# UN LUGAR PARA LA MENTE EN EL MUNDO:

# EL FIN DE UNA INTUICIÓN

**Sergio Cermeño Ainsa**

**Departamento de Filosofía**

**UNED**

**RESUMEN**

*En este trabajo me ocuparé de un aspecto central del problema de la relación entre la mente y el cuerpo, a saber, el problema de los qualia de la conciencia. Los filósofos generalmente comparten esa debilidad humana por dar explicaciones sobre aspectos del mundo que aparentan ser incomprensibles. Ese afán por saciar su curiosidad les ha llevado a cuestionarse sobre aspectos de la mente que desafían todas nuestras intuiciones. Muchas de esas explicaciones tienden actualmente a naturalizar esos aspectos aparentemente incomprensibles de nuestra mente, y en este proceso de naturalización aparece con especial fuerza una barrera que toda teoría naturalista debe superar, la de los estados cualitativos de la mente, ¿son esas propiedades cualitativas de nuestras sensaciones un obstáculo insuperable para las aspiraciones reduccionistas del funcionalismo computacional y la neurociencia? Argumentaré que no. Comenzaré con una breve introducción de donde nos encontramos en la actualidad en este aspecto clave de la mente humana, y posteriormente criticaré ciertos argumentos que se han venido elaborando con el fin de destruir las tesis materialistas y funcionalistas. Dividiré el trabajo en las siguientes seis secciones: (1) el argumento basado en la perspectiva (Nagel, 1974), (2) el argumento del conocimiento o argumento epistemológico (Jackson, 1982), (3) el argumento de los qualia invertidos y de la tierra invertida (Block, 1990), (4) el argumento de los qualia ausentes o zombis filosóficos (Chalmers, 1999), (5) el argumento de la brecha explicativa (Levine, 1983) y por último (6) el argumento de la habitación china (Searle, 1980). Mi conclusión es que ninguno de los argumentos consigue sus objetivos, que no prueban lo que pretenden probar, que los defensores de la noción tradicional de qualia, aquellos que defienden que el carácter intrínseco de nuestra experiencia consciente no es abordable científicamente, caen en una especie de ingenuidad con respecto a sus propios estados subjetivos, y que conforme vayan apareciendo un mayor número de investigaciones empíricas sobre nuestras experiencias subjetivas conscientes irán desapareciendo también todos esos misteriosos lugares donde los filósofos han tratado de encerrar los conceptos de qualia y conciencia.*

Palabras Clave: Qualia, Conciencia, Experimentos mentales, Fisicalismo, Funcionalismo.

**Introducción**

Los dolores, pensamientos, imágenes, sueños, alucinaciones, creencias, actitudes, deseos e intenciones, cuentan como fenómenos mentales, mientras que las contracciones del estómago, los latidos del corazón o los procesos nerviosos, a los que se puede atribuir una localización concreta dentro del cuerpo, figuran como no-mentales. Nuestra clasificación no plantea ninguna duda y sugiere que tenemos una intuición clara de lo que es la mentalidad, algo que tiene que ver con la no-espacialidad, quizás con la idea de que aunque desapareciera el cuerpo, los estados mentales podrían seguir perdurando de alguna manera. En un principio y de manera intuitiva, no parece que tengamos ningún problema en distinguir entre la mente y el cuerpo. No son pocas las ocasiones en las que empleamos un lenguaje mental para referirnos a determinados sucesos: a veces uno dice sentirse mentalmente agotado, cuando olvidamos las llaves de casa en el coche culpamos a nuestra mente de tal olvido, hablamos de estar mentalmente despiertos, de poner toda nuestra mente en alguna tarea importante, de tener claridad de mente, una mente brillante o incluso algunos creen tener poderes mentales. Pues bien, estas intuiciones tan comunes dentro de nuestra psicología popular son las que mantienen en vida al dualismo cartesiano, la tesis que defiende que la esencia de nuestra vida mental reside en algo *no físico.* Según el *dualismo de substancias* de Descartes la realidad se divide en dos tipos básicos de sustancias, una material, que se caracteriza por ocupar una determinada posición en el espacio, la cosa extensa, y otra inmaterial, que no tiene extensión ni posición espacial y cuya actividad principal es el pensamiento. La consecuencia más problemática de esto es que en esencia para que una persona sea lo que es no necesita más que su mente. Cuando Descartes se pregunta *"¿qué tipo de cosa soy?",* solo puede estar seguro de que es "*una cosa que piensa*", esto es, mente, entendimiento, razón[1]. Las reflexiones de Descartes dieron para mucho, la idea de que cuando uno observa algo solo puede estar seguro de que es algo que ocurre en su propia mente nos lleva a la desesperante conclusión de que nunca nadie ha visto nada material, de que si suprimiéramos lo mental del mundo no habría mundo y de ahí a los angustiosos problemas como el de la existencia del mundo externo o el de las otras mentes. Las dificultades con que se enfrenta esta posición saltan a la vista, si el dualismo mente-cuerpo es verdadero, ¿qué relación hay entre esas dos sustancias? ¿cómo es posible que la substancia mental tenga algún tipo de relación causal con la substancia física? No hubo grandes movimientos sobre este asunto durante unos cuantos siglos, parecía como si las intuiciones cartesianas hubieran dejado a los filósofos en una especie de sueño hipnótico, se aceptaba la visión tradicional porque era la que mejor casaba con nuestras intuiciones y con el panorama político-religioso de la época, pero las posiciones dualistas acababan por crear más problemas de los que solucionaban, así que a mediados del

[1] Cabe advertir del carácter lógico de dicha tesis ontológica. ¿Es posible que no tengamos un cuerpo? Quién sabe. Tal vez sí o tal vez no. ¿Es contradictorio suponer que no lo tenemos? No, no lo es. Dicho de otro modo: es posible un argumento válido que niegue la existencia del cuerpo, pero negar la existencia de la mente es autocontradictorio. Pienso que es aquí donde radica el potencial de la argumentación cartesiana.

siglo pasado el *conductismo lógico,* liderado por Gilbert Ryle (1949)[2], se propuso sacar a los filósofos del letargo cartesiano. Ryle sostiene que la filosofía de la mente descansa sobre un error colosal, un error fundamental, un *error categorial* que ha hecho que todos los esfuerzos por responder a las cuestiones sobre la mente conduzcan de manera ineluctable al sinsentido. Se trata de un uso y abuso poco juicioso del lenguaje ordinario que usamos para referimos acerca de la mentalidad. El dualismo es una ilusión que nos lleva finalmente a lo que Ryle bautizó como el "*dogma del fantasma en la máquina*", como si la mente fuera una entidad no física que desde algún lugar en el interior del cuerpo realizara diversas operaciones sobre este cuerpo y sobre el mundo en general. La idea de Ryle es que el dualismo es completamente falso por principio[3]. Al hablar de lo mental no nos referimos a nada en especial ni a ningún lugar en particular sino a determinadas habilidades, conductas, disposiciones e inclinaciones a hacer ciertas cosas y evitar otras, por lo que la teoría dualista y sus contradicciones metafísicas se disolverían evitando extraer conclusiones teóricas mediante el uso de palabras carentes de un sentido concreto. Sin embargo, el conductismo no consiguió una explicación satisfactoria de cómo elaborar una reducción adecuada de los términos mentales a términos conductuales (Hampshire, 1950), ni consiguió salir de la circularidad que supone analizar la creencia en términos de conducta, ya que para esto hay que referirse al deseo, y para analizar el deseo, hay que referirse a la creencia (Chisholm, 1957), además, deja de lado las relaciones causales entre los estados mentales y la conducta (Lewis, 1966), por ejemplo, al identificar el dolor con la disposición a la conducta de dolor, olvida el hecho de que el dolor *causa* la conducta de dolor, por no hablar de sus respuestas a fenómenos como la conciencia o los qualia, simplemente son negados, se dejan de lado los *auténticos* fenómenos mentales, no queda nada para las experiencias subjetivas, solo los patrones de conducta objetivamente observables entran dentro de lo que puede ser estudiado. Pero no es difícil imaginar un actor que pueda imitar perfectamente la conducta de alguien con dolor aunque no sienta dolor alguno (Putnam, 1963). Finalmente no hubo lugar para una posición tan estrecha y reducida de nuestros estados mentales, así que a finales de los años 50 los filósofos U. T. Place (1956), J. J. C. Smart (1958), y Herbert Feigl (1967) propusieron un acercamiento al estatuto de la mente que se ha llamado de varias maneras: teoría de la identidad, materialismo de la identidad mente-cuerpo, materialismo del estado central, fisicalismo de tipos y teoría del estado cerebral. Estos postulan que los estados mentales son idénticos a los estados del cerebro y del sistema nervioso central; las mentes son cerebros, y los contenidos de las mentes, dolores, pensamientos, sensaciones, etc. *son* (idénticos a) eventos, procesos y estados del cerebro. A diferencia del conductismo lógico, la teoría de la identidad no niega la existencia de estados mentales, ni los reduce a conductas y disposiciones, simplemente los *identifica* con estados

[2] Entre los conductistas lógicos destacan además de Ryle (1949), Carnap (1932/1933) y Hempel (1949).

[3] Estas son las cuatro objeciones principales al dualismo de substancias cartesiano, en Lycan (2009: 556): *"(1) El problema de la interacción, por supuesto. (2) Los egos Cartesianos son problemáticos, extraños y oscuros, y no son necesarios para la explicación de ningún hecho público conocido. (3) Incluso aunque sea conceptualmente inteligible, la interacción cartesiana viola las leyes de la física, particularmente la conservación de la materia-energía [Cornman 1978: 274]. (4) La teoría evolutiva es una molestia para el dualismo, ya que no tenemos idea de cómo la selección natural pudo haber producido egos Cartesianos; una substancia inmaterial no podría ser adaptativa"*.

cerebrales. Se propone una reducción inter-teórica[4] en la que un nuevo marco más amplio y con mayor capacidad explicativa basado en las ciencias del cerebro substituya al viejo e insuficiente marco mentalista. Los conceptos de sentido común como el color de una manzana, la fragancia de una flor o la intensidad del calor, deben ser suplantados por sus "*conceptos sucesores*" proporcionados por una teoría física, como la frecuencia de las ondas electromagnéticas, la estructura química de los componentes aromáticos o el movimiento molecular. Así, los predicados fenoménicos usados en la descripción de post-imágenes, sensaciones, sentimientos o emociones deben ser reemplazados por predicados neurofisiológicos (Feigl, 1967: 141-42). ¿Y si resulta que lo mental son procesos cerebrales de la misma manera que el calor es movimiento molecular o la luz ondas electromagnéticas? Esta visión soluciona muchos de los problemas con los que se encuentra el conductismo sin retroceder al dualismo, pero como es de esperar, otros problemas surgen[5]. En primer lugar, la teoría de la identidad desafía la ley de Leibniz, según la cual dos cosas idénticas deben poseer también propiedades idénticas., no resultaría complicado encontrar una propiedad de los estados cerebrales que no posean los mentales y viceversa (difícilmente encontraré mi creencia de que la tierra gira alrededor del sol en algún pliegue del lóbulo temporal de mi hemisferio cerebral izquierdo). Además el fisicalista de tipos se las tiene que ver con argumentos como la realización múltiple[6] o los argumentos modales de Kripke[7]. No tardó en surgir una nueva teoría que sin negar algunos aspectos esenciales del fisicalismo de la identidad, y siguiendo los pasos del conductismo lógico (no en vano es su heredera (Churchland, 1999: 65)), solucionó muchos de estos problemas, trató de ir más allá de estos observando los problemas surgidos desde lugares bien diferentes: se trata *del funcionalismo*[8]*.* Es la teoría según la cual las clases mentales no son ni físicas, ni biológicas, sino *funcionales*. La concepción funcionalista de la mente parece encajar a la perfección en la nueva ciencia de lo mental, pues postula un dominio autónomo y distintivo para las propiedades mentales a través de una investigación científica con independencia de las propiedades físico-biológicas, una idea que promete tanto legitimidad como autonomía para la psicología en cuanto ciencia. El funcionalismo consigue deshacerse de las limitaciones restrictivas del reduccionismo fisicalista y el conductismo lógico sin regresar a los ya desfasados dualismos, sin embargo, aunque parece tener una solución válida para las propiedades semánticas de las representaciones mentales con contenido intencional, esto es, las actitudes proposicionales, es incapaz de dar una explicación completa del contenido cualitativo de nuestras experiencias conscientes, no consigue

[4] Para una explicación más profunda sobre los problemas de la reducción inter-teórica ver, Nagel, E. (1979), principalmente el capítulo 11.

[5] La teoría de la identidad a pesar de tener una vida muy corta sentó las bases de lo que iba a venir después. De una forma u otra casi todo el mundo se quedó en el fisicalismo, sea este reduccionista o no, pero muy pocos osaron volver al dualismo de sustancias.

[6] El ya clásico argumento de la realizabilidad múltiple (Putnam, 1960, 1967a) dice que es posible considerar que los tipos de estados mentales puedan ser realizados por diferentes sistemas físicos, por lo que el fisicalismo deviene falso, o cuando menos insuficiente. En algún sentido el argumento de la realizabilidad múltiple contra el materialismo reduccionista resultó ser correcto, sin embargo, la versión materialista más amplia que aquí pretendo defender no solo sale indemne de sus acusaciones sino que queda fortalecida. Veremos que el funcionalismo es plenamente compatible con el fisicalismo reduccionista, y además que de tal compatibilidad surge una fuerte combinación de ideas que podrían acabar con todos esos misterios que giran en torno a lo mental.

[7] Los argumentos modales de Kripke surgirán varias veces a lo largo de este trabajo.

[8] El funcionalismo fue sugerido por Hilary Putnam en: "Minds and Machines" **(1960),** "The Mental Life of some Machines" **(1967a)** y "The Nature of Mental States" (1**967b**).

dar cuenta de los qualia de la conciencia. Este sigue constituyendo para el funcionalismo un auténtico problema. En este punto de la discusión, muchos, en oposición al funcionalismo y su carácter fisicista y reduccionista, propusieron una vuelta al dualismo, esta vez un *dualismo de propiedades* (Chalmers, 1999). Esta nueva versión del dualismo *no postula entidades no físicas pero mantiene que algunas de las propiedades que poseen esos objetos constituyen una clase distinta de propiedades mentales* (Bechtel, 1991*:* p 109). Se consideran propiedades no físicas en el sentido de que nunca podrán ser reducidas a las ciencias físicas ni explicadas en esos términos. Postulan la necesidad de una nueva ciencia, una ciencia exclusiva de los fenómenos mentales. En un principio, todos los fisicalismos no-reductivos, es decir, aquellos que defienden que la mente no es ontológicamente reducible a la materia, pienso que entrarían de una forma u otra dentro de esta etiqueta más amplia de dualistas de propiedades (aunque con esto no todo el mundo estaría de acuerdo). Entre estos los hay que consideran que la mente es un *epifenómeno*. Según el significado filosófico tradicional del término *si x es un epifenómeno, x es un efecto que por sí mismo no tiene ningún tipo de efecto causal sobre el mundo físico*, esto es, que algunos estados mentales y sus propiedades son subproductos de un sistema causal cerrado y estos no son causalmente reducibles a estados físicos (Jackson, 1982). Dentro de estos destaca especialmente el monismo anómalo de Davidson, que afirma que las propiedades mentales poseen una diferencia causal en el mundo, que *se resisten a ser capturados en la red nomológica de la teoría física*, no existe algo así como leyes psicofísicas (Davidson, 1980). También hay quienes defienden un emergentismo, esto es, que lo mental *emerge* de alguna forma de lo físico, que hay estructuras complejas en el mundo que emergen como nuevas propiedades irreducibles a sus componentes más básicos (Bunge, 2004, Searle, 1999). Otros apuestan por un *naturalismo biológico,* dicen que aunque los estados mentales no son ontológicamente reducibles a estados físicos en el cerebro, sí son causalmente reducibles (Searle, 1992, 1996)[9]. Incluso algunos llegan a defender un *panpsiquismo*, esto es, que toda la materia tiene algún aspecto mental, (Chalmers, 1999). En mayor o menor medida todas estas posiciones no reductivas defienden algún tipo de dualismo con respecto a lo mental y lo físico. Finalmente todo fisicalismo no-reductivo acaba aceptando un dualismo de propiedades y éste sucumbe a muchas de las objeciones formuladas para el dualismo de sustancias cartesiano, del cual finalmente no se consigue desprender[10]. La posibilidad de ser fisicalista sin ser reduccionista surge en muchos casos de una distinción entre identidad de tipos e identidad de instancias[11]:

[9] Sin embargo, Searle no se encuentra cómodo dentro de la etiqueta de dualista de propiedades, como pretende demostrar en su artículo: (Searle, 2002).

[10] El fisicalismo no reduccionista tiene serios problemas para resolver el problema de la causalidad de lo mental. Jaewom Kim ha argumentado en varios de sus trabajos amplia y considero que acertadamente sobre esa dificultad presente en este tipo de materialismo (Kim, 1989, 1992, 2002)

[11] La identidad de tipos sostiene que toda propiedad mental puede hacerse co-extensiva con alguna propiedad física, es decir, que hay una conexión perfecta entre los términos que designan estados mentales y los términos que designan estados cerebrales. Debe por tanto haber algún tipo de regularidad nomológica entre lo mental y lo cerebral. Es la reducción perfecta y la aceptada por las teorías fisicalistas. La identidad de instancias dice que todo estado mental debe ser idéntico a algún estado cerebral particular, esto es, que aunque los estados mentales poseen siempre algún soporte cerebral global este no es siempre el mismo, con lo que la posibilidad de establecer regularidades nomológicas se desvanece. El funcionalismo y los fisicalismos no reductivos sobre todo el monismo anómalo aceptan este tipo de identidad.

uno podría ser fisicalista al aceptar la identidad de instancias, pero no ser reduccionista al rechazar la identidad de tipos. Finalmente a principios de los ochenta y con la clara intención de poner solución a todos los problemas planteados por estas teorías surgió una nueva y radical forma de materialismo, el materialismo eliminativo[12]. El materialismo eliminativo se ocupó ampliamente de aquellos aspectos de la mente con carácter intencional y representacional, véanse creencias, deseos, intenciones y demás, argumentando que dichos aspectos, englobados comúnmente con el nombre de psicología popular o psicología del sentido común, suponen una teoría falsa de nuestros estados mentales. Predicen que en un futuro esta psicología popular de sentido común será abandonada, eliminada y sustituida (que no reducida) por una teoría neurocientífica completa. Lo que pretenden hacer los eliminativistas con los conceptos de nuestra psicología popular es eliminarlos, pero ¿qué sucede con nuestros estados cualitativos? ¿acaso estos también deben ser eliminados de nuestra ontología? y si es así, ¿qué es para el materialismo eliminativo el sujeto humano además de carne? Los eliminativistas en general no niegan su existencia (a excepción de Dennett (1988) quien la niega explícitamente[13], y quizás Tye (1995), aunque este no se muestre tan explícito), sencillamente dicen que una caracterización más adecuada de estos tiene cabida en una teoría neurocientífica completa (Churchland, 1985, 1989, 1999)[14]. Los defensores de la postura eliminativista no cierran las puertas a una posible reducción de determinados aspectos de la mente desde una versión algo más moderada (materialismo revisionista[15]), por lo que comparten algunos aspectos con el fisicalismo reduccionista (Churchland, 1999: 84-85), y tampoco se aferran a una hegemonía mental de lo biológico, como la defendida por Searle (1980, 2000), no tienen problemas en aceptar que otros sustratos no biológicos puedan ser portadores de estados mentales, por lo que también comparten muchos aspectos con el funcionalismo, (no en vano los Churchland (1990), Dennett (1995), Tye (1995) o Ramsey, Stich y Garon (1991), se adscriben sin reparos a algún tipo de funcionalismo además de abogar por la eliminación de ciertas nociones que consideran obstáculos filosóficos). Estos consideran que el conductismo lógico no es una aproximación completa para una explicación de la mente, pero tampoco niegan su utilidad metodológica (Dennett, 1995, 2006, 2009). Existe, pues, un continuo entre las teorías funcionalistas de la mente, empezando por aquellas que condicionan las descripciones de los roles funcionales al nivel neuronal y molecular (Bickle, 2003), hasta aquellas cuyas descripciones de los procesos mentales son mucho más abstractas, al punto en que casi ni mencionan las estructuras cerebrales en que se realizan. Ambos extremos difieren en la manera de acercarse a la comprensión de lo mental, pero

[12] Esta nueva teoría hunde sus raíces en las filosofías de Sellars (1956), Quine (1960), Rorty (1970) y Feyerabend (1963).

[13] Unos años más tarde, Dennett llegará a la conclusión de que los *qualia* no son otra cosa que *complejos de disposiciones idiosincrásicas* a reaccionar, inherentes a nuestro sistema nervioso (Dennett, 1995: 400).

[14] Esta distinción no es del todo correcta, como apunto en la nota anterior, hasta donde llego tampoco Dennett niega la existencia de dichos fenómenos mentales. Como veremos más adelante, simplemente no acepta que dichos fenómenos se caractericen por las propiedades que tradicionalmente se les atribuyen, por lo que considero que salvando ciertos matices Dennett, Churchland y Tye comparten una posición muy similar con respecto a los qualia.

[15] Reducir una teoría a otra supone aceptar que durante el proceso quizás haya que prescindir de algunos elementos de la teoría reducida porque resulten espurios, este proceso es más bien un revisionismo de una de las teorías o un proceso de adaptación a la otra.

comparten la idea general de que no hay nada de misterioso en nuestras experiencias conscientes, o al menos, nada que no pueda ser aprehendido por la investigación en ciencia cognitiva y neurociencia.

Así las cosas, aunque no parece haber un acuerdo sobre cuál es la teoría más apropiada para explicar lo mental, lo que sí parece haber es una clara tendencia a su naturalización, a ver las mentes como productos de un complejo engranaje mecánico que nada tiene de enigmático. Esto significa mostrar cómo los fenómenos mentales son fenómenos no muy distintos de otros fenómenos naturales, sean éstos biológicos o neurofisiológicos, no muy diferentes de otros fenómenos como la reproducción, el metabolismo o la auto-reparación celular. El método para ello consiste en mostrar que la ciencia que se ocupa de los fenómenos mentales, la psicología, no es esencialmente distinta de otras ciencias naturales y que las propiedades mentales dependen de, o más precisamente sobrevienen[16] a, propiedades o procesos fisiológicos, de manera análoga a como otras propiedades del ámbito natural sobrevienen a propiedades físicas más básicas. En particular, se postula que la relación entre lo mental y lo físico es una relación de dependencia, de superveniencia mediada: las propiedades mentales sobrevienen a otras propiedades más básicas, propiedades físicas de bajo nivel.

La idea que defiendo en este trabajo es que el conductismo lógico, el materialismo reduccionista, el funcionalismo y el materialismo eliminativo, es decir, todas las versiones actuales y serias del materialismo, pueden unir fuerzas, complementarse y corregirse para convertirse en algo así como un neuro-computacionalismo, un funcionalismo materialista o un funcionalismo psicofísico computacional[17] capaz de explicar en toda su amplitud todos los aspectos de la mente que necesitan ser explicados. Las herramientas ya las tenemos, solo nos queda aprender a usarlas. Ahora bien, para que esto sea una realidad, primero se debe llevar acabo la difícil tarea de explicar la naturaleza de nuestros estados mentales cualitativos, esto es, se debe dar cuenta de lo que para muchos es un plus ontológico, un ingrediente extra que queda fuera de nuestras explicaciones físicalistas y funcionalistas de lo mental. Debemos explicar que son los qualia y los estados mentales conscientes.

[16] Sobre el concepto de sobreveniencia o superviniencia se debe profundizar sin duda en los trabajos de Jaegwon Kim, por ejemplo en Kim (1982, 1990). Muy brevemente. Cuando algo superviene a otra cosa decimos que hay una relación de dependencia, co-variación, coincidencia o indiscernibilidad entre ese algo y esa cosa. O lo que es lo mismo, *mundos que son indiscernibles con respecto a sus propiedades subvenientes son indiscernibles con respecto a sus propiedades supervenientes. Mundos que coinciden con respecto a las verdades que implican sus propiedades subvenientes coinciden con respecto a las verdades que implican sus propiedades supervenientes*. (Kim, 1990: 22). Pero si esta formulación de la noción de superveniencia (global o fuerte) es correcta deberemos aceptar que si *lo mental* es instanciado como una propiedad de lo físico entonces los poderes causales d*e lo mental* deberían ser idénticos a los poderes causales de *lo físico*. En otras palabras, los estados superiores deberían heredar sus poderes causales de los estados subyacentes que los realizan. Los anti-reduccionistas dirán que los poderes causales de nivel superior son *determinados por*, aunque no idénticos (o reducibles) a, los poderes causales de nivel inferior. Pero aquí Kim se apresura a revelarnos que si por *determinados pero no idénticos* se refieren a que los poderes causales superiores son realmente nuevos, entonces una de dos o los anti-reduccionistas nos deben una explicación de tal novedad o se encuentran atrapados en una maraña de dificultades aparentemente insuperables. (Kim, 1992: 441-2). Para análisis más profundos sobre el concepto de superveniencia y su relevancia sobre la causalidad de lo mental resulta interesante leer (Putnam, 2001), especialmente la primera conferencia de la segunda parte (pags 89-110), donde se debaten las posiciones antagónicas de Kim y Davidson con respecto a este tema.

[17] De aquí en adelante me referiré como funcionalismo materialista o únicamente materialismo a esta posición híbrida.

Mi trabajo es presentar, del modo más claro que me sea posible, esa problemática ampliamente debatida en la filosofía contemporánea, la cuestión concerniente a los qualia de la conciencia. A una teoría, sea esta la que sea, que pretenda dar cuenta de todos los mecanismos y propiedades de la mente humana debemos exigirle que explique nuestras actitudes proposicionales y nuestras experiencias fenoménicas[18], pero también debemos exigirle que se comprometa a situar dichas explicaciones en un marco ontológico adecuado al mundo que conocemos y que además sea coherente con otras visiones del mundo igualmente adecuadas. Me ocuparé aquí de aquellos aspectos de la mente cuya característica esencial es según muchos la de poseer propiedades fenoménicas, como las imágenes mentales, los dolores o la inmediatez de las sensaciones, los colores, olores o sabores, también denominados «sensaciones puras», «datos sensoriales», «cualidades fenoménicas», «propiedades intrínsecas de la experiencia consciente», «contenido cualitativo de los estados mentales» y, por supuesto, qualia[19], término introducido en la jerga filosófica para hacer referencia a ese aspecto específico de nuestra vida mental[20]. Sin duda, el término qualia no pertenece a nuestro lenguaje ordinario, sin embargo posee una considerable carga teórica, se trata de uno de esos términos que pueden resultar muy útiles en ocasiones, pero también pueden generar confusiones y enredos. Su introducción en la terminología filosófica tiene como objeto esclarecer dos de los problemas centrales de la filosofía de la mente: el problema de la naturaleza de lo mental y el problema de la relación entre lo mental y lo físico (el problema mente-cuerpo). Los qualia se suelen subdividir, a su vez, en sensaciones corporales (como los dolores) y en experiencias perceptuales (como oír un sonido agudo o ver una rosa roja) que aun sin ser estrictamente intencionales poseen un objeto intencional y contenido representacional, esto es, nos sirven para representar el mundo como siendo de determinada manera. Lo que distingue a los qualia de las actitudes proposicionales es su cualidad fenomenológica, es

[18] Debe aclararse, sin embargo, que la división entre estados mentales de actitud proposicional y estados mentales cualitativos, si bien clásica, ha sido severamente criticada desde diversas perspectivas; desde autores como Dennett (1988), que simplemente niegan la existencia de estados cualitativos como tales, hasta Searle (2004), que sostiene que no puede trazarse tal distinción puesto que todo estado intencional –incluidos los estados de actitud proposicional– tienen un carácter cualitativo. Conservaremos aquí la distinción solamente por fines expositivos.

[19] Se llaman '*qualia*' a los componentes de nuestra vida mental consciente. Son unidades de sensación, subjetivas en un sentido doble: ontológico –ya que su existencia depende del sujeto que las experimenta, de forma que si el sujeto desaparece también lo hacen sus *qualia*–, y epistémico –pues constituyen formas particulares de ver el mundo–, pues son ellos los que hacen que el mundo se viva de una u otra forma. Michael Tye presenta el término de la siguiente manera: "*Los sentimientos y las experiencias pueden ser muy variados: paso la mano por un papel de lija, huelo una mofeta, siento un dolor agudo en la mano, creo ver algo de un intenso color púrpura, me enojo mucho. En cada uno de estos casos, soy el sujeto de un estado mental con características distintivas muy claras. Lo que se siente al protagonizar cada estado es específico, tiene una fenomenología determinada. La filosofía emplea el término "qualia" (en singular" quale") para referirse a nuestros aspectos fenoménicos de nuestra vida mental, accesibles por medio de la introspección. En ese sentido estándar amplio, es difícil negar que los qualia existen"* (Tye, 2005).

[20] No hay acuerdo en que clase de estados mentales son poseedores de qualia y cuáles no. Para algunos, todo estado mental tiene asociado un *quale,* incluso los deseos y creencias (Block, 1994, p. 1). Flanagan, por ejemplo, distingue también un sentido amplio de *qualia: "Las creencias, pensamientos, esperanzas, expectativas y estados de actitud proposicional en general, así como las grandes estructuras narrativas, son cualitativas (o tienen componentes cualitativos) en este último sentido arnplio."* (Flanagan, 1992: 67). Otros consideran que los deseos y creencias no tienen asociado un *quale,* aunque en algunos casos vayan acompañados de otros estados mentales (imágenes, emociones, sensaciones), que sí poseen rasgos cualitativos (Tye, 1995, p. 4). Para otros, sin embargo el concepto de qualia es confuso, y mientras no aclaremos lo que queremos decir con él, no referirá a nada en concreto y molestará más que ayudará a elaborar una teoría científica sobre nuestros estados mentales, *"lo que necesitamos no es más investigación empírica de los sitios donde se producen los efectos inferiores, sino una aclaración del concepto de qualia*" (Dennett, 2006: 120). Mi opinión es que en caso de que algo así como los qualia acaben existiendo no es muy coherente pensar que se encuentren en todo estado mental. No me queda claro en qué sentido mi creencia disposicional, esa en la que no estoy pensando ahora, de por ejemplo que las moscas son insectos posee un rasgo cualitativo, tampoco tengo claro en qué sentido mi deseo de ir a Paris, independientemente de las imágenes o recuerdos que el deseo tenga asociados, posee un rasgo cualitativo.

decir, el cómo las siente y las experimenta el sujeto de la sensación. Podemos casi asegurar que los qualia pertenecen a una etapa más primitiva de la evolución y del desarrollo individual, son pre-racionales, esto es, tenerlos no es suficiente para calificar a una criatura como racional, mientras que, cuando le atribuimos deseos y creencias a alguien, estamos tratando de darle sentido racional a su conducta. Desde la tradición filosófica, los rasgos que comúnmente se atribuyen a los qualia son los siguientes. *Carácter intrínseco.* Esto es, son no-relacionales, y por tanto ni pueden ser definidos como estados funcionales ni tienen efectos causales en la conducta. No hay nada esencial a ellos que les impida estar asociados de una u otra manera a cada uno de nosotros. Las propiedades fenomenológicas no pueden ser identificadas con propiedades funcionales ya que estas se identifican por la red causal en la que se encuentran inmersas, entonces estar en un estado cualitativo u otro no parece producir ninguna diferencia causal. Si esto es así la existencia de tal propiedad crearía un serio problema para el funcionalismo, la tesis funcionalista de que los estados mentales son identificables con estados funcionales sería, en principio, falsa. *Inefabilidad*. Esto significa que no pueden ser etiquetados a través de términos de nuestro lenguaje público. La idea es que alguien que no ha experimentado un *quale* determinado no puede comprender qué es estar en ese estado cualitativo a través de una comunicación verbal con otras personas que sí han experimentado dicho estado. No es posible describir una experiencia a quien no la ha experimentado, pues la experiencia misma constituye el significado de los estados fenomenológicos. Así, quien no ha experimentado una cierta sensación no posee el concepto de dicha sensación. *Privacidad.* Es la idea de que cada uno de nosotros tiene un punto de vista propio a partir del cual tenemos experiencia del mundo. Y este punto de vista subjetivo no puede ser descrito desde la perspectiva de la tercera persona (objetiva); por lo tanto, debe ser capturado esencialmente desde un punto de vista de primera persona. Si esto es así, el materialismo se encuentra con serias dificultades. A*cceso directo a la conciencia.* Es la idea de que los qualia son conocidos como nada es conocido en el mundo, que tenemos un acceso directo, privilegiado e infalible a nuestros propios qualia. Los mecanismos mediante los que conocemos nuestra propia experiencia consciente son diferentes a los que empleamos para observar el mundo, pues conocemos nuestra propia mente a través de la introspección.

