vez no es posible entender cómo nuestra vida mental se relaciona a la realidad física porque carecemos una comprensión mínima de aquella dimensión psicológica que intentamos incorporar al mundo estudiado por las ciencias naturales tradicionales. Numerosos escritores —entre ellos, Helen Steward, Brian O'Shaughnessy, Matt Soteriou, Tom Crowther, y

Rowland Stout— han buscado esclarecer la estructura ontológica de nuestra vida psicológica (cf. O'Shaughnessy, 1971, 1972, 2000; Steward, 1997, 2011, 2013, 2015; Stout, 1997, 2016, 2018; Soteriou, 2007, 2011, 2013; Crowther, 2009a, 2009b, 2011). La estrategia ampliamente adoptada para llevar a cabo este programa de investigación, denominada por Helen Steward la *estrategia temporal*, apunta a reconocer cómo la adopción de categorías temporales puede arrojar nuevas luces sobre la existencia de fenómenos mentales en la realidad física. En palabras de Steward (1997, 75):

> Existe espacio para discutir si acaso, y en qué sentido, los fenómenos mentales son físicos si se encuentran espacialmente localizados, y si tienen sujetos, y, en caso de tenerlos, cuáles podrían ser tales sujetos. Todas estas son preguntas sustantivas en la filosofía de la mente […] Pero no existe controversia sobre la temporalidad los fenómenos mentales —sobre el hecho de que tienen lugar en o persisten a través del tiempo—.

La estrategia temporal, en tanto, sugiere recurrir a una venerable intuición filosófica que relaciona los conceptos de mente y tiempo. Más específicamente, nos recuerda que los elementos de nuestra vida mental, no menos que aquellos de la realidad física o espacial, existe en el tiempo: una experiencia visual de una manzana dura el tiempo que un sujeto mantiene contacto visual con esta fruta; una persona puede tener una creencia política durante un número de años antes de revisarla; Juan puede amar a María durante un número de años; etc. Sobre esta intuición básica, parece entonces posible realizar una investigación de nuestra vida mental esclareciendo su estructura temporal.

En el contexto de esta estrategia, las categorías de procesos y estados han sido de especial importancia, ya que refieren a distintos esquemas temporales. Un proceso (e.g. caminar, escribir, pensar) tradicionalmente refiere a un cambio continuo que se despliega en partes o que se actualiza progresivamente, mientras que un estado (e.g. creer, conocer, amar) se entiende como la actualización de propiedades o relaciones que no toma tiempo en acontecer —esto es, existe de forma completa en cada uno de los instantes en los

cuales se obtiene— (cf. Vendler, 1957; Armstrong 1968; Taylor, 1977; Mourelatos, 1978, 1993; Gill, 1993; Steward, 1997; Rothstein, 2008; Crowther 2011). Adicionalmente, es posible relacionar ambas categorías a través de la noción de estado ocurrente, esto es, de estados constitutivamente dependientes de procesos (cf. Soteriou 2011, 2013). Para dar un solo ejemplo: una porción de agua en una olla se puede encontrar en estado de ebullición por un periodo de tiempo: dicho estado, sin embargo, depende a su vez de la existencia de ciertos procesos o movimientos moleculares que, tal vez a nivel microscópico, ocurren en la misma porción de líquido. Estas categorías proveen un modesto marco conceptual que ya permite comenzar a desplegar la estrategia temporal: dependiendo de cuál categoría temporal sea privilegiada, distintas concepciones de un determinado fenómeno mental surgirán.

En principio, resulta tentador perfilar los debates recientes sobre la ontología de la mente como tensiones entre teorías que enfatizan el carácter procesual o dinámico y aquellas que enfatizan el carácter estativo o no-dinámico de los fenómenos mentales en cuestión. Puesto que, en cuanto conscientes, ciertos fenómenos psicológicos (e.g. experiencias perceptivas, sensaciones) son intuitivamente dinámicos, es natural adoptar ontologías procesuales de tales fenómenos (cf. O'Shaughnessy, 2000; Crowther, 2009a, 2009b; Soteriou, 2011, 2013). Un punto crucial de este debate es que ambas familias de teorías parecen sugerir distintas formas de entender la noción de cambio aplicada a nuestros fenómenos psicológicos.

Proyectando la estrategia temporal hacia el futuro, también parece posible replantear el problema mente-cuerpo. Los planteamientos materialistas acerca de este último problema tradicionalmente usan categorías espaciales para iluminar la naturaleza de lo mental: por ejemplo, el principal objetivo de tales programas de investigación ha aspirado a reducir nuestra vida mentales a ítems espaciales (e.g. los sistemas nerviosos central y periférico en su totalidad o en alguna de sus partes) o a encontrar el lugar material exacto en el cual nuestras experiencias y pensamientos habitan. La estrategia temporal precisamente pone presión en asumir la espa-

cialidad de los fenómenos mentales como punto de partida para la reflexión filosófica y científica sobre la mente humana: después de todo, el debate sobre el problema mente-cuerpo aspira a determinar si, y, si es así, en qué sentido, tales fenómenos están espacialmente localizados. Teniendo esto en cuenta, resulta controversial formular el problema presuponiendo ya la categorización espacial de nuestra vida mental. Puesto que los ítems de la realidad material también existen en el tiempo, el siguiente procedimiento bipartito de investigación de la relación entre lo físico y lo mental también parece prometedor: en primer lugar, se estudia la estructura temporal de nuestra vida mental; y luego, se procede a determinar cómo dicha estructura puede relacionarse con la estructura temporal de la base material que sostiene a nuestras vidas psicológicas.

En resumen. Un debate reciente pero poco ortodoxo en la filosofía de la mente contemporánea se ha concentrado en la estructura temporal de ciertos tipos de fenómenos mentales. En esta sección, he querido sugerir que tal línea de discusión se puede reconducir a una venerable línea de reflexión filosófica sobre la mente humana, la cual sólo ha sido temporalmente interrumpida por algunas décadas de indiferencia ontológica. De esta manera, la ontología de la mente ahora en ascenso esboza el camino no transitado por filósofos ingenuamente cautivados por los encantos programáticos de las ciencias de la mente.

5. Conclusión

Durante el segundo tercio del siglo XX, la filosofía de la mente se articuló principalmente en torno a la pregunta sobre cómo la mente se relaciona a la realidad natural estudiada por disciplinas como la física, la química y la biología. En la medida en que la dilucidación de una relación entre dos tipos de elementos de distintas naturalezas parece exigir una comprensión de los elementos así relacionados, una discusión del problema mente-cuerpo intuitivamente debería presuponer una categorización ontológica de nuestros fenómenos mentales. Sin embargo, ha sido común prescindir de tal investigación ontológica. En el debate contem-

poráneo sobre el problema mente-cuerpo, filósofos de la mente de corte materialista han visto la oportunidad de poner la filosofía al servicio de las ciencias naturales: más que proponer tesis substantivas sobre la naturaleza de la realidad, la tarea del análisis filosófico contemporáneo fue entendida como aquella de articular y desactivar objeciones en contra del paradigma reduccionista que ya se posicionaba entre las ciencias cognitivas de la época. En este clima de reflexión filosófica, preguntas ontológicas o metafísicas clásicas pero fundamentales han sido implícitamente vistas como caprichos filosóficos anticuados o expresamente censuradas. De tal manera, aunque las preguntas arriba introducidas están íntimamente relacionadas entre sí, históricamente han sido discutidas de manera más o menos independiente. Como se explicó más arriba, tal estado de cosas es inadecuado desde un punto de vista metodológico y ha determinado —en mi opinión, negativamente— la discusión filosófica posterior. De la misma manera en que no parece posible entender la relación entre dos componentes sin tener una comprensión relativamente clara de cada uno de ellos por separado, es intuitivamente inadecuado pretender entender la relación entre mente y cuerpo mientras se carezca de una historia explícita acerca de qué clase de ítems los elementos de nuestras vidas mentales son. Este diagnóstico no es particularmente positivo, y por eso es probable que sea inicialmente poco popular. Sin embargo, abre nuevas perspectivas de investigación filosófica y promete hacer de la filosofía de la mente una verdadera colaboradora —no una mera sirvienta— de las ciencias de la mente.

