EL PROBLEMA MENTE – CUERPO: METAFÍSICA DE LA MENTE

Agustín Vicente
Ikerbasque Foundation for Science/ Universidad del País Vasco
agustin.vicente@ehu.es

Introducción[1]

El problema en torno a la relación de la mente, anteriormente alma o espíritu, con el mundo físico, y en particular con el cuerpo, es uno de los problemas clásicos de la filosofía. Su antigüedad explica que aún hablemos de "el problema mente-cuerpo", cuando de hecho los problemas son varios y la etiqueta confundente: hablar de "mente-cuerpo" nos induce a pensar en un problema de relación entre sustancias, mientras que actualmente se tiende a pensar que hay una única sustancia (física) que, en ciertos casos, tiene propiedades que parecen no ser físicas. Es decir, a día de hoy lo que se considera problemático no es la relación entre una cosa llamada 'mente' y otra llamada 'cuerpo', sino entre ciertas propiedades (mentales) que tienen algunos cuerpos físicos y sus propias propiedades físicas[2].

Suele hablarse de dos tipos de propiedades mentales: las propiedades representacionales o intencionales y las propiedades cualitativas[3]. Ambos tipos de propiedades son peculiares y características. Tener propiedades representacionales consiste en ser capaz de representar (en este caso, representars*e*) el entorno. Es decir,

[1] Este trabajo ha sido subvencionado por el proyecto "La naturalización de la subjetividad" (FFI2010-15717), cuyo investigador principal es David Pineda. Tengo que agradecer a bastante gente los varios años de intercambios de pareceres sobre el tipo de cuestiones que aquí se exponen, entre ellos el propio David Pineda, Manuel Pérez Otero, Manuel García-Carpintero, Agustín Arrieta, Adrián Sampedro León y Josep Lluís Prades. A Agus Arrieta y a Adrián Sampedro les agradezco también la lectura y los comentarios que ha hecho de este capítulo, y a Josep Lluís Prades el haberme invitado a participar en este volumen y su labor de coordinación.

[2] Hay notables excepciones a esta postura dominante: E. J. LOWE (2000) y L. R. BAKER (2013), entre otros, defienden un dualismo de sustancias, aunque las sustancias que distinguen, en su caso, es personas frente a cuerpos o, eventualmente, animales humanos. Su idea es que las condiciones de identidad de las personas y las de los cuerpos y/o los animales humanos, son distintas, y por tanto, personas y cuerpos y/o animales son entidades diferentes.

[3] Es posible también que existan propiedades que tengan tanto aspectos representacionales como cualitativos, es decir, que sean acerca de algo al tiempo que se sienten de determinada forma. Por ejemplo, las emociones son sensaciones, pero según muchos también tienen un contenido, ya sea sobre el propio cuerpo o sobre hechos del entorno. Algunos filósofos sostienen que todas las propiedades cualitativas son como las emociones, es decir, representacionales a la vez que cualitativas. Finalmente, también hay autores que defienden que todas las propiedades representacionales tienen aspectos cualitativos (para una introducción a la cuestión reciente de la "fenomenología cognitiva", véase JORBA, 2013, JORBA y VICENTE, 2014). Más que nada por facilitar la exposición, prescindiré de estas posibles complicaciones.

consiste en estar en ciertos estados que son acerca de otras cosas, estados que tienen un contenido que puede coincidir con la realidad del entorno o no. Un ejemplo de estado representacional es creer algo, por ejemplo, creer que se acerca un hombre. Esta creencia puede ser verdadera –de hecho, un hombre se acerca- o falsa –el hombre en realidad se aleja, lo que acerca no es un hombre, o, simplemente, no hay nada que se acerque o se aleje-. Una característica especialmente intrigante de las propiedades representacionales es que pueden ser acerca de cosas que no existen: tememos a los monstruos, creemos que los marcianos nos vigilan o soñamos con volar en caballos alados.

Tener propiedades cualitativas, por su parte, implica sentir las cosas de un determinado modo, un modo que parece que sólo nosotros podemos conocer, y además con absoluta inmediatez y sin posibilidad de error. Son, por tanto, propiedades totalmente subjetivas, y lo son en un doble sentido: por una parte, su existencia depende del sujeto que las experimenta; por otra, son el tipo de propiedades que conforman la propia subjetividad, la forma en que a un individuo particular se le presenta el mundo. Estas propiedades nos acompañan todo el tiempo: experimentamos colores, sabores, olores, miedos, dolores, placeres y frustraciones, y también tenemos experiencias asociadas con actuar en el mundo, hablar, o incluso creer o desear algo.

Como digo, estos dos tipos de propiedades son peculiares, al menos desde un punto de vista filosófico. Aunque no son exclusivas de los seres humanos, parecen ser propiedades muy diferentes de las que ocupan a las ciencias naturales, es decir, parecen muy diferentes del resto de propiedades que encontramos en el mundo natural y, en particular, del resto de las propiedades que tienen nuestros cuerpos. No hay en resto el mundo natural propiedades que sean acerca de algo y que sean susceptibles de ser verdaderas o falsas, así como tampoco hay propiedades que sean subjetivas y presenten el mundo de una determinada manera. Para nosotros mismos, seguramente no hay nada más inmediato y menos problemático que nuestras vivencias, percepciones, creencias y deseos. El problema surge cuando contemplamos estas propiedades que nos atribuimos desde el punto de vista del mundo natural. Vistas desde ahí, no podemos sino preguntarnos: ¿En qué consisten realmente estas propiedades? ¿Son realmente diferentes del resto de propiedades? ¿Cómo hacen para insertarse en el mundo natural, y en particular, en su flujo causal? Básicamente, estas son las preguntas que, a día de hoy, intentan responder los filósofos que trabajan en “el problema mente-cuerpo”.

Los problemas

Una reacción muy legítima ante todo lo anterior consiste en cuestionar la dicotomía que se establece entre las propiedades mentales y las demás. Cierto, las propiedades mentales tienen determinados rasgos característicos, pero que los tengan no implica que no sean fenómenos naturales. También los seres vivos, en cuanto tales, tienen propiedades con rasgos característicos, y las rocas tienen solidez, por ejemplo, una propiedad con rasgos característicos que no tienen otras partes del mundo natural. ¿Por qué hemos de pensar que los rasgos característicos de las propiedades mentales son diferentes, hasta el punto de creer que existe un problema de encaje que no existe en otros casos? La respuesta a esta pregunta consiste, probablemente, en invertir su formulación: parece existir, en el caso de las propiedades mentales, un problema de encaje que no existe en los otros casos, y eso es lo que nos lleva a pensar que sus rasgos característicos son especiales.

En una primera aproximación a este problema de encaje, cabe decir que mientras que no vemos excesivamente problemático calificar al resto de propiedades de "físicas", no estamos seguros de que la etiqueta sea aplicable a las propiedades mentales. No tenemos problema en afirmar que la solidez de la roca es una propiedad física, o que las propiedades, aparentemente tan *sui generis*, del mundo vivo, son propiedades físicas (aunque aquí ya habría más disensiones). Sin embargo, parece que nos cuesta pensar que las propiedades mentales puedan llamarse 'físicas'. Claro, que esto no es sino, como mucho, un síntoma. ¿Por qué el resto de propiedades nos parecen físicas y las propiedades mentales no? Para responder a esta pregunta debemos explicar, primero, qué entendemos por "físico", y segundo, cuál es la diferencia crucial que detectamos en las propiedades mentales para que nos resistamos a llamarlas 'físicas'.

El "problema del cuerpo"

La caracterización de lo físico se ha convertido en un pujante problema filosófico. Su enunciación data de algunas décadas atrás, pero el debate principal se ha producido en los últimos años (ver Montero, 1999). La empresa de dar con una caracterización de lo físico tiene como propósito sostener y explicar la intuición de que, mientras propiedades como la solidez son claramente físicas, las propiedades representacionales o las

cualitativas, tal como se nos presentan, parecen no serlo. Esta es la intuición de partida. Lo que se busca, entonces, es dar con una noción precisa de lo físico que nos permita establecer la distinción que intuitivamente establecemos, y que la justifique. Si la empresa no llega a puerto, lo correcto será revisar nuestra intuición. Parece que no podemos discutir sobre el "problema mente-cuerpo" si no resolvemos antes el "problema cuerpo" (ver Montero, 1999).