Cabe advertir que para que se den qualia parece obvio que hay que ser y estar consciente. Para que uno tenga la sensación de un color rojo intenso, de un olor nauseabundo o sienta fuertes y dolorosas punzadas en su estómago no hace falta que sea un experto en el procesamiento de la información o en la neurofisiología de la percepción visual ni olorosa, ni necesita conocer absolutamente nada sobre la estimulación de sus fibras nerviosas, pero sí es una necesidad el hecho de ser y estar consciente de que dichas sensaciones están sucediendo. Queda al margen de toda duda que hay una relación directa entre qualia y conciencia, o que incluso *el problema de los qualia* no sería sino otra expresión con la que referirnos al *problema de la conciencia* en la psicología y la filosofía contemporáneas. Se trata de un

problema inmerso en un problema más amplio, el de la conciencia. Cabe entonces esperar, que al ser los qualia esas propiedades fenomenológicas de nuestros estados mentales conscientes, todo, o al menos gran parte de lo que digamos acerca de éstos será válido también para la conciencia, pues aunque no sean términos correferenciales, sí poseen una fuerte correlación. Pero, ¿por qué dos nociones tan sumamente intuitivas y familiares se resisten a ser atrapadas por la ciencia objetiva? ¿Será por su carácter subjetivo? ¿Tiene en todo caso sentido que la ciencia objetiva se enrede en tratar de explicar lo subjetivo o es tarea imposible por pertenecer a un ámbito del que la ciencia no tiene ni tendrá nunca nada que decir? En definitiva, ¿es posible explicar la primera persona desde la tercera persona? Trataré de mostrar que estas cuestiones, son finalmente cuestiones sin sentido. Pienso que son todavía la radiación de fondo de la doctrina cartesiana, los últimos estertores de aquella defectuosa visión del mundo y el último reducto que le queda al defensor del dualismo.

Hay una serie de argumentos y ficciones filosóficas más o menos ocurrentes ideadas principalmente por los defensores de la noción tradicional de qualia que intentan mostrar que, de hecho, la tarea del funcionalismo materialista está condenada al fracaso. Los defensores de las posiciones funcionalistas y fisicalistas, reacios a aceptar tal noción de qualia, son los encargados de refutarlas, contra-argumentarlas o de una forma u otra restarles validez[21]. Estos argumentos son elaborados con la finalidad de salvar la existencia de los qualia, y rechazar el funcionalismo materialista como una posible teoría completa de nuestros estados mentales. Debe quedar claro que los filósofos que elaboran estos argumentos no están dentro de una categoría homogénea, no comparten las mismas ideas con relación a muchos de los aspectos de lo mental, lo que si comparten es la idea de que la mente no es reducible a propiedades físicas de bajo nivel, esto es, neuronas y conexiones entre estas, ni la idea de que lo mental en su totalidad sea funcionalmente aprehensible. Lo que aquí se debate es si estos argumentos son suficientemente fuertes o no para destruir las tesis funcionalistas y materialistas. Mi propósito es pues exponer los argumentos a favor de la existencia de qualia y demostrar como sucumben a las réplicas presentadas por los materialistas, para concluir que los qualia no son una piedra en el zapato del funcionalismo materialista

[21] Resulta conveniente aclarar las diferentes posiciones que adoptan los filósofos ante el problema de los qualia. Al ser estos usados como una objeción al funcionalismo materialista distinguiremos hasta un total de seis posiciones con relación a la compatibilidad o no de los qualia con el funcionalismo y el materialismo en general (la propuesta de esta clasificación está tomada de García Suarez, A. (1995)): **1**. *Incompatibilismo funcionalista* (Dennett): este niega que los *qualia* tengan ciertas propiedades de segundo orden problemáticas, propiedades que precisamente dan vida y sentido al concepto, (véase su carácter intrínseco, inefabilidad, privacidad y acceso directo a la conciencia), por lo que acaba proponiendo directamente su eliminación. **2**. *Incompatibilismo anti-funcionalista* (Block): los *qualia* son funcionalmente indefinibles y, por tanto, la validez del funcionalismo deberá restringirse. El funcionalismo es útil para explicar el contenido intencional pero no el cualitativo. **3.** *Compatibilismo funcionalista (Shoemaker):* para Shoemaker no sería posible definir funcionalmente estados cualitativos particulares aunque sí generales, lo cual abre la puerta a la posibilidad de hacer compatible el programa funcionalista con una posición realista respecto de los *qualia.* Por otra parte, P. Churchland y D. Lewis, defienden la tesis de que los *qualia* son simplemente cuestión de la sustancia física que realiza las funciones mentales, y por consiguiente no constituyen ningún problema para el funcionalismo, tampoco ven problemas Loar o Levin. **4**. *Contra el materialismo en cualquiera de sus versiones* (Nagel, Jackson, Chalmers). **5**. *El Naturalismo Trascendental* (McGinn, aunque también Levine): según la posición escéptica de McGinn, existe alguna propiedad del cerebro que explicaría la conciencia al modo naturalista, pero estamos cognitivamente cerrados a ella, del mismo modo que un niño de cinco años es incapaz de entender la teoría de la relatividad. **6**. *La subjetividad ontológica de la conciencia* (Searle): en ella estaría la diferencia que hace imposible la reducción de los fenómenos conscientes a los neurológicos y comportamentales. Searle no parece ver nada contradictorio en aceptar que nuestras experiencias conscientes son procesos neurobiológicos y a la vez ontológicamente subjetivos. Aunque esto parece seccionar el mundo en dos, el objetivo y el subjetivo, Searle se las arregla para defender esta tesis, claramente dualista, y a la vez negar la dicotomía entre lo mental y lo físico.

sino un claro ejemplo de obstáculo filosófico, un enredo que dificulta la elaboración de una ciencia rigurosa de nuestra mente. Veremos como muchos de los argumentos a favor de la existencia de qualia siguen de un modo u otro una estrategia epistémica, pues todos van de premisas epistémicas (enunciados acerca de *qué tipo* de *conocimiento* se trata, *que punto* de *vista* puede adoptarse, *que puede concebirse*) a conclusiones ontológicas (los *qualia* existen). En líneas generales se sostiene que los qualia existen porque hablar sobre un estado mental es distinto de experimentarlo, porque aparecen desde la primera persona y no desde la tercera persona, y porque podemos concebir estados mentales que no son estados funcionales y viceversa. Los qualófilos distinguen entre contenido intencional y contenido cualitativo de una experiencia: una cosa es el modo en que la experiencia representa el mundo, otra muy diferente a qué se parece tenerla[22]. Los qualia estarían constituidos por ese contenido cualitativo que (se supone que) subyace al contenido intencional. En consecuencia los qualia son lo que son porque solo pueden conocerse a través de la experiencia directa en primera persona. Ahora bien, y esto es una objeción general a todos los argumentos a favor de los qualia, hablar de un estado mental en tercera persona no implica necesariamente que ese estado *sea* diferente al experimentado desde la primera persona, y además, aunque podamos concebir aspectos mentales que no sean funcionales esto no implica necesariamente que tales aspectos existan.

He comenzado con una exposición general de donde nos encontramos en la actualidad en filosofía de la mente, destacando la fuerte tendencia hacía una naturalización de muchos aspectos que tradicionalmente no encontraban su sitio dentro de una concepción científica de lo mental. He presentado la noción de qualia y lo problemático que resulta aceptar su existencia, (tal y como es concebida por la tradición), y a la vez hacerla encajar dentro de la visión funcionalista y materialista imperante en nuestra época. La polémica en si no es otra que la de aceptar o no la existencia de determinados aspectos de nuestra mente que exceden lo puramente intencional. Una vez aclarado de que estamos hablando y en qué situación estamos con respecto a los qualia de la conciencia me propongo presentar y criticar una serie de argumentos en forma de experimentos mentales que se vienen desarrollando con el objetivo de desmontar las tesis materialistas y funcionalistas. Este trabajo consta de las siguientes seis secciones: (1) el argumento basado en la perspectiva (Nagel, 1974), (2) el argumento del conocimiento o argumento epistemológico (Jackson, 1982), (3) el argumento de los qualia invertidos y de la tierra invertida (Block, 1990), (4) el argumento de los qualia ausentes o zombis filosóficos (Chalmers, 1999), (5) el argumento de la brecha explicativa (Levine, 1983) y por último (6) el argumento de la habitación china (Searle, 1980).

[22] Un debate clave para los propósitos de este artículo es el que se da entre fenomenalistas (Block, McGinn, Loar…) y representacionalistas (Armstrong, Harman, Lycan, Tye…). Ambos están de acuerdo en que las actitudes proposicionales (deseos, creencias…) son estados mentales con contenido representacional, y por consiguiente que las explicaciones funcionalistas son adecuadas para explicar el carácter intencional de dichos estados. Sin embargo difieren en la adscripción de contenido representacional a las experiencias perceptivas y sensaciones corporales. Los primeros ven algo más en esas experiencias que contenido representacional, poseen contenido fenoménico. Los segundos abogan por lo que Lycan, un ferviente defensor de esta postura, llama "*la hegemonía de la representación*", que es *la doctrina de que lo mental y lo funcional-intencional son una y la misma cosa y que la mente no tiene propiedades distintivas que excedan sus propiedades funcionales e intencionales* (Lycan, 1996: 69).

Algunos argumentos dirigen sus ataques directamente al fisicalismo reduccionista, otros los dirigen hacia el funcionalismo y otros atacan a ambos. Mi diagnóstico es que todos estos experimentos presentan una serie de problemas que les impiden conseguir su objetivo, y mi trabajo es tratar de demostrar que esto es cierto. Muchas cuestiones quedan abiertas, no es mi propósito responder a esos interrogantes últimos sobre la mente humana, sino desbrozar el camino de malas hierbas. Creo que queda mucho por decir y muchos obstáculos por superar, pero también creo que vamos en la dirección correcta, que tenemos ante nosotros la oportunidad de despejar muchas de las incógnitas que hasta ahora se resistían a ser despejadas. Solo tenemos que mirar más de cerca y darnos cuenta de que hasta que no asumamos que muchos de nuestros conceptos más comunes nos han hecho crearnos falsas impresiones, no saldremos del laberinto en el que nosotros mismos nos hemos extraviado. El único modo de salir de tal enredo es optando por alguna de esas teorías que proponen revisiones de esos conceptos que durante demasiado tiempo nos han impedido ver con claridad. Veamos más detenidamente los principales argumentos y descubramos en que aspectos esenciales fallan.

**El argumento basado en la perspectiva**

Aun cuando llegáramos a conocer todos los hechos físicos objetivos acerca de los murciélagos (entre ellos su sistema de ecolocación), habrá siempre algo acerca de ellos que no sabremos *porque* no *somos murciélagos.* Desde la perspectiva del observador externo, hay algo en el murciélago que es inaccesible, esto es, ese carácter particular de la experiencia de ser murciélago que solo puede ser sentido desde la primera persona, es decir, siendo murciélago. La conclusión que se sigue de todo esto es que los *qualia* existen como aquellos aspectos de la experiencia que se resisten a ser expresados en términos físicos y funcionales.

Este es en líneas generales el argumento presentado por Thomas Nagel en su influyente artículo *What Is It Like to Be a Bat?*[23]. Nagel se muestra escéptico acerca de que la metodología científica sea capaz de decirnos algo acerca de la naturaleza cualitativa de nuestros estados mentales conscientes. El carácter subjetivo de la experiencia es el principal elemento a explicar en una teoría de la conciencia, y a la vez el gran obstáculo con el que se encuentra el punto de vista científico en su afán por desarrollar una ciencia sobre la conciencia. Las dificultades, para Nagel, son evidentes,

*Hay cosas del mundo, la vida y nosotros mismos que no pueden ser comprendidas adecuadamente desde un punto de vista eminentemente objetivo, por mucho que éste pueda llevar nuestros conocimientos más*

[23] Me serviré de dos trabajos de Nagel, el ya clásico artículo Nagel 1974 y su libro de 1986. La expresión "*What Is It Like to Be a Bat?"* se puede traducir al castellano de muchas maneras: ¿Cómo es ser un murciélago?, ¿Cómo que es ser un murciélago?, ¿Qué es como ser un murciélago?, ¿Qué idea tiene de sí un murciélago?, ¿Qué se siente al ser un murciélago?,

*allá del punto de donde partimos. Hay demasiadas conexiones con puntos de vista particulares, o con tipos de puntos de vista, por lo que todo intento de desarrollar un análisis completo del mundo en términos objetivos que esté totalmente libre de tales perspectivas conducirá inevitablemente a falsas reducciones o a negar por entero la existencia de ciertos fenómenos cuya realidad es irrefutable* (Nagel, 1986: 7).

Según Nagel, un ser es consciente si hay algo que es *cómo ser ese ser*; es decir, cómo parece o aparece subjetivamente el mundo desde el punto de vista mental o experiencial de esa criatura. La idea es que la perspectiva del observador, la perspectiva en tercera persona, impide el acceso a ciertos aspectos de la mente (los qualia). Sin embargo, todos los aspectos físicos de la mente pueden ser conocidos desde dicha perspectiva. Por lo que los qualia deben existir como aquellos aspectos no físicos de la mente.

Nagel afirma que ningún conocimiento adquirido desde la perspectiva de la tercera persona podrá decirnos nada sobre lo que se siente al ser un murciélago. Su argumentación implica y se sostiene sobre el siguiente bicondicional: *"un organismo tiene estados mentales conscientes si y sólo si, hay algo que lo determine a ser ese organismo, algo determinante para el organismo"* donde ese *algo determinante* es lo que Nagel llama *"el carácter subjetivo de la experiencia"* (Nagel, 1974: 436)[24].

El abismo que parece haber entre lo subjetivo (interno) y lo objetivo (externo) en el análisis de Nagel se basa en la idea de que la perspectiva del observador impide el acceso directo a ciertos aspectos de la mente. El problema es incorporar el punto de vista subjetivo a la perspectiva objetiva, y para Nagel los intentos elaborados en esa dirección que se adscriben a posiciones reduccionistas fallan porque se termina eliminando lo subjetivo, y eso es inaceptable, pues así solo se obtiene una visión incompleta del mundo[25]. Los análisis causales y funcionales, por principio, dejan algo fuera; no es que las experiencias conscientes no tengan poder causal o un rol funcional, sino que este no agota su análisis, no dicen nada sobre cómo es tener tal o cual experiencia consciente, aún queda algo por explicar, y por tanto,

*"Es inútil basar la defensa del materialismo en un análisis de los fenómenos mentales que no aborde de manera explícita su carácter subjetivo, […]. Por tanto, si no tenemos una idea de qué es el carácter subjetivo de la experiencia, no podemos saber qué debemos pedir a la teoría fisicalista. […] Si hemos de defender el fisicalismo, deben por sí mismas ofrecer una explicación física de las características fenomenológicas. Pero cuando analizamos su carácter subjetivo, tal resultado parece imposible de alcanzar".* (p 437).

[24] No encontramos el concepto de qualia en todo el artículo de Nagel, lo que si encontramos en numerosas ocasiones es la expresión *el carácter subjetivo de la experiencia*. Pienso que se refieren a lo mismo.

[25] Cabe advertir que Nagel no niega el poder de la reducción inter-teórica en otros ámbitos de la ciencia. Lo que sucede es que en el asunto de las experiencias conscientes no podemos acercarnos a la naturaleza real de la experiencia humana dejando atrás el punto de vista que de hecho constituye su esencia.

El razonamiento de Nagel, tal y como está expuesto no deja lugar a dudas, resulta extremadamente convincente, nadie puede entrar en la vida interior de nadie, nadie es capaz de experimentar lo que se siente al ser algo que no sea el mismo, no hay forma de entender lo subjetivo a partir de lo objetivo, pues el carácter *privado* de los qualia*,* hace que estos no sean accesibles desde afuera (mediante la observación) sino solo desde adentro (mediante la introspección). Veremos lo que ocurre cuando se analiza el argumento en profundidad, pues es entonces cuando las piezas no encajan tan bien como a primera vista parece. Antes, deberíamos advertir que la conclusión de Nagel no es que el fisicalismo sea falso, sino que *es algo que no podemos comprender.* El fisicalismo no puede ofrecer una explicación de por qué cuando estamos en ciertos estados físicos o funcionales tenemos las experiencias que tenemos, pero de ahí no se sigue que sea falso, sencillamente puede que nuestras limitaciones cognitivas y epistemológicas sean tales que no podamos entender que el fisicalismo sea correcto, pero de hecho puede que si lo sea (pags 446-447).[26] El argumento se podría formular así:

(1) Para saber cómo que es ser un X (un murciélago por ejemplo), hay que estar situado en la perspectiva (subjetiva) de X.
(2) El fisicalismo dice que no hay más información acerca de X que la información física descrita desde la perspectiva de la tercera persona.
(3) Ninguna cantidad de información física nos permite situarnos en la perspectiva (subjetiva) de X.

Por tanto,

A través del fisicalismo y su descripción desde la tercera persona no podemos saber todo lo que hay por saber sobre X, no es posible comprender la perspectiva subjetiva experimentada en primera persona por X. Por lo que el fisicalismo es incompleto.

El fisicalista rechaza tal conclusión porque no está dispuesto a aceptar ni (1) ni (3), pero para rechazar el argumento debe centrar sus ataques en (1), ya que (3) no es una premisa sobre la que podamos discutir, nadie sabe lo que sucedería si consiguiéramos tener toda la información física sobre algo. Lo que se discute es pues (1), que los seres humanos no somos capaces siquiera de imaginar lo que se siente al ser un murciélago debido al carácter subjetivo, privado e inefable de la experiencia. Las consecuencias de esto son al menos tres:

i) que un organismo tiene experiencias si y solo si hay algo que es como ser ese organismo
ii) que lo que una experiencia es para su poseedor no puede ser entendido por un organismo radicalmente diferente de él, y

[26] En este punto Nagel está más cerca de la idea de que hay una brecha explicativa que actualmente nos impide, y para muchos siempre nos impedirá, dar cuenta de los fenómenos subjetivos a partir de los objetivos, que de la idea de que el fisicalismo en un sentido amplio no sea verdadero. El argumento sobre la brecha explicativa (Levine, 1983) se verá con mucho más detalle más adelante. Por ahora me centraré en el carácter anti-reduccionista del argumento.

iii) que una experiencia de este tipo no puede ser aprehendida u observada desde la perspectiva de tercera persona.

Mi objetivo será rechazar (1), y mi táctica será tratar de anular las consecuencias que acabo de describir. Para ello examinaré los argumentos de Churchland contra el argumento de Nagel, estos me servirán para al menos, dudar de iii). Analizaré también, algunos fallos expositivos que pienso que son claves para que el argumento se sostenga en pie, y además mostrarán los numerosos problemas que supone aceptar ii). Finalmente me cuestionaré sobre si el *cómo qué es ser algo* teniendo una cierta experiencia implica, de hecho, que haya algo que sea como ser una cierta criatura teniendo esa experiencia, con lo que me desharé de i). La conclusión es que ninguna de las tres consecuencias de la idea sugerida en (1) se sostiene, y que por tanto hay una probabilidad muy alta de que finalmente (1) sea falso[27].

Churchland (1985) encuentra varios problemas al argumento. En primer lugar, niega que la afirmación de Nagel de que a partir de un análisis reductivo *se excluyen las características fenomenológicas de la experiencia* (Nagel, 1974: 437) sea cierta. Si aceptamos que las propiedades fenoménicas objetivas de rojez de las manzanas son idénticas a ciertas longitudes de onda en su reflectancia electromagnética, o que el calor, como una propiedad fenoménica objetiva de los objetos, es idéntico al nivel promedio de energía microscópica incorporada de dichos objetos entonces no hay razones para pensar que el análisis reductivo deje fuera dichas características fenomenológicas (Churchland, 1985: 18). Si las propiedades fenoménicas objetivas pueden ser analizadas reductivamente, ¿qué nos impide hacer lo mismo con las propiedades fenoménicas subjetivas? Nagel considera que dichas propiedades fenoménicas subjetivas son únicas y por ello inmunes a la reducción científica, la estrategia de Churchland es hacer un movimiento hacia afuera, y tratar de difuminar la (para muchos) infranqueable frontera entre lo interior y lo exterior. Afirma que el error está en haber introducido esas propiedades tan en el interior de los sujetos. Si las propiedades fenoménicas objetivas han sido analizadas reductivamente lo mismo podemos hacer con las propiedades fenoménicas subjetivas. Si podemos analizar la temperatura siendo una propiedad intrínseca de los objetos, si hemos conseguido objetivarla, ¿qué nos impide hacer lo mismo con las propiedades intrínsecas de la experiencia humana? Para Churchland no hay una diferencia esencial, son casos paralelos (p. 19).

En segundo lugar, para Churchland el argumento es un claro ejemplo de falacia intensional. Si formulamos el argumento como sigue (p 19):

(1) *Los qualia de mis sensaciones son directamente conocidos por mí, por introspección, como elementos de mi autoconsciencia.*

[27] Evidentemente si conseguimos demostrar la falsedad de (1), todo el edificio argumentativo de Nagel se vendrá abajo, pues es esa la afirmación sobre la que se sostiene.

(2) *Las propiedades de mis estados cerebrales no son directamente conocidos por mí, por introspección, como elementos de mi autoconciencia.*

*Por tanto*

(3) *Los qualia de mis sensaciones son diferentes de las propiedades de mis estados cerebrales.*

Los verbos *conocer* o *saber* introducen un contexto opaco, donde los contenidos de las representaciones mentales involucradas no están constituidos necesariamente por los objetos denotados sino por los sentidos expresados por las correspondientes expresiones del lenguaje natural empleado por el hablante. Siguiendo a Frege (1879), y su distinción entre sentido y referencia, en contextos gobernados por verbos de actitudes proposicionales las expresiones no denotan su referencia sino su sentido habitual. Es decir, que los sentidos son tan solo modos de presentación de los referentes, y es fácil observar que de las premisas acerca de modos de presentación de los objetos no es posible derivar una conclusión acerca de los objetos mismos sin exponerse a cometer una falacia, la *falacia intensional*. Compárese el ejemplo principal con el siguiente expuesto por Churchland (p 20):

*(1) La temperatura es conocida por mí, a través de mis sensaciones táctiles, como una característica de los objetos materiales.*

*(2) La energía molecular cinética no es conocida por mí, a través de mis sensaciones táctiles, como una característica de los objetos materiales.*

*Por tanto*

*(3) La temperatura es diferente de la energía molecular cinética.*

Evidentemente esa conclusión no se sigue, se está cometiendo la falacia intensional, y el ejemplo es muy similar al sugerido por Nagel. En otras palabras, el contenido de la representación mental (la temperatura o los qualia) no es el referente sino un modo de presentación del mismo. La temperatura puede ser idéntica a la energía molecular cinética aun cuando yo ignore que son idénticas de la misma forma que los qualia de mis sensaciones pueden ser idénticos a las propiedades de mis estados cerebrales aunque yo ignore que esto sea así. Por más que a Nagel le resulte tentador, afirmar que *los qualia de mis sensaciones son diferentes de las propiedades de mis estados cerebrales,* bien podría ser una afirmación falaz, pues bien podrían ser términos co-referenciales y co-extensivos. A esta objeción Nagel responderá con la siguiente versión del argumento (p 21):

*(1) Mis estados mentales son conocidos por mi mediante introspección.*

*(2) Mis estados cerebrales no son conocidos por mi mediante introspección.*

*Por consiguiente*

(3) *Mis estados mentales son distintos a mis estados cerebrales*.

Aquí Nagel consigue librarse de la falacia intensional, pues afirma que *ser conocido por mi mediante introspección* es una propiedad relacional genuina. Pero aun siendo así, la segunda premisa del argumento para un fisicalista bien podría ser falsa, si los estados mentales son idénticos a estados cerebrales son los estados cerebrales los que en definitiva estaremos conociendo por introspección, (aunque quizás no con la finura necesaria) y podremos reconocerlos mediante descripciones mentales o mediante descripciones neurofisiológicas (p 21). Es decir, el reduccionista bien podría ver esa segunda premisa como una petición de principio del antifisicalista, la cuestión es que es precisamente esa identidad entre estados mentales y estados cerebrales lo que está en juego, por lo que no podemos elaborar un argumento partiendo de la no identidad, para concluir igualmente con la no identidad entre estados. Si aceptamos una posible identidad entre estados mentales y estados cerebrales, si cuando digo que siento dolor estoy refiriéndome a lo mismo que cuando digo que tengo mis fibras C estimuladas[28], la frontera entre la primera y la tercera persona se difumina, no son dos perspectivas independientes sino dos maneras de expresar lo mismo, por lo que, al menos, deberíamos plantearnos serias dudas sobre si la experiencia en primera persona puede o no ser aprehendida desde la tercera persona. La consecuencia iii) bien podría ser ilusoria.[29]

Por otro lado, y con relación a la consecuencia ii) pienso que debemos centrarnos en la exposición del argumento, pues este lleva adheridas una serie de ambigüedades que no podemos dejar de analizar. En primer lugar, no queda claro si los límites de la privacidad son individuales o pertenecen a cada especie. La interpretación general del argumento de Nagel es que existen qualia comunes a una especie determinada, en su caso la de los murciélagos[30]. Pero esto nos compromete a pensar que mis qualia son similares a los de un habitante nativo de la Polinesia por pertenecer ambos a la misma especie ¿Acaso puedo saber cómo que es ser un habitante nativo de la Polinesia? ¿Puede saber un vidente como que es ser un ciego? Más aún, un humano ciego de nacimiento ¿es capaz de saber cómo qué es ser un ser humano? Todo esto genera múltiples dificultades. Nagel debería decirnos cuales son los límites en los que se encuentran esos puntos de vista que comparten los individuos que tienen un sistema perceptual similar, si el límite lo colocamos en las especies nos encontramos con los numerosos problemas que genera este concepto, pero si colocamos los límites de la privacidad en los individuos, entonces yo no puedo saber cómo que es ser nadie excepto yo, y esto nos conduce directamente al clásico problema de las otras mentes, y finalmente a

[28] Nagel, desde luego insiste, aunque no es capaz de decirnos por qué, en que modelos des-subjetivizantes como el del *calor/movimiento de las partículas* no tiene ninguna aplicación en el caso que nos ocupa.

[29] Nagel propone finalmente desarrollar una fenomenología objetiva que describa*, al menos en parte, el carácter subjetivo de las experiencias en forma comprensible para los seres incapaces de tener esas experiencias* (Nagel, 1974: 449), aunque no ofrece ninguna pista de cómo hacerlo. Finalmente esto se queda en una mera propuesta.

[30] Según Nagel *la distancia entre uno mismo y otras personas y otras especies puede encontrarse en cualquier parte a lo largo de un continuum* (Nagel, 1974: 442), es decir que a mayor diferencia entre especies menor comprensión sobre que se siente al ser ellos. Sin embargo Nagel no deja claro donde termina un punto de vista y comienza otro.

una solución solipsista[31]. En segundo lugar, y como hace notar el propio Nagel (p 438) en su argumento elige deliberadamente al murciélago por ser un mamífero, y por tanto parecerse lo suficiente a nosotros como para que exista la convicción de que *sin duda los* murciélagos son criaturas conscientes. Pero el lugar a donde pretende llevarnos su análisis no debería impedirnos rebasar la barrera de los mamíferos y cuestionarnos igualmente por como qué es ser un gusano, una ameba o incluso una célula procariota. La cuestión aquí es que si nos saltamos esa barrera y hay algo qué es como ser una ameba por ejemplo, tal y como está expuesto el argumento, deberíamos reconocer que hay un punto de vista desde la ameba, esto es, deberíamos concederle a la ameba la facultad de ser consciente, pero, ¿acaso las amebas tienen sentimientos? ¿y subjetividad? Si colocamos la barrera de la conciencia en lo vivo, nos encontramos con el problema de definir tal concepto, ¿hay algo que es como ser un virus?, pero aún podemos llevar la barrera más allá, pues no queda claro en qué sentido debemos atribuir conciencia a una célula procariota y no a un meteorito. Este ir más allá implica plantearse el panpsiquismo, la radical y controvertida visión de que lo mental esta por doquier, está distribuido por todo el universo.[32] Pero si la conciencia es algo que está por doquier, y los qualia son esos aspectos esenciales de la experiencia consciente, entonces no podremos negar que las piedras, las cafeteras y las sillas poseen qualia. Definitivamente, hay algo aquí que no estamos haciendo bien. Será mejor que nos echemos a un lado, retrocedamos unos cuantos pasos y volvamos a repasar las cuentas, hay alguna incógnita que no hemos despejado bien o que tenía más valor del que le hemos dado. Aceptar el panpsiquismo y el solipsismo es un altísimo precio a pagar para que una experiencia sea para su poseedor algo que no puede ser entendido por un organismo diferente de él. Definitivamente, es un precio tan alto que aceptar ii) trae consigo demasiadas dificultades. Pero aún hay otra dificultad mayor, ¿acaso yo mismo se cómo qué es ser yo en este preciso momento o como qué fue ser yo la semana pasada? Tengo serias dificultades para responder a esta pregunta. Aquí el defensor de los qualia apelaría a la inefabilidad de la experiencia fenoménica en primera persona como uno de sus rasgos más destacados, pero yo no creo que mis dificultades para responder a esa pregunta provengan de esa supuesta inefabilidad sino de la defectuosa perspectiva desde la que está formulada[33]. Esto nos lleva al último punto, i) *que un organismo tiene experiencias si y solo si hay algo que es como ser ese organismo*[34].