Referencias bibliográficas

Armstrong, D. (1968). *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Block, N. (1978). Las Dificultades del Funcionalismo. En E. Rabossi (ed.), *Filosofía de la Mente y Ciencia Cognitiva* pp. 105-142. Barcelona: Paidos.

Broad, C.D. (1937). *The Mind and its Place in Nature*. London: Rutledge and Kegan Paul.

Chalmers, D. (1996/1999). *La Mente Consciente*. Barcelona: Gedisa.

Crowther, T. (2009a). Watching, sight, and the temporal shape of perceptual activity. *Philosophical Review*, 118(1): 1-27.

Crowther, T. (2009b). Perceptual Activity and the Will . En L. O'Brien y M. Soteriou (eds.), *Mental Actions*, pp. 173-191. Oxford: Oxford University Press.

Crowther, T. (2011). The Matter of Events. *The Review of Metaphysics*, 65(1): 3-39.

Davidson, D. (1970). Eventos Mentales. En D. Davidson (ed.), *Ensayos sobre Acciones y Sucesos*. Barcelona: Crítica.

Dennett, D. (1969/2010). *Content and Consciousness*. Milton Park: Routledge.

Dennett, D. (1995). *La Conciencia Explicada: Una teoría interdisciplinar.* Barcelona: Paidós.

Guttenplan, S. (1994). *A Companion to the Philosophy of Mind*. Cambridge: Blackwell.

Dennett, D. C. (1991/2013). *Intuition pumps and other tools for thinking*. New York: WW Norton & Company.

Descartes, R. (1642/1977). *Meditaciones Metafísicas*. Madrid: Ediciones Alfaguara.

Descartes, R. (1999). *Correspondencia con Isabel de Bohemia y otras cartas*. Barcelona: Alba Editorial.

Farrell, B. A. (1950). Experience. *Mind*, 59(234): 170-198.

Feigl, H. (1967). *The mental and the physical: The essay and a postscript*. Minneapolis: University of Minnesota Press.

Gill, K. (1993). On the metaphysical distinction between processes and events. *Canadian Journal of Philosophy*, 23(3): 365-384.

Kim, J. (1988/2002). El Problema Mente-Cuerpo tras Cincuenta Años. *Azafea*, 4: 45-63.

Kripke, S. (1981./1995). *El Nombrar y la Necesidad*. México: Instituto de Investigaciones Filosóficas UNAM.

Lewis, D. (1980). Mad Pain and Martian Pain. En N. Block (ed.), *Readings in the Philosophy of Psychology*, vol. I, pp. 216-222. Cambridge: Harvard University Press.

Mourelatos, A. P. (1978). Events, processes, and states. *Linguistics and philosophy*, 2(3): 415-434

Mourelatos, A. P. (1993). Aristotle's kinesis/energeia distinction: A marginal note on Kathleen Gill's paper. *Canadian Journal of Philosophy*, 23(3): 385-388.

O'Shaughnessy, B. (1971). The Temporal Ordering of Perceptions and Reactions. En F. Sibley (ed.), *Perception: A Philosophical Symposium*, pp. 139-187. London: Methuen.

O'Shaughnessy, B. (1972). Processes. *Proceedings of the Aristotelian Society*, 72: 215-240.

O'Shaughnessy, B. (2000). *Consciousness and the World*. Oxford: Clarendon Press.

Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology*, 47: 44-50.

Putnam, H. (1967/1981). *La Naturaleza de los Estados Mentales*. México: Instituto de Investigaciones Filosóficas UNAM.

Putman, H. (1988/1990). *Representación y realidad.* Barcelona: Gedisa.

Rothstein, S. (2008). *Structuring Events. A Study in the Semantics of Lexical Aspect*. Malden: Blackwell Publishing.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Ryle, G. (1954). *Dilemmas*. Cambridge: Cambridge University Press.

Ryle, G. (1956). Sensation. En G. Ryle (ed.), *Collected Papers, vol. 2*., pp. 349-362. London: Hutchinson.

Searle, J. R., Willis, S. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University press.

Shoemaker, S. (2003). *Identity, Cause, and Mind: Philosophical Essays*. Oxford: Clarendon Press.

Skinner, B. F. (1951/2014). *Science and Human Behavior*. Cambridge: The B.F. Skinner Foundation.

Smart, J. J. (1959). Sensations and brain processes. *The Philosophical Review*, 68(2): 141-156.

Soteriou, M. (2007). Content and the Stream of Consciousness. *Philosophical Perspectives*, 21: 543-568.

Soteriou, M. (2011). Occurrent perceptual knowledge. *Philosophical Issues*, 21: 485-504.

Soteriou, M. (2013). *The mind's construction: The ontology of mind and mental action*. Oxford University Press.

Steward, H. (1997). *The Ontology of Mind*. Oxford: Clarendon Press.

Steward, H. (2011). Perception and the Ontology of Causation. En N. Eilan, H. Lerman, y J. Roessler (eds.), *Perception, Causation and Objectivity: Issues in Philosophy and Psychology*, pp. 139-160. Oxford: Oxford University Press.

Steward, H. (2013). Processes, continuants, and individuals. *Mind*, 122(487): 781-812.

Steward, H. (2015). What Is a Continuant? *Aristotelian Society Supplementary Volume*, 89(1): 109-123.

Stout, R. (1997). Processes. *Philosophy*, 72(279): 19-27.

Stout, R. (2016). The Category of Occurrent Continuants. *Mind*, 124(496): 41-62.

Stout, R. (2018). *Process, Action and Experience*. Oxford: Oxford University Press.

Taylor, B. (1977). Tense and continuity. *Linguistics and philosophy*, 1(2): 199-220.

Van Fraassen, B. C. (2002). *The Empirical Stance*. New Haven: Yale University Press.

Vendler, Z. (1957). Verbs and times. *The philosophical review*, 66(2): 143-160.

Wittgenstein, L. (1953). Investigaciones Filosóficas. En L. Wittgenstein (ed.), *Tractatus Logico-Philosophicus.* Madrid: Tecnos.