Quizás la primera idea que viene a la cabeza sobre qué quiere decir que ciertas propiedades sean físicas es que son propiedades que maneja la ciencia llamada 'física'. Esta idea es claramente incorrecta: la física no menciona propiedades químicas, ni geológicas, ni biológicas. Puede enmendarse la idea si la complicamos un poco: cabe decir, por ejemplo, que son físicas tanto las propiedades que menciona la física como las que son reducibles a propiedades mencionadas por la física. Llamaremos a esta definición la 'definición reductivista'. La *definición antirreductivista* es más liberal: sostiene que entra dentro de lo físico no sólo lo anterior, sino también aquellas propiedades que *dependen* de las propiedades que maneja la física, es decir, aquellas propiedades que no se pueden dar de no darse algún sustrato de propiedades estrictamente físicas y que necesariamente se dan siempre que se da una determinada propiedad estrictamente física (entendiendo por "estrictamente físicas" las propiedades mencionadas por la física)[4].

Ninguna de estas definiciones es enteramente satisfactoria. La definición antirreductivista resulta demasiado liberal: al fin y al cabo, incluso las propiedades mentales parecen depender de las físicas[5]. Es decir, la definición puede ser tan liberal que no nos permita trazar la frontera que queremos trazar entre "lo físico" y "lo mental". Por otra parte, la liberalidad de la definición antirreductivista nos fuerza a incluir entre lo físico algunas otras propiedades que podemos considerar dudosamente físicas. Por ejemplo, hay quien sostiene que existen propiedades que no causan nada, propiedades *epifenoménicas*. Es difícil pensar que haya propiedades físicas de este tipo, y es difícil pensar que estemos dispuestos a llamar 'físicas' a propiedades de esta clase. Sin

[4] Estas definiciones de "lo físico" tienen su origen en el debate sobre el fisicismo, que es un debate ontológico, esto es, un debate acerca de lo que existe y lo que no, y en particular acerca de si existe sólo lo físico o hay algo más. He cambiado ligeramente los términos del debate, presentándolo más bien como una discusión en torno a la noción de "lo físico" por motivos expositivos. No creo que el hacerlo genere problemas de comprensión para quien conozca el debate ontológico, ni para quien vaya a conocerlo.

[5] Como veremos, cabe poner en cuestión que las propiedades cualitativas dependan en el sentido considerado de las propiedades físicas. Aun así, sólo con el hecho de que las propiedades representacionales sí dependen de lo físico, tenemos un problema de delimitación.

embargo, si estas propiedades –de existir- dependieran de propiedades físicas, la definición antirreductivista de lo físico nos obligaría a considerarlas físicas.

No obstante, la definición reductivista tampoco es perfecta y arroja resultados dudosos en algunos casos. Por ejemplo, pensemos en la doctrina emergentista. Los emergentistas mantienen que, cuando el mundo estrictamente físico se hace lo suficientemente complejo, aparecen o "emergen" propiedades nuevas con poderes causales distintivos. La definición reductivista nos obliga a considerar que estas propiedades no son físicas (ya que no son reducibles a propiedades estrictamente físicas). Pero quizás esta idea nos genere cierta incomodidad: no podemos dejar de pensar que hay un sentido de "lo físico" que abarca a este tipo de propiedades.

Como en muchos otros casos, el problema puede radicar en que nuestros conceptos intuitivos son demasiado vagos y polisémicos. Contrariamente a lo que podamos creer, los conceptos no tienen definiciones. Desde la filosofía platónica se ha intentado, en vano, dar con definiciones de conceptos como los de bondad, justicia, belleza, verdad o conocimiento. Un simple argumento de inducción a partir de los fracasos pasados debería hacernos pensar que la tarea de capturar en una definición nuestra idea de "lo físico" es más que complicada. La definición, por tanto, tiene que ser estipulativa, y eso significa que o bien optamos por una definición restrictiva, que dejaría fuera propiedades que podríamos ver como "físicas", o bien apostamos por una definición liberal, con el riesgo de que no ejerza su labor de filtrado. Entre estas dos opciones, la más razonable es la primera: si buscamos explicarnos por qué creemos que las propiedades mentales tienen un encaje problemático en lo físico, lo principal es asegurar que nuestra definición de lo físico respeta *a priori* la demarcación entre propiedades mentales y propiedades físicas. No parece que la definición antirreductivista cumpla con este cometido, mientras que, dado que *a priori* las propiedades mentales no parecen reducibles a propiedades estrictamente físicas, la definición reductivista sí lo cumple. Nótese que esto no quiere decir que estemos asumiendo, incurriendo en una petición de principios, que las propiedades mentales no son físicas. La definición restrictiva nos ayuda a marcar una frontera, que intuitivamente existe, entre lo mental y lo físico[6]. Pero no dictamina que lo mental no es físico: existe

[6] Es razonable pensar que esta distinción intuitiva está anclada en nuestro desarrollo psicológico: ya los bebés de pocos meses parecen concebir el mundo en términos de objetos físicos *vs* agentes intencionales (ver, BLOOM, 2004, CAREY, 2009).

la posibilidad de que descubramos ciertas cosas acerca de lo mental que nos hagan pensar que, en realidad, lo mental *sí* es reducible a lo físico, y por tanto, que la distinción que fundamentamos en nuestras intuiciones (y que la definición reductivista explicita) es infundada.

De modo que tenemos que son propiedades físicas o bien las propiedades que maneja la física o bien las que se pueden reducir a éstas. ¿Es satisfactoria esta definición? Ya hemos dicho que es restrictiva, y algunos pensarán que lo es inmotivadamente. Pero, dejando esta cuestión a un lado, existen dos problemas que hacen que no podamos darnos por satisfechos: en primer lugar, no está claro de qué hablamos cuando hablamos de *reducibilidad*; en segundo, simplemente, no podemos anclar nuestra definición a lo que diga la física actual.

En Pineda (este volumen) encontrará el lector la orientación necesaria para ver que existen diferentes modos de entender esta relación, que aparejan distintos modos de entender nuestra definición de lo físico. Por no reiterar, aquí me centraré en el segundo de los problemas, el que toca a la caracterización de la física. He dicho que no podemos anclar nuestra definición de lo físico a lo que diga la física actual. Tal como se ha enunciado nuestra definición, es la física actual la que dicta qué propiedades hemos de considerar físicas: son físicas o las propiedades que maneja la física o las que son reducibles a éstas. El problema es que, muy probablemente, la física actual es falsa, y por tanto es posible que muchas de las propiedades que esta física maneja ni siquiera existan. Un argumento general que sostiene esta afirmación aparentemente radical es el llamado 'argumento de la meta-inducción pesimista' (ver Laudan, 1981). El argumento es simple: todas las teorías físicas anteriores han resultado ser falsas. Luego probablemente las actuales también lo sean. Pero además hay un argumento particular que se aplica a la física contemporánea, y es que se sostiene sobre dos teorías (la cuántica y la relativista) que resultan ser contradictorias. En conclusión: caracterizar lo físico en términos de la física contemporánea no es en absoluto una buena idea.

En 1980, Hempel propuso un dilema al filósofo fisicista que sostiene que sólo existen las entidades que postula la física. Acabamos de toparnos con la primera rama del dilema: si esa física de la que hablamos es la física contemporánea, entonces el fisicismo es falso, pues muy probablemente lo que realmente existe *no* son las entidades que postula la física contemporánea. La segunda rama del dilema sostiene que si, en

lugar de apostar por la física contemporánea, apostamos por la física verdadera, entonces el fisicismo se convierte en una doctrina vacía. Como no sabemos qué tipo de entidades manejará la física verdadera, no estamos diciendo nada sobre qué cosas existen y qué cosas no. Es más, la doctrina corre el riesgo de ser trivial. El filósofo fisicista quiere sostener que no hay propiedades irreductiblemente mentales. Sin embargo, según Chomsky (1995), dado que no sabemos nada de esa hipotética física verdadera, podemos pensar que entre las propiedades que maneje estén propiedades irreductiblemente mentales. A lo largo de la historia, la física ha ido ampliando "su negocio": en su origen estudiaba los fenómenos mecánicos, después incorporó los electromagnéticos, más tarde los atómicos… Quizás llegue el día en que vuelque su interés en los fenómenos mentales e incluso que no vea otra forma de hacerlo más que incorporando a las propiedades mentales tal cual. La posibilidad, sin duda, parece remota, pero dada nuestro estado de ignorancia en lo que respecta a una eventual futura física verdadera, quizás no podemos descartarla. El problema, entonces, es que el fisicismo se vuelve trivial, pues resultaría ser verdadero incluso si resultara haber propiedades irreductiblemente mentales. Llevado a nuestra discusión, el problema se convierte en que corremos el riesgo de estar llamando 'físicas' a las propiedades mentales.