[31] De hecho la afirmación de Nagel de que hay alguna propiedad esencial y privada en nuestro pensamiento conduce inevitablemente al solipsismo.

[32] Para una detallada discusión sobre el panpsiquismo: Chalmers (1999: cap.8); Nagel (1979), y sobre todo véase la visión de Strawson en Freeman (2006) o las réplicas del propio Strawson en Strawson (2006).

[33] Si en vez de preguntarme como que es ser yo ahora, me preguntan que estoy viendo a través de mi órgano visual, o que estoy escuchando a través de mi órgano auditivo, entonces puedo responder que estoy experimentando visualmente la pantalla de mi ordenador en la parte central de mi fóvea y el resto de mi habitación en la periferia, y auditivamente experimento el repiqueteo de las teclas de mi ordenador. Sin duda estas son las experiencias de las que soy consciente en este momento. No cabe duda de que muchas otras, a pesar de ser captadas por mis sentidos, no han pasado el umbral de mi conciencia (si es que esto se puede decir así). Quizás mientras conscientemente experimentaba esto, un perro ladraba en la calle, una mosca se posó sobre mi brazo o alguien puso en marcha un vehículo, pero no era ahí donde mi atención estaba dirigida.

[34] Generalmente Nagel suele marcar una diferencia clara entre los organismos y lo que no lo son, dice *que un organismo tiene experiencias si y solo si hay algo que es como ser ese organismo,* sin embargo no queda claro hasta qué punto su generalización no es aplicable por ejemplo a robots o a partes de un organismo, ¿acaso no hay algo que es como ser un cerebro? Además un organismo bien puede pasar toda su vida sin tener experiencias conscientes (en coma, por ejemplo) sin dejar de ser un organismo. (Snowdon, 2010: 11).

El análisis de la expresión *"como que es ser una cierta criatura*" es de suma importancia, pues es la base sobre la que se sostienen muchos de los argumentos anti-fisicalistas, y por supuesto el de Nagel. Para muchos, la idea de ser poseedor de experiencias conscientes puede ser perfectamente caracterizada en términos de la expresión *"hay algo que es como ser una cierta criatura*", esta parece ser la llave que abre la puerta que nos lleva a la explicación de las experiencias conscientes[35]. Ahora bien, también existe la posibilidad de que haya algo ilusorio en nuestro uso del lenguaje cuando afirmamos de forma reiterativa que "*hay algo que es como ser algo"*, los murciélagos no tienen ni idea de cómo que es ser un murciélago ni se cuestionan sobre ello, ellos no comparten ideas ni experiencias, nosotros sí, a nosotros el lenguaje nos permite ir más allá de nosotros mismos, pero a veces esa peculiaridad de nuestro lenguaje nos hace perder la perspectiva de lo que estamos viendo y nos lleva por sendas que acaban en puntos muertos. ¿Qué queremos decir cuando decimos que hay algo que es como ser algo? ¿hay, realmente algo que es como tener una experiencia? Y si es así, ¿significa esto que entonces hay algo que es como ser un murciélago? ¿hay algo que es como ser usted o como ser yo? Se trata pues de destripar la noción de *como que es ser X,* y veremos como no resulta nada obvio lo que queremos decir con ella. Aunque para algunos es imprescindible (Nagel, 1974; Chalmers, 1999; Janzen, 2011), para otros es un eslogan sin sentido (Hacker, 2002), es trivial (Snowdon, 2010), no es una amenaza para el fisicalismo (Nagasawa, 2003)[36], o no es posible analizarla fuera de consideraciones existenciales (Evnine, 2008). Desarrollaré brevemente los argumentos de Evnine y Snowdon, y veremos si es necesario mantener la noción de *como que es ser X* como una noción clave en la explicación de las experiencias conscientes o resulta más sensato y coherente prescindir de ella de una vez por todas[37].

La cuestión es, ¿qué queremos decir cuando decimos que *un organismo tiene experiencias si y solo si hay algo que es como ser ese organismo?* Parece claro que debe haber algún tipo de relación entre *como que es ser algo* y las diferentes modalidades sensoriales de ese algo, por lo que cabe distinguir entre la noción de (1) *como que es ser C* (donde C es un cierto tipo de criatura) y la noción (2) *como que es E* (donde E es tener un cierto tipo de experiencia a través de una modalidad sensorial concreta). La primera se refiere al sujeto de la experiencia, la segunda a la experiencia del sujeto. No hay duda de que ambas nociones están relacionadas, para que haya algo que sea como tener un cierto tipo de experiencia debe haber algo que sea como ser un cierto tipo de criatura, lo que no queda tan claro es el tipo de relación que hay entre

[35] Quizás el primero que empleo la expresión fue Wittgenstein, posteriormente Farrell y Sprigge la usaron en un contexto similar y finalmente Nagel hizo uso de ella dotándola de un poder especial en la explicación de las experiencias conscientes. Desde entonces ha pasado del uso al abuso en muy poco tiempo, de ahí la necesidad de aclarar exactamente a que se refiere.

[36] Nagasawa hace una curiosa relación entre los argumentos de Nagel y el argumento Tomista respecto a la naturaleza de la omnipotencia divina, una cuestión que, a primera vista, no tiene relación con el argumento de Nagel. Su conclusión es que a pesar de la imposibilidad de obtener un conocimiento acerca de lo que se siente al ser un murciélago, esto en ningún caso refuta el fisicalismo. Para una crítica a esta línea argumental véase Gorman (2005).

[37] Es importante advertir que el sentido que debemos dar al verbo *ser* en la frase *que es como ser X* es el de subjetividad, en un principio se trata de evitar darle un sentido existencial, pues no tendría lugar en el contexto que aquí se discute, el análisis es sobre *que es como ser el sujeto de una experiencia o que es como ser un sujeto consciente,* hablaremos del sujeto en sí no de su existencia.

ellas.[38] Evnine encuentra hasta cuatro maneras de analizar la noción (1) en base a su relación con la noción (2). La primera es a través de un análisis minimalista, como el que parece hacer Nagel cuando dice que *el hecho de que un organismo tenga experiencias conscientes significa, básicamente, que hay algo que es como ser ese organismo* (Nagel, 1974: 436). Este análisis sugiere que hay una equivalencia analítica entre las expresiones (1) y (2) a través de la expresión *C tiene experiencias conscientes*, donde tener experiencias conscientes significa que hay algo que es como ser un cierto tipo de criatura. Pero en realidad aquí estamos pasando de una expresión sin compromiso ontológico, *C tiene experiencias conscientes*, a otra con compromiso ontológico, *hay algo que es como ser C,* de ahí que la noción (1) adquiera un compromiso ontológico mínimo[39]. Si admitimos que los humanos tenemos experiencias conscientes deberemos aceptar, aunque sea mínimamente cierto, que hay algo que es como ser un humano. Ahora bien, ¿es suficientemente fuerte dicho compromiso como para aceptar tal equivalencia?, y por tanto, ¿es lícito hacer aseveraciones filosóficas serias a partir de transformaciones de este tipo?, más aun, ¿debemos comprometernos con ellas hasta el punto de aceptarlas como una pieza fundamental en la explicación de las experiencias conscientes? Pienso que al comprometernos con formas de discurso que contienen entidades ontológicamente mínimas corremos el riesgo de cometer errores desde el inicio, errores que sin duda arrastraremos a lo largo de todo el discurso. En definitiva, no es filosóficamente prudente que una teoría que trate de dar cuenta de nuestras experiencias conscientes se sostenga sobre una expresión cuyo compromiso ontológico es mínimo, por lo que esta opción pierde interés filosófico (Evnine, 2008: 191). Una segunda opción es ver (1) como una expresión primitiva, esto es, negar su equivalencia con cualquier otra expresión, y verla como una especie de zumbido ontológico que acompaña a la existencia como un C (p. 191), claro que para que esto sea así se ha de sentir dicho zumbido ontológico, y Evnine no lo siente, yo tampoco. Estos dos análisis extremos de la noción (1) no parecen ser satisfactorios, el primero legitima la existencia de que haya algo que sea como ser algo, pero pagando el alto precio de privarla de interés filosófico, el segundo mantiene el interés filosófico, pero el precio a pagar es el de una existencia altamente improbable (p 191). Pero aún podemos hacer análisis no tan extremos entre *como que es ser C* y *como que es E* sin que colapse la primera noción en la segunda. Una posible relación entre ambas nociones es por extrapolación. Para que C esté relacionado por extrapolación con las diferentes modalidades de E se han de obtener relaciones en virtud de algún rasgo intrínseco de los E, y además C no puede ser idéntico a

[38] También podríamos diferenciar entre (1) *como que es ser C* (donde C es un cierto tipo de criatura) y (1´) *como que es ser X* (donde X es un miembro de C). En un principio para saber cómo que es ser C deberíamos saber cómo que es ser todos los X que lo componen, para saber cómo que es ser un murciélago deberíamos saber cómo que es ser cada uno de los murciélagos, sin embargo este sentido de lo que se siente al ser un C carece de interés teórico en esta discusión, bien podría haber algo que se siente al ser X sin ser relevante para lo que se siente al ser C (Evnine, 2008: 188). Esta distinción es interesante, pero no es útil para nuestros propósitos, así que tomaremos las nociones de C y X como lógicamente distintas, (1) es diferente de (1´). Lo que me interesa es (1) *como que es ser C* (donde C es un cierto tipo de criatura).

[39] Cuando hablamos de entidades ontológicamente mínimas nos referimos a que sentencias del tipo ‘‘Fido es un perro’’, carentes de compromiso ontológico, pueden ser pleonásticamente transformadas en otras como ‘‘Fido tiene la propiedad de ser un perro”, adquiriendo así un cierto compromiso ontológico, aunque mínimo. Algo similar sucede en la equivalencia que sugiere Nagel cuando pasa de *C tiene experiencias conscientes* a *hay algo que es como ser C*. Se trata de creaciones de nuestras prácticas lingüísticas y conceptuales a partir de las cuales sacamos una propiedad de la nada, suministrando a una expresión un valor ontológico del que previamente carecía. Para un profundo análisis sobre las entidades ontológicamente mínimas, Thomasson (2001) o Schiffer (1996).

la suma de todos los E. En este caso *como que es ser C* se extrapola a partir de *como que es E*. E equivale a todas y cada una de las modalidades sensoriales de C, por lo que *como que es ser C* resulta ser algo que tienen en común todas las posibilidades existentes de *como que es E*. Así que si queremos conocer *como que es ser C* deberemos encontrar un elemento común a todos los *como que es E* (p 192-193). En el caso de los murciélagos de Nagel tenemos ese elemento común, viven en un mundo percibido a través de un sonar o un sistema de eco-localización, por lo que en un principio, conociendo como que es percibir el mundo desde ese sonar sabríamos como que es ser un murciélago[40]. Pero nótese que el hecho de que ese elemento sea común a todos los murciélagos no implica que no pueda ser un modo de percepción disfrutado por otras criaturas. Lo qué se siente al ser un ser humano debe tener alguna relación con lo que es ver, oír, oler, y así sucesivamente; pero tal y como está definido el método de extrapolación, no habría ningún problema en que se diera la misma relación con lo que es percibir a través del sonar, a pesar de que este no sea un modo de percepción que se encuentre en los humanos (a fin de cuentas, un humano con un sonar no dejaría por ello de ser un humano). En resumen, el método de extrapolación sería coherente si lo que se siente al ser un C es lo mismo para todos los valores de C, independientemente de las diferencias en sus capacidades de percepción, y esto sería incompatible con las afirmaciones de Nagel de que no podemos saber cómo que es ser un murciélago porque no podemos saber cómo que es percibir a través de un sonar (p 193). Aun concediendo que yo sé lo que es ver, oler, oír, etcétera, no puedo trasladar este conocimiento a cualquier cosa que pueda contar como conocimiento de lo que significa ser un ser humano, por lo que tampoco el método por extrapolación nos es de mucha ayuda. Pero aún podemos ver una última opción, una relación agregativa, esto es, *como que es ser un C* es un agregado de todos los como que es E asociados con todas las modalidades perceptuales de C. Pero obsérvese que si aceptamos esto debemos aceptar que al identificar *como que es ser un C* con la suma de todos los *como que es E* para cada modalidad sensorial de E perteneciente a todos los C, entonces un individuo que pertenece a los C y que no sepa como que es una determinada E tampoco sabrá *como que es ser un C* (p 195). O lo que es lo mismo, una persona ciega de nacimiento no sabe, o al menos tiene un conocimiento parcial de lo que es ser un ser humano, pues entre sus modalidades perceptuales falta un E, el de la visión. Tampoco este parece ser un método muy útil. Finalmente, y tras varias vueltas más, Evnine parece rendirse, no hay forma de extraer un significado claro de la noción (1) *como que es ser C* en su relación con la noción (2) *como que es E,* uno debe o permanecer en el escepticismo con respecto a estas nociones o finalmente introducir elementos existenciales, aunque al introducir dichos elementos, la noción de lo que se siente al ser un C ya no será especialmente relevante para el estudio de la conciencia tal y como se viene desarrollando dicho proyecto filosófico (p 201).

[40] Según Nagel no podemos saber cómo que es ser un murciélago porque no podemos saber cómo que es percibir a través de un sonar (Nagel, 1974: 438), pero, ¿esto significa que si consiguiéramos crear un sistema de eco-localización e implantárselo a un ser humano conseguiríamos saber cómo que es ser un murciélago?

Ahora, diré algo sobre el análisis realizado por Snowdon (2010). Este analiza ambos lados del bi-condicional de Nagel ligeramente modificado: (1) *un suceso E es una experiencia si y solo si hay algo que es como ser una entidad que está sometida a o que disfruta de E.* Lo importante no es que Nagel acepte esta afirmación acerca de la experiencia como una verdad necesaria sino que la acepta como (2) *una verdad básica o fundamental*. Un rápido vistazo a (1) nos permite ver que podemos hablar acerca de cómo que fue una experiencia para un sujeto solamente en el caso de que las experiencias estén relacionadas con determinados sucesos, y es precisamente el caso, las experiencias, como cualquier otra cosa, deben tomar alguna forma, y esta no es otra que la de los sucesos que las componen. Toda experiencia está compuesta por unos determinados sucesos, acontecimientos o hechos, no hay más. Si E es un estado consciente entonces E debe tener alguna propiedad de la que finalmente debemos poder hablar. La lectura de izquierda a derecha del bi-condicional es correcta, pero completamente trivial, no nos dice nada que no sepamos, las experiencias son sucesos y hay entidades que las padecen o las disfrutan. Pasemos entonces a la lectura de derecha a izquierda. Aquí nos damos cuenta de que podemos ver lo que es un evento E como para un sujeto aun cuando no haya ninguna razón para pensar que E sea una experiencia. Del mero hecho de que un suceso tenga alguna propiedad no se deduce que se trate de un estado consciente. Si esto es así entonces no hay razones como para pensar que para que los E sean una experiencia hay algo que deba ser como someterse o disfrutar de los E (p 14-15). Los sueños son experiencias pero no sucesos, las operaciones quirúrgicas son sucesos pero el paciente anestesiado no los está experimentando. Si me preguntan ¿qué soñaste ayer?, puedo responder que soñé que volaba, y que fue una experiencia increíble, hubo una experiencia, pero desde luego no sucedió ningún suceso. Si me preguntan ¿cómo fue tu operación de apendicitis?, puedo responder que no sentí nada, estaba anestesiado, no experimenté nada, no fui consciente de ello, pero de hecho sí hubo tal suceso. Lo que indica esto es que puede haber sucesos que no sean experimentados aunque haya algo que sea como ser una entidad sometida a dicho suceso, y que además, como en el caso de los sueños, que puede haber algo que sea como una entidad que tenga experiencias que no sean sucesos. En resumen, Snowdon concluye que el bicondicional "*un suceso E es una experiencia si y solo si hay algo que es como ser una entidad que está sometida o dispone de E"* es trivial si va de izquierda a derecha y falso si va de derecha a izquierda, es decir que no es relevante vaya en la dirección que vaya, y por tanto que como (1) o es trivial o es falso, (2) *que (1) es una verdad básica o fundamental*, difícilmente será cierto, por lo que ese aspecto básico o fundamental del eslogan de Nagel no será ni tan básico ni tan fundamental (p 22-23). Finalmente nos advierte de que su aceptación incondicional supone correr el altísimo riesgo de aceptar como fundamental algo que muy probablemente no lo sea.[41]

[41] A una conclusión similar llega Hacker (2002) con su análisis. Concluye que las sentencias del tipo *"hay algo que se siente al ser un murciélago"*, "*hay algo que se siente al ser un ser humano"* o "hay algo que se siente al ser yo" no están usadas correctamente, las cuestiones del tipo *"¿cómo es para un X ser un X?"* son reiterativas, las cuestiones del tipo *"¿qué es como ser un X?*", pueden ser respondidas en forma de

Me gustaría decir algo sobre el cómo de *como que es ser un murciélago.* En un principio no parece que ese *como* tenga una función comparativa, no es un *como* como el de la expresión *me siento como aquella vez que me tocó la lotería.* Nagel no nos dice que tener una experiencia consciente se parezca o sea similar a algo. Pero si su *cómo* no tiene una función comparativa, ¿qué función tiene? Cuando por ejemplo un amigo me dice que su vecino se ha comprado un coche nuevo yo puedo preguntar: ¿cómo es el coche?, este es un uso del *cómo* no comparativo, no estoy preguntando a que se parece sino como es, y él me puede responder que es grande, deportivo, rojo o como sea. Lo importante en todo esto es que cuando la gente habla de *cómo* que fue tener una determinada experiencia si no encontramos en ese *como* elementos comparativos debemos suponer simplemente que fue de una manera o de otra. Si esto es así y el *cómo* de Nagel es como el que acabo de describir, no hay nada misterioso, sencillamente sería posible responder a la pregunta como que es ser un murciélago de la misma manera que es posible responder a como que es subirse a una montaña rusa, o como que fue ganar la lotería o como que es tener cualquier experiencia. Ahora bien, alguien nos podría objetar que una cosa es como que son las experiencias para un sujeto y otra bien distinta como que es ser el sujeto de las experiencias. La cuestión es que para responder a que es ser el sujeto de las experiencias no es necesario centrarse en las experiencias.

La expresión *¿cómo que es ser un ser humano?*, significa lo mismo que *¿cómo qué es para un ser humano ser como un ser humano?* solo que es una expresión reiterativa, nada que no sea un ser humano puede ser un ser humano. La pregunta ulterior sería *que es como ser un ser consciente de que es*, o aún más, *que es como ser un ser consciente de que es consciente de que es…* y así sucesivamente, pero esto es sumamente reiterativo, es una regresión infinita, y no solo eso, resulta que es así por detrás, pero también por delante, podemos estirarla hacia el otro lado diciendo *que es ser como ser un ser consciente de que es* o *como es como ser un ser consciente de que es*… puede ser estirada hasta el infinito por ambos lados, con lo que se convierte en una noción que se escapa por el abismo, acaba perdiendo el sentido porque finalmente no se agarra a nada, no queda claro a que se refiere, nos deja en una especie de limbo, un bucle del que difícilmente podemos escapar. Parece, que después de todo, de nuevo, el lenguaje nos ha jugado una mala pasada, no hay ningún sentido en el que estas expresiones consigan aprehender algo, no hay forma de encontrarles un referente claro, son expresiones vacías. ¿Tiene sentido entonces cuestionarse sobre como que es ser algo? Es posible que no, es posible que cuestionarse sobre como que es ser un murciélago tenga el mismo sentido que cuestionarse sobre como que es ser yo si hubiera nacido ayer, o como que es ser un dinosaurio, o como que es ser Francia, o como que es ser un universo. Queda mucho trabajo por hacer, mucho más que profundizar en el análisis de la cuestión que se hace Nagel en el título de su célebre artículo. No creo que los análisis expuestos aquí contra dicha noción sean concluyentes, pero espero cuando menos haber dejado en el aire aunque solo sea las dudas suficientes como para al menos sacar a

las características distintivas de la vida de los Xs, y cuestiones como "*¿cómo que es para mí ser yo?" o "¿cómo que es para mí ser un ser humano?"* es igualmente ilícita. Una crítica reciente y muy interesante de los análisis de Snowdon y Hacker en Stoljar (2014).

dicha noción del preeminente lugar en el que algunos la han colocado, las expresiones *como que es ser C*, *que se siente al ser C* o *como que es tener la experiencia E*, no me parecen del todo inteligibles.

Para acabar me gustaría decir algo más general acerca del sugestivo argumento de Nagel. El objetivo de Nagel es elaborar una pregunta retórica que nos obligue a ejercitar nuestra imaginación, y vaya si lo consigue, es el suyo sin duda uno de los artículos más citados de la literatura existente sobre el tema[42]. Lo que trata de hacernos ver es que la realidad de las experiencias conscientes colapsa sobre sí misma en el mismo momento en el que estas aparecen. Este es el juego de espejos, un hábil juego de palabras e imágenes confusas que la redundante sentencia de Nagel provoca en el lector más sugestivo. Sin embargo, en vez de marear al lector con endiablados juegos de palabras, lo que debería hacer el amante de los qualia es tratar de responder a las múltiples cuestiones que se amontonan en su escritorio y que a mi parecer no debería pasar por alto: ¿cómo puedo conseguir, desde mi fenomenología en primera persona saber en qué estado cualitativo me hallo? ¿puedo experimentar cientos de quale diferentes al mismo tiempo? Si observo un círculo azul y me imagino que estoy viendo un cuadrado amarillo, ¿qué quale estoy experimentando? ¿El hecho de que tú me digas que estas experimentando un quale concreto ante un estímulo determinado influirá en el quale experimentado por mí ante el mismo estímulo? ¿Cuántos quale existen? ¿Qué cantidad de estados fenoménicos debe poseer un individuo para ser considerado como un ser consciente o, al menos, fenoménicamente relevante? No hay respuestas convincentes a estas preguntas porque, en definitiva, es más coherente renunciar a la idea generalizada, pero confusa, de que se siente algo al ser un murciélago, y admitir sencillamente que estos animales, así como nosotros, hacemos un uso funcional, adecuado o no, de la información espectral que a través de nuestros órganos sensoriales recibimos.[43]

## El argumento del conocimiento

Casi una década después de la ocurrencia de Nagel, Frank Jackson ideó un experimento mental algo diferente pero con consecuencias similares[44]. El experimento pretendía destruir la tesis de que todo lo que podemos conocer acerca del mundo se reduce completamente a un conocimiento de lo físico. Es un argumento contra el fisicalismo. Dice así:

[42] Pienso que una de las razones por las que el lema de Nagel ha llamado tanto la atención es porque no lleva incrustado ningún término filosóficamente robusto, no contiene términos como por ejemplo qualia, conciencia o subjetividad, su lema es lo suficientemente cotidiano y convincente como para parecer evidente.

[43] Si el lector no se ha sentido persuadido por los argumentos esgrimidos contra Nagel debería continuar leyendo la siguiente sección, pues el argumento del conocimiento es primo hermano de este, todo lo que sirve para uno sirve para otro, mi objetivo hasta ahora no ha sido otro que el de generar una cierta duda. A partir de ahora trataré de que las dudas se disipen, claro está, hacia mi lado.

[44] La diferencia principal que consigo ver entre los argumentos de Nagel y Jackson es que mientras Nagel se sitúa en el nivel ontológico, esto es, habla de la perspectiva en primera persona en virtud de lo que somos (seamos murciélagos, humanos o lo que sea), Jackson, como veremos, se sitúa en el nivel epistemológico, esto es, se cuestiona más bien por el tipo de conocimiento que somos capaces de adquirir.

*"Mary es una brillante científica que, por alguna razón, se ve obligada a investigar el mundo desde una habitación en blanco y negro por medio de un monitor en blanco y negro. Se especializa en la neurofisiología de la visión y adquiere, supongamos, toda la información física que se puede obtener acerca de lo que ocurre cuando observamos tomates maduros, o el cielo, y empleamos términos como 'rojo', 'azul', y otros colores. Mary descubre, por ejemplo, qué combinaciones de longitudes de onda del cielo estimulan la retina, y cómo esto produce, a través del sistema nervioso central, la contracción de las cuerdas vocales y la expulsión de aire de los pulmones necesarias para producir el enunciado 'El cielo es azul' […] ¿Qué ocurrirá cuando Mary salga de la habitación blanca y negra o se le entregue un monitor a color? ¿Aprenderá algo? Parece obvio que aprenderá algo acerca del mundo y de nuestra experiencia visual en él. Entonces es innegable que su conocimiento previo fuera incompleto. Sin embargo tenía toda la información física. Por lo tanto, hay más por saber, y el fisicalismo es falso".* (Jackson, 1982: 130)[45].

Se trata de un argumento de tal amplitud que ha provocado una literatura prácticamente inabarcable, se ha reinterpretado de múltiples maneras[46], se ha analizado desde numerosos puntos de vista y se ha formalizado de todos los modos posibles. Me quedaré con la formulación que elabora el propio Jackson (Jackson, 1986: 293):

*(1) Mary (antes de su liberación) conoce todo lo que físicamente hay por conocer acerca de las otras personas.*

*(2) Mary (antes de su liberación) no conoce todo lo que hay por conocer sobre las otras personas (porque aprende algo acerca de ellos tras su liberación).*

*Por consiguiente,*

*(3) Hay verdades acerca de las otras personas (y acerca de sí misma) que escapan a la versión fisicalista.*

Las réplicas al argumento se dividen en dos grupos: los hay que dicen que Mary no aprende nada cuando sale de su cautiverio monocromático (Dennett, 1988, 2006; Hardin, 1988; Foss, 1989 y Churchland, 1985), estos tienen serios problemas para aceptar la segunda premisa porque tienen serias dudas de que la primera nos la podamos llegar siquiera a imaginar, ambas premisas son problemáticas y evidentemente la

[45] Otra forma de decirlo: *"Dime todo lo físico que se pueda decir acerca de lo que sucede en un cerebro vivo, de qué tipo son sus estados, cuál es su rol funcional y su relación con lo que ocurre en otras ocasiones y en otros cerebros, y así sucesivamente cosas por el estilo, y supongamos además que yo soy lo suficientemente listo como para hacer que todo ello encaje: no me habrás dicho nada del carácter doloroso del dolor, ni de lo que se siente con un picor, ni del tormento de los celos, ni tampoco nada en absoluto de las experiencias características de gustar un limón, oler una rosa, oír un fuerte sonido o mirar el cielo*" (Jackson, 1982: 127).

[46] Aparte de Fred y Mary de Jackson, véase por ejemplo: el experimento mental acerca de Mark de Nagasawa (2002), de Marianna de Nida-Rümelin (1998), de Jones de Horgan (1984), de Harpo de Robinson (1982), del Dennett's Restaurant de Zoglauer (1999), etc. Antes del experimento mental de Jackson se generaron otros experimentos con la finalidad de negar el fisicalismo, véanse: el experimento del arcángel matemático de Broad (1925: pp. 71-2), el experimento del marciano visitante de Farrell (1950: pp. 183), el experimento acerca de la carencia del repertorio de experiencias terrícolas familiares de los marcianos de Feigl (1967: 139-40). Una excelente ampliación acerca de la relación de estos y otros experimentos mentales en Stoljar- Nagasawa (2003).

conclusión no se sigue. Y también los hay que dicen que aun en el caso de que Mary aprendiera algo nuevo, esto no es razón suficiente para refutar el fisicalismo (Churchland, 1985; Lewis, 1983, 1990; Nemirow, 1980, 1990; Tye, 1986; Horgan, 1984; Lycan, 1990; Loar, 1990; Van Gulick, 1985, 1993; Levin 2008; Papineau 2002, 2007 y muchos otros). Estos deben aceptar la primera premisa, si no quieren acabar donde los primeros y aceptan con importantes matices la segunda premisa, matices que les llevará a rechazar la conclusión. Diré algo sobre los primeros y posteriormente me centraré en el segundo grupo de réplicas.

Aunque la mayoría de filósofos conceden que Mary aprenderá algo cuando salga de su cautiverio, tanto Paul Churchland como Daniel Dennett han llegado a rechazar que esto sea necesariamente así. Según Churchland la afirmación de que *Mary no puede imaginar cómo sería una explicación relevante, a pesar de su detallado conocimiento neurocientífico, y que por consiguiente tiene que estar perdiéndose cierta información crucial* (Churchland, 1985: p 25), bien podría ser falsa. La cuestión aquí es que nadie sabe lo que sucedería en caso de que la primera premisa del argumento de Jackson fuera verdadera, pues nadie hasta el momento ha conseguido tener un conocimiento neurocientífico completo ni de nuestra visión ni de ninguna otra cosa. Podríamos imaginar que en este caso Mary aprendería incluso a conceptualizar su propia vida interna incluyendo su propia introspección (p 25). Afirmar por tanto que un conocimiento neurocientífico completo no proporcionaría a Mary toda la información existente, es en este caso tan gratuito como afirmar lo contrario, pues en definitiva nadie hoy día puede saber con certeza lo que sucedería si se diera el caso. En la misma línea que Churchland, y casi como único apoyo a esta argumentación, Dennett afirma también que la premisa clave es que *ella posee toda la información física* (Dennett, 1995: 408), algo difícil de imaginar y que por tanto nadie se molesta en hacerlo en toda su amplitud, nadie profundiza en las posibles consecuencias que se derivan de tan (en apariencia) inocente premisa. Quizás se nos olvida que lo que hoy sabemos sobre la neurofisiología del color es solo la punta de un inmenso iceberg, quizás nuestra imaginación no da todavía como para comprender la magnitud de lo que significa poseer toda la información física de algo tan complejo como nuestra neurobiología de la visión. Es posible, y nadie puede objetar en contra de esto (aunque tampoco nadie sea capaz de probarlo), que Mary no aprendiera absolutamente nada. A fin de cuentas no solo sabe mucho acerca de la neurofisiología del color (que es lo que retiene el lector del argumento aun cuando lo lee detenidamente), para que el argumento sea válido deberemos de considerar lo que se dice, y lo que se dice es que Mary lo sabe TODO, y saberlo todo acerca de la neurofisiología de la visión, es algo que sin duda supera los límites de nuestra imaginación, la del lector, la mía, la de Jackson y la de cualquier neurofisiólogo especializado en la visión humana[47]. Sin embargo, Jackson insiste en que su argumento del conocimiento es válido ya que sus premisas son plausibles, *pero indemostrables* (Jackson, 1986: 294). A mí esto me parece suficiente

[47] Esta idea está perfecta y elocuentemente expuesta en Dennett (1995: 408 y ss. y 2006: 123 y ss.)

como para aceptar que no es un argumento con el que se concluya definitivamente que el materialismo sea falso, no es legítimo llegar a la verdad únicamente partiendo de la plausibilidad[48]. Sobre este argumento de Churchland y Dennett sencillamente no hay una contrarréplica eficaz pues es un argumento que se sitúa en el mismo nivel de plausibilidad que el de Jackson. No hay forma de demostrar que sea de una manera o de la otra, se trata de un experimento mental, y como tal, de él solo podemos extraer suposiciones, no verdades irrefutables. Lo único que puede hacer Jackson es insistir en que su argumento es filosóficamente plausible, pero en ningún caso definitivo. Aún quedaría mucho por hacer para refutar definitivamente el materialismo.