Sobre el Autor

Sebastián Sanhueza es profesor del Instituto de Filosofía de la Universidad de Concepción. Realiza investigación en filosofía de la mente y epistemología con un enfoque específica en los problemas de la percepción. Contacto: ssanhue@gmail.com

Capítulo 7

La mente agencial: elementos para una teoría de las atribuciones de agencia mental

Pablo López-Silva; Andrea Arancibia; Gabriel Cordero; Leonardo Henríquez

Resumen

El concepto de atribución de agencia mental refiere al acto mediante el cual un sujeto asigna agencia a un pensamiento en primera persona. Esta noción nace en filosofía en el contexto de la discusión sobre la forma en que los humanos asignan agencia a sus propios movimientos corporales. Con el fin de contribuir al desarrollo de la discusión en el plano mental, este capítulo describe las dos ofertas teóricas más populares en el debate motor y analiza la forma en que éstas se han adaptado a la dimensión mental. Luego, se elabora una crítica sistemática a la estrategia paralelista que subyace a tal adaptación y, finalmente, se intenta clarificar el *explananda* para una posible teoría sobre las atribuciones de agencia mental mediante un análisis de las principales propiedades fenoménicas que podrían estar asociadas a la agencialidad del pensamiento.

Palabras Clave: Atribuciones de agencia, agencia motora, agencia mental, delirios.

1. Introducción

En casos normales, los seres humanos no tienen grandes dificultades para distinguir experiencialmente entre aquellos movimientos corporales generados voluntariamente, de aquellos iniciados involuntariamente. Piensa cuando estás escribiendo algo en un

papel, y contrástalo con el caso de que alguien mueva tus dedos y brazos imitando el acto de escribir sin que tu hagas ningún esfuerzo. En filosofía de la mente y ciencias cognitivas se sugiere que el proceso mental a la base de tal distinción experiencial es la denominada 'atribución de agencia', esto es, el acto de atribuir el inicio de un movimiento corporal a un agente específico (Pacherie, 2008; Gallagher, 2000; 2007; Marcel, 2003)[1]. Se sugiere, pues, que aquellos movimientos que cuenten con una *auto-atribución* de agencia motora serían experimentados como voluntarios, mientras que aquellos que externalicen la agencia del movimiento en cuestión serían experimentados como involuntarios[2].

Independiente de los múltiples debates que rodean la noción de atribución de agencia motora, algunos autores sugieren la existencia de un concepto similar en el caso cognitivo. Intentando mapear las diferencias experienciales entre la aparición de pensamientos voluntarios e involuntarios en el flujo de la conciencia (asumiendo que tal distinción no es controversial), se ha indicado que, a la base de tal diferencia experiencial estarían las denominadas atribuciones de agencia mental, i.e. el acto de atribuir el inicio o producción de un pensamiento o tren de pensamientos a un agente específico (Campbell, 1999; Stephens & Graham, 2000; Proust, 2009; Gallagher, 2014). Ahora bien, las discusiones acerca de la naturaleza y arquitectura de las atribuciones de agencia motoras y cognitivas están lejos de estar resueltas. Es más, mucho de la obscuridad del debate original – el motor – es reproducida en el debate cognitivo. Esto, entre otras razones, porque el debate cognitivo siempre ha sido tratado como una mera nota al pie del debate motor, lo que ha producido no solo problemas de superficialidad e inespecificidad, sino que también, debates incluso sobre

---

[1] De ahora en adelante utilizaremos 'experiencial' y 'fenomenológico' como términos intercambiables, al entender 'fenomenológico' en su acepción de 'fenomenalidad' o como aquellos estados que poseen una 'fenomenología.

[2] Estamos conscientes de que esto puede leerse como una hiper-simplificación del debate pero diversos aspectos de este serán aclarados a lo largo del texto.

si es posible aplicar las ideas del contexto motor al contexto cognitivo de la forma en que se ha hecho hasta la fecha (Proust, 2009; López-Silva, 2019).

Con el fin de contribuir al desarrollo de la discusión en el plano cognitivo, en este capítulo intentaremos describir las dos ofertas teóricas más populares en el debate motor y analizaremos como éstas se han adaptado al debate cognitivo. Luego, elaboraremos una crítica sistemática a la estrategia paralelista que subyace a tal adaptación. Una vez clarificados los problemas de este marco de referencia metodológico, intentaremos describir diversas propiedades fenoménicas vinculadas a la experiencia de ser un agente, tanto para el dominio motor, como para el mental, para así poder clarificar algunos de los aspectos exclusivos de cada dimensión. Esto último, no solo se realizará para demostrar la existencia de componentes exclusivos en el caso cognitivo que lo hacen merecedor de un tratamiento diferenciado y específico, sino que, y sobre todo, para clarificar el *explananda* para una posible teoría sobre las atribuciones de agencia mental.

2. Arquitectura y naturaleza de las atribuciones agencia motora

2.1 Modelo *top-down* o de narradores

El modelo *top-down* postula que la atribución de agencia motora se produce al llevar a cabo procesos cognitivos de orden superior - juicios de agencia (UP) -, que nos dicen si un evento, ya sea un pensamiento o movimiento (down), son consistentes con la narración que tenemos de nosotros mismos (ver Bayne y Pacherie, 2007). Por 'narración', autores como Graham & Stephens (2000) refieren al set de expectativas que un sujeto P tiene de sí mismo en un contexto C específico, y de la auto-imagen de P en C. Sobre esto, Graham & Stephens (1994) indicarán que:

> [W]hether I take myself to be the agent of a mental episode [or physical movement] depends upon whether I take the ocurrence of this episode tobe explicable in terms of my underlying intentional states (p. 94).

Para el enfoque top-down, la auto-atribución de agencia se produce luego de evaluar la ocurrencia de los propios pensamientos y movimientos – respectivamente - retrospectivamente en contraste con tal narración. La idea es que si la realización de un movimiento o la aparición de un pensamiento en el flujo de la conciencia de un sujeto en consistente con tal narración, los sujetos terminarán auto-atribuyendo tales eventos motores y mentales. Como consecuencia de este acto, los sujetos podrían, finalmente, experimentar una sensación de agencia motora o mental. En este sentido, cualquier sensación agencial asociada a la actividad motora y/o cognitiva en este modelo es de 2do orden i.e. es un subproducto de procesos cognitivos superiores.

Como podemos observar, dentro de este modelo, la naturaleza de las atribuciones de agencia es inferencial, de forma tal que su producción va desde el *background* de conocimiento previo sobre la realidad y el sujeto (top), hacia la evaluación de la agencia que éste mismo tiene acerca de sus propios pensamientos o movimientos (down). Intentando clarificar el modelo top-down, Graham y Stephens (1994) indican que:

> The 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 (p.102)

Desde las palabras de los autores, finalmente se puede sugerir que, dentro de este modelo, la sensación de agencia tiene un carácter evaluativo, inferencial y retrospectivo como imposición de la categoría de 'agencia' a los pensamientos o movimientos a los cuáles se tiene acceso en primera persona, los cuales estarían basados en información privilegiada respecto de su ocurrencia que les daría sentido.

2.2 Modelo *bottom-up* o comparadores

El enfoque *bottom-up* o de 'comparadores' indica que la sensación y la atribución de agencia se producen en procesos de primer orden asociados a características intrínsecas de las experiencias relevantes. Con esto, este modelo se separa del enfoque *top-down* que enfatiza la dimensión cognitiva-evaluativa y retrospectiva del proceso. La base de las atribuciones de agencia mental y motora es primordialmente experiencial ya que, se sugiere, se producirían en los procesos neurales responsables de las acciones motoras (Gallagher, 2007)[3].