En los últimos tiempos buena parte de los autores que se ocupan de este problema han optado por renunciar a entrar en el dilema propuesto por Hempel. Esto supone abandonar el intento de caracterizar lo físico únicamente en términos de las ciencias físicas, sean las contemporáneas o las futuras o hipotéticas. Hay varias propuestas que siguen esta línea, y lo que todas ellas tienen en común es recurrir a un mismo elemento negativo, o de oposición, en su definición. La más simple de todas las caracterizaciones propuestas es la que se limita a establecer esa oposición (ver Montero y Papineau, 2005): lo físico, según esta idea, es simplemente lo que no es irreductiblemente mental. Esta nueva vía, llamada "la vía negativa", no sólo despeja el horizonte de problemas como el de Hempel, sino que acomoda de la mejor manera posible (por ser la más directa) la intuición de que las propiedades mentales se diferencian de las físicas. Además, nos permite prescindir de debates engorrosos, como el de si las propiedades biológicas son o no reducibles a las estrictamente físicas. Si decimos que una propiedad es física sólo si o bien es estrictamente física o bien es reducible a propiedades estrictamente físicas, surge la cuestión de si las propiedades biológicas son físicas, dado

que hay autores que sostienen que las propiedades biológicas (o muchas de ellas, al menos) no son reducibles a propiedades estrictamente físicas. Si resultara que no lo son, entonces nuestra caracterización de lo físico no incorporaría a las propiedades biológicas, un resultado contraproducente en este contexto. La "vía negativa" nos permite dejar al margen toda la discusión acerca de la relación entre lo estrictamente físico y lo biológico, colocando a este segundo mundo de propiedades en el lado en que seguramente queremos que esté.

La "vía negativa" es incapaz de realizar otras funciones que podría realizar una definición de lo físico con más contenido. Según la exigua definición que nos proporciona la "vía negativa" resultarían ser parte de lo físico, de existir, cosas tales como la divinidad, las energías positivas o las influencias astrológicas. Dado que, según la definición, todo lo que no es mental es físico, tendríamos que considerar a estas entidades más o menos estrafalarias (insisto, de existir) como entidades físicas. Podemos considerar el problema superfluo, dado que nos creemos en disposición de asegurar que tales entidades no existen (aunque la cuestión de la divinidad sería algo peliaguda), pero lo interesante es que no podemos cimentar nuestra creencia en un argumento metafísico. Una definición lo físico con contenido, por ejemplo, la definición reductiva, nos permite elaborar una tesis fisicista que también tendrá contenido, a saber: no hay en el mundo más que las entidades que postula la física o las que son reducibles a ellas. Armados con esta tesis sustantiva, podemos excluir que existan influencias astrológicas, irradiaciones de energías positivas, hombres invisibles o dioses griegos. Sin embargo, la definición que nos brinda la "vía negativa" sólo nos permite sostener que en el mundo no hay más que entidades no mentales.

Tampoco nos servirá la "vía negativa" para afrontar otro tipo de debates. Por ejemplo, hemos visto que hay autores que mantienen que hay diferencias importantes entre lo biológico y lo físico (en el presente contexto: entre lo biológico y lo estrictamente físico). Si queremos hacer el debate inteligible tenemos que dar con una definición de lo físico (o lo estrictamente físico) que podamos oponer a lo biológico. Como es obvio, la "vía negativa" no nos sirve. Quizás podría servir una "vía negativa" revisada y adoptada a ese debate, en la que lo físico se oponga a lo biológico. El problema es que la revisión trae consigo invitados indeseados al mundo físico: las propiedades mentales tendrían que ser consideradas físicas.

Sin embargo, a mi modo de ver el mayor problema de la "vía negativa" reside en la que parece ser su mayor virtud, esto es, que, a diferencia de la definición reductivista –excluyente- y de la antirreductivista –excesivamente incluyente-, la "vía negativa" sí recoge la distinción intuitiva entre lo físico y lo mental. Esta es una virtud sólo aparente: la definición negativa realmente no recoge o acomoda esa distinción; más bien, la transcribe de modo literal. Al adoptar la "vía negativa" nos estamos declarando incapaces de hacer algo que queríamos hacer, a saber, motivar nuestra distinción. No nos da ninguna idea de por qué las propiedades mentales (irreductibles) pueden ser diferentes del resto de propiedades del mundo natural. La "vía negativa" no nos dice nada acerca de las otras propiedades del mundo natural ni de sus relaciones. Las propiedades biológicas serán físicas independientemente de si son reducibles a las propiedades estrictamente físicas o no lo son. Supongamos que no lo son. Entonces surge la pregunta: ¿por qué, siendo irreductibles tanto las propiedades biológicas como las mentales, unas han caer de un lado y otras de otro? Como respuesta, sólo podemos invocar nuestra intuición pre-filosófica[7].

El problema de la representación

Hemos dicho que para poner en pie el problema cuerpo-mente necesitamos primero tener una idea de lo que es y lo que no es físico. Esa cuestión está de momento abierta, y no está claro, desde la perspectiva explorada hasta ahora, qué es lo que hace diferentes y especiales a las propiedades mentales. Pero cabe decir que usualmente se considera problemático que no sean reducibles a propiedades físicas. Desde luego, está claro que el problema mente-cuerpo se disolvería si consiguiéramos mostrar que las propiedades mentales son explicables en los términos acuñados por otras ciencias. Y también está claro que tenemos un problema mente-cuerpo si resulta que las propiedades que manejan las distintas ciencias son reducibles a un vocabulario más básico y que las propiedades mentales no lo son. Lo más habitual ha sido pensar que las ciencias naturales como un todo exhiben algún tipo de unidad, y por lo tanto, que resolver el problema mente-cuerpo exige integrar a las propiedades mentales en esa "unidad de las

[7] Esta crítica se aplica a la versión más simple de la "vía negativa". Hay versiones con más contenido que disponen de los medios para responder a este tipo de cuestiones. Véase, por ejemplo, la definición "futurista" de WILSON (2006) o la mereológica de PINEDA (2006). Para una crítica de cualquier definición que incluya un elemento negativo, ver VICENTE (2011).

ciencias". Este supuesto es el que ha alimentado el programa que se ha dado en llamar *la naturalización* de lo mental.

El objetivo del programa "naturalizador" es mostrar que tanto las propiedades representacionales como las cualitativas no son sino propiedades físicas. Como digo, parece la manera más convincente de resolver el problema mente-cuerpo. El programa ha atravesado dos estadios temporalmente definidos: en la década de los ochenta y principios de los noventa, los esfuerzos se centraron principalmente en la naturalización de las propiedades representacionales. A mediados de los noventa, sin embargo, hizo su entrada estelar en el mundo intelectual la consciencia, y el foco de los filósofos "naturalizadores" se volvió hacia las propiedades cualitativas.

Las propiedades y estados representacionales, como se ha dicho, se caracterizan por ser propiedades y estados que pueden ser evaluados en términos como los de verdad y falsedad, o acierto y equivocación. La creencia de que se acerca un hombre puede ser verdadera o falsa; la percepción de un objeto moviéndose en una trayectoria de zig-zag puede ser verídica o ilusoria; la intención de atrapar un pez que corre río abajo puede ser tener éxito o fracasar, y así sucesivamente. Todos estos estados, cuyos componentes son propiedades representacionales, tienen un contenido, son acerca de algo, algo que puede de hecho ocurrir en el mundo, pero que también puede no hacerlo.

El propósito de los filósofos "naturalizadores" ha consistido en mostrar que este rasgo característico de las propiedades representacionales es explicable en términos que no son, desde el punto de vista naturalista, sospechosos. Los principales exponentes de esta empresa son Fred Dretske (1988), Jerry Fodor (1987) y Ruth Millikan (1984), y las nociones no sospechosas con las que han trabajado son dos: la de relaciones y leyes causales y la de función biológica.

Exponer con detalle las propuestas de estos tres filósofos requeriría un capítulo específico. Aquí me conformaré con explicarlas un tanto superficialmente y remitir al lector a lo que creo son buenas introducciones a las tres teorías[8]. Dretske, Fodor y Millikan comparten la idea de que el cerebro es, básicamente, un órgano representacional, esto es, un órgano que fundamentalmente representa el entorno, y que

[8] Las teorías de Dretske, Fodor y Millikan son las que más atención han atraído, pero no son las únicas. Se pueden encontrar buenas introducciones a estas y otras teorías en RUPERT (2008) y en ADAMS y AIZAWA (2010).

utiliza estas representaciones para guiar las interacciones del sujeto con el entorno[9]. No hay un acuerdo entre los tres acerca del formato de estas representaciones: Fodor capitanea la idea de que son parte de un "lenguaje del pensamiento", mientras que Millikan parece preferir pensar en términos de algo semejante a mapas. Sin embargo, los tres coinciden en proponer que estar en un cierto estado mental con capacidad representacional consiste en tener un cierto tipo de relación con una representación mental. La cuestión a explicar, entonces, se convierte en cómo explicar que las representaciones mentales tengan contenido o sean acerca de algo. Fodor y Dretske tienen en común el uso de nociones causales, mientras que Dretske y Millikan comparten el invocar en su análisis el término de función teleológica.