También Churchland considera el siguiente problema: el argumento podría cometer el mismo error si en vez de ser construido como una refutación al materialismo se construyera como una refutación al dualismo. Si Mary lo supiera todo acerca de una supuesta sustancia mental (Mary es una ectoplasmóloga), y lo supiera todo acerca de los procesos ectoplasmáticos que subyacen a la visión, (es decir, que lo supiera todo sobre los qualia), aún le quedaría algo por saber sobre la sensación de rojez, por lo que tampoco el dualismo sería una postura adecuada para explicar los fenómenos mentales, (Churchland, 1985: 24-25). Jackson rechaza esta réplica porque toda explicación de cómo son los qualia del color sin haber sido experimentados directamente en primera persona no puede ser explicado sino en términos fisicalistas. Además un argumento contra el dualismo solo puede ser construido si la primera premisa del argumento del conocimiento sustituye todo el repertorio sobre el materialismo por todo el repertorio sobre el dualismo, y esto a Jackson no le parece plausible (Jackson, 1986: 294-295). Pero si según Jackson uno no puede saberlo todo acerca del color sin haber visto nunca ninguno, tampoco podrá saberlo todo acerca del color aunque haya visto en primera persona todos los colores existentes. Aquí no entiendo en qué sentido Jackson ve plausible conocer todo sobre el materialismo y no sobre la parte no materialista del supuesto dualismo, en caso de que esto realmente fuera algo. De hecho, Jackson, con su argumento postula la existencia de ese algo, ¿qué problema habría en conocer todo lo cognoscible acerca de esa substancia no física? Aun con todo Churchland ofrece este argumento como subsidiario, y en su contrarréplica (Churchland, 1989: 4) insiste en que a pesar de todo, el argumento del conocimiento funciona tanto contra el materialismo como contra el dualismo de substancias, solo debemos substituir la primera premisa del argumento por la siguiente: (1) *Mary se convierte en una experta sobre una* (supuesta) *ciencia ectoplasmática* (se entiende como esa substancia no física sugerida por el dualismo) *de la naturaleza humana*[49].

Por mi parte considero que este primer grupo de argumentos no conducen a nada, todo permanece en el aire, no podemos saber con certeza lo que sucedería en caso de que alguien lo supiera todo acerca de

[48] El de la plausibilidad o la obviedad es sin duda uno de los principales problemas de los argumentos basados en experimentos mentales indemostrables, no podemos llegar a verdades a partir de ellos, ¡Hay tantas cosas plausibles y aparentemente obvias, pero falsas!

[49] Una explicación más extensa de este argumento en Nagasawa, 2002.

nuestra fisiología del color, tanto unos como otros nos dejan pendientes de lo que acontecerá en un futuro, pero de momento nos mantenemos en la incerteza. Además, no sirve de nada sugerir que tampoco el dualismo sale indemne del argumento de Jackson ligeramente modificado, hay argumentos mucho más poderosos para refutar el dualismo[50]. En este punto de la discusión, ya nada queda por hacer, nada más podemos esperar de la filosofía, solo queda cambiar de rumbo y ver qué podemos hacer si elegimos caminos alternativos.

Esos caminos alternativos son precisamente los que transitan la mayoría de filósofos. Mediante la posibilidad de que Mary no aprenda nada tras su cautiverio en ningún caso se llegará a una refutación definitiva del argumento de Jackson, si seguimos ese camino la partida queda en tablas, nadie gana, nos topamos ante un callejón sin salida. Debemos entonces, para poder continuar debatiendo, dirigirnos hacia otro lugar, conceder que Mary aprende algo nuevo al salir de la habitación, pero debemos cuestionarnos acerca de la naturaleza de lo que aprende, y no necesariamente deberemos extraer las radicales conclusiones a las que llega Jackson. No admitiremos, en principio, que esto sea una refutación definitiva del materialismo. La cuestión aquí se centra en el tipo de conocimiento que adquiere Mary tras su liberación. ¿Conocerá Mary nueva información, nuevos hechos o nuevas proposiciones acerca del mundo o simplemente adquirirá una nueva forma de conocer el mundo? ¿Se trata de un *saber que* o de un saber cómo? Tanto Nemirow como Lewis han explotado esta posibilidad. Piensan que el argumento de Jackson falla precisamente en este punto, Mary no adquiere un conocimiento proposicional sino una nueva destreza, una habilidad nueva para imaginar, reconocer e identificar colores (hipótesis de la habilidad[51]). Se trata de un *saber cómo* y no de un *saber qué*[52]. Lo que hace Lewis es un análisis exhaustivo de la hipótesis de la información fenoménica, para concluir que *si hay algo así como información fenoménica, esta no es independiente de la información física* (Lewis, 1983: 511). Mary solamente adquiere ciertas habilidades prácticas, *una habilidad para adquirir una información si se le proporciona otra información. Sin embargo, la información adquirida no es fenoménica, por lo que la habilidad para adquirir una información no es lo mismo que la información en sí misma* (p. 515), se trata de otra habilidad*, la habilidad para imaginar experiencias relacionadas nunca antes experimentadas* (p. 516). El error de Jackson es que al equiparar el acto de imaginar la experiencia de una cualidad con el acto de aprehender intelectualmente la cualidad misma no diferencia esos dos modos diferentes de conocer, confunde aprender un modo diferente de representación con un conocimiento nuevo sobre un hecho concreto (Nemirow, 1980: 492)[53]. Pero aceptar

[50] Pienso que Churchland ofrece este argumento como una forma más de refutar el dualismo de sustancias, pues no dice nada contra el argumento antifisicalista de Jackson. Argumentos mucho más potentes contra el dualismo de sustancias en la nota 18 de este trabajo.

[51] Algunos analizan esta nueva habilidad en términos de conocimiento directo (acquaintance): Conee (1994), Balog (2012b) y también Nemirow (1990).

[52] Esta distinción se la debemos a Ryle en Ryle (2005, cap 2).

[53] Según Nemirow si mantenemos esta distinción desaparece el problema: "*La ecuación de habilidad (ability equation) sugiere un análisis en términos de habilidad. La expresión 'X sabe cómo visualizar el rojo' o bien debe reemplazar o bien puede ser usada para parafrasear a 'X sabe a qué se parece la experiencia de ver rojo'. Este análisis desmitificaría la sub-expresión 'a qué se parece la experiencia de ver rojo' (what the experience of seeing red is like)" (*Nemirow, 1980: 494).

la hipótesis de la habilidad implica que Mary no adquiere nueva información acerca del mundo, implica que saber cómo qué es tener una experiencia es lo mismo que saber cómo imaginar tener esa experiencia, e implica finalmente que no es correcto analizar el acto de imaginarse teniendo una experiencia de un hecho como la aprehensión intelectual de dicho hecho (Nemirow, 1980: 495). Pues bien, todo esto es algo que muchos filósofos no están dispuestos a aceptar, así que aún daremos una vuelta de rosca más al argumento.[54]

¿Qué sucedería si en vez de suponer que Mary aprende una nueva habilidad, supusiéramos que lo que aprende es nueva información sobre hechos y proposiciones que ya conocía previamente? ¿Salvaríamos en este caso el fisicalismo? Algunos autores (Churchland, Tye y Horgan) conceden que hay un aprendizaje nuevo, que Mary adquiere nueva información, pero advierten que ni aun así quedaría refutado el fisicalismo, pues esa información no implica el aprendizaje de nuevos hechos o proposiciones sino el de un nuevo modo de aprehenderlas. A diferencia de la posición anterior, aquí no hablaremos de la adquisición de una nueva habilidad sino que nos cuestionaremos acerca del tipo de conocimiento que emplea Mary ante un mismo hecho. El razonamiento es que Mary adquiere ciertos hechos y proposiciones *de una forma diferente* de la que ya posee (Churchland, 1985: 24, Tye, 1986: 13). Es decir que Mary ha estado en el mismo lugar todo el tiempo, pero lo ha visto de formas diferentes, esto es, adquiere al salir de su habitación por introspección lo que previamente solo sabía indirectamente por inferencia (Van Gullick, 1993: 141). Consideremos la siguiente formulación del argumento de Jackson (Churchland, 1985: 23)[55]:

*(1) Mary sabe todo lo que se puede saber acerca de los estados cerebrales y sus propiedades.*

*(2) No es el caso que Mary sepa todo lo que se puede saber acerca de las sensaciones y sus propiedades.*

*Por consiguiente, y por la ley de Leibniz,*

*(3) Las sensaciones y sus propiedades no son idénticas a los estados cerebrales y sus propiedades.*

Churchland advierte que el "*saber acerca de"* en la primera premisa es un "*saber acerca de"* diferente del de la segunda. El conocimiento en (1) es una cuestión acerca de haber aprendido ciertas sentencias y proposiciones incluidas en textos sobre neurociencia, mientras en (2) se trata de un asunto acerca de tener ciertas representaciones pre-lingüísticas de rojez, verdor o lo que sea a través de una representación sensorial o discriminación sensorial (p 23). La diferencia radica en la naturaleza del conocimiento no en lo conocido en sí mismo, por lo que el argumento se podría formular como sigue (p 24):

[54] Para críticas de la hipótesis de la habilidad véase, Lycan, 1996; Loar, 1997; Kuna 2004, Alter 1998; Chalmers 1999; McConnell 1994; Robinson 1993; Conee 1994 y véase especialmente Tye 2004 para una aceptación de tal hipótesis, pero una aclaración sobre sus insuficiencias explicativas acerca del tema que nos ocupa. Argumentos a favor de tal hipótesis los encontramos además de en los citados, Lewis (1983, 1990) y Nemirow (1980, 1990), en Dennett (1995) o Meyer (2001).

[55] El caso es muy similar al del argumento de Nagel visto en el apartado anterior.

(1) *Mary ha aprendido un set completo de proposiciones verdaderas acerca de los estados cerebrales de las personas.*

(2) *Mary no tiene una representación de rojez en su medio de representación pre-lingüística para variables sensoriales.*

*Por consiguiente, y por la ley de Leibniz,*

(3) *La sensación de rojez es diferente de cualquier estado cerebral.*

Pero esta conclusión no necesariamente se sigue de esas premisas, pues desde el fisicalismo se puede admitir que uno pueda tener el conocimiento de una sensación independientemente del conocimiento científico que uno puede adquirir de ella. El cerebro emplea diferentes modos de representación, pero de ahí no se sigue que existan diferencias en lo representado. El argumento de Jackson falla de nuevo en el *conocer qué,* pero esta vez es más bien una nueva forma de *conocer acerca de*. La historia cambia cuando las diferencias se presentan únicamente en el modo de acceso a los hechos del mundo. En una línea similar se sitúa Horgan cuando nos invita a ver que *es enteramente posible que haya sentencias que expresen ontológica pero no explícitamente la información física* (Horgan, 1984: 150) y viceversa[56]. La diferencia entre poseer información explicita u ontológica es aquí la clave, esto es, que *aunque Mary, previamente a su primera experiencia de color, tuviera explícitamente un stock completo de la información física acerca de los procesos de la visión humana, no es legítimo inferir a partir de esto que tenga un stock ontológicamente completo de toda la información física* (151), por lo que tampoco en este caso concluiríamos que el fisicalismo es una posición insuficiente, pues la diferencia reside de nuevo en nueva información de un mismo hecho y no de un hecho diferente. También en este punto Tye[57] se muestra convencido de que Mary aprenderá algo al salir de su cautiverio, pero no aprenderá nada acerca de nuevos hechos, hechos de un tipo diferente de aquellos que ya conocía anteriormente. Es cierto que adquiere *experiencias* de un cierto tipo que no poseía previamente, es cierto que adquiere *información introspectiva* sobre ciertos hechos, pero no es cierto que aprenda nuevos hechos acerca del mundo, simplemente adquirirá conocimiento de un modo diferente al que ya poseía. Antes de ser liberada Mary solo sabía sobre ciertos hechos obtenidos a través de un conocimiento externo por la información adquirida en libros que contienen, eso sí, todo lo que se puede conocer sobre la neurofisiología del color, y al salir de la habitación

[56] Para Horgan hay dos sentidos relevantes de información física: "*Sea S una oración que expresa información acerca de procesos de una cierta clase específica, tales como procesos perceptuales humanos. Diremos que S expresa información explícitamente física en caso de que S pertenezca a, o se siga de, una explicación teóricamente adecuada de esos procesos. Y diremos que S expresa información ontológicamente física en caso de que (i) todas las entidades referidas o cuantificadas en S sean entidades físicas, y (ii) todas las propiedades y relaciones expresadas por los predicados en S sean propiedades y relaciones físicas. Así, la información explícitamente física se expresa en el lenguaje abiertamente fisicalista, mientras que la información ontológicamente física puede expresarse en otras clases de lenguaje – por ejemplo, el lenguaje mentalista*" (Horgan, 1984: 150).

[57] Tye desarrolla un ejemplo similar al de Mary de Jackson, se trata de Jones. Ciertamente debemos agradecer a Tye este ejemplo, pues al introducir el caso de una persona ciega de nacimiento que recupera la vista por medio de una operación quirúrgica nos evitamos decenas de discusiones metodológicas sobre la situación de Mary en el interior de su habitación y su empobrecido entorno visual. Sin embargo continuaré con Mary por el bien de la claridad de la argumentación.

sabe algo más, algo más sobre los mismos hechos a través de su conciencia introspectiva. Hasta ese momento Mary simplemente era una ignorante en cuanto al acceso a determinados hechos a través de su contenido fenoménico, a partir de ese instante aprende como son las cosas fenomenológicamente (Tye, 1986, p. 9-10). Pero insistamos, se trata de una nueva forma de conocimiento, se trata del conocimiento de un tipo de experiencia en particular, pero no se trata del conocimiento de un nuevo hecho (p. 13). Según estos autores, finalmente, Mary, al salir de la habitación, no se sorprende en absoluto al percibir la rojez de los tomates o el azul del cielo, no hay nada nuevo en su repertorio, con su vasta imaginación, su conocimiento completo sobre las bases neurofisiológicas del color y todo lo que le han contado acerca de *cómo qué* es percibir un color, en todo momento se da cuenta de que no aprende nada que no hubiera aprendido ya, no hay nada nuevo, ya había estado ahí, solo aprende una nueva forma de adquirir el mismo conocimiento que ya poseía[58].

Pero aún podríamos hacer un análisis diferente del nuevo aprendizaje obtenido por Mary. Volveremos a dar una vuelta de rosca más al asunto. Lycan, Van Gulick, de nuevo Tye y sobre todo Loar, sugieren lo siguiente: lo que Mary adquiere al salir de la habitación no es tan solo una habilidad nueva, ni es solo un modo diferente de conocer lo que ya conocía, lo que adquiere Mary es conocimiento sobre nuevas proposiciones que previamente desconocía. Sin embargo, cabría distinguir diferentes usos en la noción de proposición, distinción que nos permitirá encontrar una salida para salvar el fisicalismo. Esto es lo que persigue Van Gulick cuando distingue entre proposiciones con modos de individuación de grano-fino y de grano-grueso. Cuando hablamos de modo de individuación de grano grueso nos referimos a una caracterización tosca de las proposiciones, se trata de funciones con valor de verdad en cualquier mundo posible. En este modo de individuación, la proposición *5+7 es igual a 12* es la misma que *1444 es el cuadrado de 38*, y la proposición *el agua hierve a 100$^{o}$C* es la misma que e*l $H_2O$ hierve a 100$^{o}$C*, todas ellas son verdaderas en cualquier mundo posible. Sin embargo, uno puede hacer una caracterización más fina de las proposiciones considerándolas entidades estructuradas compuestas de conceptos que deben a su vez encajar para que dos proposiciones sean idénticas (Van Gulick, 1993: 141-142). Es decir, uno puede individualizar las proposiciones más finamente distinguiendo entre ellas los términos de su estructura intensional interna, más específicamente, en términos de conceptos incorporados que determinan respectivamente sus funciones de grano grueso (Van Gulick, 2004: 20-21). La cuestión es que según Van Gulick Mary adquiere nuevas proposiciones en ese modo de individuación de grano fino, pero al salir de su habitación no aprende ningún valor de verdad dado en todo mundo posible como los expuestos anteriormente, sino que lo que aprende es parte de la estructura interna de dichos valores de verdad, por lo que, si esto es así, el fisicalismo no quedaría refutado. Algo similar sugiere Tye cuando dice que Mary (antes de ser liberada) no sabe cómo qué es tener la experiencia de rojo, no conoce el concepto

[58] Las principales críticas a esta posición las encontramos en Jackson (1986); Lewis (1990) y Graham y Horgan (2000).

fenoménico del estado de experimentar el rojo, después de todo nunca ha experimentado la sensación de ver algo rojo. Y no lo sabe por dos razones. Primero porque carece del concepto fenoménico (predicativo) de rojo, y segundo porque no es capaz de aplicar este concepto fenoménico (indexical) al color representado por la experiencia de rojo. Pero no hay nada de tipo no conceptual que Mary no sepa. Lo que no conocía y ahora conoce (al salir de su habitación) es una proposición de grano-fino bajo el cual caen determinados conceptos fenoménicos. Lo que aprende es una determinada concepción de unos hechos determinados, sin embargo, no adquiere ningún estado de cosas del mundo objetivo, esto es, no adquiere ningún conocimiento de grano-grueso, no adquiere ningún hecho no conceptual de contenido ampliamente físico (Tye, 1995: 173-174). También Lycan nos cuenta algo similar, afirma (tras nada menos que nueve argumentos en favor de que Mary aprende algo más que una simple destreza) que Mary adquiere un nuevo conocimiento tras su liberación. Para Lycan "*S sabe lo que se siente al ver rojo*" significa algo así como que "*S sabe que se siente Q al ver rojo*" donde Q es la cualidad fenoménica en cuestión. Cuando Mary experimenta el color rojo incorpora semánticamente un término mental primitivo de primer orden, término que desempeña un rol conceptual en el lenguaje interno de Mary, y que es único en el sentido en que no hay otro cuya representación sea funcionalmente similar. De algún modo el término introspectivo (fenoménico) no es sinónimo de ninguna expresión primitiva en ningún lenguaje dado (Lycan 1996: 101), por lo que también aquí adquiere relevancia el modo de presentación del conocimiento adquirido, parece que no hay problema en admitir que Mary adquiere un nuevo conocimiento, un conocimiento de carácter fenoménico que a su vez cae bajo algún concepto fenoménico, pero eso sí, los hechos que componen dicho concepto no son tan desconocidos para ella, como a algunos les gusta creer.

Pero el filósofo que ha llegado más lejos con esta *estrategia fenoménico-conceptual* es sin duda Brian Loar. Para Loar el fisicalista debe aceptar la intuición de Jackson: definitivamente, Mary, antes de ser liberada, no sabe lo que es experimentar un color incluso aunque sepa todos los hechos físico-funcionales relevantes que hay por saber, Mary aprende algo nuevo acerca de nosotros, aprende un nuevo hecho o verdad acerca del mundo, pero solo podemos hacer esta concesión desde una lectura opaca[59] de "*Mary aprende que nosotros tenemos tales y tales experiencias de color"* y en las correspondientes lecturas de "*aprende un nuevo hecho o verdad acerca de nosotros"* (Loar, 1997: 598). Esto es, el conocimiento concebido en términos físico-funcionales no puede bastar a priori para el conocimiento concebido en términos fenoménicos, sin embargo, esto no implica que las cualidades fenoménicas sean algo diferente a las cualidades físico-funcionales, pues podemos aceptar que conceptos fenoménicos y conceptos físico-funcionales simplemente introducen diferentes modos de presentación de los mismos estados. La clave en el argumento de Loar es la distinción entre conceptos y propiedades. No se pretende negar el poder de la intuición anti–fisicalista, sino dar una explicación de la relación entre conceptos fenoménicos y

[59] En contextos opacos las oraciones vienen regidas por un verbo de actitud proposicional, (creer, desear, pensar, suponer…), y nada nos puede garantizar su verdad.

propiedades físicas que alivie esta "*ilusoria intuición metafísica*" (p. 598). Para Loar el argumento del conocimiento depende de dos asunciones, primero, la ya citada independencia conceptual entre los conceptos fenoménicos y los conceptos físico-funcionales, que una vez llegados hasta aquí deberemos aceptar, y segundo, lo que llama la *premisa semántica,* que deberemos rechazar (p. 600)[60]. Por premisa semántica entenderemos lo siguiente: *una declaración de identidad de propiedad que une conceptualmente conceptos independientes es verdadera sólo si al menos un concepto identifica la propiedad a la que se refiere connotando una propiedad contingente de esa propiedad* (p. 600). Pero esto significa que si dos tipos de propiedades son idénticas, debe poderse saber a priori. O recíprocamente, si dos tipos de propiedades no están relacionados a priori, entonces no pueden ser idénticas. Esto nos lleva a afirmar que la única manera de tener en cuenta el estado a posteriori de una verdadera propiedad de identidad es que uno de los términos exprese un modo contingente de presentación, pero si un concepto fenoménico puede identificar una propiedad física directa o esencialmente (y no mediante un modo contingente de presentación), y aún ser conceptualmente independiente de todos los conceptos físico-funcionales que refieren a la misma propiedad física, entonces el argumento de Jackson deja de ser eficaz (p. 600). Es precisamente en este punto donde el argumento del conocimiento falla, no tiene en cuenta esta distinción entre individuación de conceptos e individuación de propiedades.

Loar sugiere que una vez que reconozcamos las características especiales de los conceptos fenoménicos ya podremos *tomar la intuición fenoménica con su valor nominal, aceptando los conceptos introspectivos* (fenomenales) *y su irreductibilidad conceptual, y al mismo tiempo tomando las cualidades fenoménicas como idénticas a las propiedades físico-funcionales de la especie previstas por la ciencia del cerebro contemporánea* (p. 597). Resumiendo: los conceptos fenoménicos son conceptos de reconocimiento directo, esto es, *S posee un concepto fenoménico C de la experiencia E solo si S tiene ciertas disposiciones a reconocer, discriminar e identificar E, en base a su introspección.* Dos conceptos pueden ser conceptualmente independientes y a la vez ser co-extensivos (véase agua y $H_2O$ o dolor y estimulación de las fibras C), por lo que aunque un concepto fenoménico pueda ser diferente de un concepto físico-funcional esto no significa que las propiedades fenoménicas sean algo diferente de las propiedades físico-funcionales. Por tanto, el error del argumento de Jackson es que al mezclar conceptos y propiedades comete el error de llegar a conclusiones metafísicas a partir de contextos opacos (p. 598). Admitiremos la intuición de Jackson, pero no aceptaremos su radical conclusión, porque aun en el caso de que Mary aprenda algo al salir de su cautiverio, no es cierto que aprenda hechos cuyas propiedades sean distintas de las de los hechos que poseía en todo su repertorio de conocimientos anterior a su liberación.

[60] Según Loar hay dos tesis funcionalistas: *"que todos los conceptos de estados mentales son conceptos funcionales, y que todas las propiedades mentales son propiedades funcionales. La primera la rechazo, pues acepto la intuición antifisicalista: estoy de acuerdo con el antifisicalista en que los conceptos fenoménicos no pueden ser capturados en puros términos funcionales. Sin embargo, nada en filosofía impide a las propiedades fenoménicas ser propiedades funcionales*" (p. 613).

Desde estas posiciones se acepta un monismo en cuanto a la ontología de la conciencia y un dualismo en cuanto a los conceptos que usamos para referirnos a esas entidades (Papineau, 2002: 47). Se trataría de un caso estándar de correferencia en el que dos conceptos distintos tienen una misma referencia, pero esto no es problemático para el fisicalismo, se puede aceptar incluso un relativismo conceptual y al mismo tiempo reconocer la sustancia física como única sustancia en el mundo. Nada malo hay en que el fisicalista acepte una realidad conceptual dual compuesta por dos modos diferentes de referirnos a las propiedades del mundo, llamémoslas fenoménicas o físico-funcionales. Lo que el fisicalista no puede estar dispuesto a aceptar es algo diferente a un monismo ontológico.

Hasta donde puedo ver, esta última opción presentada por Loar (1997) y apoyada por Lycan (1990), Tye (1995) y VanGulick (1993) es la más elaborada de todas las que tratan de deshacerse de las conclusiones anti-fisicalistas del argumento de Jackson en base al análisis de los conceptos fenoménicos[61]. Un defensor de los conceptos fenoménicos debería en principio aceptar los siguientes criterios:

El concepto C es un concepto fenoménico solo si:

1. Existe alguna experiencia fenoménica tipo E, y alguna propiedad P, de tal manera que la experiencia dada cae bajo E en virtud de su relación con P.

2. C se refiere a P.

3. Bajo circunstancias normales, un ser humano puede poseer C solo si ha tenido alguna vez una experiencia de tipo E.[62]

El defensor del argumento del conocimiento debe reconocer la existencia de conceptos fenoménicos. El hecho de que Mary adquiera conocimiento de un nuevo hecho o una proposición nueva implica que antes de su liberación, hay algún concepto que no podía incorporar a su repertorio. Además, si queremos aceptar que hay algún tipo de conocimiento concerniente a hechos acerca de la visión humana que desconoce antes de su liberación y que aprende al salir de su cautiverio, y aceptamos los criterios expuestos anteriormente sobre lo que es un concepto fenoménico, entonces el contenido de lo que incorpora Mary a su repertorio implica necesariamente la existencia de un concepto fenoménico.

El crítico del argumento del conocimiento, esto es, el defensor del fisicalismo del tipo de los que aceptan que Mary aprenderá algo tras su cautiverio, también parece requerir de la existencia de conceptos fenoménicos. Como hemos visto anteriormente podemos incluso aceptar toda la argumentación de

[61] No entraré en las diferencias entre estos autores, que las hay, principalmente con respecto a los diferentes usos de nociones como concepto fenoménico, carácter fenoménico o la relación entre ambas. Lo importante es señalar que sus posicionamientos con respecto al tema que nos ocupa son lo suficientemente similares como para colocarlos en el mismo bando. Todos ellos basan sus análisis en la noción de concepto fenoménico, todos admiten que Mary aprende algo al salir de su cautiverio, rechazan que su aprendizaje sea el de una simple habilidad, y no ven en el argumento ningún perjuicio para el fisicalismo.

[62] Criterios aceptados en (Tye 1995, p. 169) (Papineau 2007, pp. 126-7) o (Harman 1990, p. 670) por ejemplo.

Jackson pero no seguir su conclusión al distinguir entre presentaciones de grano fino y de grano grueso y aceptar que es posible que varios contenidos, varios conceptos, refieran a un mismo hecho o a una misma propiedad. Todo esto implica la existencia de conceptos fenoménicos que caen dentro de los criterios expuestos anteriormente.

Parece entonces que postular la existencia de conceptos fenoménicos es una estrategia atractiva tanto para el defensor como para buena parte de los críticos del argumento del conocimiento. Pero, la noción de concepto fenoménico es, de hecho, una noción que pide a gritos una aclaración, necesitamos saber dónde se encuentran sus límites, por lo que aún nos podríamos cuestionar sobre qué sucedería si negáramos la existencia de dichos conceptos. ¿Y si resulta ser una de esas nociones que tanto gusta a los filósofos dar vueltas y más vueltas pero que finalmente acaban por no referirse a nada concreto? ¿Qué sucedería si negáramos su existencia? Pues que quizás el argumento del conocimiento se derrumbaría, y se llevaría con él a todos los argumentos que siguen la estrategia fenoménico-conceptual. No entraré aquí en esta discusión, pues va más allá de mis expectativas en este apartado, sin embargo, es una posibilidad que no quería dejar de comentar.[63] En caso de que todo esto fuera cierto y no existiera algo así como conceptos fenoménicos, ¿en qué situación nos encontraríamos con respecto al argumento de Jackson? Solo veo dos posibilidades, o aceptar la estrategia de Dennett y compañía, esto es, Mary no aprende nada, porque no hay nada que aprender, o la de Lewis y Nemirow, esto es, que solo aprende una habilidad nueva, ni nuevos hechos, ni nuevos conceptos. Pero aunque este último argumento fallara, la única salida que le queda al argumento del conocimiento para refutar definitivamente el fisicalismo pasa por aceptar que Mary aprende algo nuevo al salir de su cautiverio, que lo que aprende no es un mero *saber cómo* sino un *saber qué*, que es un aprendizaje de nuevos hechos y proposiciones acerca del mundo que no poseía previamente y que esos nuevos hechos y proposiciones de ninguna manera son individuaciones de grano fino sino nuevos hechos acerca del mundo en un modo de individuación de grano grueso. Y no solo eso. En el supuesto caso de que Mary al salir de su habitación se encontrara con un conocimiento sobre propiedades no físicas del mundo, éstas estarían pidiendo a gritos una explicación, y hasta el momento, nadie ha sido capaz de dar ni una sola pista de qué tipo de propiedades son esas. El argumento del conocimiento se está viniendo abajo, demasiadas suposiciones, ahora solo falta darle el golpe definitivo. El propio Jackson parece tomar conciencia del lio que ha montado, se da cuenta de que lleva demasiado tiempo encerrado en la habitación con Mary y de que con su ocurrencia ha creado un monstruo que aún hoy día, más de treinta años después, permanece vivo, finalmente, los argumentos fisicalistas le están

[63] Al que le interese profundizar en este tema haría bien ojeando los siguientes artículos: A favor de la existencia de conceptos fenoménicos Balog, (1999, 2012a); Loar, (1997); Lycan, (1996); Papineau, (1993, 2007); Tye, (1995, 2000); Perry, (2001); y en contra Prinz (2007); Ball (2009); Tye, (2009) o Fazekas (2011) quien no niega su existencia pero tampoco les concede un lugar preeminente en la explicación de los qualia. (A juzgar por los años de las citas encontradas la idea de que no hay algo así como conceptos fenoménicos está cobrando bastante fuerza). Cambios de parecer tan bruscos como el que realiza Tye es digno de ser analizado, cuando alguien cambia de opinión tan bruscamente, o no estaba tan seguro de lo que decía o ha encontrado algo digno de ser escuchado. Las críticas a la estrategia de los conceptos fenoménicos serán analizadas más adelante, cuando veamos el argumento de la brecha explicativa.

cerrando el cerco hasta tal punto que no hay forma de salir de ahí. Así es que Jackson parece alistarse a las filas del bando contrario, se hace fisicalista. Pero lejos de seguir el camino seguido por el aluvión de ataques recibidos tras la publicación, en 1982, de su polémico artículo, lo que hace es sugerir una nueva manera de adherirse el fisicalismo, sigue una estrategia bien distinta. El cambio de parecer de Jackson pasa por aceptar la teoría representacional fuerte de la experiencia sensorial consciente (TRF)[64]. Esta es, para Jackson, la única vía posible para desactivar de manera tajante el argumento del conocimiento. El argumento pasa por la idea de que estamos bajo los influjos de una ilusión acerca de la naturaleza de la experiencia de color, y que esa ilusión es la que despierta en nosotros esa intuición epistémica que nos impide deducir *el como que es tener una sensación de color* desde una explicación meramente física del mundo (Jackson, 2003: 426). La respuesta del fisicalista a la intuición epistémica debería autorizarnos a ver como la naturaleza de la experiencia de color podría seguir a priori una explicación física de como es el mundo, y la explicación representacional de la experiencia sensorial encuentra las herramientas necesarias para hacerlo (p 431). Pero, ¿qué hay de especial en la representación cuando algo se ve o se siente de determinada manera? Jackson ve hasta cinco rasgos distintivos en los que la experiencia sensorial *representa* las cosas tal y como son (pp 436-439):

Primero, la riqueza de la representación. La experiencia visual representa como las cosas son aquí y ahora en términos de color, forma, localización, extensión, orientación y movimiento.