La idea del modelo *bottom-up* es que, a lo largo de su desarrollo, los sujetos almacenan distintos predictores que le permiten prever las consecuencias de sus posibles movimientos corporales. El trabajo de los predictores es el de ser comparados con las consecuencias directas de los movimientos (*feedback* directo) y, cuando estas consecuencias se condicen con la predicción, entonces, se envía una señal aferente que indicará que el movimiento fue generado por uno mismo; cuando no hay coincidencia, los movimientos se experimentarían de forma involuntaria (Bayne y Pacherie, 2007)[4]; este enfoque propone que las señales enviadas generan una experiencia fenoménica de agencia intrínseca a los movimientos voluntarios (Gallagher, 2007) y por esto, se asume que la fenomenología de los movimientos voluntarios es fundamentalmente activa por lo que la dirección causal de tal atribución va desde la experiencia (bottom) hacia la atribución (up).

Sobre este enfoque, Gallagher (2014) indica que:

---

[3] Este será la base de una serie de problemas para el enfoque al ser aplicado al contexto cognitivo ya que no es claro que tales procesos neuronales sean los mismos en forma y categoría en el caso de la formación de pensamientos.

[4] Esta idea será otra fuente de problemas al ser aplicada al caso cognitivo ya que no es claro que el concepto de 'comparador' aplique a la formación de un pensamiento.

> On a bottom-up account, attributions of agency typically depend on the first-order experience of agency. If I walk over to open the door, I experience myself initiating the movement, that is, my movement is accompanied by an implicit prereflective SA for this movement; if I am then asked, did I open the door, I can correctly attribute agency to myself. On this view, problems with SA would show up initially at the level of first-order experience, likely due to disruptions in neuronal processes (p.4)

Ahora bien, el enfoque indica que los pensamientos que serán finalmente auto-atribuidos gozarán de esta categoría en virtud de la existencia de una sensación de agencia mental intrínseca a éstos (Gallagher, 2017). Los pensamientos voluntarios poseen una fenomenología activa i.e. se sienten como algo que uno hace, y desde esta información fenoménica surgen instancia de orden superior como los juicios o creencias de agencia mental. Así, en estricto rigor, dentro del enfoque *bottom-up*, un pensamiento o movimiento es autoatribuido agencialmente simplemente porque este se experimenta como algo que uno ha creado. En ambos casos, la atribución de agencia es informada por una sensación de agencia – motora o mental – intrínseca al estado mental en cuestión.

3. El problema del marco de referencia: críticas al paralelismo motor-cognitivo

Tal como se puede observar, los dos modelos descritos en la sección anterior serán la base de las explicaciones disponibles en la literatura para darle sentido a la pregunta por la naturaleza y arquitectura de las atribuciones de agencia mental (ver por ejemplo, Gallagher, 2007; 2017). Lo que se puede ver en esta estrategia es que las aplicaciones de los modelos motores al debate cognitivo se han realizado mediante el establecimiento de un paralelismo directo entre la naturaleza y la arquitectura de la actividad motora y la actividad cognitiva. Es más, a la base de tal estrategia metodológica está la idea de que los pensamientos pueden ser tratados como un tipo de acción motora (Feinberg, 1978; Schmahmann, 2004; Ito, 2008).

Ahora, si bien es evidente que cada modelo posee claros problemas internos (ver Proust, 2009; Vosgerau & Voss, 2014; López-Silva, 2019), en esta sección nos dedicaremos a examinar *el problema del marco de referencia*, esto es, las diversas dificultades que emergen del paralelismo establecido entre la actividad motora y cognitiva al momento de elaborar enfoques para explicar la arquitectura y naturaleza de las atribuciones de agencia mental.

Uno de los primeros puntos problemáticos de la estrategia paralelista emerge al observar la existencia de evidencia que indicaría que los sustratos neurológicos de las enfermedades interpretadas en términos de 'errores' en la atribución de agencia en el caso motor y en el caso cognitivo son distintos. Esto viene a ser crítico para todo el proyecto paralelista, ya que, al observar diferencias morfofuncionales demasiado específicas, no existirían buenas razones para confiar en modelos creados específicamente para explicar síndromes motores para aplicarlos a alteraciones cognitivas, esto, porque tales diferencias se pueden traducir en diferencias etiológicas, fenomenológicas, y por lo tanto, arquitectónicas y, tal vez, de naturaleza.

Este problema se evidencia de forma más clara al revisar la evidencia empírica disponible para los delirios de inserción del pensamiento y en el síndrome de la mano ajena, ambos síndromes interpretados como alteraciones del acto de atribuir el inicio de un pensamiento (delirio de inserción del pensamiento) o un movimiento corporal (síndrome de la mano ajena) a un agente específico respectivamente (Mullins & Spence, 2003; Assal, Schwartz & Vuilleumier, 2007). En el primer caso (cognitivo), en un experimento realizado por Walsh, Oakley, Halligan, Mehta y Deeley (2015) en donde se inducía a personas a estados de delirios de inserción de pensamiento mediante hipnosis, se notaron ciertas áreas cerebrales con 'activación reducida'. Ejemplo de esto fueron (i) el giro temporal superior bilateral y giro temporal centro-derecho, áreas relacionadas con el procesamiento y escritura del lenguaje y el desarrollo de la comunicación (Howard, Volkov, Mirsky, Garell, Noh, Granner, et al., 2000; Liebenthal, Binder, Spitzer, Possing, y Medler, 2005). (ii) Los ganglios basales subcorticales y estriado,

áreas que permiten a los sistemas envueltos en el procesamiento de la gramática y lingüística acceder al sistema motor frontal que media la escritura (Anderson, Saver, Tranel, & Damasio, 1993; Duffau et al., 2002 en Walsh et al., 2015). (iii) El tálamo, el cual está relacionado con la integración de aspectos cognitivos y motores en la producción del lenguaje (Hebb & Ojemann, 2013 en Walsh et al., 2015) y, para finalizar, (iv) el giro occipital inferior derecho y giro occipital superior izquierdo que se encargan (entre otras áreas) del procesamiento visual. Ahora bien, en el caso del síndrome de la mano ajena, se analizó el reporte de un paciente de 70 años que desarrolló el síndrome luego de un derrame cerebral. Mediante imágenes de resonancia magnética funcional, se pudieron evidenciar ciertas diferencias en sus movimientos voluntarios y los movimientos que pertenecen al síndrome para lograr especificidad y discriminación metodológica (Assal, Schwartz, y Vuilleumier, 2007). Al explorar los movimientos pertenecientes al síndrome se evidenció un aumento en la actividad del área motora primaria, que es el área que se cree responsable de los movimientos que tienen la característica de ser involuntarios o inconscientes. Además, esto estaba acompañado con una lesión ocurrida en el lóbulo parietal (por el derrame cerebral), generando así una falta de retroalimentación propioceptiva, la cual puede explicar porqué el paciente era comúnmente inconsciente de los movimientos de su mano izquierda por la falta de comunicación, ya que esta área (lóbulo parietal), tiene el rol crucial de generar imágenes conscientes de acción y mantener representaciones motoras internas, las cuales son necesarias para las autoatribuciones de agencia motora (Assal, Schwartz, y Vuilleumier, 2007). Si bien necesitamos generar mayores análisis paralelos, la relevancia de las alteraciones propioceptivas relacionadas con las alteraciones de la agencia motora, hacen altamente implausible la aceptación de paralelismo motor-cognitivo, esto, ya que los pensamientos, o incluso toda la actividad cognitiva, de tenerlo, no poseería un feedback que sería propioceptivo, o no al menos de la forma en que los modelos motores necesitan que sea para poder ser aplicados a los casos cognitivos.