De acuerdo con la "vía causal", para que una representación sea acerca de algo tiene que existir una relación causal (que no tiene por qué ser directa), sostenida por al menos una ley, entre la representación y ese algo. Por ponerlo en una forma burda, para que una representación sea acerca de los hombres, ha de cumplirse que sólo los hombres causen la activación de la representación, y que lo hagan todos. Se dice entonces que la representación #*hombre#* indica o lleva la información de la presencia de hombres en el entorno. Esta idea, tal como está, es claramente incorrecta. Hay cantidad de cosas –maniquíes, mujeres, gatos que pisan una rama- que pueden activar mi representación #*hombre#* sin que ésta deje de ser acerca de los hombres. En estos casos, decimos que el acto de representación es erróneo, pero eso no quiere decir que la representación pierda su contenido: sigue siendo acerca de los hombres.

De modo que la condición de causalidad ha de complementarse con alguna otra. Fodor opta por introducir una segunda cláusula: para que una representación *R* sea acerca de *x,* ha de cumplirse (a) que los *x*s causen la activación de *R,* y (b) que para cualquier *y* distinto de *x* que cause la activación de *R,* la ley causal que conecta a *y* con *R* depende (asimétricamente) de la ley causal que conecta a *x* con *R*. Yendo a nuestro ejemplo: mi representación #*hombre#* es acerca de los hombres porque, primero, los hombres la activan, y segundo, cualquier otra cosa que la active, lo hace justamente porque los hombres la activan. Es decir, si no fuera porque hay una ley causal que conecta a los hombres con #*hombre#*, los gatos que pisan ramas, los percheros que

[9] En los últimos tiempos, hay más y más autores que se inclinan por intentar explicar la cognición y la acción sin apelar a las representaciones: ver VAN GELDER (1995), CALVO GARZÓN (2008), CHEMERO (2009).

acarician espaldas, o los maniquíes que simulan hombres, no causarían en ningún caso la activación de #*hombre#*.

La manera en que Drestke afronta la situación es bien diferente. Su propuesta consiste en recurrir a la noción de función teleológica. Según esta propuesta, para que *R* represente o sea acerca de *x*, además de la condición causal ha de cumplirse que *R* haya sido reclutada o tenga la función de indicar *x*. Pensemos en la siguiente situación: unos hombres se acercan a un animal. En el animal se activa un conjunto de neuronas. Los hombres le acercan comida al animal, y hacen lo mismo en días sucesivos. Esto hace que el grupo de neuronas que se activaron ante la presencia del hombre establezcan conexiones con algunos grupos de neuronas motoras, responsables de que el animal, a partir de cierto momento, exhiba un patrón de respuestas tolerantes y amistosas ante la presencia de los hombres. En ese momento, podemos decir que el grupo de neuronas original ha sido reclutado, y que ha adquirido una doble función: por una parte, ha adquirido la función de activar un cierto tipo de respuestas; por la otra, y más importante por lo aquí nos concierne, ha adquirido también la función de indicar la presencia de los hombres. En ese momento podemos decir que la representación tiene un contenido determinado, es acerca de algo. También podemos decir que cuando la representación se activa con, en lugar de hombres, el silbido del viento, el acto de representación es erróneo, puesto que no cumple su función.

Ciertamente, el uso de la noción de función es muy acertado: si entendemos que la clave de la idea de representación es que ésta puede ser errónea, la única noción que tenemos en las ciencias naturales que incluya la posibilidad de que algo falle o yerre es la de función teleológica. En efecto, asignar una función a algo es decir que ese algo puede cumplir con su función o no hacerlo, en cuyo caso decimos que no ha funcionado bien, que ha fallado o que su comportamiento ha sido errado.

La forma en que Millikan utiliza la noción de función para explicar la de representación difiere de la de Drestke, sobre todo porque no incluye ningún elemento causal, aunque también tiene algunas semejanzas con la teoría de Dretske (y bastante más complicaciones). Según Millikan, un "productor" de representaciones tiene que cumplir dos condiciones: en primer lugar, tiene que producir signos que guarden una similitud estructural con aquello que representan, i.e., el sistema representacional tiene que ser isomorfo a lo representado. En segundo lugar, los signos que forman parte del sistema representacional tienen que tener algún papel funcional: tienen que tener una

función. La función de un sistema representacional es, en general, la de ayudar a que quien (o lo que) "consume" la información realice funciones que sirven para que ambos productor y consumidor de la representación prevalezcan o perduren. Es decir, el productor de la representación está "diseñado" para ajustarse al entorno y suministrar al consumidor el tipo de información que éste necesita para que ambos realicen su función cooperativa.

Por ejemplo, podemos pensar en que una abeja es un productor de representaciones, cuyos consumidores son otras abejas. La abeja produce representaciones que están diseñadas para ajustarse al entorno de acuerdo con las necesidades de sus consumidores, por ejemplo, la de encontrar néctar. Si la abeja cumple con su función y los consumidores son capaces de encontrar el néctar, la probabilidad de que la especie perviva se incrementa notablemente. Los productores y los consumidores de las representaciones, no obstante, pueden ubicarse en el mismo organismo: el cerebro de una rana produce representaciones, que son "consumidas" por su aparato digestivo. El buen funcionamiento del sistema representacional de la rana ayuda a que su aparato digestivo también funcione bien, y de esa forma a que la rana continúe viviendo[10].

Ninguno de estos tres programas naturalizadores es satisfactorio (ni tampoco lo son las propuestas más recientes). Hay múltiples problemas, específicos de cada uno de ellos, que parecen no poder resolver. Entrar en esa casuística nos llevaría demasiado lejos, de modo que me conformaré con comentar algunos problemas que los tres tienen en común. El primer problema es el de las representaciones de entidades abstractas. Como se puede observar, este tipo de enfoques se centra en la explicación de la representación de propiedades perceptibles del entorno. Sin embargo, una buena parte de nuestros pensamientos es acerca de cosas abstractas como la democracia, la justicia, o, sin irse muy lejos, los números. El segundo problema es el de las entidades no existentes. En muchas ocasiones, cuando se habla de lo especiales que son las propiedades representacionales se dice que son las únicas propiedades en cuya individuación pueden entrar entidades inexistentes: hemos dicho que las propiedades representacionales son acerca de cosas; lo más curioso, para muchos, es que pueden ser acerca de cosas que no existen (unicornios, poderes telepáticos, los Reyes Magos…).

[10] Para una excelente discusión sobre el programa de Millikan, véase ARTIGA (2013).

Explicar cómo las representaciones mentales pueden tener este tipo de contenidos no es tarea fácil, y no parece que ninguno de los tres programas expuestos tenga una explicación sencilla a mano, dado el lugar donde han puesto el foco.

Estos dos problemas son problemas "clásicos" que perviven desde que se empezó a hablar de la representacionalidad o *intencionalidad*. Son problemas que cabe calificar de "grandes problemas", y por ello quizás el filósofo naturalizador puede pedir paciencia: de momento, puede decir, vamos a explicar los casos más sencillos y después ya afrontaremos los complicados. La cuestión es que no puede adivinarse de momento cómo, con los mimbres que se proporcionan, se pueden construir explicaciones satisfactorias de los casos complicados. La estrategia general de comenzar con lo más básico siempre está en el filo de la navaja: por una parte, es una estrategia sensata, pero por la otra, mientras no incluya al menos una explicación programática que nos haga ver cómo los casos complicados son simplemente eso, casos complicados, en lugar de casos de naturaleza diferente, lo que se nos pide es, básicamente, fe. Cuando la fe decae, decae con ella la percepción de continuidad y se abre un hueco entre los casos explicables y sencillos y los otros. Cuando esto ocurre, la estrategia ha fallado. Es como querer explicar la capacidad de volar a partir de la de saltar: primero se salta, luego se salta un poco más, y en algún momento, como es bien sabido, ya se está volando.