Segundo, la representación es inextricablemente rica. Somos capaces de aislar conceptualmente cada uno de los rasgos de un objeto aun cuando no seamos capaces de fragmentarlos en nuestra experiencia visual, cuando decimos que *X es rojo* estamos usando sentencias para representar algo acerca del color mientras permanecemos en silencio acerca de su forma, y lo mismo sucede cuando decimos que *X es circular*. Sin embargo, en nuestra experiencia perceptual no hay forma de separar la rojez de la circularidad.

Tercero, la representación es inmediata.

Cuarto, hay un elemento causal en el contenido. La percepción es la representación del mundo en interacción con quien lo representa. Cuando escucho un sonido detrás de mí y a mi izquierda, mi experiencia representa el sonido como procedente de esta ubicación. Sentir algo es sentirse parte de su contacto con el cuerpo.

Y por último, la experiencia sensorial desempeña un papel funcional distintivo en la mediación entre un estado de creencia y otra. No es en sí un estado de creencia, pero representa lo que determinará una función que irá de unos estados de creencia a otros.

[64] Es la visión según la cual estar en un estado fenoménico es representar propiedades objetivas, donde las propiedades objetivas así como la representación en sí misma son accesibles mediante una explicación fisicalista.

Finalmente Jackson acaba reconociendo que tras su liberación Mary adquiere estados representacionales con todas estas propiedades. El haber confundido propiedades intensionales con propiedades instanciadas nos ha llevado a cometer el error de pensar que Mary tras su liberación había adquirido un nuevo hecho acerca del mundo (p. 440). Pero no es así. Desde una visión representacional de la experiencia sensorial consciente es posible dar cuenta de lo que le sucede a Mary desde un punto de vista meramente físico y funcional.[65]

Como hemos podido ver, el argumento del conocimiento ha sido atacado por todos los flancos, incluso finalmente hasta por su creador, sin embargo se han ofrecido otros argumentos anti-fisicalistas muy sofisticados, muchos de ellos inspirados en el argumento de Jackson. Estos se basan en la concebibilidad. Se trata de argumentos como los del espectro invertido, la tierra invertida o la posibilidad lógica de zombis filosóficos, o lo que es lo mismo argumentos en base a la posibilidad de qualia invertidos o qualia ausentes. Los veremos en los siguientes apartados.

**Qualia invertidos**

Hace ya más de tres siglos John Locke propuso un argumento que constituye uno de los más destacados obstáculos para los actuales planteamientos funcionalistas y materialistas en filosofía de la mente: el experimento mental del *espectro invertido*[66]. En líneas generales es la idea según la cual un estado mental M1 puede ser concebido como distinto de M2 aun cuando no haya diferencias funcionales entre ambos; por tanto, los estados mentales son distintos de los estados funcionales ya que hay un aspecto mental cualitativo que no permite identificar los estados mentales con estados funcionales.

Actualmente existen dos versiones del experimento: la intersubjetiva y la intrasubjetiva. Según la versión intersubjetiva, dos individuos idénticos conductual y funcionalmente podrían tener experiencias cromáticas sistemáticamente invertidas a pesar de reaccionar de forma idéntica ante los mismos estímulos. Por ejemplo, ante la pregunta acerca del color de la hierba ambos responden verde, a pesar de que uno la ve roja y el otro verde. La idea es que a pesar de que dos personas sean exactamente iguales en su constitución física y en sus disposiciones para actuar, no hay manera de verificar (ni de rechazar) la idea

[65] Para una crítica de la estrategia de Jackson véase Alter (2006) o Van Gulick (2009).

[66] Locke, J. (1690): *Ensayo sobre el entendimiento humano*. México, D. F.: Fondo de Cultura Económica, 2005. Libro II, XXXII, nº 15. El caso que comenta Locke dice así: *"que la idea que la violeta produce en la mente de un hombre a partir de su vista fuese la misma que se produce en la vista del otro por una caléndula, y viceversa*". Locke afirma además que esto jamás podría conocerse: *"Porque como todas las cosas que tuvieran la misma textura de una violeta producirían de un modo constante la idea que uno de esos hombres denominaría azul, y aquellas otras que tuvieran la textura de una caléndula igualmente provocarían la idea que constantemente ha denominado amarillo, con independencia de las apariencias que tuviesen, podría distinguir, a partir de su utilización, con la misma constancia, las cosas que tuvieran esa apariencia, del mismo modo que podría entender y significar esas distinciones señaladas por las palabras azul y amarillo, como si las ideas que en su mente recibe de esas dos flores fueran exactamente las mismas que las de las mentes que otros hombres reciben" (ibid.).* Lo que Locke está apuntando es la dificultad existente a la hora de determinar si dos sujetos conscientes en condiciones normales perciben la misma experiencia subjetiva ante la misma sensación.

de que se encuentran en posesión de estados cualitativos diferentes. En la versión intrasubjetiva, la inversión fenoménica se produce en un mismo individuo: un buen día uno se despierta por la mañana y comprueba que *ve* de color rojo las mismas cosas que el día anterior eran de color azul; abre la ventana y comprueba que el cielo tiene un extraño color rojo, que la manzana que hay sobre su mesa la ve de color azul y que el resto de los colores le *parecen* diferentes, en definitiva, que sus experiencias cromáticas están invertidas. Tras un proceso de adaptación comienza a invertir el nombre de los colores tal y como se presentan en su experiencia subjetiva, y hace exactamente lo mismo con el resto de sus experiencias cromáticas y su conducta verbal sobre las mismas, es decir termina por decir que el cielo es azul aunque lo vea rojo y las manzanas rojas aunque las vea azules. El resto de sus disposiciones reactivas se normalizan, al igual que su conducta verbal, tras una etapa de adaptación. Ambas versiones se basan en la intuición de que es concebible una inversión fenoménica manteniéndose constante el contenido funcional e intencional del sujeto o de los sujetos. Se concluye que las representaciones mentales tienen algo más que contenido intencional, tienen contenido cualitativo, pues la inversión de dicho contenido cualitativo es capaz de convivir con un isomorfismo funcional. Cabe advertir que la inversión intrasubjetiva conduce finalmente a la intersubjetiva, *si la inversión intrasubjetiva es posible, la inversión intersubjetiva también debe ser posible* (Shoemaker, 1982: 359). Si suponemos que todo el mundo tiene la misma percepción del color y alguien es sometido a inversión cromática, asumiendo que nadie más lo ha sido, parece evidente pensar que la experiencia de color de ese alguien será radicalmente diferente de la de los demás, por tanto nos encontraremos también ante un caso de inversión intersubjetiva. Sin embargo, la inversión intersubjetiva no implica inversión intrasubjetiva, pues bien podría suceder que dos individuos nacieran ya con sus qualia invertidos y que ninguno de ellos lo supiera.[67] Sea como fuere, el objetivo de estos argumentos es hacer evidente que lo fenoménico no puede reducirse a lo funcional, y en definitiva que no es posible capturar a los *qualia* a través de una explicación funcional, por lo que el funcionalismo y por extensión el materialismo son insuficientes para explicar la mente humana[68]. Pero si esta insuficiencia explicativa ha de ser admitida por la posibilidad o concebibilidad del argumento, podríamos admitir también la posibilidad de que dos estados sean funcionalmente equivalentes y que uno de ellos tenga rasgos fenoménicos y el otro carezca de ellos. Este sería el argumento de los qualia ausentes, desarrollado por Block (1978). Imaginemos que se conecta a todos los habitantes de China mediante equipos de radio

[67] La introducción de la inversión intrasubjetiva supuso cambios sustanciales en el argumento original. Si mediante la inversión intersubjetiva no es posible verificar la inversión del espectro cromático, pues nunca podremos saber si mis experiencias de color son idénticas a las tuyas, la intrasubjetiva si permite dicha verificación, ya que es a partir de un momento t que las experiencias cambian, hay un antes y un después en la inversión, por tanto podrían ser verificadas por el mismo individuo mediante introspección.

[68] Ahí va una construcción del argumento del espectro invertido para la versión anti-funcionalista:

- P) Es posible que dos sujetos observando un objeto rojo, uno con su espectro invertido y el otro no, sean funcionalmente iguales. Por consiguiente lo mental no superviene a la organización funcional:
- C) Por tanto, el funcionalismo es falso.

Lo mismo funcionaría contra el fisicalismo:

- P) Es posible que dos sujetos observando un objeto rojo, uno con su espectro invertido y el otro no, sean físicamente iguales. Por consiguiente lo mental no superviene a la composición física
- C) Por tanto, el fisicalismo es falso.

de tal modo que simulen la organización funcional que presenta un cerebro, asignándole a cada individuo una función concreta dentro del sistema representado. Imaginemos que se les proporciona una entrada sensorial a la que responden funcionalmente con una salida a partir de la computación pertinente de los datos manejados. El sistema de los ciudadanos de la nación china cumplirían, según el experimento mental, los requisitos que el funcionalismo estipula para la atribución de estados mentales, pero de ellos es obvio que estaría ausente cualquier rasgo fenomenológico[69] (Block, 1978: 279 y ss.). Desde la perspectiva funcionalista, concluiríamos que la experiencia consciente no introduciría diferencias de ningún tipo, pero esta sería, para muchos, una conclusión absurda. La conclusión de Block es que, hoy por hoy, los qualia quedarían por completo fuera del dominio de la psicología. Pero aquí entramos en terreno peligroso, evidentemente nadie sabe lo que sucedería en el caso de que se diera algo así como la conexión entre los millones de ciudadanos de la nación china, si de ahí surgiera de verdad una mente sería algo asombroso, (no creo que a nadie le resultara algo siquiera inteligible), pero lo curioso es que no deja de ser igual de asombroso el hecho de que de un grupo de miles de millones de neuronas interconectadas y tras varios millones de años de evolución, surja algo así como una mente consciente.[70]

El mismo Block, desde su posición realista y cuasi-funcional[71], ha ofrecido un nuevo argumento en favor de la existencia de un aspecto no intencional de la mente basado en el experimento mental de la Tierra Invertida (Block, 1990)[72]. Es el inverso del espectro invertido, es decir, un caso de contenido intencional invertido y contenido cualitativo idéntico. Es un argumento en el que Block da una vuelta de rosca más al argumento del espectro invertido con la intención de anular definitivamente la pretensión de que el funcionalismo constituya una teoría que agota por completo la explicación de la mente humana. Veámoslo. Block nos invita a imaginar que a una de dos hermanas gemelas genéticamente idénticas se le colocan unas lentes que invierten el color y es enviada a la Tierra Invertida (un lugar idéntico a la Tierra pero con los espectros de color invertidos). Aquí encontramos una doble inversión: en primer lugar, los colores están invertidos respecto de los colores de la Tierra (el prado es rojo, la sangre es verde, etc.); y en segundo lugar, los significados de los términos de color están también invertidos respecto de los significados en la tierra (en el lenguaje Invertido, "rojo" significa verde, ''verde'' significa rojo, etc.). Ambas

[69] El caso es similar al de la posibilidad lógica de zombis (aunque como veremos hay diferencias importantes). Para una réplica interesante de este argumento véanse entre otros Van Gullick (1993: 145 y ss.), y Shoemaker (1981). Estos serán tratados en profundidad en el siguiente apartado.

[70] Churchland y Churchland (1981) advierten de que los argumentos de la nación china de Block necesitarían un sistema que manejara alrededor de $10^{30.000.000}$ de entradas en la retina y un número aún mayor de estados internos en el cerebro.

[71] Realista porque se compromete con la existencia de rasgos mentales intrínsecos de nuestra experiencia sin cometer *la falacia de intencionalizarlos*, y cuasi-funcional porque cree que el contenido cualitativo de la experiencia no es funcionalmente caracterizable (Block, 1990: 58).

[72] Los experimentos mentales de la Tierra Gemela fueron introducidos por primera vez por Putnam (1975) para defender una posición externista respecto del significado lingüístico. Poseen un elevado poder explicativo ya que una Tierra Gemela es un lugar contrafáctico idéntico a nuestra Tierra donde, *caeteris paribus*, suele existir una pequeña diferencia en relación al significado de algo concreto. En el ejemplo original de Putnam, por ejemplo, la diferencia crucial, que permite afirmar la posición externista del significado, es una diferencia en la composición química del agua que, sin embargo, no tiene relevancia a la hora de que los hablantes puedan referirse a ella de modo normal. Es crucial, también, la existencia de un *döppelganger*, esto es, un doble idéntico molécula por molécula a un sujeto original (Óscar Gemelo, en el ejemplo de Putnam).

inversiones tienen como consecuencia la cancelación del efecto de las lentes, de modo que durante un tiempo, la experiencia y la conducta de la gemela en la Tierra Invertida es idéntica a la experiencia y la conducta de su hermana en la Tierra. Por consiguiente, frente a los respectivos prados, ambas tienen una sensación de verde y dicen "prado verde", haciendo referencia, en ambos casos, a los prados verdes de la Tierra. Pero, cincuenta años después, la gemela enviada a la Tierra Invertida ha empezado a hablar el lenguaje de la Tierra Invertida, por lo que su emisión "prado verde" no hace referencia al prado verde de la tierra sino al prado rojo de la Tierra Invertida. Es entonces, cuando aunque sus respectivas experiencias continúen siendo idénticas respecto de su carácter cualitativo, pues ambas tendrán un *quale* verde, una se referirá a los prados verdes de la Tierra y la otra a los prados rojos de la Tierra Invertida, es decir, diferirán en su contenido intencional. Véase que el experimento puede formularse también para el caso intra-subjetivo comparando las sucesivas experiencias de una misma persona antes y después de recibir las lentes invertidoras del color y ser enviada a la Tierra Invertida. De este modo, se llegará a la misma conclusión: después de haber vivido en la Tierra Invertida durante cincuenta años, el factor intencional del contenido de la experiencia habrá cambiado, sin variación cualitativa. La conclusión es que hay dos aspectos diferentes de la mente, el aspecto cualitativo y el funcional, y el aspecto cualitativo no puede ser reducido a -o identificado con- el aspecto funcional. En definitiva, Block no tiene ninguna duda de que el funcionalismo psicofísico computacional no ha sido aún capaz de dar con una caracterización de los qualia en términos de "programa" de la mente. El argumento llega a la misma conclusión que los anteriores pero modificando, o mejor dicho, invirtiendo las premisas: un estado mental M1 puede ser concebido como idéntico a M2 aun cuando haya variación funcional entre ambos. Por tanto, dado que podemos concebir un aspecto funcional invertido sin concebir un aspecto cualitativo invertido, los aspectos funcionales son distintos de los cualitativos, y por consiguiente el funcionalismo no es capaz de explicarlo todo sobre la mente y por extensión el materialismo es falso.

Tenemos pues sobre la mesa los siguientes casos: (i) que dos sistemas conscientes pueden compartir la misma organización funcional y, sin embargo, tener sus experiencias invertidas el uno con respecto del otro (qualia invertidos intersubjetivamente); (ii) que un mismo sistema consciente en dos momentos distintos puede experimentar una inversión de sus experiencias (qualia invertidos intrasubjetivamente); (iii) que dos sistemas sean funcionalmente diferentes, y sin embargo, tengan la misma experiencia consciente (tierra invertida); y (iv) que dos sistemas fueran funcionalmente idénticos y, sin embargo, uno de ellos careciera de experiencias fenoménicas (qualia ausentes, zombis).

¿Qué puede decir el funcionalista ante las conclusiones a las que llega tan extravagante estrategia? Una primera impresión es que los del bando contrario, los antirreduccionistas, los antimaterialistas, los qualófilos, consideran los argumentos del espectro invertido tan relevantes que no han tenido reparo alguno en generar una serie de fantasías hasta tal punto delirantes y retorcidas que difícilmente puede uno

creer que las cosas pudieran llegar, ni por asomo, a ser así. Pero bueno, los experimentos mentales estrambóticos parecen haber sido su estrategia más fructífera, por lo que no nos queda otra opción que entrar en el juego y ver qué podemos hacer para desenredar todo este lío. Aunque quizás, si queremos entrar verdaderamente en el juego, lo que deberíamos hacer es enredarlo todavía más, desencajar y volver a encajar una a una las piezas del rompecabezas y solo después, ver los errores cometidos y valorar los resultados. Pero, antes que nada deberíamos cuestionarnos sobre la validez de los argumentos basados en los experimentos mentales que aquí se sugieren, ¿de verdad pueden existir dos personas (o una misma) que a pesar de ser funcionalmente idénticas tengan experiencias subjetivas y fenomenológicas invertidas respecto del mismo objeto intencional?[73] Los qualófilos no dudan en responder afirmativamente, no ven obstáculo alguno para que esto pudiera suceder, es concebible, pero ¿qué tipo de vínculo hay entre que algo sea *concebible* y que sea *posible*?[74] Claro está que dependiendo de cómo respondamos a esto, nuestras conclusiones tenderán a ir hacia un lado o hacia otro. Por otro lado, si adoptamos una perspectiva materialista, ¿qué sentido tendría aceptar este argumento como válido si no hay ni la menor prueba de que dos personas física y funcionalmente idénticas pudieran tener experiencias fenomenológicas sustancialmente diferentes? ¿Acaso alguien ha conseguido replicar un humano (o un animal) con la finura necesaria como para establecer que esto es cierto? Muchos filósofos sostienen que la inversión del espectro cromático sin una diferencia comportamental o funcional no es ni lógica ni conceptualmente posible, pero no hay ningún argumento decisivo en favor de esta imposibilidad. Lo que aquí pretendo es dejar claro que el argumento del espectro invertido y sus múltiples y disparatadas variantes se basan en una intuición, una intuición en la que se visualizan los qualia bajo el trasnochado modelo de los teóricos de los datos sensoriales (Harman, 1990: 49)[75]. Parece, por tanto, haber muchas dudas acerca de *lo que se aparece como contenido cualitativo de los estados mentales.* Algunos parecen creer que hay algo de

[73] No son pocos los defensores de la posibilidad del espectro invertido véase Chalmers (1999); Nida-Rumelin (1996); Hoffman (2006); Block (1990, 2007b). Por ejemplo, Nida-Rumelin (1996) sugiere que podría ser que hubiera casos de visión en color pseudonormal, casos en los que los conos R tengan un patrón de respuesta usualmente asociado a los conos G y los conos G al patrón de respuesta de los conos R. Ambos casos son las causas de ceguera al color rojo y verde. Pero si estas ocurrieran juntas a un mismo sujeto, este sería conductualmente similar a un sujeto normal, aunque sus experiencias de color estarían invertidas (pp. 146-148). También Hoffman (2006) sugiere que las experiencias de color y las relaciones funcionales pueden ser representados matemáticamente, y que dichas experiencias de color se pueden permutar manteniendo las relaciones funcionales constantes, contradiciendo así el funcionalismo. Argumentos contra la posibilidad del espectro invertido en Sundstrom (2002); Harman (1990, 1996); Dennett (1988, 1995, 2006); Hardin (1997).

[74] Que la concebibilidad sea un indicio de posibilidad es algo más que dudoso, la concebibilidad es una noción epistemológica y la posibilidad una noción ontológica, y hacer aseveraciones ontológicas en base a nociones epistemológicas es como mínimo arriesgado. Pero aunque admitamos que todo lo concebible es posible, esto supondría un alto coste en nuestros esfuerzos por conocer el mundo. Quizás aquí deberíamos introducir la probabilidad como variable reguladora. Por ejemplo, podemos concebir una mosca de diez millones de toneladas, por lo tanto sería posible, pero ¿es probable? Es muy poco probable que algo así exista. Por otro lado hoy día podemos concebir cosas que hace mil años no eran concebibles ¿acaso esto significa que hace mil años no eran posibles y ahora sí? No queda claro. Más sobre la relación entre concebibilidad y posibilidad en Hill, C. (1997) y Yablo, S. (1993), y en el siguiente apartado de este trabajo.

[75] Los teóricos de los *datos sensoriales* sostienen que la conciencia del color de un objeto está mediada por el conocimiento de una característica intrínseca de una representación perceptual. Lo que nos sugiere Harman es: "*Cuando Eloise tiene un árbol ante ella, los colores de su experiencia son todos experimentados como propiedades del árbol y sus alrededores. Nada es experimentado como una propiedad intrínseca de su experiencia. No experimenta ninguna propiedad de algo como propiedad intrínseca a su experiencia. Y esto es así para cualquiera. No hay nada de especial en la experiencia visual de Eloise. Cuando ves un árbol, no experimentas ninguna propiedad como propiedad intrínseca de tu experiencia. Mira un árbol y trata de volver tu atención a las propiedades intrínsecas de tu experiencia visual. Yo digo que lo que encontrarás es que solo las propiedades que presenta el árbol, incluidas sus propiedades relacionales, son las propiedades a las que puedes prestar atención"* (Harman 1990, 39).

misterioso en todo ello, que es algo que va más allá de relaciones entre inputs sensoriales, otros estados mentales y ciertos outputs conductuales, les gusta la idea de que haya algo así como cualidades intrínsecas de la experiencia, no aprehensibles por métodos científicos. Pero también hay otros que creen que hay algo de ilusorio en todo esto, que tenemos acceso a las propiedades de los objetos representados pero no a las propiedades de la experiencia de representación de dichos objetos, esto es, que cuando tenemos la experiencia visual de algo rojo, somos conscientes de un objeto rojo pero no de la propia experiencia de que estamos viendo algo rojo. Pero también hay quien piensa que estas son vías muertas, que no conducen a ningún lugar donde las cuestiones aparezcan con la suficiente claridad como para dar la cuestión por zanjada, por lo que otra estrategia es dar al qualófilo lo que pide, aceptar la existencia de qualia y aceptar la posibilidad de la inversión del espectro, pero no por eso rechazar el funcionalismo materialista.

Pasemos pues por alto las objeciones sobre la posibilidad o imposibilidad del argumento, y vayamos directamente a él. Veamos si es posible aceptar lo que nos proponen dichos experimentos, pero no aceptar sus conclusiones. Aceptaremos la intuición y su correspondiente petición de principio, esto es, que los colores son propiedades intrínsecas de la experiencia, aceptaremos el compromiso ontológico con los *qualia*, y aceptaremos la posibilidad por concebibilidad de una inversión espectral, sea esta inter o intra-subjetiva, pero no aceptaremos que esto suponga renunciar al funcionalismo materialista como una teoría general de la mente. Trataremos de mostrar la idea de que a pesar de todo, si existe algo así como los qualia, y aunque sea posible una inversión de estos, hay una salida para aceptar que nuestras experiencias conscientes finalmente (y necesariamente) supervienen a la constitución física y funcional de los individuos. Esta es la estrategia que sigue Sydney Shoemaker. Shoemaker es un realista con respecto a los qualia, pero digámoslo así, un realista moderado. Hemos de conceder que los *qualia* existen, que la inversión del espectro cromático es posible, pero que no hay nada en todo ello que sea incompatible con el funcionalismo y el materialismo en general, es decir, que los qualia, si existen, pueden ser definidos funcionalmente. Veamos como argumenta esa posición intermedia. En primer lugar dice que para que haya inversión cromática debe haber una manera de mapear determinados tonos de color en determinados tonos de color, esto es, debe haber algo que haga que cuando yo veo rojo tu veas su complementario, el azul. Pero dentro de la gama de colores perceptibles por el ojo humano encontramos diferentes brillos, sombras, saturación, tonalidades, etc., debe haber entonces multitud de mapas diferentes en los que cada color encuentre su complementario[76]. Esto pudiera ser una dificultad para la inversión cromática, sin embargo, a pesar de que nuestro espacio de color sea asimétrico (Hardin, 1987), es posible que existan criaturas que se caractericen por tal simetría cromática, y por lo tanto, criaturas

[76] Por ejemplo, acabo de decir que el complementario del rojo era el azul, pero esto es falso, el complementario del rojo es el cian, que es una mezcla entre azul y verde, pero ¿qué tonalidad de azul y de verde?, ¿con que matices?, ¿con que brillo?, ¿y qué hay de las sombras? Ahí está la dificultad, un espectro invertido en sentido estricto debería encontrar su complementario exacto, y esto es sumamente improbable.

cuya experiencia de color sea invertible. Para Shoemaker la mera posibilidad de la existencia de dichas criaturas es suficiente para que el problema filosófico de la inversión del espectro cromático se tome en serio (Shoemaker, 1982: 366-367). En segundo lugar, y como ya he dicho más arriba, es importante advertir que si hay inversión del espectro intrasubjetiva, entonces hay también inversión del espectro intersubjetiva, aunque también admitiremos que podría haber inversión intersubjetiva sin inversión intrasubjetiva (p. 359). Lo importante es darse cuenta de que no es posible demostrar la existencia de inversión intersubjetiva a través de informes conductuales (algo que sí sucede con la intrasubjetiva), para que se dé deberemos exigir que las dos personas tengan la misma estructura, los mismos *espacios cualitativos*, y que ante un mismo objeto emitan los mismos juicios de similitud cromática, juicios del tipo *A es más similar a B que a C* (p. 361-362). Algo parecido sucedería con la versión intrasubjetiva, solo que en este caso compararíamos directamente los informes de un mismo sujeto antes y después de la inversión. Si admitimos esto, según Shoemaker no hay razones para negar la posibilidad de que se produzca una inversión del espectro cromático, pero aceptar esto nos lleva directamente al clásico problema de las otras mentes, ¿cómo puedo saber si mis experiencias de color son idénticas o diferentes a las de otras personas? Llegados a este punto, y en tercer lugar, Shoemaker nos sugiere que distingamos en la expresión "*parece el mismo*" un sentido intencional y un sentido cualitativo. Cuando decimos que antes y después de la inversión del espectro, y tras la etapa de adaptación, alguien dice ver el mismo color en un objeto, estamos empleando un sentido intencional de "*parece el mismo*", pero el hecho de que haya sufrido una inversión en su espectro hace que en realidad no "*parezca ser el mismo"* antes que después de la inversión, este sería el sentido cualitativo. Por tanto, las cosas parecen ser diferentes en el sentido cualitativo, pero no en el sentido intencional (p. 365)[77]. Es precisamente ese sentido intencional uno de los puntos clave, pues es en el que Shoemaker logra ver una compatibilidad entre la inversión del espectro y el programa funcionalista. No hay razones, en principio, para pensar que no seamos capaces de descubrir si las experiencias de color entre dos seres humanos son cualitativamente comparables, esto es, que sean similares o diferentes bajo las mismas condiciones, para ello solo tendríamos que encontrar la realización de los qualia en el cerebro y determinar las similitudes fisiológicas relevantes entre los cerebros de los diferentes seres humanos (p. 378)[78]. Ahora bien, y en cuarto y último lugar, si las diferencias o similitudes cualitativas son detectables esto significa que el sujeto posee una creencia cualitativa asociada con la experiencia además de una disposición para realizar informes verbales sobre ella, y tanto dichas creencias como dichas disposiciones juegan un rol importante en el análisis funcional, por lo que mientras los estados cualitativos particulares no pueden ser funcionalmente definidos las *clases* de dichos estados

[77] Shoemaker ve posible una inversión del espectro en el sentido de cómo qué es para nosotros ver un determinado color, en el sentido de como que es para mí ver un tomare rojo que es como para ti ver un tomate verde, pero no en el sentido de que los tomates rojos nos aparecen visualmente a nosotros tal y como lo hacen.

[78] Véase que haciendo esta distinción Shoemaker cree evitar el problema de las otras mentes. Es posible conocer los qualia de los otros en el sentido intencional, pero no podemos *ver* desde donde los otros *ven,* luego en ese sentido cualitativo no es posible. Hasta donde llego, esto no soluciona el problema sino que lo desplaza, o como mucho lo soluciona solo a medias.

cualitativos, los estados cualitativos generales, si pueden serlo (Shoemaker, 1975: 308-309) ¿Qué significa todo esto? Pues que a través del análisis funcional yo no puedo revelar mis qualia pero si el carácter fenoménico de mis experiencias. Es posible explicar funcionalmente la propiedad de tener carácter cualitativo pero no la creencia de que uno está en un estado cualitativo particular. Por ejemplo, puedo decir que mi experiencia de un tomate rojo tiene un cierto carácter fenoménico, es decir, que tengo una representación del tomate como teniendo una cierta propiedad fenomenal, como el resultado de tener un quale particular asociado con la experiencia de enrojecimiento del tomate. La propiedad fenomenal está determinada por el quale al cual está asociado, pero el quale y la propiedad son fenomenalmente distintos. Las propiedades fenomenales son propiedades de los objetos, y los qualia son propiedades de la experiencia. Tenemos acceso directo a dichas propiedades fenomenales, y a través de ellas un acceso indirecto a los qualia.

Así es como Shoemaker se opone a la concepción de sentido interno del autoconocimiento introspectivo, pero a la vez defiende la visión de que los estados sensoriales y perceptuales tienen ciertas características no representacionales (qualia) que determinan qué es como tenerlos. Este es el camino que sigue Shoemaker para refutar la tesis intencionalista negadora de los qualia y a la vez no entrar en conflicto con su objetivo, que no es otro que el de compatibilizar la existencia de los qualia con el funcionalismo[79]. La cuestión es que Shoemaker otorga algo al funcionalista y algo al anti-funcionalista, con este último coincide en que hay un cierto aspecto de los qualia que es funcionalmente indefinible, con el primero, en que las condiciones de similitud e identidad y la propiedad de tener un cierto carácter cualitativo si son funcionalmente definibles. Shoemaker cree que esta es una posición intermedia satisfactoria. Ahora bien, ¿no significa todo esto que en última instancia hay un cierto aspecto de nuestras experiencias conscientes que permanece fuera de nuestro alcance, un aspecto al que no es posible darle forma? Creo que en su intento de reconciliación Shoemaker aun deja demasiado espacio al anti-funcionalista. Aunque parece claro que se adscribe a las filas del funcionalismo, finalmente se deja abierta alguna puerta, algo creo que se le escapa, pues parece dar a entender que aunque podamos explicar más de lo que los anti-funcionalistas dicen, aún queda algo misterioso en el hecho de que seamos capaces de tener experiencias conscientes. Pienso que el misterio es que no tenemos ni idea de cómo dar cuenta de esa supuesta conciencia inmediata y directa de las cualidades intrínsecas de nuestra experiencia. Shoemaker únicamente le quita algo de terreno al anti-funcionalista, dejando más espacio al funcionalista, pero como sucede muy a menudo con las estrategias conciliadoras acaba por no contentar a nadie. En definitiva, en Shoemaker, el misterio no deja de serlo porque finalmente no consigue dar cuenta de eso que nos dice que existe, necesitamos pues decantarnos por uno de los otros dos bandos.