El segundo problema surge cuando, en el afán de aplicar su estructura explicativa al caso mental, el paralelismo cognitivo-motor barre con todas las particularidades del caso mental. Esto provocaría que cualquier enfoque basado en esta estrategia no esté realmente capturando el fenómeno en cuestión, sino que simplemente sea una mala copia permitida por el filtro de tal paralelismo. Por ejemplo, la arquitectura sensorial del *feedback* motor que permite las atribuciones de agencia motora hace a este elemento imprescindible en cualquier enfoque al fenómeno tanto en su trayectoria normal como anormal. Sin embargo, en el caso mental no contamos con nada parecido, pues bien ¿por qué tendríamos que confiar en un modelo que tiene como eje central un elemento que ni siquiera es observado en el caso mental para explicar, justamente, el caso mental? En estricta relación con esto, existen marcadas diferencias fenomenológicas entre la experiencia consciente de un movimiento voluntario y la voluntariedad experimentada en algunos episodios cognitivos. Claramente la voluntariedad en el pensamiento tiene que ver con sensaciones de control, monitoreo, consistencia y fluidez y todo esto hace que tal experiencia general de voluntariedad sea más flexible. Sin embargo, la voluntariedad motora no parece tener tal característica y es más, parece ser robustamente distinguible desde un punto de vista experiencial, lo cual parece ser explicado por su arquitectura sensorial. Pues bien, diferencias fenomenológicas hacen plausible hipotetizar diferencias etiológicas fundamentales. Sin embargo, el paralelismo motor-cognitivo no sería capaz de integrar este asunto a su ecuación explicativa, lo cual es otra razón más para desconfiar de él.

El último problema que identificamos en la estrategia paralelista tiene que ver con la existencia de ciertas distinciones fenomenológicas que el modelo motor no lograría incorporar desde el caso cognitivo. En el caso de la autoatribución de agencia motora, se pueden identificar dos tipos de movimientos, a saber, los voluntarios y los involuntarios. Los últimos se diferencian de los primeros porque éstos carecerían de autoatribución de agencia. Esto parece ser diferente en el caso de las autoatribuciones de agencia mental, ya que, mientras algunas personas reportan sus pensamientos

como algo 'que hacen' (carácter voluntario) y otras reportan su actividad cognitiva como algo que meramente 'les pasa' (carácter involuntario), , en ambos grupos observamos autoatribuciones de agencia de tales pensamientos. Esta observación hace posible hipotetizar que la misma 'sensación' de agencia en el caso cognitivo no tendría la relevancia explicativa que posee en el caso motor, lo cual es una buena razón para pensar en un proyecto alternativo. Esta idea sintetiza varios de los elementos discutidos anteriormente. No solo el hecho de que existan diferencias empíricas y fenomenológicas claras entre los casos motores y cognitivos, y por lo tanto, probablemente también existan diferencias etiológicas, sino que también el hecho de que existen distinciones fenomenológicas relevantes dentro del caso cognitivo que no parecen ser consistentes con los enfoques motores hacen necesaria la tarea de repensar la plausibilidad de la estrategia paralelista a la hora de darle sentido al fenómeno de las atribuciones de agencia mental.

4. Hacia una clarificación de algunos elementos para una teoría de las atribuciones de agencia mental

La existencia de claros problemas conceptuales, empíricos y metodológicos en el paralelismo a la base de las explicaciones de la naturaleza y arquitectura de las atribuciones de agencia mental hace plausible – y necesaria - la exploración de caminos conceptuales y metodológicos alternativos. Sin embargo, esta tarea no ha sido realizada en la literatura actual, por lo que el desafío no solamente está abierto, sino que casi inexplorado.

Creemos que una forma promisoria para pavimentar caminos explicativos que respeten las particularidades de ambas dimensiones de las atribuciones de agencia implica la descripción de los aspectos fenomenológicos más fundamentales asociados a aquellas experiencias que son finalmente referidas por los sujetos como 'agenciales'. La idea es describir la fenomenología de la agencia mental y revisar similitudes y diferencias con la fenomenología de la agencia motora, y con esto, lograr generar grados de diferenciación mínimos para la elaboración de enfoques exclusivos en

el caso mental. Esta tarea asume la idea de que existirían distintas instancias en las cuáles un sujeto podría sentirse más o menos como el agente de ciertos estados motores y cognitivos y que, por lo tanto, una sensación de agencia robusta no es la única forma en que la agencia se podría instanciar tanto en los contextos motores y cognitivos. Por consiguiente, es necesario explicitar otras posibilidades fenomenológicas que podrían ser parte del fenómeno agencial general en ambos casos. Este análisis, finalmente, ayudaría a especificar parte del *explananda* de una teoría para las atribuciones de agencia mental y los diversos componentes a los cuáles tal teoría podría echar mano a la hora de proponer un concepto de 'agencia'.

4.1 Sensación de propiedad motora y mental

La discusión respecto de la existencia de una *sensación de agencia* asociada a los estados motores y cognitivos surge en complemento a la descripción de la denominada *sensación de propiedad*; propiedad supuestamente asociada a todos los estados mentales conscientes, o por lo menos, a aquellos con características paradigmáticas normales (Zahavi, 2005; cf. López-Silva, 2017)[5]. En

[5] Para Gallagher (2000; 2007) la sensación de agencia tiene que ver con la conciencia que tenga el sujeto de ser la causa de sus propias acciones. Pacherie (2007) indica que la sensación de agencia depende del sentido que tiene el agente de que él es el autor de determinada acción. Por su parte, Moore (2016) propone que la sensación de agencia refiere al sentimiento de control sobre nuestras acciones y sus consecuencias y Chambon, Wenke, Fleming, Prinz y Haggard (2013) sugieren que ésta sensación tiene que ver con el sentimiento de poder controlar un evento externo mediante nuestras propias acciones. En todos estos casos, la sensación de agencia es una propiedad fenomenológicamente saliente. Sin embargo, existen claros desacuerdos respecto de su naturaleza. Para los defensores de teorías del tipo top-down, la sensación de agencia sería una mera consecuencia residual de un proceso fundamentalmente cognitivo-retrospectivo. Para aquellos que defienden teorías del tipo bottom-up, la sensación de agencia sería una propiedad experiencial de primer orden que informa estados mentales superiores.

el caso motor, la idea es que sujetos con desórdenes que afectan la producción de ciertos movimientos corporales y, por lo tanto, a las atribuciones de agencia motora, podrían carecer de una sensación de agencia la cual potenciaría externalizaciones de movimientos observados en tales casos (Gallagher, 2007, 2010, 2014; Kang, Im, Shim, Nahab, Park, Kim, Kakareka, Miletta y Hallett, 2015). Proust (2009) refiere a este fenómeno de la siguiente forma:

> Although they have normal proprioceptive and visual experience while acting (and therefore, a preserved sense of ownership) they often feel that someone else is acting through them (they present a disturbed sense of agency) (p.254).

Pero a su vez, la idea es que, en cada caso, los sujetos que poseen errores en la atribución de agencia de movimientos corporales sienten que es *su propio cuerpo* el que está siendo el lugar y medio donde ocurren tales sucesos. Por lo tanto, si bien su cuerpo (o una parte de su cuerpo) no posee agencia (entre muchas otras cosas), sigue siendo su cuerpo, y a esto último se le denominará sensación de propiedad u *ownership* (Gallagher, 2014).