El tercer problema de estos ejemplos del "proyecto naturalizador" es que, contrariamente a lo que parece y a lo que pretenden, sus credenciales naturalistas son cuestionables. Como se ha dicho, la teoría de la dependencia asimétrica de Fodor pretende hacer uso sólo de leyes causales. Sin embargo, muchos autores (ver, por ejemplo, Burge, 2010) encuentran poco naturalista hablar de unas leyes que dependen asimétricamente de otras y de la invocación a situaciones contrafácticas que esto supone: ¿qué ciencia natural hace uso de herramientas semejantes? No parece que realmente estemos utilizando las nociones y herramientas que nos proporcionan otras ciencias para explicar la capacidad representacional. En este respecto, las teorías de Dretske y Millikan parecen teorías realmente naturalizadoras. No obstante, ambas tienen el mismo problema. Es razonable pensar que algunas de las representaciones que manejamos tienen el contenido que tienen por razones evolutivas. Esto es ciertamente plausible en el caso de muchos contenidos perceptuales, pero también hay razones para pensar que algunos conceptos básicos, como el de objeto o el de agente (ver, p.e. Carey, 2009), o incluso algunos conceptos numéricos (Margolis y Laurence, 2008), pueden ser

innatos, fruto de la selección natural. Sin embargo, es obvio que la mayor parte de nuestros conceptos son aprendidos en el marco del desarrollo cognitivo individual. Teorías como la de Dretske y la de Millikan no especifican cuál es el mecanismo que hace que nuestras representaciones representen lo que representan: una representación puede tener el contenido que tiene bien sea porque así lo ha fijado la selección natural o bien porque la representación ha sido "reclutada" en un proceso de aprendizaje.

Originalmente, Drestke sostenía que sólo el aprendizaje es capaz de reclutar representaciones para indicar la presencia de ciertas cosas en el entorno, mientras que Millikan mantenía que sólo la selección natural puede generar funciones, y entre ellas las funciones representacionales. Con el tiempo, ambos han hecho sitio en sus teorías a la selección natural y al aprendizaje. La cuestión que surge entonces es: ¿no es problemático invocar el aprendizaje en una teoría naturalizadora? Aprender es un proceso cognitivo (mental): ¿cómo puede una teoría realmente naturalizadora incluir, entre sus mecanismos de reclutamiento, un mecanismo que es mental? A partir del momento en que el aprendizaje se invoca, pero no se explica en términos naturalistas, las teorías no pueden seguir siendo consideradas teorías naturalizadoras[11,12].

El problema de la sensación

Cualquiera diría que las sensaciones (sabores, olores, dolores, placeres) constituyen lo más primitivo y menos complicado del mundo mental. Posiblemente, cabría decir que su naturaleza no es estrictamente mental; más bien, parecen parcialmente corpóreas. Parecen, además, inmediatas y "brutas", en el sentido de que nos conectan directamente con el mundo, sin análisis ni filtros. Por así decirlo, nos dan el mundo. En sintonía con todo esto, las sensaciones son relativamente fáciles: las tienen muchos otros seres, incluso algunos bastante básicos. En definitiva, las sensaciones son, por así decirlo,

[11] Dretske aspiraba a explicar el aprendizaje en términos naturalistas al restringirlo al aprendizaje asociativo. Sin embargo, parece claro hoy en día que la capacidad explicativa del aprendizaje asociativo es escasa, y que la mayor parte de los procesos de aprendizaje implica el concurso de facultades cognitivas.

[12] BURGE (2010) ofrece una crítica interesante y novedosa a los programas naturalizadores, y es que borran la distinción entre sistemas representacionales y sistemas meramente sensoriales como pueden ser el de un paramecio o, sin ir tan lejos, el de una lombriz de tierra. Burge sostiene que la idea fundadora del programa naturalizador, esto es, que la noción de representación es misteriosa a menos que sea reducible, es un error: la noción de representación es tan, o tan poco, problemática como cualquier otra noción de las que usa la ciencia. Para un respuesta, véase VICENTE (2012)

poco cognitivas y poco "nuestras". Así que si algo puede ser explicado en términos naturales, el candidato son ellas.

Sorprendentemente, la mayor parte de los filósofos piensa que las sensaciones, también llamadas "propiedades cualitativas" o simplemente "*qualia*", es lo más misterioso que hay. La razón es que, junto con los anteriores, tienen algunos rasgos que realmente las hacen diferentes de todo lo demás. En primer lugar, son entidades esencialmente subjetivas: sólo las conoce quien las instancia. Además, mientras el conocimiento del mundo parece ser inferencial –unos fotones impactan contra tu retina y acabas creyendo que se acerca un hombre- el conocimiento de las sensaciones es directo, inmediato e infalible: mientras que puedo estar muy convencido de que se acerca un hombre, y sin embargo equivocarme, si siento un dolor, sé que lo siento y no hay lugar a la equivocación.

Estos dos rasgos de las sensaciones son los responsables de que éstas hayan acaparado el interés de los filósofos en los últimos años, hasta el punto de considerar el reto de explicarlos *el* problema de la consciencia o incluso *el* problema mente-cuerpo. De ellos se deriva, por ejemplo, el conocido como "el argumento del conocimiento" en favor del dualismo (ver Maxwell, 1968, Nagel, 1974, Jackson, 1986) que, muy brevemente, dice lo siguiente: imaginemos que lo sabemos absolutamente todo sobre el mundo físico/natural, es decir, hemos conseguido llevar las ciencias naturales hasta su frontera. Aún así, hay algunas cosas que no podemos saber. Por ejemplo, no podemos saber cómo experimentan el mundo los murciélagos y no podemos saber en qué consiste ver, en el sentido de sentir, el color rojo si nosotros mismos, por lo que fuera, no lo vemos. En definitiva, las sensaciones y las experiencias no son cognoscibles más que *en primera persona,* y en eso difieren radicalmente del resto de las entidades que pueblan el mundo. Como se puede ver, esta diferencia en la forma de conocer las sensaciones se debe a que éstas son, como se ha dicho, esencialmente subjetivas.

Otro rasgo peculiar de las sensaciones, derivado de los anteriores, es que diríamos que son pura apariencia. Las cosas del mundo se nos presentan de formas particulares: la esquina de una casa se me presenta de una forma cuando la veo desde enfrente, de otras cuando me sitúo en un lado o en otro o si la veo desde arriba, de otra forma si paso por ella la mano, etc. En estos casos, podemos separar la apariencia de la realidad, y pensamos que la realidad es la causa de las apariencias. También podemos hacerlo en los casos en los que hacemos una cierta abstracción de las apariencias particulares. Por

ejemplo, podemos decir que, aunque la forma en que se nos presenta un tigre varía de un encuentro momentáneo al siguiente, el tigre tiende a presentársenos como un objeto que se mueve autónomamente con una cierta cadencia, que su cuerpo tiene una cierta composición y que sus colores son el negro y el amarillo, dispuestos en rayas verticales no uniformes. Este "modo de presentación" del tigre, que abstrae de los modos particulares, sigue siendo una forma en que se nos presenta algo que existe independientemente de nosotros, es decir, sigue habiendo una diferencia entre apariencia y realidad. Sin embargo, en el caso de las sensaciones no parece que podamos hacer esa distinción: las sensaciones son *pura apariencia*. El dolor que siento no parece ser sino la manera en que lo siento. Tiene una causa –el hombre que se acercaba me ha golpeado con un martillo- pero esa causa no es el dolor ni forma parte de él. Esta peculiaridad de las sensaciones hace que no podamos considerarlas idénticas a ninguna propiedad física, esto es, que las tengamos que considerar *irreducibles*. El argumento, expuesto por primera vez por Saul Kripke (1972) es el siguiente:

Pensemos en un enunciado de identidad verdadero como 'Anakin Skywalker es Darth Vader'. Aparentemente, este enunciado es verdadero, pero sólo contingentemente. Es decir, nos parece que Anakin Skywalker podría no haber sido Darth Vader. Sin embargo, esto es un error: dado que Anakin Skywalker *es* Darth Vader, no hay forma en que Anakin Skywalker pudiera no haber sido Darth Vader. Lo que sí pudiera haber ocurrido es que alguien que fuera idéntico en su apariencia y buena parte de su historia a Anakin Skywalker no hubiera sido la misma persona que alguien que hubiera sido idéntico en su apariencia y buena parte de su historia a Darth Vader. Es decir, creemos que Anakin Skywalker podría no haber sido Darth Vader porque creemos que ser Anakin Skywalker consiste en, por ejemplo, ser un brillante Jedi apadrinado por Obi-Wan-Kenobi. Pero esto es lo que sabemos, o creemos saber, de Anakin Skywalker: es el modo en que pensamos sobre Anakin Skywalker. Ser Anakin Skywalker, claramente, no es eso (de hecho, Anakin Skywalker podría no haber llegado a ser un Jedi, y seguiría siendo Anakin Skywalker). De modo que una vez que distinguimos entre apariencia y realidad vemos que el enunciado 'Anakin Skywalker es Darth Vader', de ser verdadero, es *necesariamente* verdadero. Si realmente mantenemos al margen las dos formas diferentes que tenemos de pensar en la persona, parece que la cosa es clara: Anakin Skywalker (o sea, Darth Vader) no podría haber sido diferente de sí mismo.