[79] La visión reconciliadora de Shoemaker supone defender un *intencionalismo débil*, esto es, que todos los estados mentales son intencionales, pero que algunos también poseen propiedades no intencionales (los qualia), o lo que es lo mismo, que el carácter fenoménico de la experiencia no se agota en su intencionalidad (Shoemaker, 1991; Peacocke, 1983). Los *intencionalistas fuertes* por el contrario, piensan que las únicas propiedades de la experiencia que existen son propiedades intencionales (Tye, 1995; Lycan, 1996; Dretske, 1995).

Veremos que dice Harman (1990, 1996) al respecto.[80] Consideremos un caso hipotético de espectro invertido intersubjetivo. Cuando un sujeto con su espectro cromático normal observa un tomate maduro, hay un procesamiento perceptivo a través de una representación perceptual del tomate, incluyendo, claro está, una representación de su color. Dicha representación le sirve al sujeto como orientación en su medio ambiente, esto es, como su creencia acerca de los tomates, en particular, su creencia acerca de su color, y expresa su creencia sobre el color del tomate mediante la expresión *"es de color rojo*". Lo mismo sucede con el sujeto con el espectro invertido, su percepción del tomate es una representación perceptual del color del tomate que utiliza como su creencia sobre el color del tomate y la expresa con las mismas palabras, "*es rojo*". Ahora bien, en un caso normal de percepción, no puede haber distinción entre cómo se ven las cosas y cómo se cree que son, el cómo se ven las cosas se da gracias al contenido de la propia representación perceptual que a su vez se utiliza como una creencia acerca de cómo las cosas se cree que son. Pero el argumento del espectro invertido dice que hay diferencias en la experiencia de color en cada sujeto, así que si todo funciona en ellos de forma normal, se deduce que deben tener diferentes creencias sobre el color de los tomates maduros, uno debería creer que son rojos y el otro que son verdes. Por lo que si asumimos que tienen las mismas creencias, al menos uno de ellos tendrá una representación perceptual que no concuerda con su creencia acerca del color de los tomates (Harman, 1990: 47). Lo interesante en este análisis es que Harman introduce la relación entre el funcionamiento de los estados perceptuales y las creencias del sujeto. Esta interpretación del argumento del espectro invertido tiene fuertes consecuencias ya que si los dos sujetos se refieren a lo mismo cuando usan las expresiones de color, entonces o uno de ellos no está usando correctamente las palabras para expresar sus creencias sobre el color o bien uno de ellos no está usando su representación en la percepción de los colores como sus creencias sobre el color. En cualquier caso, aquí estamos ante un fallo del argumento (p. 48). Recordemos que según la hipótesis del espectro invertido dos sujetos uno de ellos con su espectro de color invertido expresarán las mismas creencias acerca del color de, por ejemplo, los tomates maduros (dirán que los tomates son rojos), y esto es así a pesar de que sus representaciones perceptivas difieran en su contenido. Pero esto significa que en alguno de ellos hay una disonancia entre su creencia y su representación perceptual.[81] La visión de Harman es muy clara, su propuesta es que las propiedades fenoménicas no son otra cosa que propiedades representacionales, por lo que la experiencia se agota en su contenido representacional. No hay tal cosa

[80] Una vez aquí pienso que ha llegado el momento de hacer un balance general de donde estamos. Shoemaker es de los pocos que defienden un realismo moderado con respecto a los qualia, cree que los qualia existen, y que el carácter fenoménico de nuestras experiencias no es otra cosa que una propiedad representacional. Block y Chalmers entre otros, defienden un realismo con respecto a los qualia, los qualia existen y además el carácter fenoménico de nuestras experiencias es algo más, o algo diferente de una propiedad representacional. Y por último, Harman, Dennett, Tye y algunos más, defienden una postura representacionalista, no existen los qualia, solo el carácter fenoménico de nuestras experiencias, y este no es otra cosa que una propiedad representacional. Sigamos.

[81] Debemos asumir que hay una representación perceptual normal, esto es, que las cosas rojas aparezcan como rojas equivale a lo normal, que las cosas rojas aparezcan como verdes es lo que no es normal. Asumiremos que esto es así, aunque si aceptamos el argumento del espectro invertido bien podría ser que a la mitad de la población se le aparezcan los objetos rojos como rojos y a la otra mitad como verdes, ¿de qué color se nos presentan entonces los tomates maduros? Asumir esto es necesario si queremos evitar la idea de que estamos inmersos en un engaño general y radical.

como cualidades intrínsecas de la experiencia, no existen algo así como los qualia tal y como están definidos por la tradición de los datos sensoriales, y por tanto el funcionalismo es suficiente para explicar la experiencia consciente, el resto es pura ficción[82].

Harman y Shoemaker coinciden en que para dar cuenta de lo que llaman *reacciones subjetivas*[83] hace falta una explicación más detallada y profunda del color, coinciden también en que la experiencia perceptual tiene carácter representacional, y en que el color está representado por nuestras experiencias visuales como una característica de los objetos externos y no como una característica de nuestra experiencia. También coinciden en que el color se experimenta más como una cualidad básica que como un conjunto de propiedades causales, esto es, que somos conscientes introspectivamente de los contenidos de presentación de nuestra experiencia, que no hay algo así como una *pintura mental* que establece dicho contenido. Y además están de acuerdo en que dos individuos con sus espectros cromáticos invertidos, en caso de que esto finalmente resultara posible, observan los mismos objetos aunque perciban propiedades diferentes, a pesar de que como perceptores normales deberían percibir los mismos objetos con los mismos colores. En lo que no parecen estar de acuerdo es en que Harman, finalmente, no encuentra coherente el argumento del espectro invertido, mientras que Shoemaker ya hemos visto que sí (Shoemaker, 1996: 55).

La razón por la que Harman no encuentra coherente la versión de Shoemaker del argumento del espectro invertido es porque detecta un problema de circularidad. La cuestión es: un objeto pongamos por caso rojo es percibido por S1 como R y por S2 como V, parecería que dicho objeto puede ser a la vez R (rojo) y V (verde) en el mismo lugar, al mismo tiempo y de la misma manera. Pero como esto no es posible, se sigue que o la experiencia de S1 o la de S2 o las dos son en algún aspecto no verídicas, de alguna manera debe haber alguna representación incorrecta del objeto percibido. Shoemaker soluciona este problema suponiendo que las propiedades R y V son *propiedades relacionales*[84] (las acaba llamando *propiedades fenoménicas*), de modo que algo puede ser R a una persona sin ser R a la otra, lo que sería incompatible es decir que R y V son iguales a una misma persona, en el mismo lugar, en el mismo momento y de la misma forma, pero no a dos sujetos distintos o a un mismo sujeto en distintos momentos. Ahora bien, ¿qué es para un objeto ser R en un determinado sujeto S1? Shoemaker (1994a) considera dos posibilidades. En primer lugar, podría ser que un objeto es R para una persona dada S1 si y sólo si la experiencia perceptual de S1 representa el objeto como R, o en segundo lugar, podría ser que un objeto es R para una persona

[82] Harman se propone distinguir entre las propiedades de los objetos de la experiencia y las propiedades de la experiencia de los objetos, y para ello asume el principio de la diafanidad de la conciencia, según el cual cuando intento concentrarme en los qualia acabo concentrándome en las propiedades representacionales, soy consciente de las propiedades de los objetos, su color, forma etc., pero no soy consciente de las propiedades de mi experiencia (Harman, 1990: 35).

[83] Por reacciones subjetivas Harman entiende aquellas impresiones subjetivas de un perceptor normal cuando observa los colores: cómo el color se ve. Una persona ciega de nacimiento podría aprender que los objetos tienen colores y podría en cierto sentido tener reacciones subjetivas al color, pero no es ese el tipo de reacciones subjetivas a las que Harman se refiere. (Harman, 1996: 1).

[84] Por más que lo intento no consigo ver con claridad en qué sentido del término relacional dice Shoemaker que R y V son propiedades relacionales, (lo explica en Shoemaker, 1994b: 297-298). De todos modos seguiré con el argumento, la circularidad está a punto de llegar.

dada S1 si y sólo si el objeto tiene una tendencia a proporcionar a S1 experiencias perceptuales que representan a ese objeto como R. En el primer caso, los objetos son R para S1 si y sólo si S1 los experimenta como R. En el segundo caso, los objetos pueden ser R para S1 incluso si S1 no los experimenta actualmente como para producir experiencias relevantes en las condiciones adecuadas. Y aquí llega la circularidad. Si Shoemaker está en lo cierto, para entender lo que es el concepto R, tenemos que entender como los objetos son R para alguien, pero a fin de entender como los objetos son R para alguien, tenemos que entender lo que es el concepto R, y según Harman, no hay forma de salir de ahí (Harman, 1996: 12-13)[85]. Considero que la dificultad en el planteamiento de Shoemaker, y en general en aquellos que aceptan la posibilidad del espectro invertido y la existencia de los qualia, es que al sugerir que una experiencia tiene una cierta propiedad fenoménica intrínseca que es la responsable de su propia representación, y al sugerir que el modo en como las cosas se perciben posee una cierta independencia de nuestras disposiciones reactivas, lo que en realidad parecen sugerir es que hay *alguien* en el interior del cerebro que surge cuando la información proveniente del medio consigue superar un cierto umbral, el de la conciencia y que además ese alguien incluso puede actuar con independencia a como lo hace el cerebro. Sencillamente, aunque no lo digan abiertamente, sus hipótesis parecen postular entidades intermedias no menos fantasmales que las sugeridas por el dualismo cartesiano.

Finalmente me gustaría sugerir un par de dificultades que creo que deberían ser tenidas en cuenta por los defensores del argumento del espectro invertido. Pienso que el periodo de adaptación es el aspecto clave para que el experimento adquiera sentido, pero también es la parte débil del argumento. Parece ser que ha de haber un período de adaptación en el que el sujeto cuyos qualia han sido invertidos se adapta a las nuevas circunstancias, y finalmente acaba recibiendo qualia de un color pero informando de un color distinto al que de hecho está percibiendo. La cuestión es que esa adaptación a la nueva conducta *bien pudiera implicar reinversión funcional* a través de la adquisición de un nuevo lenguaje para el color, un lenguaje donde verde significa rojo y rojo significa verde, por lo que el contenido funcional estaría tan invertido como el contenido cualitativo, y por tanto no habría diferencia alguna entre ambos. Además en este tipo de casos, se debería no solo tener en cuenta el contenido representacional de la experiencia concreta, pues esta no depende sólo de la distribución funcional de los sujetos perceptores sino, además, y sobre todo, de las características externas de los objetos percibidos y de la peculiar historia cognitiva de cada individuo,[86] que permiten que la percepción del objeto concreto pueda ser entendida como correcta en condiciones normales (Tye 2000: 104-116). El hecho de que cada individuo sea poseedor de una historia cognitiva concreta me sugiere una reflexión que pienso que todo defensor de este tipo de argumentos debería plantearse. Si hay ciertos rasgos asociados a ciertas experiencias, pongamos por caso traumáticas,

[85] Shoemaker se defiende de tal circularidad en Shoemaker, 1996.

[86] Por ejemplo, Hardin (1990) dice que con la edad el cristalino se va amarilleando, lo que altera nuestra manera de percibir los colores primarios.

¿serán esos rasgos también transferidos a las supuestas inversiones experimentadas por el individuo? Un ejemplo: si tengo asociado un recuerdo traumático al color azul turquesa porque cuando era niño vi morir a mi perro ahogado en un mar azul turquesa, ¿ese recuerdo se mantendrá en ese quale o no cuando sufra una inversión de mi espectro cromático del azul turquesa, pongamos por caso al rojo? No resulta sencillo responder a esta cuestión. Por un lado no es muy probable que una experiencia traumática se asocie solamente a un único quale, la mente humana no es tan sencilla, pero el ejemplo es imaginable, concebible y por tanto, para muchos qualófilos posible, más aún, podríamos asociar la experiencia traumática a miles de quales diferentes e invertirlos todos y nuestra cuestión no dejaría de tener sentido.[87]

Mi conclusión es que si son necesarios argumentos de este tipo para poder dar cuenta de la existencia de los qualia volvemos de nuevo a argumentos peligrosamente circulares, si necesitamos a los qualia para reconocer la posibilidad de los argumentos del espectro invertido, y a los argumentos del espectro invertido para reconocer la existencia de los qualia nos vemos inmersos en un bucle del que difícilmente se puede salir. Ni queda claro lo que los qualia son, ni queda clara la posibilidad de que los casos ofrecidos por el experimento sean casos posibles. Por otro lado, ni aunque aceptemos, como sugiere Shoemaker, que el funcionalismo es compatible con la existencia de los qualia y con el argumento del espectro invertido en todas sus múltiples versiones, conseguimos dar cuenta de todo aquello con lo que un qualofilo se debería comprometer. Por lo que de momento, y siguiendo a Harman y Dennett, me mantengo escéptico acerca de la existencia de algo así como experiencias subjetivas intrínsecas, inefables, privadas y directamente accesibles a la conciencia, esto es, me mantengo escéptico acerca de la existencia de qualia tal y como son definidos por la tradición.

Veremos ahora que sucedería si en vez de invertir los espectros cromáticos de los sujetos nos planteamos la idea de que hubiera una persona física y funcionalmente igual a mí pero sin experiencias conscientes, son los qualia ausentes o zombis filosóficos. Veremos cómo se las arreglan este tipo de argumentos, basados también en la posibilidad por concebibilidad, para convencernos de la existencia de esa misteriosa *sustancia* tan familiar y a la vez tan escurridiza. Veremos que los zombis traen consigo consecuencias algo diferentes de las que hemos visto hasta ahora con el espectro invertido. Los zombis, vienen bien armados.

**Qualia ausentes, zombis y el problema difícil de la conciencia**

[87] Una reflexión similar la he leído en Dennett, aunque no consigo recordar donde. Desde luego esto no soluciona ninguno de los problemas que tenemos sobre la mesa, pero no toda estrategia filosófica supone solucionar problemas, a veces también es provechoso buscar obstáculos a las posiciones defendidas por los del bando contrario, y pienso que este es uno de los serios. El hecho de que todos los argumentos que desarrollan estos temas lo hagan a través de un quale aislado pienso que es un obstáculo y quizás un error, ¿acaso podemos aislar al perceptor y al rojo del tomate de todas las circunstancias que les rodea? Dejaré aquí esta reflexión.

Ya vimos en el anterior apartado el experimento mental sugerido por Block de los qualia ausentes en el que se conectaba causalmente a todos los habitantes de China mediante equipos de radio con el fin de simular una organización funcional similar a la que presenta un cerebro. Parecerá *obvio*[88]*,* dice Block, que habrá una ausencia total de estados fenoménicos[89]. Pues bien, una versión algo diferente de este experimento ha sido explorada por los filósofos que simpatizan con la noción de qualia, principalmente por David Chalmers. Esta vez hablaremos de la posibilidad lógica de zombis (filosóficos)[90].

El debate en torno a la posibilidad de "zombis" (*personas* que se comportan como nosotros pero que carecen de toda experiencia consciente) fue iniciado por el filósofo Robert Kirk (1974). La definición de un zombi, del gemelo zombi de Chalmers, es esta:

*"Esta criatura es idéntica a mí molécula por molécula e idéntica en todas las propiedades de bajo nivel postuladas por una física terminada, pero carece por completo de experiencia consciente […]. Es físicamente idéntico a mí y podemos también suponer que está inmerso en un ambiente idéntico. Con seguridad será idéntico a mí funcionalmente: procesará la misma información, reaccionará de un modo similar a las entradas, sus configuraciones internas se modificaran en forma apropiada y como resultado su conducta será indistinguible de la mía […]. [E]stará despierto, será capaz de informar del contenido de sus estados internos, capaz de concentrar su atención en diversos lugares, etc. Solo que nada de ese funcionamiento estará acompañado por alguna experiencia consciente real. No habrá ninguna experiencia fenoménica. No existe una experiencia de ser como un zombi."* (Chalmers, 1999: p. 133-134).

Un zombi, por tanto, por definición, no sería más que un conjunto funcional de materia (orgánica o inorgánica) que tan sólo parece actuar del modo en que cualquiera de nosotros actuaría cuando tenemos alguna experiencia consciente pero que, sin embargo, no dispone realmente de aquello que, parece, caracteriza a nuestra conciencia: carecería de *qualia*, de contenido fenoménico cualitativo y consciente. Obsérvese que desde un punto de vista externo, desde la tercera persona, Chalmers y su gemelo zombi son indiferenciables, ningún observador externo sería capaz de apreciar ni la menor diferencia entre ellos, pero obsérvese también que además Chalmers le concede a su gemelo no solo la capacidad de tener estados internos con contenido, le concede también la capacidad de informar acerca de ellos, y sin

[88] Estas obviedades obligan a uno a ponerse alerta. Que algo sea o parezca obvio no significa que sea verdadero sino que es fácilmente digerible, por lo que conviene prestar especial atención a sus implicaciones. (Consejo de Dennett; 2006: p 131).

[89] Las mejores críticas que he encontrado contra los qualia ausentes de Block son las de Shoemaker (1981), Tye (2006) y Van Gullick (1993).

[90] A pesar de sus semejanzas (los dos tratan sobre la ausencia de experiencias internas), estos dos experimentos parecen diseñados para combatir ideas diferentes. Me explico. La idea de la nación china de Block está diseñada para atacar al funcionalismo, pues nos dice que difícilmente de una configuración de este tipo surgirá una mente (Block, 1978), mientras la idea de los zombis está diseñada para demostrar la irreductibilidad de la experiencia consciente y de ahí concluir que el materialismo reduccionista es falso. Chalmers no tiene ningún problema en aceptar que la unión de diferentes materiales, sean latas de cerveza, pelotas de ping-pong o el conjunto de la población China, con la configuración adecuada producirá una mente. No en vano, Chalmers defiende una tesis que combina el funcionalismo con un dualismo de propiedades, la llama funcionalismo no-reductivo, la experiencia consciente bien podría surgir de la organización funcional, pero no ser un estado funcional (Chalmers, 1999: 317). La idea de los zombis tal y como la entiende Chalmers nos sugiere la mera posibilidad de que un ser así sea concebible, si esto es así, entonces los zombis, dice, nos demostraran que la experiencia consciente no es reducible a los componentes físicos del sistema, según él, necesitamos ampliar las barreras impuestas por la ciencia para atrapar explicativamente a la experiencia consciente.

embargo el gemelo zombi de Chalmers, según Chalmers, no es capaz de apreciar todo ello, carece de conciencia, no tiene qualia. El argumento de la concebibilidad de los zombis es como sigue:

1. Los zombis son concebibles.
2. Lo que es concebible es posible.
3. Los zombis son posibles.[91]

Cabe advertir que Chalmers diferencia dos tipos de materialismo, el tipo A y el tipo B. Los que se adhieren al tipo A sostienen que la conciencia (en caso de que algo así exista), superviene lógicamente a lo físico. Los duplicados físicos y funcionales que carecen del tipo de experiencia que nosotros tenemos son inconcebibles, por lo que no reconocen la premisa (1) del argumento, para ellos, los zombis no son concebibles, y la explicación de la conciencia no es otra cosa que la explicación de la realización de las diversas funciones cognitivas. Los materialistas de tipo B no aceptan que la conciencia sea lógicamente superveniente a lo físico, sostienen que no existe ninguna implicación a priori de lo físico en lo fenoménico, pero aceptan el materialismo, aceptan que los zombis son concebibles, pero no que sean metafísicamente posibles, estos tienen problemas para aceptar (3) porque no ven claro que (2) sea completamente cierta[92] (Chalmers, 1999: 218).

Mi estrategia para este apartado será la siguiente. En primer lugar trataré de dejar claro que el hecho de que algo sea concebible no necesariamente lo hace posible, es decir atacaré la segunda premisa del argumento (sobre este asunto he dicho algo en el apartado anterior, ahora profundizaré en él), en esta línea se encuentran Botterell (2001), Balog (1999), Brueckner (2001), Hill (1997), Levine (2001) o Yablo (1999). En segundo lugar, expresaré mis dudas sobre si el que haya algo así como mi gemelo zombi es algo verdaderamente concebible, esto es, atacaré la primera premisa, entre los autores que emplean esta estrategia están otra vez Boterell (2001), Dennett (1995), Kirk (1996), Marcus (2004), Stoljar (2001), Tye (2006) o Shoemaker (1999). En tercer lugar, tratare de mostrar que ni aunque los zombis fueran lógica y metafísicamente posibles por ser concebibles el argumento anti-fisicalista de los zombis o de los qualia ausentes saldría victorioso, o lo que es lo mismo, que ni aun aceptando los argumentos por concebibilidad, aunque los zombis fueran posibles, deberíamos rechazar el fisicalismo, aquí tenemos a Stalnaker (2002), Hawthorne (2002) o Braddon-Mitchell (2003). Y en cuarto lugar analizaré brevemente el argumento de los anti-zombis de Frankish (2007). Finalmente diré algo sobre un tema estrechamente relacionado con este, el del problema difícil de la conciencia, argumentaré que no hay tal problema.

---

[91] Un argumento similar pero algo más extenso dice así: (1) Los zombis son concebibles, (2) si los zombis son concebibles, entonces los zombis son metafísicamente posibles, (3) si los zombis son metafísicamente posibles, entonces la conciencia no es física, (4) por tanto, la conciencia no es física (Chalmers, 2002b: 249).

[92] Defienden posiciones de tipo A, Armstrong, (1968); Dennett (1995); Dretske (1995); Kirk, (2005, 2008); Lewis, (1966, 1990); Roshental (1996) o Ryle (2005). Y posiciones de tipo B, Hill, (1997); Levine, (1983); Loar, (1990; Lycan, (1995, 1996); Papineau, (2002); Tye (1995) o Horgan (1984).

El asunto de los zombis ha traído consigo un especial interés por la relación entre lo concebible y lo posible, tanto el argumento de los qualia invertidos como el de los qualia ausentes se sostienen sobre una relación fuerte entre concebibilidad y posibilidad, por lo que antes de entrar en si los zombis son concebibles o no, me gustaría decir algo sobre la espinosa cuestión metafísica de si la concebibilidad implica de hecho posibilidad, sobre la dudosa relación entre que algo sea concebible y que sea posible. Es importante comenzar diferenciando entre posibilidad lógica, posibilidad metafísica y posibilidad física. La posibilidad lógica (conceptual) incluye cualquier proposición que por pura lógica pueda quedar abierta sin importar lo imposible que pudiera parecer, es enteramente conceptual (es lógicamente posible que existan seres que alcancen una velocidad superior a la velocidad de la luz, aunque no es lógicamente posible concebir un círculo cuadrado). La posibilidad metafísica es aquella posibilidad lógica sustentada por la naturaleza de todas las cosas que existen o podrían haber existido, se trata de una posibilidad que exige una cierta *coherencia conceptual,* algo es metafísicamente posible si y sólo si hay algún orden alternativo de las cosas dentro del cual existiría ese algo (es metafísica y lógicamente posible que existan los números o las leyes de la naturaleza aunque no sea posible encontrarlos físicamente, sin embargo no es metafísica aunque si lógicamente posible un universo sin gravedad o un mundo en el que el agua no contenga hidrógeno). Y las posibilidades físicas (o naturales, empíricas o nomológicas) son las posibilidades lógicas y metafísicas sustentadas únicamente por las leyes físicas de la naturaleza (es física, metafísica y lógicamente posible que existan seres extraterrestres, pero no es físicamente posible que estos seres alcancen una velocidad superior a la de la luz, a menos que nuestras leyes de la física estén completamente equivocadas). Adviértase que todo mundo físicamente posible ha de ser metafísica y lógicamente posible, que todo mundo metafísicamente posible es lógicamente posible aunque no es necesario que sea físicamente posible y que todo mundo lógicamente posible no es necesario que sea ni física ni metafísicamente posible. En cualquier caso, podemos imaginar multitud de mundos lógicamente posibles, menos mundos metafísicamente posibles y muy pocos o quizás un solo mundo física o naturalmente posible. La posibilidad lógica es demasiado amplia, y la posibilidad física demasiado restrictiva, aquí nos interesa la posibilidad metafísica.

Una vez aclarado qué tipo de posibilidad es la que entra en juego ya estamos en disposición de explorar la relación entre lo concebible y lo metafísicamente posible. Hace ya unos cuantos siglos Hume afirmó que *es una máxima establecida en metafísica que todo lo que el espíritu concibe claramente incluye la idea de una existencia posible o, en otras palabras, que nada de lo que imaginamos es absolutamente imposible* (Hume, 2001: 41). Esto es, si es concebible que *p*, entonces es posible que *p.* Muchos de los argumentos elaborados en base a la concebibilidad suponen aceptar esta relación fuerte entre que algo sea concebible y que sea posible, pero pasar de la concebibilidad a la posibilidad supone pasar primero de la cuestión epistémica sobre qué es eso que podemos llegar a concebir, a la cuestión modal sobre que es aquello que

es posible o necesario y por último a la cuestión metafísica sobre la naturaleza de las cosas que hay en el mundo (Chalmers, 2002b: 145). Este paso de cuestiones y términos epistémicos a cuestiones y términos ontológicos es ya de por sí problemático, no es que sea ilícito, pero debe hacerse con sumo cuidado. Por ejemplo, es concebible pero no posible que el agua no sea $H_2O$, es concebible pero no posible que la estrella matutina no sea la estrella vespertina, pero además, es concebible y por tanto sería posible que la tierra fuera el centro del universo y sin embargo no lo es, es concebible y por tanto sería posible que las sirenas o los unicornios existieran, y sin embargo no existen. A donde pretendo llegar es a que si aceptar la posibilidad en base a la concebibilidad supone correr riesgos, llegar a la verdad en base a una posibilidad obtenida a través de la concebibilidad es correr más riesgos todavía.

Sobre lo concebible podemos encontrar una noción fuerte, según la cual que algo sea concebible significa que no sea contradictorio (esta sería la versión de Hume, sería concebible que hubiera un mundo en el que los bebés escribieran tratados sobre problemas de lógica modal, pero no es concebible que algo sea y no sea al mismo tiempo, como tampoco es concebible que un matemático encuentre el mayor de los números primos), aunque hay otra noción más suave, según la cual lo concebible es aquello que es compatible con todo lo que sabemos (es concebible que exista dolor sin activación de las fibras C, pero no sin alguna activación, hasta donde sabemos, no podemos concebir un ser inmaterial con dolor). Chalmers (2002b) trata de desambiguar el concepto y para ello lo disecciona hasta encontrar las siguientes dimensiones independientes de lo concebible (p. 147-160):

- Concebibilidad prima facie vs ideal: S es prima facie concebible para un sujeto cuando S es concebible para ese sujeto en una primera impresión. Es decir, después de cierta consideración el sujeto encuentra que S pasa las pruebas como criterios para la concebibilidad. Y por otro lado, S es ideal si es idealmente concebible cuando es concebible a través de una reflexión racional. A veces sucede que S es prima facie concebible para un sujeto, pero esta concebibilidad a primera vista se ve socavada por una mayor reflexión que muestra que las pruebas como criterios para lo concebible no han sido superadas. En este caso, S no es idealmente concebible[93].
- Concebibilidad positiva vs negativa: según la noción negativa S es concebible cuando S no es descartable, S es concebible cuando no está descartado por lo que se sabe, o por lo que uno cree. Por el contrario la noción positiva requiere que uno pueda formar algún tipo de concepción positiva de una situación en la que S sea el caso. Uno puede colocar las variedades de concebibilidad

[93] Aquí no queda claro en qué lugar deberíamos colocar la barrera para considerar que algo sea concebible prima facie o ideal, quizás no haya más que una sola concebibilidad ideal, pero en todo caso, ¿bajo qué criterios debemos considerar que han sido superadas las pruebas que nos permiten afirmar que algo es idealmente concebible? Al considerar que algo es idealmente concebible cuando se ha obtenido tras una reflexión racional perfecta y al aplicar los criterios antes mencionados a dicha reflexión, nos damos cuenta de lo difícil que resulta concebir una reflexión perfecta o ideal, hasta quizás podríamos considerar que es algo inconcebible. Resulta hasta paradójico tratar de concebirnos concibiendo idealmente algo.

positiva bajo la amplia rúbrica de la imaginación, en este caso concebir positivamente una situación es en cierto sentido imaginar una configuración específica de objetos y propiedades.

- Concebibilidad primaria vs secundaria: decimos que S es primariamente concebible cuando es concebible que S sea realmente el caso. Podemos decir que S es secundariamente concebible cuando es concebible que S podría haber sido el caso. Esto corresponde a dos formas diferentes de pensar en posibilidades hipotéticas, la concebibilidad primaria es siempre una cuestión a priori, no es un conocimiento empírico sino un razonamiento apriorístico sobre las diferentes formas de cómo el mundo podría ser, y la secundaria estaría en el modo subjuntivo, esto es, como formas contrafácticas no de cómo el mundo puede ser, sino de cómo podría haber sido.

Después de analizar todas las combinaciones posibles entre los diferentes tipos de concebibilidad y posibilidad, Chalmers encuentra la siguiente vinculación de implicancia a la que llama Racionalismo Modal Débil (RMD)[94], como la más plausible[95]:

(i) La concebibilidad ideal primaria positiva implica posibilidad primaria.[96]

Recordemos que *p* es una concebibilidad ideal cuando *p* es concebible a través de una reflexión racional ideal; que *p* es una concebibilidad primaria cuando es concebible que *p* sea realmente el caso y que *p* es una concebibilidad positiva cuando uno puede imaginar que *p*. Y en cuanto a la posibilidad, decimos que p es una posibilidad primaria si la intensión primaria es cierta en algún mundo posible, es decir, si *p* es verdad en un mundo considerado como real[97].

[94] Además del Racionalismo Modal Débil, tenemos el Racionalismo Modal Fuerte donde la concebibilidad negativa implica posibilidad y el Racionalismo Modal Puro donde la concebibilidad positiva implica concebibilidad negativa y ambas implican posibilidad. Resumiendo, la versión débil da acceso sólo a los mundos posibles, la versión pura a *todos* los mundos posibles, y la fuerte dice que podemos conocer todas las verdades acerca de estos mundos posibles (Chalmers, 2002b: 191).