En el caso de los pensamientos, la idea es similar para algunos autores (Gallagher, 2012; 2014; Zahavi, 2005; 2015; Billon, 2013; López-Silva, 2017). Existirían pacientes con delirios que parecen indicar que, aunque ciertos pensamientos han sido introducidos en sus mentes, tales pensamientos siguen ocurriendo en *sus* mentes aún cuando no se identifican como los autores de éstos, y por consiguiente, tales pensamientos estarían acompañados de una sensación de propiedad (Zahavi, 2005; Carruthers, 2012). Sobre esto, Proust (2009) propone lo siguiente:

> Another frequent delusion, however, is still more intimately associated with self knowledge: Patients experience ´thought insertion'; they complain that some of their thoughts are in their minds (and, to this extent, are experienced subjectively), but at the same time are not their in the agentive sense; they speculate retrospectively that someone else has inserted them 'into their heads', making them think these ideas (p.254).

Según lo observado, se puede indicar que es posible tener una sensación de propiedad sin necesariamente tener sensación de agencia, pero el caso inverso no parece ser del todo claro y aún se encuentra bajo discusión (López-Silva, 2018; Zahavi, 2019). Ahora, pasemos a nuestra discusión principal.

4.2 Sensación de control

Para Pacherie (2007), *la sensación de control* es uno de los aspectos fenomenológicos más fundamentales asociados a los movimientos voluntarios y, por lo tanto, una de las instancias más claras de agencialidad. La sensación de control puede ser descompuesta en dos tipos de experiencias distintas. Por un lado, en la sensación de estar en control de nuestras acciones y de que todo sucede como uno esperaba que sucediera. Por otro lado, en la sensación de que uno tiene que ejercer el control para poder mantener un 'programa de acción' estable a pesar de los inconvenientes y, por lo tanto, alude al tipo de control que uno ejerce para alcanzar ciertas metas predefinidas. Además de esto, Pacherie (2007) indica que la sensación de control se manifestaría en tres dimensiones: sensación de control motor (que tiene como objeto el propio cuerpo), de control situacional (que tiene como objeto las circunstancias que rodean una acción) y de control racional (que tendría como objeto las razones e intenciones para un movimiento).

¿Existe algo como una sensación de control en el caso mental? Intuitivamente, la respuesta parece ser positiva, aunque creo que es necesario clarificar las diferencias. Si bien su naturaleza y estructura representacional podría ser diferente a la del caso motor, todos parecemos experimentar una sensación de control cuando, por ejemplo, intentamos concentrarnos en un cierto tren de pensamiento más que en otro para llegar a una respuesta específica. Lo mismo ocurre cuando intentamos razonar voluntariamente sobre cierto asunto. En ambos casos, el rol de la atención y las habilidades de concentración podrían alimentar la sensación de control. Ahora bien, es importante señalar una diferencia fundamental. Mientras que en el caso motor aquello que es controlado es un

movimiento específico y, por lo tanto, el contenido de tal sensación podría tener la forma [Yo controlo M - donde M es un movimiento específico], en el caso cognitivo, aquello que es controlado no es el contenido del pensamiento en sí, sino que las condiciones que permiten su emergencia, lo que abre una posibilidad de distinción entre el *proceso de pensar P* (*thinking*) y un *pensamiento* P (*thought*). Por esto podríamos tener sensación de control incluso en casos donde evocamos un contenido equivocado. Es más, suena poco intuitivo indicar que uno puede 'controlar' el contenido de un pensamiento propiamente tal. Sin embargo, las condiciones que propician la emergencia de un pensamiento específico, y no de otro frente a una tarea parecen poder coloreados con distintas tonalidades agenciales. Esto, claramente, debería ser considerado por las teorías de atribuciones de agencia en el caso mental.

4.3 Sensación de intencionalidad causal

En el caso motor se ha sugerido la existencia de una *'sensación de* intencionalidad causal', la que referiría a la experiencia de que uno causa un efecto mediante mis acciones (Wegner, 2000; Aarts, Custers y Wegner, 2005; Wohlschläger et al. 2007; Pacherie, 2008). Para Wohlschläger et al., (2007) esta sensación también estaría asociada a los momentos en que observamos a otras personas haciendo determinadas acciones, las cuáles parecen ser causadas por nosotros, por ejemplo, como cuando asustamos a alguien y tal persona sale corriendo. En el caso mental, una situación potencialmente comparable podría darse cuando nuestro pensamiento expresado verbalmente hace que otro cambie su conducta u opinión, sin embargo, esto parece alejarse del alcance del tipo de propiedades de los pensamientos propiamente al referirse, eventualmente, al análisis retrospectivo que podemos hacer de las consecuencias de nuestros pensamientos en el mundo.

La intencionalidad causal unida al conocimiento sobre el inicio de un movimiento generaría la *sensación de iniciar una acción*, que sería diferente a la sensación de control, ya que uno podría sentir que está en control de movimientos que no fueron iniciados por

uno, como, por ejemplo, cuando controlo el movimiento de una pelota que alguien ha hecho rodar previamente. La conexión intención-conocimiento parece darse entre los 80 y 200 msgs. antes del inicio observado del movimiento y, correspondería al tiempo de preparación potencial de un movimiento. Esto, sería una señal que refiere a la preparación de un programa motor específico que se llevaría a cabo y también tendría una expresión fenomenológica, como por ejemplo, cuando entra nuestro superior a la oficina y uno se prepara para levantarse y darle la mano y éste nos dice que no nos paremos. El momento de 'preparación' se evidencia en una breve sensación de tensión muscular en las partes corporales fundamentales para la propiciación del movimiento en cuestión. En el caso del dominio mental también podría identificarse una *sensación de iniciar un pensamiento* o *tren del pensamiento*. Sin embargo, esto no implicaría la activación de un plan de pensamiento con expresión fenomenológico, sino que a la experiencia de haber dado 'el puntapié' inicial a la emergencia de diversos pensamientos que componen un tren de pensamiento específico, como por ejemplo, cuando estamos intentando pensar en las cosas que debemos empacar para un día de *trekking*. Ahora bien, la sensación de inicial un movimiento parece ser fundamental para la producción de una auto-atribución de agencia motora (Frith, 2000). Sin embargo, esta característica parece ser prescindible en el caso cognitivo, esto, por dos razones: (i) porque es difícil hablar de una intención de pensar P sin ya estar en el estado de pensar P, por lo que cualquier posición que utiliza la noción de intención en el caso cognitivo parece carecer plausibilidad, y; (ii) porque simplemente no es posible pensar en la experiencia de 'prepararse' para pensar P, sin que eso ya implique pensar P.

4.4 Sensación de esfuerzo

Para Bayne & Montague (2011) *la sensación de esfuerzo* tiene que ver con la experiencia de necesitar invertir voluntad y energía en el inicio de acciones que se llevarán a cabo. Esta sensación puede generar ciertas confusiones al momento de interpretar su significado, pues al hablar de esfuerzo, no necesariamente se refiere a la percepción de tensión muscular, agotamiento físico, elevación del

ritmo cardíaco, etc., sino que se asocia a una *sensación de hacer algo voluntariamente con niveles altos de esfuerzo.* Por esto, este tipo de sensación es separable de la sensación de iniciar un movimiento ya que no siempre el inicio de nuestros movimientos corporales está acompañado por altos niveles de esfuerzo.