Vayamos ahora al caso de las sensaciones y la razón por la que, supuestamente, no pueden ser idénticas a propiedades físicas. Durante buena parte del siglo XX se sostuvo una teoría mente-cuerpo conocida como "la teoría de la identidad de tipos". Lo que caracterizaba a esta teoría era la afirmación de que cualquier propiedad mental es contingentemente idéntica a una propiedad neurofisiológica. Así, se propuso que el dolor no era sino la activación de las fibras C –su correlato neurofisiológico-, aunque podría haber sido idéntico a algún otro proceso cerebral. Pues bien, el primer paso del argumento de Kripke en contra de esta teoría consiste en negar que una identidad como la propuesta pueda ser contingente: un enunciado como 'el dolor es la activación de las fibras C' es un enunciado del tipo '*a* es *b*', igual que 'Anakin Skywalker es Darth Vader'. Por lo tanto, si es verdadero, tiene que serlo necesariamente.

El segundo paso del argumento dualista de Kripke se centra en mostrar que el enunciado 'el dolor es la activación de las fibras C' no puede ser necesario. Su idea (muy cartesiana) es que podemos concebir el dolor y la activación de las fibras C separadamente. Esto es, nos parece que hay mundos posibles en los que a un individuo se le activan las fibras C sin que sienta dolor y mundos en los que un individuo siente dolor sin que ni siquiera tenga fibras C. En principio, esto no parece significativo: también parece que podemos concebir mundos en los que Anakin Skywalker y Darth Vader son personas diferentes. Sin embargo, sí es significativo, porque hay una diferencia entre ambos casos: si Anakin Skywalker es quien es, y Darth Vader es quien es, entonces no podemos realmente concebir que Anakin Skywalker sea diferente de Darth Vader. Sin embargo, si el dolor que siento es lo que es y la activación de las fibras C es lo que es, aún así puedo concebir que no sean iguales. La razón última de la asimetría entre un caso y otro es que el dolor no tiene más propiedades que las aparentes: su esencia es ser la sensación que es. De aquí se sigue que podamos imaginar el dolor –siendo lo que es- sin la compañía de las fibras C, mientras que no podemos imaginar a Anakin Skywalker –siendo lo que es- siendo diferente de Darth Vader (no teniendo la apariencia que tiene Darth Vader –siendo malo, y todo eso-, sí, pero esa no es la cuestión).

Así que, finalmente, dado que la identidad entre el dolor y las fibras C no es necesaria, hemos de concluir que no hay tal identidad.

Como se puede comprobar, el argumento no es sencillo –aunque creo que lo es (relativamente) para quien dispone de ciertas herramientas de filosofía del lenguaje que

he optado por no dar por supuestas-. A mi entender, es el argumento dualista de mayor calado, y no es fácil darse por satisfecho con las respuestas que cabe encontrar en los trabajos más recientes, la mayor parte de las cuales siguen la línea conocida como la "estrategia de los conceptos fenoménicos"[13]. La idea básica de esta estrategia es que las diferencias entre lo físico y los *qualia* que anteriormente hemos localizado en el terreno de las propiedades son en realidad diferencias entre sistemas de conceptos. En el mundo no vamos a encontrar más que propiedades físicas. Sin embargo, esas propiedades físicas pueden conceptualizarse de modos diferentes: como lo que son y, al menos en algunos casos, como lo que nos parecen. Un determinado estado de mi cerebro lo puedo conceptualizar como la activación de las fibras C –lo que es-, pero también como cómo se me presenta, esto es, como un dolor.

Esta estrategia general, aparentemente, está en disposición de responder tanto al argumento del conocimiento como al argumento kripkeano. Si el conocimiento total del mundo físico no basta para conocer el mundo experiencial es porque sólo teniendo ciertas experiencias se pueden tener ciertos conceptos. Si alguien es incapaz de sentir dolor, entonces sólo podrá "ver" el dolor como la activación de las fibras C, y no podrá verlo como la sensación de dolor: carecerá del concepto *recognitivo* de dolor, el que nos presenta a la activación de las fibras C como una determinada sensación desagradable. Pero eso será lo único que le faltará por conocer: un concepto, no una realidad. Es decir, la estrategia de los conceptos fenoménicos incide en la diferencia entre apariencia y realidad, sosteniendo que también en el caso de las sensaciones existe esa diferencia. De este modo, impide también que el argumento de Kripke llegue a su conclusión: el argumento requiere que concibamos las sensaciones como propiedades cuya esencia es ser ellas mismas. La estrategia de los conceptos fenoménicos intenta resistirse justamente en ese punto: las sensaciones no son propiedades, sino formas en que se nos presentan otras propiedades.

El problema de esta estrategia general es que usualmente cuando establecemos la distinción entre las propiedades y los modos en los que éstas se nos presentan tenemos que recurrir a propiedades fenoménicas: diferenciamos entre lo que es el calor –el

[13] Muchos otros filósofos fisicistas optan por otra línea (que también puede convertirse en una variante de la estrategia de los conceptos fenoménicos): la de considerar que los *qualia* son propiedades representacionales. Así, la sensación de rojez representa la rojez del mundo; el dolor representa el daño en alguna parte del cuerpo, etc. Existe un debate sobre si los *qualia* pueden realmente reducirse a representaciones. BLOCK (2003) sostiene que lo interesante de los *qualia* no es lo que éstos representan –si es que tienen contenido representacional- sino la manera en que lo representan. Lo interesante de la sensación que llamamos "orgasmo" no es, en absoluto, su contenido representacional –la eyaculación, pongamos por caso

movimiento de las moléculas- de cómo éste se nos presenta diciendo que una cosa es el calor y otra la sensación que el calor produce en nosotros. En el caso del dolor, parece que tendríamos que decir que una cosa es el dolor –la activación de las fibras C- y otra cómo se nos presenta: como sensación de dolor. Pero entonces parece que, de nuevo, tenemos dos cosas y no dos conceptos: la propiedad física y la sensación. Por esta razón, los partidarios de la estrategia de los conceptos fenoménicos se ven obligados a sostener que los conceptos fenoménicos son un tipo especial de conceptos cuya particularidad reside en que no presentan a las propiedades a las que remiten a través de otras propiedades. Pero esto puede ser fácil de enunciar pero muy difícil de explicar[14].

El problema de la causación mental

Si echamos la vista atrás, vemos que, como anunciamos, "el" problema mente-cuerpo se resuelve en unos cuantos problemas diferentes y específicos: tenemos el problema del cuerpo, el problema de la representación y el problema de la sensación. El problema de la representación y el de la sensación se considerarían solucionados si fuéramos capaces de mostrar que tanto las propiedades representacionales como las cualitativas no son sino propiedades de las que manejan otras ciencias. Si no somos capaces de hacerlo, dado que tampoco hemos sido capaces de resolver satisfactoriamente el problema del cuerpo, siempre nos va quedar un cierto sentimiento de ambivalencia: por un lado, parece que, efectivamente, las propiedades mentales son algo especial, incluso muy especial; por otro, no podemos explicar cuán especiales son, puesto que, al no disponer de un corte claro entre lo físico y lo demás, no podemos ni siquiera decir razonablemente que las propiedades mentales parecen ser no físicas[15].

El problema de la causación mental viene a complicar todo esto bastante más. Básicamente, el problema nos va a decir que si las propiedades mentales son, efectivamente, irreducibles, entonces su eficacia causal es dudosa. Es decir, si las propiedades mentales no son propiedades físicas, la idea de que nos movemos en el mundo en virtud de lo que pensamos, sentimos y queremos se vuelve problemática. Esto es, sin duda, grave: no podemos dejar de creer que somos agentes que se desenvuelven

[14] Para una revisión crítica de la estrategia, véase CHALMERS (2006). Para una defensa, DÍAZ LEÓN (2010).
[15] Si sólo hubiera una ciencia, o si estuviera claro que todas las ciencias son, en algún sentido, sólo una, la irreductibilidad de las propiedades mentales las haría únicas. Pero no está claro que las ciencias exhiban la unidad requerida.

en el mundo a partir de lo que sienten, planean, razonan, creen, desean y quieren. Es más, seguramente nos tenemos a nosotros mismos como "agentes especiales", agentes diferentes de los demás: tenemos información privilegiada acerca de nuestros propios estados mentales, cierta capacidad de decisión o control sobre nuestros propios procesos mentales y sus resultados, e incluso nos parece que somos "motores inmóviles", causas incausadas. La experiencia de agencialidad, que incluye todo esto, es insobornable e insoslayable. ¿Cómo puede ser que, si las propiedades mentales son irreducibles, toda esta rica experiencia resulte ser ilusoria? El problema es severo, y va más allá de la notoria incompatibilidad entre la "imagen científica" del mundo y su "imagen manifiesta", distinción que se aplica a, por ejemplo, la forma en que la física de partículas ve una mesa y la forma en que la vemos habitualmente. En el caso de la agencia, el choque que se produce entre lo que creemos ser, es más, entre lo que sabemos de nosotros mismos, y lo que el problema de la causación mental –inducido por la "imagen científica"- nos hace concluir, es catastrófico.