[95] El razonamiento es el siguiente: la concebibilidad prima facie es una guía imperfecta para la posibilidad, o al menos, más imperfecta que la ideal, que por otro lado es la que nos interesa. La concebibilidad secundaria puede ser una buena guía para la posibilidad, pero está, en general, estructurada a posteriori, por lo que prescindiremos de ella. La concebibilidad primaria es una guía pobre para la posibilidad secundaria, pero puede ser una buena guía para la posibilidad primaria, es decir, la posibilidad de satisfacer una intensión primaria. Por tanto la concebibilidad primaria ideal es la que cuenta. La tesis más plausible es que la concebibilidad ideal positiva primaria de S implica la posibilidad (de un mundo que satisface la intensión primaria de S). Y quizás: la concebibilidad ideal negativa primaria de S implica posibilidad (de un mundo que satisface la intensión primaria de S). Para una explicación más detallada de estas implicancias véase Chalmers, (2002b).

[96] *P* es una posibilidad primaria si su intensión primaria es cierta en algún mundo posible, es decir, si *p* es verdad en un mundo considerado como real. Más concretamente, *p* es una posibilidad a priori porque es independiente de nuestro conocimiento de ella; por ejemplo, que el agua sea ese líquido claro y bebible que ocupa la mayor parte de la superficie terrestre sería una posibilidad primaria, pues sería así incluso aunque nosotros no hubiéramos existido. Mientras que p es una posibilidad secundaria si su intensión es secundaria, esto es, no está determinada a priori sino a posteriori, por ejemplo, que el agua sea $H_2O$ es una posibilidad secundaria, pues depende enteramente de que nosotros le asignemos esa etiqueta, es decir, de cómo las cosas resultan ser en un mundo concreto, en este caso, el nuestro. (Mas sobre este asunto en Chalmers, (1999: 88-100) y por supuesto en Kripke, (1995)).

[97] Hay varios autores que ven el RMD implausible. Por ejemplo, Mizrahi y Morrow (2015) sostienen que el RMD no es lo suficientemente débil como para ser correcto. Elaboran dos argumentos por reducción al absurdo que parten de la asunción de que el RMD es verdadero y acaban llegando a las siguientes contradicciones: que la concebibilidad es y no es una evidencia conclusiva para la posibilidad, y que el RMD es falso y verdadero al mismo tiempo (p. 3-8) No desarrollaré aquí los argumentos. Evidentemente hay premisas en ellos sobre las que los partidarios de la RMD tendrían mucho que objetar, pero en caso de ser correctos tendríamos que concluir que la concebibilidad ideal primaria positiva no implica necesariamente posibilidad metafísica primaria. El RMD entra en contradicción cuando se le hace pasar por sí mismo, ¿o acaso no podemos concebir ideal, primaria y positivamente un mundo en el que el RMD sea falso?

La principal objeción a (i) viene de la mano de Kripke (1995). Recordemos que Kripke demostró en su análisis sobre las verdades necesarias a posteriori que existe una clase de enunciados necesariamente verdaderos cuya verdad no es cognoscible a priori, por ejemplo, el enunciado "*el agua es $H_2O$*". No podemos saber a priori que esto sea cierto, pues el agua bien podría estar compuesta de por ejemplo XYZ, no obstante, dado que el agua es $H_2O$ en el mundo real será $H_2O$ en todos los mundos posibles, y por tanto "*el agua es $H_2O$*" será una verdad necesaria a pesar de su naturaleza a posteriori. En términos de concebibilidad, que sea concebible que el agua no sea $H_2O$ no implica que sea posible, pues el descubrimiento científico a posteriori que identifica agua con $H_2O$ es verdadero en toda situación posible en la que los mismos agentes dicen "*agua*" ante la misma materia líquida. De igual modo, y en términos de zombis, que sea concebible la existencia de zombis no implica que sea posible, la correlación entre conciencia y estados mentales pudiera ser una necesidad a posteriori, esto es, en caso de que lo que concibiéramos cuando concebimos un zombi fuera precisamente un ser con estados mentales pero sin estados fenoménicos, esto no nos garantiza que los zombis sean posibles, pues esta relación bien pudiera ser finalmente una necesidad a posteriori. Para salvar este obstáculo Chalmers apela a la diferencia entre intensión primaria y secundaria, pues la necesidad a posteriori Kripkeana surge cuando la intensión secundaria en un enunciado respalda una proposición necesaria, pero la intensión primaria, que es las que a Chalmers le interesa, no (Chalmers, 1999: 98). Esto quiere decir que hay un carácter bidimensional de la referencia, y que la intensión primaria y la secundaria que a su vez coinciden con los aspectos a priori y a posteriori del significado respectivamente, marcan la diferencia esencial entre lo que debemos explicar. Es decir, que es la intensión primaria la que captura independientemente de factores empíricos los conceptos que refieren a fenómenos naturales, por lo que es esta la que nos interesa. Ahora bien, esto nos obliga a dividir también la concebibilidad en dos tipos, por ejemplo "*el agua es XYZ*" es un enunciado concebible en un sentido porque hay al menos un mundo en el que dicho enunciado podría ser verdadero (siempre y cuando sea evaluado de acuerdo con su intensión primaria), pero inconcebible en otro, pues no hay ningún mundo concebible en el que este enunciado sea verdadero (evaluado con la intensión secundaria). El caso es que la primera concebibilidad implica posibilidad primaria y la segunda posibilidad secundaria, y debemos tener cuidado de no mezclarlas (Chalmers, 1999: 102). Así es como Chalmers consigue deshacerse de las implicaciones de las teorías de Kripke, pero lo hace pagando el precio que supone diferenciar un tipo de concebibilidad que implica posibilidad lógica y otro que implica posibilidad metafísica[98]. La concebibilidad implica posibilidad en el primer caso, aquí los zombis son concebibles y por

[98] En cierto sentido pienso que Chalmers minimiza la importancia de la distinción entre posibilidad lógica y metafísica, más aun, entiendo que no las diferencia, que presupone que en cuanto a los conceptos fenoménicos la intensión primaria y la secundaria son la misma, por ejemplo, dice que *los mundos metafísicamente posibles son solo los mundos lógicamente posibles (y que la posibilidad metafísica de los enunciados es posibilidad lógica con un giro semántico de los enunciados)* (1999: 66) o más adelante explica que *la distinción citada frecuentemente entre la posibilidad lógica y la posibilidad metafísica que se origina en los casos de Kripke […] no es una distinción en el nivel de mundos, sino, cuanto más, una distinción en el nivel de enunciados* (p. 102) y en otro lugar, unos años más tarde insiste en que *incluso si no todos los mundos imaginables son metafísicamente posibles, necesitamos un concepto modal racional ligado a la coherencia racional o concebibilidad para analizar mejor los fenómenos en cuestión. Podríamos llamar a esa noción de posibilidad, posibilidad lógica.* Y continua con el siguiente

tanto posibles, es decir, en el sentido en el que la posibilidad es entendida como posibilidad primaria y por tanto a priori, pero no en el segundo caso en el que la concebibilidad, como ya demostró Kripke, no siempre implica posibilidad. Pero un fisicalista siempre podrá insistir en que solo la posibilidad secundaria, a posteriori de la existencia de mundos zombi implicaría que el fisicalismo es falso, ya que no hay forma de demostrar que el fisicalismo sea falso a priori. Necesitaríamos una explicación que fuera más allá de la concebibilidad primaria, necesitaríamos que Chalmers demostrara que el asunto que tenemos delante, el de si es concebible un ser físicamente idéntico a nosotros pero sin estados fenoménicos, es aprehendido únicamente en base a que la concebibilidad ideal, primaria y positiva de la existencia de zombis implica la posibilidad primaria de su existencia y no de otro modo, pero adviértase que un zombi en un mundo zombi podría elaborar los mismos argumentos, con premisas verdaderas, y extraería una conclusión falsa (ya que el materialismo si sería cierto en un mundo zombi). Definitivamente pasar de lo concebible a lo posible, y defender la tesis de que el materialismo requiere una vinculación a priori es sumamente problemático (Balog, 1999)[99]. Chalmers pretende llevarnos a un mundo donde la afirmación "*el agua es XYZ",* como primariamente concebible, nos garantiza la posibilidad de la existencia de un mundo en el que existe algo llamado XYZ, pero no un mundo en el que aunque XYZ se comporte y actúe como el agua, sea agua, al menos, tal y como la conocemos, y por extensión, también pretende llevarnos a un mundo en el que cabe la posibilidad de que existan zombis, aunque no a un mundo en el que estos sean tal y como en nuestro mundo los imaginamos. ¿Es concebible que la tierra gire en una órbita fuera de la zona habitable? Sí, es concebible, no hay nada conceptual o lógicamente inconsistente en dicha concepción, ¿es posible? No, no es posible, si fuera posible la tierra no estaría habitada y por tanto no habría nadie capaz de concebirlo. No hay duda de que la segunda premisa del argumento de los zombis es sumamente problemática, y no hay duda de que aquel que se sienta atraído por las tesis fisicalistas no va a renunciar a ellas a partir de argumentos tan endebles.

---

ejemplo: *incluso si son metafísicamente imposibles todos los mundos con diferentes leyes de la naturaleza, todavía será tremendamente útil tener un espacio más amplio de mundos lógicamente posibles (o entidades similares del mundo) con diferentes leyes, para ayudar a analizar y explicar las hipótesis e inferencias de un científico que investiga las leyes de la naturaleza* (Chalmers, 2002b: 189). Lo cierto es que es una cuestión abierta la de si esta distinción es o no pertinente, no es fácil delimitar la posibilidad metafísica, aunque sospecho que es algo más que un simple *giro semántico* de la posibilidad lógica, la cuestión es que hay un número abismal de mundos lógicamente posibles porque nuestra imaginación nos permite concebir incontables mundos diferentes, pero tal flexibilidad nos obliga a ponernos alerta, el espacio entre la posibilidad lógica y la meramente física es demasiado amplio.

[99] Merece la pena analizar algo más detenidamente el argumento de Balog (1999). Parte de las siguientes asunciones, cada una de ellas para cada una de las premisas: 1) Que mi zombi y yo compartimos todos nuestros estados intencionales excepto aquellos que implican conceptos fenoménicos, 2) que aquellos conceptos de mi zombi que corresponden a mis conceptos fenoménicos se refieren a algún estado físico de mi zombi, y 3) que la aprioricidad de los pensamientos superviene a los roles conceptuales de sus conceptos constituyentes (y relacionados). Efectivamente, si admitimos que un zombi tiene todo lo que nosotros tenemos, incluso estados intencionales, aunque no posea estados fenoménicos, entonces un zombi en un mundo zombi podría elaborar el siguiente argumento (los conceptos fenoménicos del zombi se marcarán con +):

(1) Si el fisicalismo es cierto, entonces para cualquier verdad T, los enunciados de la forma si K entonces T son verdades conceptuales.
(2) Hay alguna declaración Q+ verdadera en el sentido en que un estado fenoménico se produce (el eliminativismo acerca de los estados fenoménicos es falso).
(3) Si Q+ es una declaración fenoménica+, entonces "K entonces Q+" no es una verdad conceptual.

Y por tanto,

(4) El fisicalismo es falso.

Pero si el argumento de Chalmers es verdadero también lo será este, pero en un mundo zombi el fisicalismo ha de ser verdadero, luego hay algo en el argumento de Chalmers y en este, que no funciona (Balog, 1999: 513).

Hasta ahora hemos expresado serias dudas sobre lo que se dice en (2) *que la concebibilidad implica posibilidad*, pero supongamos que no hubiera dudas, que (2) es una premisa verdadera, y que rechazamos la posición contraria a (2) que defiende el materialista de tipo B, ¿significará esto que el argumento de los zombis es un argumento plausible? No, aún podemos atacar (1), que *los zombis son concebibles*, si no lo son, tampoco serán posibles independientemente de que (2) sea cierta o no, nos habremos convertido así en materialistas de tipo A con respecto a los zombis. Una primera idea es que postular mundos idénticos al nuestro pero sin ingredientes no físicos extra implica aceptar que vivimos en un mundo en que esos ingredientes no físicos existen, y esto, además de ser claramente una petición de principio, no es ni lógica ni metafísicamente posible. Además, el argumento de la concebibilidad del tipo utilizado por Chalmers presupone que tenemos un conocimiento adecuado y completo de los propiedades y estructuras microfísicas de la experiencia consciente, sin embargo, actualmente no hay razón alguna para pensar que nuestro conocimiento de la física de la conciencia sea completo, por lo que todo intento de concebir mundos posibles en los que supongamos ese conocimiento completo están condenados a una regresión infinita. El argumento de Chalmers visto desde este punto de vista acaba siendo poco sólido (Bailey, 2007: 13). Otra idea es que al imaginarnos un zombi, nos imaginamos un ser vacío por dentro, sin interioridad, un ser del que no se puede decir que haya algo que sea como ser ese ser, efectivamente, no hay nada que sea como ser un zombi porque no hay nada dentro de él, por tanto, lo que nos imaginamos es nada, y ahí está el problema, ¡que no hay nada que imaginarse! Quien dice poder imaginarse un zombi, no está imaginando algo, esta imaginando nada. Solo podemos atribuir inconsciencia a seres que sean o hayan sido conscientes, sencillamente, no podemos inferir que sea un acto mental coherente el hecho de que uno atribuya a las personas conscientes la propiedad de la inconsciencia a partir del hecho de que la inconsciencia es concebible (Marcus, 2004: 484-485). También podemos imaginar cómo hacen Van Gullick (1999) o Dennett (1995), un argumento análogo al de la concebibilidad de zombis pero con el vitalismo. Imaginemos un vitalista del siglo XIX que pretende ofrecer un experimento mental para probar que la vida es algo que va más allá de lo físico. Diría por ejemplo que puede concebir una criatura similar a las criaturas de su tiempo en todos los aspectos físicos pero sin la habilidad de la reproducción, y por tanto que la habilidad de reproducirse es algo distinto y separado de cualquier propiedad física. Pero hoy en día sabemos que la conclusión del vitalista es falsa, la habilidad de reproducirse no está ni por encima ni por debajo de lo físico, o al menos no más que la liquidez lo está del agua. ¿Dónde ha ido mal el argumento del vitalista? El problema es que no tiene un concepto adecuado ni de la reproducción ni de la estructura física total de un organismo físico, y por tanto falla a la hora de formular una teoría adecuada de como esos conceptos pueden ir juntos. No tiene ni idea de cómo la reproducción implica la replicación y transferencia de la información genética, ni de cuáles son las bases bioquímicas de tal proceso, ni de como la información genética puede ser codificada por secuencias de ADN y ARN. Lo que era concebible desde la perspectiva del vitalista no lo es actualmente en cuanto a su posibilidad lógica en cuestiones de

reproducción y estructura física. El vitalista podía unir su concepto de la estructura física total de un organismo con la negación de su concepto de la habilidad de reproducirse sin generar contradicción, pero dada la radical incompletitud de sus conceptos acerca de la naturaleza de ambos fenómenos nada hay en la argumentación del vitalista que se sostenga. Cuando se trata de explicar la vida es la realización de las diversas funciones lo que debe explicarse, por lo que concebir la posibilidad lógica de que la reproducción o la vida sean hechos ulteriores no explicables por la concepción de los hechos físicos es claramente insostenible. Y por analogía, lo mismo podríamos decir sobre el asunto de los zombis (Van Gulick, 1999: 19)[100].

Quizás deberíamos admitir que si se lograra afirmar la verdad metafísica de la existencia de zombis, entonces (pero tan sólo entonces), el fisicalismo correría el peligro de ser falso, pero la mera posibilidad conceptual de la existencia de duplicados físicos sin propiedades fenomenológicas no implica por sí sola su necesidad metafísica. Recordemos de nuevo que según la teoría de la referencia directa (propuesta por Putnam 1975 y Kripke 1995), los designadores (nombres que refieren a objetos del mundo real) refieren a sus objetos de manera rígida (es decir, de una manera necesaria, dirigida directamente a aquello que los hace ser lo que son y no otra cosa) a través de todos los mundos posibles. Y el conjunto de todos los mundos posibles también incluye aquellos mundos conceptualmente imaginables, por tanto, no podríamos concebir un mundo donde el agua dejara de ser $H_2O$. Pues lo mismo ocurre con los conceptos fenoménicos. Parece ser que aquello que caracterizaría a los conceptos fenoménicos es, precisamente, su ser fenoménicos, esto es, entre otras cosas, su carácter introspectivo, pero parece que no sería posible concebir, como tampoco lo es en el caso del agua, un mundo que sea posible, aunque sea de manera conceptual, donde existan conceptos fenoménicos y éstos no sean introspectivos, del mismo modo que no es concebible un mundo en el que el agua sea como el agua pero no sea $H_2O$. Así, vemos que la posibilidad conceptual de la existencia de zombis es ilusoria, porque no existe ni tan siquiera la posibilidad conceptual de la existencia de duplicados que aunque parezcan tener conceptos fenoménicos realmente carezcan de fenomenología. Una de dos, o los duplicados físicos que parecen tener fenomenología realmente la tienen (porque si actúan como teniendo experiencias fenoménicas será porque realmente las tienen) o debemos negar la posibilidad de la existencia de zombis (dado que es necesario que todo aquel ser que parezca tener fenomenología, realmente tenga acceso introspectivo). La conclusión es que finalmente, la existencia de zombis es metafísicamente imposible por mucho que sea conceptualmente posible (Tye 1995: 194-200). En definitiva, si un observador externo es capaz de realizar inferencias sobre lo que está sucediendo en el interior del gemelo zombi de Chalmers, entonces no hay diferencia alguna entre Chalmers y su

[100] Algo similar pero con su estilo más irónico nos cuenta Dennett: "*Todo esto está muy bien; todos esos detalles sobre el ADN, las proteínas y todo lo demás, pero puedo imaginar la posibilidad de descubrir una entidad que fuera y actuara como un gato, de la sangre que circulase por sus venas al ADN de sus «células», pero que no estuviera vivo. (¿Puedo realmente? Por supuesto: ahí está, maullando, y después Dios me murmura al oído, «¡no está vivo! ¡No es más que un no-sé-qué mecánico hecho de ADN!». Y yo, en mi imaginación, Le creo.)"* (Dennett, 1995; 295).

gemelo zombi, pues tal y como ese zombi está definido no solo posee estados internos sino que en cierta forma debería *ser consciente* de ellos para poder ser capaz de informar acerca de ellos, ¿o acaso alguien es capaz de relatar lo que sucede en su interior sin ser consciente de lo que sucede en su interior? El gemelo zombi de Chalmers es capaz de decirnos que está disfrutando de una hermosa puesta de sol mientras saborea gustosamente un helado al lado de su hermosa novia zombi con la que desea pasar el resto de su vida y tener media docena de niños zombis, sin embargo, nada de esto es cierto, sabemos que nos está mintiendo, que en verdad, no está sintiendo absolutamente nada de lo que dice. O algo se me escapa, o existe alguna contradicción, una disonancia entre la definición de lo que es un zombi y lo que no es, o el mismo Chalmers es un zombi y el mismo no lo sabe, o todos somos zombis sin saberlo, o la fantasía zombi se encuentra tan lejos de ser filosóficamente útil que quizás debería ser desechada para siempre.

Otros autores han sugerido la posibilidad de que los zombis sean concebibles, pero afirman que no por eso debemos rechazar el fisicalismo. Esta posibilidad ha sido analizada de manera independiente por Stalnaker (2002), Hawthorne (2002) o Braddon-Mitchell (2003). Se trata del análisis condicional de los conceptos fenoménicos, la idea es que la concebibilidad de zombis es compatible con el fisicalismo si aplicamos a los conceptos fenoménicos una determinada estructura condicional a priori verdadera. Dicha estructura es la siguiente:

(AC1) Si el mundo contiene estados no físicos de algún tipo relevante, entonces nuestros conceptos fenoménicos se refieren (rígidamente) a esos estados.

(AC2) Si el mundo es meramente físico, entonces nuestros conceptos fenoménicos refieren (rígidamente) a los estados físicos que juegan algún tipo de papel funcional relevante.

Veamos cómo los condicionales proporcionan al fisicalista una respuesta al argumento de los zombis. En primer lugar, el análisis condicional parece explicar la intuición zombi: la posibilidad de zombis no puede descartarse a priori. Si el mundo real contiene estados fenoménicos no físicos (AC1), los zombis son posibles, se trataría de un duplicado físico de nuestro mundo ausente de aspectos fenoménicos. Por otro lado, si el mundo real es meramente físico (AC2), entonces los zombis no son posibles, en este caso nuestros conceptos fenoménicos denotan algo físico, y por tanto cualquier duplicado físico de un humano también será un duplicado fenoménico. Pero como no podemos saber a priori qué clase de mundo es el mundo real, no podemos descartar la posibilidad de zombis a priori; es por eso que los zombis son concebibles. En segundo lugar, el análisis condicional bloquea el argumento zombi atacando su segunda premisa, recordemos (2) *si los zombis son concebibles, entonces son posibles.* Si el mundo real contiene estados fenoménicos no físicos, entonces los zombis son posibles, mientras que si el mundo real es meramente físico, entonces los zombis no son posibles. De ello se desprende que no podemos inferir la posibilidad metafísica de zombis a partir de su concebibilidad. El análisis condicional ofrece una explicación

de por qué falla la segunda premisa del argumento zombi, y al mismo tiempo nos permite sostener que existe, en general, una estrecha relación entre lo concebible y lo posible. Los conceptos fenoménicos poseen una estructura condicional de este tipo, y parece plausible que la concebibilidad implique la posibilidad de proposiciones que no contengan tales conceptos (Haukioja, 2008: 146-147). Con el análisis condicional sucede algo así (Majeed, 2014: 239-241):

1) (AC1) y (AC2) son verdaderas.

2) Si (AC1) y (AC2) son verdaderas, los mundos zombi sólo son metafísicamente posibles si hay estados fenoménicos no físicos en el mundo real.

3) Los mundos zombi son concebibles sin importar si hay o no estados fenoménicos no físicos en el mundo real.

4) Por tanto, en el caso en que concebimos mundos zombis cuando el mundo real carece de estados fenoménicos no físicos, estaríamos concibiendo un escenario que es metafísicamente imposible.

5) Por tanto, la concebibilidad de mundos zombis no implica su posibilidad metafísica.

Pensemos en el dolor. Si resulta que hay estados de dolor no físicos, como indicaría el primer condicional, entonces el dolor se referirá a esos estados no físicos. Pero si resulta que no hay tales estados de dolor no físicos, no concluiríamos de inmediato que no hay dolores, por el contrario, llegaríamos a la conclusión de que "dolor" debe referirse (o supervenir) a algo físico. La reflexión sobre estos casos parece mostrar que nuestros conceptos fenoménicos, en cierto modo, podrían preferir referentes no físicos, pero en ausencia de tales referentes, los referentes físicos lo harían adecuadamente (Hawthorne, 2002: 26). En definitiva, el análisis condicional de los conceptos fenoménicos pretende dar al fisicalista una forma de entender los conceptos fenoménicos que les permita aceptar la intuición zombi, aceptar que la concebibilidad es generalmente una buena guía para la posibilidad, y sin embargo rechazar la conclusión de que los zombis son metafísicamente posibles. [101]

Por último hay otro argumento al que deberían responder los partidarios de la aventura zombi: el de los anti-zombis (Frankish, 2007). Aquí la estrategia es bien diferente, ni se ataca (1) ni se ataca (2), lo que se ataca es el argumento en su conjunto, se trata de un argumento similar al de los zombis solo que con anti-zombis. Un anti-zombi es un ser que es un duplicado físico de un ser humano, que habita en un universo que es físicamente un duplicado del nuestro, y que además tiene exactamente las mismas experiencias

[101] Criticas del análisis condicional en Alter (2007), Yetter-Chappell (2013) y Chalmers (2010). Alter por ejemplo presenta tres objeciones: primero, afirma que los condicionales empleados en el análisis no son a priori, en segundo lugar, dice que debe haber algo mal en esta estrategia, ya que podríamos usar un argumento similar para rechazar el propio análisis condicional, y en tercer lugar dice que el análisis condicional no puede explicar la auténtica intuición zombi. No entraré en los detalles de estas objeciones. Hay una contrarréplica de las objeciones de Alter en Haukioja (2008).

fenoménicas que nosotros tenemos. En el mundo anti-zombi, la conciencia es un fenómeno físico, ya que superviene metafísicamente a las propiedades microfísicas del mundo, por lo que en el mundo anti-zombi el fisicalismo es verdadero. Para un fisicalista no hay diferencia alguna entre nosotros y los anti-zombis, entre nuestro mundo y el mundo anti-zombi. El argumento dice así (p. 653):

(1) Los anti-zombis son concebibles.

(2) Si los anti-zombis son concebibles, entonces los anti-zombis son posibles.

(3) Si los anti-zombis son posibles, entonces la conciencia es física.

(4) Por tanto, la conciencia es física.[102]

La estrategia más común del defensor del argumento zombi es rechazar (1), que los anti-zombis sean concebibles, esto es, que sea concebible un mundo enteramente formado por propiedades físicas y en el que se den experiencias conscientes. Pero lo mismo podría sugerirse en sentido contrario, las mismas razones tenemos para rechazar que sea concebible un mundo en el que hay ciertas propiedades no físicas que son las causantes de que haya experiencias conscientes. Finalmente ambas premisas parecen estar a la par en la medida en que ambas nos obligan a concebir la ausencia de algo, propiedades fenoménicas en el caso de los zombis, propiedades no físicas en el caso de los anti-zombis. Pero los dos argumentos no pueden ser verdaderos al mismo tiempo, ya que si los hechos puramente físicos son los que hacen que los anti-zombis sean conscientes, entonces los mismos hechos físicos exactamente iguales en los zombis los harían conscientes también, y por tanto no serían zombis después de todo. Así es como el argumento anti-zombi acaba neutralizando al argumento zombi, pero no porque uno sea correcto y el otro no, sino porque ambos por la misma razón deben ser defectuosos.[103] La cuestión es que si rechazamos el argumento zombi en base a la premisa (1) también nos veríamos obligados a rechazar los anti-zombis por una razón muy similar. Ahora bien, si no podemos hacer nada con ninguna de las premisas (1) de ambos argumentos porque están formuladas a la par, y además, la verdad de esta premisa está estrechamente conectada con la verdad de la premisa (2), de la cual ya hemos expresado nuestras dudas, esto me hace pensar que, o hay algo que está mal en lo que decimos cuando queremos decir zombi y cuando queremos decir anti-zombi, o hay algo que está mal en ambos argumentos tomados en su conjunto. Ciertamente, no parece que todo esto tenga ningún sentido, finalmente se ha convertido en un círculo vicioso del que no queda claro cómo podríamos salir.

[102] Véase que el argumento es idéntico al citado en la nota 91 de este trabajo y tomado de (Chalmers, 2002b: 249) solo que con anti-zombis.

[103] Argumentos similares en Marton (1998) y Sturgeon (2000). Sturgeon por ejemplo argumenta que es posible que existan zombis, pero que es igualmente posible que los zombis sean imposibles. Son conclusiones contradictorias, su conclusión es que las inferencias sobre la concebibilidad-posibilidad deben ser, de algún modo, defectuosas.

Con este experimento mental hemos llegado *casi* a un punto de no retorno en la paranoia colectiva que afecta a determinados filósofos en su ansiosa obsesión por dar a esos fenómenos cualitativos conscientes un lugar determinado en una misteriosa naturaleza situada en algún lugar que poco tiene que ver con la naturaleza conocida de las propiedades físicas (esperemos no rebasar la barrera del casi). Pero lo que es verdaderamente asombroso es que ni ellos mismos saben cuál es ese lugar que realmente les corresponde. No he encontrado hasta ahora ningún argumento positivo convincente de lo que son los qualia de la conciencia, ni de una explicación de estos en los términos en los que están definidos en la introducción de este artículo, sin que al final se llegue a algún tipo de entidad misteriosa o inexplicable. Los qualófilos se limitan a decirnos lo que los qualia no son, donde no están, y que métodos no son buenos para captarlos, pero no tienen ni la menor idea de lo que son, de donde están, ni de como captarlos. Permanecen tan fieles a su propia intuición que no parece que pase por su cabeza la posibilidad de que una vez más, nuestras intuiciones nos hayan podido jugar una mala pasada.

El asunto de los zombis (así como el de los murciélagos, espectros invertidos y habitaciones monocromáticas) está estrechamente ligado a otra intuición que el propio Chalmers ha explotado hasta la saciedad, y debemos decirlo, con un éxito notable, se trata del *problema difícil* de la conciencia. Según Chalmers parece haber dos problemas en la explicación de nuestros estados mentales conscientes, cualitativos, intrínsecos, inefables, subjetivos y aprehensibles solo desde el punto de vista de la primera persona, son el problema fácil y el problema difícil de la conciencia. El problema fácil es fácil no porque sea precisamente fácil sino porque es imaginable, concebible, posible y muy probable que la ciencia cognitiva y las neurociencias lleguen a explicar satisfactoriamente esos aspectos de la mente que implican a su funcionamiento computacional y a sus mecanismos estructurales neurológicos, se trata de: *la habilidad de discriminar, categorizar, y reaccionar a estímulos medioambientales, la integración de la información a través del sistema cognitivo, la capacidad para informar de los estados mentales, la habilidad para acceder a nuestros propios estados internos, la focalización de la atención, el control deliberado del comportamiento o la diferencia entre la vigilia y el sueño* (Chalmers, 1995: 201). Todos estos son problemas susceptibles de ser explicados por una ciencia de la conciencia que ya tenemos sobre la mesa, no haría falta nada más, solo investigación. Ahora bien, el verdadero problema, el problema difícil, el problema duro, es aquel que involucra a la experiencia en sí misma, cuando pensamos o percibimos algo hay algún tipo de procesamiento de la información, pero también un aspecto subjetivo, un algo que es como ser ese algo (Nagel, 1974). El problema difícil es difícil porque agota toda explicación física y funcional, ¿cómo es posible que a partir de procesos físicos en el cerebro tengamos experiencias conscientes subjetivas? (Chalmers, 1995, 1999, 2002a). El problema es tan difícil que la única respuesta posible es para Chalmers aceptar un tipo de dualismo al que llama *dualismo de las propiedades fundamentales,* en el que ve las propiedades mentales conscientes como constituyentes básicos de la

realidad, al mismo nivel que las propiedades físicas fundamentales, como la carga electromagnética, la masa o el espacio-tiempo. Estos pueden interactuar de modos causales y legaliformes con otras propiedades fundamentales, como son las de la física, pero ontológicamente su existencia no depende ni se deriva de otras propiedades (Chalmers, 1999). Así como todos los objetos de la física y de las ciencias naturales pueden ser analizados en términos de estructuras y sus relaciones, la conciencia, entendida en el sentido de Chalmers, es algo que esta fuera de dichas estructuras y relaciones[104].