Para el caso mental Otto, Zijlstra y Goebel (2014) postulan la existencia de una *sensación de esfuerzo mental* la cual se experimentaría en diversas tareas tales como recordar un evento pasado con gran dificultad, planificar unas vacaciones cuando nos cuesta mucho planificar algo, o, nuevamente, recordar las cosas que debo llevar a una tarde de *trekking*, cuando no tengo idea de trekking. Lo que tienen en común todas estas tareas es que el sujeto siempre debe esforzarse más de lo normal por encontrar la respuesta a una tarea cognitiva específica o resolver un problema específico, esto, ya sea por cansancio o por falta de disponibilidad de ciertos recursos cognitivos. Por lo tanto, la sensación de esfuerzo mental implicaría una sensación subjetiva de gasto de energía y tensión en el sujeto con la esperanza de producir un pensamiento que actúe como la respuesta para la tarea cognitiva en cuestión.

Es importante notar que la existencia de una sensación de esfuerzo no es necesaria para una autoatribución de agencia mental. Pensemos en el caso de los denominados pensamientos repentinos (unbidden thoughts; Franfkurt, 1976). Este tipo de pensamientos emergen en el flujo de la conciencia de forma repentina y, al aparecer, sin coherencia con el tren de pensamiento que es foco de la atención de un sujeto en ese momento. En sentido amplio, la emergencia de este tipo de pensamientos no está caracterizada por ningún tipo de sensación de esfuerzo ni control. Sin embargo, a pesar de esto, los pensamientos repentinos son atribuídos autorreferencialmente, lo que parece abrir una diferencia importante en la forma que tendría este proceso en el caso mental a diferencia del caso motor.

4.5 Sensación de agencia extendida

En el plano de lo motor es posible también hablar de una *sensación de agencia extendida* cuando un objeto diferente al propio cuerpo es parte fundamental de la consecución de una tarea física como, por ejemplo, en el manejo de un mouse en un computador. Si uno mueve el mouse y el puntero de este - en un intervalo pequeño de tiempo - también se mueve en la pantalla y en la dirección correcta, y a partir de esto uno experimenta un grado de agencia mediado por el mouse. En el dominio mental, no existe en la literatura actual una forma homóloga a este tipo de agencia, pero podríamos hipotetizar que algo similar podría darse cuando un sujeto utiliza una calculadora para resolver un problema matemático complejo. Es tal caso, el hecho de introducir los dígitos correctamente, junto otras acciones voluntarias, genera una sensación de agencia al observa el resultado lanzado por la calculadora. En tal caso, la calculadora es parte del proceso de resolución del problema matemático, y por consiguiente, la agencia también está extendida en su uso.

4.6 Sensación de coherencia y sensación de fluidez

Finalmente, tanto en el dominio motor como en el cognitivo se ha hablado de *sensaciones de coherencia* y fluidez. En el dominio motor, ambas sensaciones parecen estar asociadas al monitoreo de la ejecución de una acción específica en curso. Si este monitoreo es consistente con la predicción de como un movimiento *debería ser,* el sujeto podría experimentar una sensación de coherencia al observar cómo su acción se ajusta a esta predicción. Ahora, si tal movimiento posee altos niveles de eficacia e implica bajos niveles de control y esfuerzo, este movimiento también podría estar acompañado por una sensación de fluidez (ver Chambon y Haggard, 2012).

Ahora bien, en el dominio de lo mental, estas sensaciones han sido tratadas de forma superficial y, casi siempre, han sido entendidas a partir del análisis del *contenido* de un pensamiento (Galla-

gher, 2007, 2012). Sin embargo, creemos que esto es un error. La sensación de coherencia y fluidez, proponemos, no son propiedades que pertenecen al *qué* del pensamiento propiamente tal i.e. al contenido del pensamiento, sino que al *cómo* i.e. a la forma en que tales contenidos emergen en el flujo de la conciencia de un sujeto. No existe mucha coherencia de contenido entre pensar que debo C: ir a comprar vino y queso con el pensamiento de que P: debo terminar la postulación para un proyecto de investigación. Sin embargo, en nuestro diario vivir podemos comenzar pensando C y terminar en P pasando por múltiples otros pensamientos, y con todo esto, tener una robusta sensación de coherencia y fluidez. En este sentido, al igual que en el caso motor, al parecer las sensaciones de fluidez y coherencia surgen cuando el paso de un pensamiento a otro es monitoreado y el sujeto tiene razones suficientes como para comprender la forma en la cual se pasó de un pensamiento a otro. Así, al tener claridad sobre el camino recorrido, los pensamientos podrían ser experimentados como consistentes, aunque su contenido no tenga relación en términos de contenido, y si tal recorrido es realizado sin altos niveles de esfuerzo mental, el monitoreo de un cierto tren del pensamiento podría estar también acompañado de una sensación de fluidez, propia de cuando, por ejemplo, ejecutamos una tarea mental en la cual ya tenemos un grado de experticia considerable.

5. Conclusiones

Con el fin de contribuir al desarrollo de la comprensión del concepto de atribución de agencia mental, en este capítulo hemos analizado los diversos problemas que emergen de la adopción de una estrategia paralelista al intentar aplicar las explicaciones del caso motor al caso cognitivo libremente. Estas críticas han motivado la idea de que un primer paso imprescindible para explorar la naturaleza de las atribuciones de agencia mental implica la descripción de los aspectos fenomenológicos asociados a la agencialidad en el pensamiento. Así, mediante la realización inicial de este análisis hemos intentado clarificar el *explananda* para una posible

teoría sobre las atribuciones de agencia mental ya que, sin duda, todas las potenciales teorías explicativas que intenten darle sentido al fenómeno deberán establecer relaciones claras entre los aspectos fenomenológicos descritos en este capítulo. Estamos conscientes de que existe mucho trabajo por hacer, sin embargo, esperamos haber contribuido con la construcción de los primeros pasos para un proyecto de investigación que respete las particularidades del análisis de la agencia en nuestra vida cognitiva.

Agradecimientos

PLS agradece al 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 por proveer y promover espacios para pensar y re-pensar varias de las ideas que llevaron a la producción del presente trabajo. Además, quisiera agradecer a Jöelle Proust, Tim Bayne y a mis estudiantes de Seminario de investigación 2018 y 2019 por las interesantes discusiones en torno a la agencia mental.

Referencias bibliográficas

Aarts, H., Custers, R., y Wegner, D. (2005). On the inference of personal authorship: Enhancing experienced agency by priming effect information. *Consciousness and Cognition. 14*(3), 439-458.

Assal, F., Schwartz, S., & Vuilleumier, P. (2007). Moving with or without will: functional neural correlates of alien hand syndrome. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *62*(3), 301-306.

Bayne, T., y Pacherie, E. (2007). Narrators and Comparators: the architecture of agentive self-awareness. *Synthese. 159.*, 478-479.

Montague, M., & Bayne, T. (2011). Cognitive phenomenology: an introduction. *Cognitive phenomenology*.

Billon, A. (2013). Does consciousness entail subjectivity? The puzzle of thought insertion. *Philosophical Psychology*, 26(2), 291-314.

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

Carruthers, G. (2012). A Metacognitive model of the sense of agency over thoughts. *Cognitive Neuropsychiatry. 17*(4). 291-314.