El problema de la libertad surge cuando intentamos hacer sitio a nuestra experiencia de libre determinación en un mundo que es o determinista o indeterminista. El problema de la causación mental es anterior y más básico. En este caso, parece que no podemos hacer sitio ni siquiera a la idea de que nuestros estados mentales –estén éstos bajo nuestro control o no- tienen potencia causal. Se dice que la física contemporánea ha mostrado que todo efecto físico (i.e., evento físico causado) tiene una causa suficiente que es igualmente física. No se trata de una ley que haya descubierto y a día de hoy figure en los libros de texto, pero parece que los físicos realmente se sorprenderían si se vieran obligados a recurrir a alguna otra ciencia para explicar un suceso de lo que consideran su ámbito. En este caso, el sentido de 'físico' que se emplea es restringido, y aunque sigue siendo difícil de definir, podemos tomar este principio de cierre causal como un principio de la física contemporánea que se aplica a esta misma física (ver Vicente, 2006). Bien, el caso es que parece que algunas de las cosas que ocurren en el mundo físico, algunos de sus cambios, tienen causas mentales: mi mano se mueve y alcanza un papel que, acto seguido, comienza a bajar hacia la calle mecido por el viento. ¿Qué ha causado este proceso? Razonablemente, mi intención de tirar el papel a la calle. ¿Cómo puede ser, si, como se ha dicho, todo efecto físico tiene una causa física suficiente?

El problema de la causación mental suele presentarse como la colisión entre varias proposiciones que creemos verdaderas:

(1) El mundo físico está causalmente cerrado: todo efecto físico (i.e., todo evento físico que tiene una causa) tiene una causa física suficiente y completa;

(2) Los estados mentales, sean representacionales o cualitativos, causan cambios en el mundo físico;

(3) No puede haber dos causas independientes y suficientes para un mismo suceso, salvo en casos de sobredeterminación causal;

(4) No hay sobredeterminación causal mente-física;

(5) Los estados mentales no son idénticos a estados físicos.

Como se puede ver, el problema consiste en conciliar (1) y (2). El papel de (3) y (4) es el de cerrar las puertas a una posible conciliación: la de que causas mentales y físicas sean ambas efectivas. (3) nos dice que usualmente no tenemos dos causas independientes para un mismo suceso, aunque hace una salvedad: los casos de sobredeterminación. Durante mucho tiempo se ha pensado que la extinción de los dinosaurios se debió al impacto de un meteorito. Ahora se dice que pudieron ser dos, pero que quizás cualquiera de ellos habría bastado. Si fuera así, la extinción de los dinosaurios estaría causalmente sobredeterminada. Pero los casos de sobredeterminación causal son raros. Que los efectos físicos que provocamos estuvieran sobredeterminados por causas mentales y físicas significaría que, o bien hay constantes coincidencias en la producción de cada uno de nuestros movimientos, o bien hay algún tipo de armonía –puede ser una ley- que conecta lo físico y lo mental, haciendo que causas mentales y físicas se *co-instancien*. Pero cualquiera de estas dos opciones tiene un aire demasiado extraño. De modo que parece razonable creer que (4) es verdadera.

La proposición (5), por su parte, viene a decirnos que el problema existe sólo si las propiedades mentales no son reducibles a propiedades físicas. Ciertamente, el problema de la causación mental se desvanece si resulta que las propiedades mentales no son sino propiedades físicas. La cuestión es que no parecen serlo. Incluso muchos filósofos naturalizadores afirmarían que las propiedades mentales no son idénticas a las propiedades físicas que se mencionan en el problema de la causación mental. Pensemos en el dolor: el dolor, se dice, no puede ser idéntico a la activación de las fibras C ya que animales que no tienen fibras C parecen sentir dolor. El dolor, como cualquier otra propiedad mental, parece que es *múltiplemente realizable*. Esto no quiere decir que el dolor, o cualquier otra propiedad mental, no sea una propiedad física en un sentido laxo, pero sí quiere decir que, por regla general, las propiedades mentales no son idénticas a las propiedades físicas con las que compiten en el problema de la causación mental.

De modo que parece que nos vemos abocados bien a negar que el mundo físico es causalmente completo, bien a considerar que la causalidad mental es ilusoria. Nótese que para enfrentarnos a este dilema no necesitamos saber qué es lo físico: nos basta con saber que hay otro tipo de causas, razonablemente llamadas 'físicas', que causan lo mismo que supuestamente causan los estados mentales. Nótese también que este tipo de problema no es exclusivo de la causación mental: la causación biológica, por ejemplo, es igual de problemática, dado que para cualquier causa biológica en que pensemos, parece haber una causa de un nivel más básico que causa los mismos efectos que queremos explicar apelando a la explicación biológica.

¿Qué cabe hacer ante este dilema? Ciertamente, lo más recomendable sería poder eludirlo. Sin embargo, no parece que se pueda hacer. Muchos autores sostienen que sólo se percibe una colisión entre la causalidad mental y la física desde una determinada teoría de la causalidad. Si uno adopta una teoría contrafactualista de la causalidad (ver capítulo), por ejemplo, no se produce la colisión. En efecto, si pensamos que '*a* causa *b*' quiere decir que si *a* no se hubiera producido, *b* tampoco lo habría hecho, cabe afirmar que los estados mentales *tienen* potencia causal: si no hubiera tenido un dolor de cabeza, no habría ido al baño a por una aspirina. Dado que el contrafáctico es cierto, la causalidad mental queda vindicada. Es decir, una teoría contrafactualista de la causalidad nos dice que, efectivamente, tenemos dos causas para un mismo suceso, puesto que tampoco habría ido al baño de no estar en el estado cerebral que estaba.

Sin embargo, no creo que este resultado nos deje completamente tranquilos: en primer lugar, porque aunque la teoría contrafactualista nos reafirme en la idea de que hay dos tipos de causas en juego, seguimos pensando que es raro que las haya. En segundo lugar, porque da la impresión de que lo que consigue la teoría contrafactualista no es lo que buscamos. Lo que buscamos es hacer buena nuestra experiencia de agencialidad o al menos una parte mínima de ésta: dejamos de lado cosas como la libertad o el control, y nos centramos en la capacidad que tienen nuestros estados mentales de producir cambios en el mundo. Pues bien, diríamos que hay una conexión íntima entre lo que pensamos y sentimos y lo que hacemos, una conexión que la traducción en términos contrafácticos no recoge: simplemente, lo que experimentamos no es una dependencia contrafáctica entre nuestros estados mentales y ciertos efectos. Kim (2007) sostiene que lo único que nos dejaría satisfechos sería poder mostrar que nuestros estados mentales son capaces provocar cambios en el mundo mediante la

transferencia de energía. Quizás esto es excesivo, pero es una idea de causalidad que se acerca más a la implícita en nuestra experiencia como agentes. Sin embargo, si realmente la idea de que nuestros estados mentales causan cambios en el mundo se puede poner en términos de transferencia de energía, el problema de la causación mental es todo lo severo que puede ser. Si el problema de la causación mental consiste en cómo explicar que algo que no es físico es sin embargo capaz de transferir energía al mundo físico, entonces podemos colocarlo a la par del problema de cómo explicar que somos causas incausadas. Y si realmente nuestra experiencia como agentes nos presenta a nosotros mismos como causas incausadas capaces de transferir energía, quizás sea más acertado investigar de dónde procede esa ilusión que intentar buscarle asiento en el mundo[16].

Conclusión

El llamado 'problema mente-cuerpo' se resuelve en al menos cuatro problemas: el problema del cuerpo (o quizás mejor, de lo físico), el problema de la representación, el problema de la sensación y el problema de la causación mental. El problema del cuerpo, o de lo físico, surge en el momento en el que intentamos hacer precisa la distinción que establecemos entre lo mental y lo físico, es decir, en el momento en que intentamos explicar en qué consiste exactamente el problema mente-cuerpo. Se resume en que, aparentemente, no contamos con una noción precisa de lo físico que contraponer a "lo mental". Este problema no sería importante si pudiéramos mostrar que los rasgos típicamente mentales, esto es, los rasgos esenciales de las propiedades mentales, son explicables en términos de nociones de las conocidas como "ciencias naturales". Si fuéramos capaces de hacer tal cosa, que la noción de lo físico sea más o menos esquiva sería irrelevante, pues no habría nada de misterioso en las propiedades mentales, con lo que no habría razón para precisar la contraposición entre lo mental y lo físico.

Sin embargo, el problema de la representación y el de la sensación parecen mostrar que las propiedades mentales no son reducibles a propiedades postuladas por

[16] El psicólogo WEGNER (2002) es el principal exponente de la idea de que nuestra experiencia como agentes es ilusoria. Según él, la psicología está en disposición de mostrar este carácter ilusorio de lo que él llama 'la voluntad consciente', así como de explicar cuál es su fundamento psicológico y su función. Esta provocativa tesis se inscribe dentro de una tendencia en psicología (no freudiana) que sostiene que el individuo sabe muy poco de lo que realmente acontece en su vida mental y, en general, de sí mismo (ver WILSON, 2002). CARRUTHERS (2007) es, entre los filósofos, quien con más simpatía ve las tesis de Wegner.

las ciencias naturales. Esto, por sí mismo, no quiere decir que las propiedades mentales sean, efectivamente, misteriosas. El problema de la representación, por ejemplo, simplemente nos devuelve al punto de partida: nos parece que hay dos grandes tipos de propiedades en el mundo, pero nos preguntamos si las cosas son como nos parecen o no, y si nuestra distinción intuitiva está fundamentada. El problema de la sensación, sin embargo, sí parece mostrarnos que hay algo realmente diferente y en cierto modo misterioso en el reino de lo mental.

No obstante, el problema más acuciante es el problema de la causación mental, pues consiste en que, aparentemente, nuestra rica experiencia agente es irreconciliable con lo que nos dicen las ciencias naturales. Es decir, el problema de la sensación nos revela que un aspecto muy característico y querido de nuestra vida mental –la idea de vivir sin sensaciones es aterradora- tiene un encaje difícil en el mundo natural tal y como lo describen las ciencias. Pero incluso si acabáramos convencidos de que el encaje es imposible, lo que nos tocaría, a lo sumo, es convertirnos en dualistas. El alcance del problema de la causación mental, sin embargo, es mayor: si, efectivamente, deseos, creencias, temores e intenciones no son, en cuanto tales, causalmente eficaces, prácticamente cuanto pensamos de nosotros mismos como agentes resulta ser falso y nuestra experiencia una ilusión.

Desde que, a mediados de los setenta del XX, re-emergiera la filosofía de la mente, especialmente dentro de la filosofía analítica, los progresos en la comprensión del "problema mente-cuerpo" han sido evidentes. No obstante, y como ocurre siempre, o casi siempre, en filosofía, los progresos no han sido espectaculares. Sin embargo, el trabajo que se hace hoy en día en cada uno de los problemas aquí reseñados es, en muchos casos, excelente.

Bibliografía

ADAMS, F. Y AIZAWA, K. (2010): "Causal Theories of Mental Content", *Stanford Encyclopedia of Philosophy*: http://plato.stanford.edu/entries/content-causal/

ARTIGA, M. (2013): *A Naturalistic Theory of Intentional Content*. Tesis. Universitat de Girona.

BAKER, L. R. (2013) *Naturalism and the First-Person Perspective*, Oxford University Press, New York.

BLOCK, N. (2003): "Mental Paint", en M. Hahn y B. Ramberg (eds.), *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, MIT Press, Cambridge, MA, pp. 165-201.

BLOOM, P. (2004): *Descartes' Baby: How The Science Of Child Development Explains What Makes Us Human*, Basic Books.

BURGE, T. (2010): *Origins of Objectivity*, Oxford University Press, New York.

CALVO GARZÓN, F. (2008): "Towards a General Theory of Antirepresentationalism", *British Journal for the Philosophy of Science*, 59, pp. 259-292.

CAREY, S. (2009): *The Origin of Concepts*, Oxford University Press, New York.

CARRUTHERS, P. (2007): "The Illusion of Conscious Will", *Synthese*, 159, pp. 197–213.

CHALMERS, D. (2006): "Phenomenal Concepts and the Explanatory Gap", en T. Alter y S. Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press, Oxford, pp. 167-195.

CHEMERO, A. (2009): *Radical Embodied Cognitive Science*, MIT Press, Cambridge, MA.

CHOMSKY, N. (1995): "Language and Nature", *Mind*, 104, pp. 1-61.

DÍAZ LEÓN, E. (2010): "Can Phenomenal Concepts Explain the Epistemic Gap?", *Mind* 119, pp. 933-51.

DRETSKE, F. (1988): *Explaining Behavior: Reasons in a World of Causes*, MIT Press, Cambridge, MA.

FODOR, J. (1987): *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT/Bradford, Cambridge, MA.

HEMPEL, C. (1980): "Comments on Goodman's Ways of Worldmaking", *Synthese* 45, pp. 193-200.

JACKSON, F. (1986): "What Mary Didn't Know", *Journal of Philosophy,* 83, pp. 291-295. (Trad. cast. en M. Ezcurdia y O. Hansberg (eds.), *La Naturaleza de la Experiencia. Vol. 1 Sensaciones*, Instituto de Investigaciones Filosoficas, UNAM, 2003).

JORBA, M. (2013): *Cognitive Phenomenology: A Non-Reductive Account*. Tesis. Universitat de Barcelona.

JORBA, M. y VICENTE, A. (2014): "Cognitive Phenomenology, Access to Contents, and Inner Speech", *Journal of Consciousness Studies*, 21, pp. 74–99.

KIM, J. (2007): "Causation and Mental Causation", en B. McLaughlin y J. Cohen (eds.), *Contemporary Debates in Philosophy of Mind*, Blackwell, Oxford, pp. 227-243.

KRIPKE, S. (1972): "Naming and Necessity", en D. Davidson y G. Harman (eds.), *Semantics of Natural Language*, Dordrecht: Reidel, pp. 253-355. (Trad. cast.: *El Nombrar y la necesidad*, UNAM, México).

LAUDAN, L. (1981): "A Confutation of Convergent Realism", *Philosophy of Science*, 48, pp. 19-49.

LOWE, E. J. (2000): *An Introduction to the Philosophy of Mind*, Cambridge University Press, Cambridge.

NAGEL, T. (1974): "What Is It Like to be a Bat?", *The Philosophical Review*, 83, pp. 435-450. (Trad. cast. en M. Ezcurdia y O. Hansberg (eds.), *La Naturaleza de la Experiencia. Vol. 1 Sensaciones*, Instituto de Investigaciones Filosoficas, UNAM, 2003).

MARGOLIS, E. y LAURENCE, S. (2008): "How to Learn the Natural Numbers: Inductive Inference and the Acquisition of Number Concepts", *Cognition*, 106, pp. 924–939.

MAXWELL, N. (1968): "Understanding sensations", *Australasian Journal of Philosophy*, 46, pp. 127–146.

MILLIKAN, R. (1984): *Language, Thought and Other Biological Categories*, MIT Press, Cambridge, MA.

MONTERO, B. y PAPINEAU, D. (2005): "A Defence of the *Via Negativa* Argument for Physicalism", *Analysis*, 65, pp. 233-237.

MONTERO, B. (1999): "The Body Problem", *Noûs*, 33, pp. 183–200.

PINEDA, D. (2006): "A Mereological Characterization of Physicalism", *International Studies in the Philosophy of Science*, 20, pp. 243-266.

RUPERT, R. (2008): "Causal Theories of Mental Content," *Philosophy Compass*, 3, pp. 353–80.

VAN GELDER, T. (1995): "What Might Cognition Be, If not Computation?", *The Journal of Philosophy*, 91, pp. 345–381.

VICENTE, A. (2006): "On the Causal Completeness of Physics", *International Studies in the Philosophy of Science*, 20, pp. 149-171.

VICENTE, A. (2011): "Current Physics and "the physical"", *British Journal for the Philosophy of Science*, 62, pp. 393-416.

VICENTE, A. (2012): "Burge on Representation and Teleological Function", *Thought: A Journal of Philosophy*, 1, pp. 125-133.

WEGNER, D. (2002): *The Illusion of Conscious Will*, Cambridge University Press, Cambridge, MA.

WILSON, J. (2006): “On Characterizing the Physical”, *Philosophical Studies*, 131, pp. 61-99.

WILSON, T. (2002): *Strangers to Ourselves*. The Belknap Press of Harvard University Press.