Chambon, V., Wenke, D., Fleming, S., Prinz, W., y Haggard, P. (2013). An online neural substrate for sense of agency. *Cerebral Cortex. 23*(5) 1031-1037.

Feinberg, I. (1978). Efference copy and corollary discharge: Implications for thinking and its disorders. *Schizophrenia Bulletin, 4*(4), 636-640.

Frankfurt, H. (1976). Identification and externality. In A. O. Rorty (Ed.), The identities of persons (239–251). Berkeley: University of California Press.

Frith, C. D., Blakemore, S. J., and Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philos. Trans. R. Soc. Lond. B Biol. Sci. 355*(1404), 1771–1788. doi: 10.1098/rstb.2000.0734

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

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

Gallagher, S. (2012). Multiple Aspects in the sense of agency. *New Ideas in Psychology. 30*(1), 15-31.

Gallagher, S. (2014). Relations between agency and ownership in the case of schizophrenic thought insertion and delusions of control. *Review of Philosophy and Psychology. 6*(4), 865-879.

Gallagher, S. (2017). *Enactivism Interventions: rethinking the mind.* Oxford: Oxford University Press.

Haggard, P., y Chambón, V. (2012). Sense of control depends on fluency of action selection, not motor performance. *Elsevier. 125*(3), 441-451.

Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience*, 9(4), 304–313.

Kang, S., Im, C., Shim, M., Nahab, F., Park, J., Kim, D., Kakareka, J., Miletta, N., Hallett, M. (2015). Brain Networks Responsible for Sense of Agency: An EEG Study. Plos One. 10(8), 1-16.

López-Silva, P (2017). Me and I are not friends, just Acquaintances: On thought Insertion and Self-Awareness. *The Review of Philosophy & Psychology.* DOI: https://doi.org/10.1007/s13164-017-0366-z

López-Silva, P. (2018). Mapping the psychotic mind: A review on the subjective structure of thought insertion. *Psychiatric Quarterly*, *89*(4), 957-968.

López-Silva, P (2019-En Prensa). ¿De quién son éstos pensamientos? Examinando el enfoque top-down de las atribuciones de agencia mental. *Tópicos*.

Marcel, A. (2003). The sense of agency: Awareness and ownership of action. In J. Roessler & N. Eilan (eds.), Agency and Self-Awareness: Issues in Philosophy and Psychology. Oxford, Clarendon Press.

Moore, J. (2016). What Is the Sense of Agency and Why Does it Matter? *Frontiers in Psychology. 7*(1272), 1-9.

Mullins, S., y Spence, S. (2003). Re-examining thought insertion. Semi-structured literature review and conceptual analysis. *British Journal of Psychiatry. 182*(4), 293-298.

Otto, T., Zijlstra, H., Goebel, R. (2014). Neural correlates of mental effort evaluation - involvement of structures related to self-awareness. *SCAN.* 9(3), 307-315.

Pacherie, E. (2007). The sense of control and the sense of agency. *Psyche*, *13*(1), 1-30.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, *107*(1), 179-217.

Proust, J. (2009). Is there a sense of agency for thoughts? En L. O'Brien, M. Soteriou (eds.), *Mental Action*s, pp. 253-279. Oxford: Oxford University Press.

Schmahmann, J. D. (2004). Disorders of the cerebellum: ataxia, dysmetria of thought, and the cerebellar cognitive affective syndrome. *Journal of Neuropsychiatry and Clinical Neuroscience*, 16(3), 367–378.

Stephens, G. L., y Graham, G. (2000). *Philosophical psychopathology: Disorders in mind. When self-consciousness breaks: Alien voices and inserted thoughts.* Cambridge, The MIT Press.

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

Walsh, E., Oakley, D., Halligan, P., Mehta, M., y Deeley, Q. (2015). The functional anatomy and connectivity of thought insertion and alien control of movement. *Cortex, 64*, 380–393.

Wenzlaff, R., y Wegner, D. (2000). Thought Suppression. *Annual Review of Psychology. 51*, 59-91.

Zahavi, D. (2005). *Subjectivity and Selfhood: Investigating the first-person perspective*. Cambridge, MA: The MIT Press.

Sobre los autores

Pablo López-Silva es psicólogo por la Pontificia Universidad Católica de Valparaíso, Master in Research y PhD in Philosophy por la Universidad de Manchester, Reino Unido. Actualmente se desempeña como profesor adjunto en la Escuela de Psicología de la Universidad de Valparaíso, Chile. Su campo de investigación es la filosofía de la mente, la psicopatología y la psicología filosófica.

Andrea Arancibia, Leonardo Henríquez y Gabriel Cordero participan en la versión 2019 del Seminario de Investigación de la Escuela de Psicología de la Universidad de Valparaíso, Chile.

S | T

La Serie Selección de Textos es una producción editorial del Centro de Estudios en Filosofía, Lógica y Epistemología (CeFiLoE), del Instituto de Filosofía de la Universidad de Valparaíso, Chile. Nace en el año 2013 con el propósito de abrir un espacio a los autores para la publicación de libros y capítulos en el área de la filosofía y disciplinas afines. Todos los trabajos son sometidos a arbitraje a doble ciego (double blind review).

La Serie es dirigida por el profesor Juan Redmond y es editada por Rodrigo López Orellana y Jorge Budrovich. Su Comité Científico lo componen destacados académicos nacionales e internacionales, cuya responsabilidad es asegurar la calidad de las publicaciones.

Sus objetivos generales son: i. ofrecer publicaciones académicas de calidad científica; ii. proporcionar a la comunidad de académicos y estudiantes un medio de publicación sin fines de lucro; y iii. publicar libros que sean accesibles para todos, sin un costo asociado.

Volumen 1 - 2013
Ciencia, Tecnología e Ingeniería. Reflexiones filosóficas sobre problemas actuales
Editores: Carlos Verdugo S. & Juan Redmond C.

Volumen 2 - 2013
Amauta y Babel. Revistas de disidencia cultural
Editores: Osvaldo Fernández D. & Patricio Guitiérrez D. & Braulio Rojas C.

Volumen 3 - 2015
Conceptos y lenguajes, en ciencia y tecnología
Editores: Guillermo Cuadrado & Juan Redmond & Rodrigo López O.

Volumen 4 - 2015
Hacer filosofía con niños y niñas. Entre educación y filosofía
Editores: Juan Estanislao Pérez & Juan Pablo Álvarez & Claudia Guerra A.

Volumen 5 - 2015
Estudios y preludios. Contribuciones a la filosofía desde Valparaíso
Editores: Jorge Budrovich Sáez & Rodrigo López Orellana

Volumen 6 - 2015
De camino a la filosofía. Sobre el aprendizaje de la filosofía escribiendo
Editores: Juan Redmond & Rodrigo López O. & Jorge Budrovich S.

www.selecciondetextos.cl

S | T

La Serie Selección de Textos es una producción editorial del Instituto de Filosofía de la Universidad de Valparaíso, Chile. Nace en el año 2013 con el propósito de abrir un espacio a los autores para la publicación de libros y capítulos en el área de la filosofía y disciplinas afines. Todos los trabajos son sometidos a arbitraje a doble ciego (*double blind review*). Sus objetivos generales son: ofrecer publicaciones académicas de calidad científica; proporcionar a la comunidad de académicos y estudiantes un medio de publicación sin fines de lucro; y publicar libros que sean accesibles para todos, sin un costo asociado ni retención de derechos de autor.

ISBN: 978-956-402-647-3