Es muy posible que el problema difícil detectado por Chalmers no sea la formulación de un auténtico problema sino una estrategia más para rechazar el funcionalismo fisicalista, una estrategia que sin duda viene apoyada por los argumentos anti-fisicalistas analizados a lo largo de este trabajo, por eso, todo defensor de este tipo de experimentos mentales encontrará razonable la división sugerida por Chalmers entre problemas fáciles y difíciles. Por un lado están todas esas estructuras neurológicas y su funcionamiento computacional, y por el otro, un único problema, el problema difícil, explicar la experiencia consciente, el como que es ser algo, la primera persona, etc. Pero, ¿qué tipo de racionalidad se emplea para hacer una división de este tipo? ¿cómo puede alguien afirmar que aun solucionando todos los problemas fáciles aun quedaría algo por solucionar? ¿acaso sabemos todo lo que debemos saber acerca de esos problemas supuestamente fáciles? Afirmar que hay un problema difícil es en el mejor de los casos una petición de principio, ni Chalmers puede demostrar que tal problema exista ni nosotros podemos demostrar que no, que todo lo que hay por saber acerca de nuestras experiencias conscientes nos lo proporcionaran la ciencia cognitiva y la neurociencia en su completo desarrollo. La cuestión es: ¿Está realmente bien definido el problema difícil? No encuentro razones suficientes como para pensar que hay un problema difícil porque no encuentro razones suficientes como para pensar que el fenómeno definido por tal problema exista. Argumentaré que no hay un problema difícil después de todo.

Hay al menos tres líneas de investigación desde donde abordar críticamente la cuestión del problema difícil de la experiencia consciente y cualquiera de ellas es capaz de proporcionar respuestas satisfactorias a la intuición planteada por Chalmers: el eliminativismo[105], el reduccionismo fuerte[106] y el reduccionismo

[104] Obsérvese que si aceptamos la intuición de Chalmers, si aceptamos que la conciencia es una propiedad fundamental de la naturaleza, a la altura del tiempo, el espacio o la masa, la idea del panpsiquismo (ya comentada anteriormente) no sería tan descabellada. En un sentido amplio de conciencia, esta es experiencia y es información, toda experiencia exige información, la experiencia esta por doquier y también la información, por tanto, también la conciencia como propiedad fundamental de la naturaleza la podemos encontrar por todas partes. Nadie podría negar, ni el propio Chalmers lo hace, que esta visión, es extremadamente especulativa, yo encuentro al menos dos objeciones. Primero, colocar la conciencia a la altura de las propiedades fundamentales de la naturaleza significa o bien que la conciencia siempre ha estado ahí o bien que es un fenómeno de surgimiento tardío, la primera opción exige una concepción tan basta de la noción de conciencia que es dudoso que sea esta la que está en juego, y en el segundo caso, en caso de que la conciencia fuera la hermana menor de las propiedades fundamentales tendríamos que solucionar los problemas de cuando surgió, a fin de que y a partir de qué. Y segundo, ¿es la conciencia experiencia e información?, si esto es así debemos exigir una explicación coherente de qué tipo de relación hay entre esos tres conceptos.

[105] El eliminativismo sostiene que no hay ningún problema con la experiencia consciente porque no existe tal cosa. Es una posición defendida por Rey (1983), Dennett (1978, 1988, 1995), Churchland (1981) o Wilkes (1984). Pero esto puede parecer absurdo: ¿cómo se puede negar que existe la experiencia consciente? La conciencia podría ser la única cosa que es cierta en nuestra epistemología. Pero los eliminativistas sostienen que la conciencia es una construcción filosófica defectuosa, y por tanto problemática, por lo que podría ser rechazada sin caer en el absurdo. Es la definición de la experiencia consciente como algo no funcional lo que hace que el problema difícil sea científicamente

débil[107]. Sin embargo, mi estrategia aquí será algo diferente, argumentaré que el problema difícil de la experiencia consciente surge de un error categorial, esto es, que no hay diferencia entre los problemas fáciles y el difícil, ya que este último es una ilusión causada por la pseudo-profundidad que con frecuencia acompaña a todo error categorial. Me explicaré. Un error categorial se produce cuando se intenta aplicar una categoría conceptual a un problema determinado o a un objeto, pero esa categoría conceptual no pertenece al problema u objeto en cuestión. Por ejemplo, si alguien me preguntara por la composición química del número pi, mi respuesta sería que la composición química no es relevante en el contexto matemático en el que se desenvuelve todo lo concerniente al número pi, ya que pi es un número irracional, una constante matemática que se emplea frecuentemente en matemáticas, física e ingeniería, que se simboliza por la decimosexta letra del alfabeto griego (π) y cuyo valor es 3,14159265358979…, valor que equivale a la relación entre la longitud de una circunferencia y su diámetro, pero nada hay de relevante en la composición química de tal número. Relacionar el número pi con una determinada composición química es sin duda un claro ejemplo de error categorial. Lo mismo sucede con el problema difícil de Chalmers. En primer lugar, para responder al *porqué* de las cuestiones sobre la experiencia consciente deberíamos tratarla como un fenómeno biológico, son cuestiones que conciernen directamente a la biología evolutiva. Así como para explicar la circulación de la sangre un biólogo evolutivo debe tener alguna historia que contar sobre cómo evolucionó el corazón, cuando hablamos de la experiencia consciente algo nos tendrá que decir sobre su aparición en un determinado linaje de homínidos. Esta es sin duda una cuestión primordial para la biología evolutiva. Esto no significa que debamos esperar una respuesta a corto plazo, pues posiblemente no llegue nunca, pero sí que la respuesta al *porqué* llegará en todo caso de una biología evolutiva madura. Las cuestiones históricas sobre los rasgos de comportamiento son muy difíciles de abordar, sobre todo cuando hay pocas (quizás ninguna) otra especie con la que compararnos adecuadamente y además por razones obvias el registro fósil no puede decirnos mucho al

irrelevante. En este sentido la conciencia podría ser eliminada de nuestra ontología. Si eso ocurre, obviamente, no hay ningún problema difícil de qué preocuparse. Sin conciencia, no hay problema.

[106] El reduccionismo fuerte, desde enfoques funcionalistas, dice que la conciencia no es más que un proceso funcional. La versión más elaborada de este punto de vista es la hipótesis del "espacio de trabajo global", que sostiene que los estados conscientes son estados mentales disponibles para el procesamiento de una amplia gama de sistemas cognitivos (Baars 1988, 1997; Dehaene y Naccache 2001). Este espacio de trabajo global puede ser funcionalmente caracterizado y puede ser objeto de interpretación neurológica. Versiones de este tipo de reduccionismo funcionalista pero con un enfoque más neurocientífico en Churchland, P.S. (1986), Crick (1994), y Koch (2004). Otro enfoque fuertemente reductor conocido como "representacionalismo de primer orden" sostiene que los estados conscientes son estados que representan el mundo de determinada manera (Dretske 1995, Tye 1995, 2000).

[107] El reduccionismo débil, en contraste con la versión fuerte, sostiene que la experiencia consciente es un fenómeno básico, y que no es posible descomponerlo en elementos simples no conscientes. Se mantiene la opinión de que podemos identificar la conciencia con propiedades físicas si la teoría más parsimoniosa y productiva apoya dicha identidad (Block 2002, Block y Stalnaker 1999, Hill 1997, Loar 1997, Papineau, 2002, Perry 2001). Es más, una vez que la identidad ha sido establecida, no hay más explicación. Las identidades no tienen una explicación: una cosa simplemente es lo que es. Preguntar cómo podría ser que el agua es $H_2O$, una vez que tenemos la presentación más parsimoniosa de los hechos, es ir más allá de todo cuestionamiento significativo, y lo mismo vale para la identidad de los estados conscientes con estados físicos. Pero aún queda la cuestión de por qué la afirmación de la identidad de los reduccionistas es tan contraria a la intuición. Aquí el reduccionismo débil apela a la "*estrategia de los conceptos fenoménicos*" (PCS) (Stoljar, 2005). La PCS sostiene que el problema difícil no es el resultado de un dualismo de hechos, fenoménicos y físicos, sino más bien un dualismo de conceptos, un concepto físico explicable desde la tercera persona, y otro fenoménico explicable desde la primera persona. Pero debido a las diferencias subjetivas en estos modos de acceso conceptual, la conciencia no parece ser intuitivamente física, sin embargo, una vez que entendemos las diferencias entre los dos conceptos, no hay necesidad de aceptar esta intuición.

respecto. En segundo lugar, la respuesta al *cómo* la conciencia fenoménica es posible es una cuestión para la ciencia cognitiva y la neurobiología. De nuevo, si nos cuestionamos sobre cómo funciona el corazón, la respuesta nos la da la anatomía y la biología molecular, no veo ninguna razón por la que las cosas deban ser diferentes en el caso de la conciencia. Una vez que hayamos respondido al cómo y al porqué de la conciencia, ¿qué más hay que decir? Chalmers, Nagel y otros dirán que *"todavía no nos han dicho lo que se siente al ser un murciélago, o un ser humano o un zombi",* pero lo que nos debemos cuestionar es si tiene sentido preguntar sobre el cómo y el por qué en cualquier otro sentido que no sean los que acabamos de discutir. Por supuesto una explicación no es lo mismo que una experiencia, pero ahí se encuentra el error categorial, se trata de dos categorías completamente independientes, como la composición química y el número pi. Es obvio que no puedo experimentar lo que es ser tú, pero puedo tener potencialmente, una explicación completa de cómo y por qué se puede ser tú.

Aunque hay razones suficientes como para pensar que el problema difícil, como un problema ontológico de la experiencia consciente, es una ilusión teórica, hay un aspecto de dicho problema que conviene observar más profundamente, se trata del aspecto conceptual, la amplia brecha explicativa que parece abrirse entre nuestras descripciones físicas y funcionales del mundo y el mundo de nuestras sensaciones subjetivas, la distancia que parece haber entre sentir un dolor y tener estimuladas las fibras C. Esta distancia entre la explicación de la experiencia y la experiencia en sí ha hecho que muchos filósofos se cuestionen si es ahí donde se encuentra el auténtico problema difícil. Para muchos es un camino intransitable, una brecha sobre la que no es posible tender puentes de unión, para algunos incluso se trata de dos realidades diferentes y diferenciadas, y para otros, entre los que me incluyo, sigue siendo la misma ilusión sólo que vista desde otro ángulo, no hay tal brecha, hay dos modos distintos de hablar de lo mismo, pero solo uno de ellos posee la fuerza explicativa necesaria como para establecer enunciados verdaderos. Veámoslo.

**La brecha explicativa**

---

*"El sentimiento de un abismo infranqueable entre la conciencia y el proceso cerebral: ¿cómo es que esto no surge en las consideraciones de la vida ordinaria? Esta idea de una diferencia de clase va acompañada de un ligero mareo […] ¿Cuándo ocurre este sentimiento en el presente caso? Ocurre cuando, por ejemplo, dirijo mi atención de un modo particular a mi propia conciencia y, asombrado, me digo a mí mismo: ¡se supone que ESTO ha de ser producido por un proceso en el cerebro! […] ¡Con toda seguridad ésta es la cosa más extraña que puede haber!"* (Wittgenstein, 1958/1988). *"¿Cómo es posible que los estados conscientes dependan de estados en el cerebro? ¿Cómo puede la fenomenología del color surgir de la húmeda materia gris?"* (McGuinn, 1989). No son pocos los filósofos que piensan que cuestiones como estas ponen en un

serio apuro al funcionalismo materialista. Argumentan que no hay forma de cerrar la brecha abierta entre el cerebro material y el mundo de la experiencia consciente (Levine, 1983), y que el materialismo está lejos de dar una respuesta satisfactoria al llamado *problema difícil* de porqué los cerebros ocasionan, producen, causan o dan lugar a algo tan familiar pero tan misterioso como los qualia de la conciencia (Chalmers, 1996).

Hay un conjunto de argumentos que explotan la existencia de una *brecha explicativa* (Levine, 1983), un hueco, un espacio intransitable entre la descripción física del mundo y el mundo de las sensaciones subjetivas. La idea es que la ciencia natural nunca encontrará las herramientas conceptuales necesarias para explicar la conciencia fenoménica, ya que entre las teorías científicas sobre el mundo natural y el mundo de las experiencias conscientes se abre una brecha conceptual que no será posible cerrar. No tenemos, ni tendremos, la más mínima idea de cómo rellenar el abismo que se abre entre los disparos de las fibras C y nuestra sensación de dolor. El "*argumento de la brecha explicativa*" desarrolla como cierta la idea de que las teorías fisicalistas y funcionalistas son incapaces de establecer una explicación del carácter cualitativo de los estados mentales, es decir, que simplemente partiendo de las propiedades físico-funcionales de los estados sensoriales no podemos dar cuenta en absoluto de su contenido fenomenológico. En definitiva, hay algo inexplicable en los estados mentales conscientes a partir de estados puramente físicos.

Aunque parezca un tanto paradójico, las conclusiones a las que llega Levine provienen de una argumentación contra los argumentos modales y anti-reduccionistas elaborados por Kripke[108] en su ya clásico *El nombrar y la necesidad*. La argumentación de Levine es un intento de transformar el argumento ontológico de Kripke en un argumento epistemológico. De hecho, para Levine, ni el argumento de Kripke ni los basados en la posibilidad por concebibilidad (véanse el espectro invertido, la tierra invertida o la posibilidad lógica de zombis, ya vistos más arriba), ni los basados en el conocimiento o la perspectiva (véase Mary de Jackson y los murciélagos de Nagel) son suficientes como para establecer tesis ontológicas, pues pertenecen estrictamente al terreno epistemológico.

Recordemos de nuevo que según el argumento modal de Kripke todos los enunciados de identidad entre nombres propios y términos generales verdaderos son (metafísicamente) necesarios, porque se considera que tales expresiones son designadores rígidos (términos que denotan a un mismo objeto en todos los mundos posibles) de ciertas entidades. Al analizar tales enunciados, se encuentra una diferencia esencial entre los enunciados de identidad pertenecientes a las ciencias físicas tales como (1) *el calor es el*

[108] Ya hemos visto como uno de los primeros y más contundentes ataques contra el fisicalismo de tipos fue desarrollado por Saúl Kripke a través de sus consideraciones sobre las propiedades modales de los enunciados de identidad (lo he dicho en la introducción de este trabajo). En el ensayo "*Identidad y Necesidad*" (Kripke, 1978), muestra la manera en que estas consideraciones muestran que no existen enunciados de identidad contingentes. La teoría kripkeana ha servido frecuentemente como apoyo a las posiciones dualistas y no reduccionistas, ya que da cabida a elementos de clase natural permitiendo la existencia de elementos no reductibles físicamente.

*movimiento de las moléculas* y los enunciados de identidad psicofísicos como (2) *el dolor es la activación de las fibras C*. La cuestión es que según Kripke (1) es verdadera pero (2) es falsa. La oración (1) es necesariamente verdadera puesto que el calor no puede existir sin el movimiento de las moléculas. No puede haber un mundo donde haya calor sin movimiento de moléculas porque cuando imaginamos un mundo así, lo que imaginamos no es realmente calor sino alguna otra cosa con sus mismas propiedades fenomenológicas. Por tanto, la identificación del calor con movimiento de moléculas, es una identificación necesaria. No sucede lo mismo en enunciados de identidad como (2), pues podemos concebir un mundo donde haya sensación de dolor sin que se activen las fibras C. No es concebible un mundo donde exista algo con todas las propiedades fenomenológicas de las sensaciones de dolor sin ser realmente sensaciones de dolor. Esta es la diferencia esencial. En el caso de las entidades psicológicas, la manera en que el fenómeno se presenta no puede ser diferente del fenómeno mismo. Además, según Kripke al *fijar la referencia de un término*, tanto en el caso de *calor* como en el de *sensación de dolor*, la referencia se fija apelando a la manera en que el calor y el dolor son sentidos respectivamente. En el primer caso, se trata de una propiedad *accidental o contingente* del calor, es decir, que para que algo sea calor pero no sea movimiento de moléculas, ese algo tiene que tener las mismas propiedades fenomenológicas del calor, pero no será calor. En el segundo caso la manera en que el dolor es sentido es una propiedad *esencial* de la sensación de dolor, pues no es posible separar la sensación de dolor de sus propiedades fenomenológicas, es decir, no puede haber algo que tenga las mismas propiedades fenomenológicas de las sensaciones de dolor sin ser una sensación de dolor. Por tanto (2) no es necesariamente verdadera, y todas las identidades psicofísicas como (2) pudieran ser falsas, y en consecuencia, la teoría de la identidad psicofísica, que no es otra cosa que un conjunto de enunciados de este tipo, es falsa.

El argumento puede ser reformulado contra el funcionalismo y en defensa de la existencia de los *qualia.* Supongamos que cierta teoría funcionalista de la mente implica un enunciado como (3) *tener una sensación de rojo es estar en un estado funcional F.* Si (3) es verdadera, es necesariamente verdadera, por lo que no puede haber un mundo posible en el que (3) sea falsa. Sin embargo, es concebible una situación en la que un individuo esté en el estado funcional *F* sin tener la sensación de rojo, (pensemos en un robot o un zombi). Por tanto, (3), y todos los enunciados funcionalistas de identidad como (3), son falsos (Levine, 1983: 355)[109]. Así, Kripke nos dice que podemos imaginar que tengamos dolor sin que haya descarga en las fibras C o que tengamos la sensación de rojo sin estar en un estado funcional determinado. Pero según Levine esto es imaginable desde un punto de vista estrictamente epistemológico, haría falta un argumento mucho más fuerte para situamos en el punto de vista ontológico.

[109] Los estados funcionales son más abstractos que los estados físicos, y su realización es posible en una amplia variedad de componentes físicos, por esto la intuición de que los dolores pueden existir sin el disparo de las fibras C se explica mejor en términos de la realizabilidad múltiple de los estados mentales.

Según Levine, el argumento de Kripke está construido sobre las siguientes afirmaciones: (1) que los enunciados de identidad que usan designadores rígidos en ambos lados, en caso de ser verdaderos, son necesarios y verdaderos en todos los mundos posibles, y (2) es concebible que los enunciados de identidad psicofísicos sean falsos en algún mundo posible, y por tanto no son necesarios y tampoco verdaderos (Levine, 1983: 354). La idea de Levine es que (2) está basada en una intuición insuficiente como para dar sustento a una tesis metafísica y que más bien sustenta una tesis epistemológica. Así, si según Kripke podemos imaginar que tenemos dolor sin que se dé descarga en las fibras C, esto es imaginable desde un punto de vista estrictamente epistemológico, haría falta un argumento mucho más fuerte para situamos directamente en lo ontológico. Por consiguiente, de tal interpretación del argumento de Kripke no se sigue que el funcionalismo y el fisicalismo sean falsos, aunque Levine advierte que sí se sigue un problema diferente.

El problema es que según Levine mientras (1) ofrece explicaciones satisfactorias de todo el fenómeno (por ejemplo que "*el fenómeno que experimentamos a través de nuestras sensaciones de calor o frío, responsable de la expansión o contracción del mercurio de los termómetros, que causa que algunos gases asciendan y otros desciendan etc., es el movimiento de las moléculas"* (Levine, 1983: 355)), en (2), las explicaciones funcionalistas y fisicalistas, (como por ejemplo que al cortarnos un dedo se produce la excitación de ciertas terminaciones nerviosas que a su vez excitan las correspondientes fibras C, y que éstas despiertan ciertos mecanismos que generan determinados efectos en nuestro sistema nervioso que son los causantes del dolor) son insuficientes, pues aún queda por explicar *el carácter cualitativo,* el por qué el dolor se siente como se siente. Nada en el disparo de las fibras C explica las propiedades fenoménicas del dolor, de hecho, la conexión entre los disparos de las fibras C y nuestra identificación del dolor es completamente misteriosa, de ahí su irreductibilidad y de ahí la brecha explicativa (Levine, 1983: 357). En palabras de Levine: "*es la no-inteligibilidad de la conexión entre el sentimiento de dolor y su correlato físico, la que subyace a la contingencia aparente de esta conexión*" (Levine 1983, p. 359). El asunto de la brecha puede formularse entonces de la siguiente manera: si no hay nada en la activación de las fibras C que explique por qué tenerla lleva consigo ese carácter cualitativo, entonces inmediatamente se vuelve imaginable que haya activaciones de las fibras C sin sentir dolor y viceversa. No puede decirse lo mismo en el caso del calor y el movimiento de moléculas: cualquier rasgo que haya que explicar del calor es explicado por el movimiento molecular, nada se queda fuera. Para determinar si una reducción explica lo que pretende reducir debemos comprobar si el fenómeno que ha de ser reducido es explicado por el fenómeno reductor, y Levine está seguro de que mientras esto ocurre con la teoría termodinámica, no sucede en absoluto en una teoría físico-funcional de los qualia. No obstante, una de las consecuencias más notables del argumento de Levine favorables a los planteamientos materialistas es esa matización sobre el tipo de problema que los qualia representan para el funcionalismo y el fisicalismo en general: los

argumentos que explotan la concebibilidad lejos de habitar en el terreno metafísico, se encuentran, más bien, y en sentido estricto, en el terreno epistemológico[110]. Esto es algo que debe quedar bien claro: la brecha explicativa es, en todo caso, un fenómeno epistémico o conceptual (existen los conceptos de la experiencia consciente por un lado y los conceptos de la neurociencia por el otro), pero carente de consecuencias metafísicas (no hay dos sustancias en el mundo), el argumento de la brecha explicativa no demuestra la existencia de una brecha en la naturaleza sino de nuestra comprensión de ella[111].

Las cuestiones aquí son: ¿existe esa llamada brecha explicativa?, y si es así, ¿es posible tender puentes entre nuestras teorías funcionalistas y materialistas de la mente y nuestras experiencias conscientes, o debemos reconocer que identificar nuestras teorías psicofísicas con los hechos brutos es algo ininteligible, que hay un vacío explicativo que no tenemos idea de cómo llenar? Podríamos responder que tal brecha, en el sentido de un espacio intransitable, sencillamente no existe, que puede que exista como un espacio aun no transitado, pero no como un espacio imposible de transitar, aunque también podemos argumentar que no existe tal brecha porque no es cierto que existan ambos lados de la brecha, que no es cierto que haya dos formas de ver el mismo mundo sino una, que la brecha explicativa se esconde tras las reminiscencias de un dualismo mente-cerebro tan incrustado en nuestra intuición que nos está costando demasiado tiempo y esfuerzo abandonarlo, en definitiva, que es posible que estemos inmersos en un profundo error, y que en realidad, no debemos esperar a que nada intersecte. Tenemos un total de cuatro ángulos diferentes desde los que contemplar el asunto de la brecha[112]:

1) La brecha explicativa existe, y es insalvable, las explicaciones puramente físicas no son, y según algunos no serán, nunca suficientes, (Levine, 1983, 1993, 2001; McGinn, 1989; Jackson, 1982, 1986; Nagel, 1974; Sturgeon, 2000; Searle, 1992; Chalmers, 1999, 2007; Kripke, 1995)[113]

---

[110] Aunque algunos neo-dualistas más que ver esto como una ventaja para el materialista sacan a relucir un optimismo exacerbado esgrimiendo que cuanto más fuerte sea nuestra premisa epistemológica mayor será la esperanza de obtener una conclusión metafísica, esto es, pasan de la brecha epistemológica a la ontológica argumentando que si no podemos explicar la experiencia consciente, los qualia, en términos de procesos físicos entonces la experiencia consciente, los qualia, no pueden ser procesos físicos, o como hemos visto en los argumentos por concebibilidad, que si uno puede inferir que puede haber un mundo físico exactamente igual al nuestro pero sin experiencias conscientes, entonces es concebible que esto sea metafísicamente posible, y por tanto el fisicalismo como tesis ontológica sería falso. Véanse Chalmers, 1999, o Foster, 1989, 1991.

[111] Lo que esto supone es que tal brecha, en el sentido entendido por Levine, no destruye las tesis fisicalistas en sentido estricto, pues el fisicalismo es una tesis ontológica. Recordemos. Es la posición que defiende que todos los hechos acerca del mundo están determinados en virtud de los hechos físicos, el mundo es enteramente físico, solo la sustancia física existe, otra cosa es cómo nosotros concibamos ese mundo. No hay nada en el argumento de Levine que sea contrario al fisicalismo. Como tesis ontológica el fisicalismo sale indemne porque permanece neutral en este juego. Ahora bien, los fisicalistas tampoco son un grupo homogéneo, los hay que se comprometen exclusivamente con la tesis ontológica, esto es, que existe una relación de dependencia ontológica entre las entidades empíricas y las entidades físicas, y los hay que se comprometen con una versión más fuerte, una versión epistémica, estos no solo se comprometen con lo anterior, también con la existencia de una relación de dependencia epistémica entre los conceptos empíricos y los conceptos físicos y funcionales.

[112] En realidad no resulta fácil elaborar una clasificación exacta de donde situar a cada autor con respecto al asunto de la brecha, por ejemplo yo coloco a Nagel en el grupo 1) debido a que parece albergar pocas esperanzas de que la conciencia sea finalmente reducida a procesos físico-funcionales, pero quizás sería más correcto colocarlo en el grupo 2) o a Van Gulick lo coloco en el grupo 2) pero bien pudiera ser un fuerte representante del grupo 3) o incluso del 4). La clasificación, llegados a este punto, es más bien orientativa.

[113] Los defensores de 1) son un grupo heterogéneo. Algunos dicen que la brecha explicativa es insalvable, las experiencias y sentimientos tienen cualidades irreductiblemente subjetivas que van más allá de lo físico, la historia físico-funcional es incompleta. Se trata finalmente de una brecha epistemológica que subyace a lo que en realidad es una brecha ontológica (Jackson 1982, Chalmers 1999, Kripke 1995). Otros afirman que las cualidades fenoménicas son irreductiblemente subjetivas, aunque esto sea compatible con su ser físico (Searle 1992). Otros sostienen que la brecha explicativa podrá salvarse algún día, pero actualmente carecemos de los conceptos necesarios para que las

2) La brecha podrá algún día cerrarse, pero en la actualidad carecemos de los conceptos necesarios como para unir ambas perspectivas, la objetiva y la subjetiva, (Hardin, 1987; Clark, 1993; Van Gulick, 1993).
3) La brecha conceptual existe, pero en ningún modo afecta al fisicalismo, hay un dualismo conceptual, pero un monismo ontológico. (Aquí se sitúan todos los fisicalistas de tipo B, son los proponentes de la estrategia de los conceptos fenoménicos, Loar, 1997; Balog, 2012a; Díaz-León, 2010; Hill, 1997; Hill y McLaughlin, 1999; Tye, 1999; Perry, 2001; Papineau, 2002; Stoljar, 2005 o Levin, 2008)
4) No hay tal brecha explicativa, es una ilusión. (Aquí tenemos a los fisicalistas de tipo A, Dennett, 1995; Dretske, 1995; Wilkes, 1988; Rey, 1995 o Pauen, 2011).

Ante la afirmación 1) nada podemos hacer, si adoptamos la idea de que hay una brecha insalvable que separa nuestras concepciones científicas de las fenomenológicas estamos perdidos, nunca obtendremos una explicación completa de cómo y porqué la mente hace lo que hace. La vía pesimista es una vía muerta porque postula que hay hechos del mundo físico que debido a la constitución de nuestras capacidades cognitivas no podemos ni podremos aprehender. Para estos, la mente humana está constituida de tal manera que se dirige únicamente a objetos de naturaleza espacial, pero la conciencia es no espacial por lo que cuando trata de conocerse a sí misma no lo puede hacer. Pero obsérvese que esto no significa que la conciencia no esté relacionada con lo físico sino que estamos cognitivamente cerrados al conocimiento de ese vínculo que debe existir entre lo mental y lo físico. En definitiva, 1) exige que ampliemos los límites de la ciencia hasta un lugar en el que quepan conceptos que refieren a entidades que parecen situarse más allá de lo físico, y esto para el fisicalista, es completamente inaceptable, además de innecesario. Veremos entonces que sucede con 2), 3) y 4).

El defensor de 2) acepta la existencia de una brecha explicativa que separa los conceptos fenoménicos y los físicos, considera que hay un vacío explicativo entre los procesos neuronales que subyacen a la acción de mirar una manzana roja y el hecho de que dicha manzana se nos aparezca como roja y no como amarilla. ¿Por qué es roja y no amarilla?, si no conseguimos responder a esto entonces hay, en algún sentido, un fallo en la explicación de la propiedad fenoménica. Ahora bien, lo que el defensor de 2) no acepta es que esa brecha sea de una naturaleza tal que no sea posible cerrar en un futuro, el defensor de 2) cree, en definitiva, que debe haber leyes puente que atraviesen la barrera que separa los conceptos fenoménicos

explicaciones de las perspectivas subjetivas y objetivas se junten. Puede que los estados fenoménicos acaben siendo físicos, pero en este momento no hay una concepción clara de cómo esto podría ser, quizás una ampliación de nuestras concepciones científicas darían cabida a dichos conceptos (Nagel 1974). Otros insisten en que las experiencias y los sentimientos son tan parte del mundo natural como la vida física, la fotosíntesis o el ADN, sólo que con los conceptos que tenemos y los que somos capaces de formar, estamos cognitivamente cerrados a una explicación completa de la estructura de nuestra mente. Hay una explicación, pero está necesariamente más allá de nuestra comprensión cognitiva (McGinn, 1989). Todos ellos coinciden en que una teoría que pretenda ser explicativamente eficaz debe proporcionar una explicación del (presunto) hiato existente entre los fenómenos físicos y no-físicos que acontecen en el mundo.

de los físicos. Clark (1994) por ejemplo, sugiere en primer lugar que debemos afrontar el problema no como una cuestión particular de en qué condiciones un estímulo se nos presenta como rojo, sino como una cuestión de identidad o de similitud cualitativa, la cuestión no es porque percibimos esa manzana de color rojo, sino porqué la percibimos roja y no amarilla (Clark, 1994; 9). Según Clark cuando adoptamos esta posición el problema se convierte en una cuestión física y funcionalmente tratable. Podemos describir las condiciones en que los estímulos son cualitativamente diferentes o similares. Si conocemos la absorción precisa de los espectros de los conos que se encuentran en la retina del observador, entonces las predicciones de la mezcla de colores y su combinación se pueden cuantificar. Podemos predecir cómo cambiarán la longitud de onda de esos estímulos si la iluminación de fondo se mueve, si se desplaza en intensidad o si el observador se adapta o no a los cambios, y podemos saber si hay o no cambios en la temperatura corporal del sujeto. Podemos explicar por qué individuos con anomalías en su percepción del color perciben diferentes estímulos ante un mismo objeto, y podemos en definitiva dar cuenta de la brecha si analizamos los efectos de contraste, la constancia de matiz y brillo o los efectos de los cambios de iluminación (p. 9). Ahora bien, aun pudiendo explicar todo esto, ¿quedará algo más por explicar acerca de los qualia? ¿quedará algo más que unir para que la brecha desaparezca? Para Clark la respuesta es no, el resto es pura ilusión, las cualidades sensoriales pueden explicarse completamente dando cuenta de las relaciones de similitud y diferencia dentro de los espacios cualitativos. En la misma línea, VanGulick (1993) propone el siguiente argumento como una reconstrucción del argumento de la brecha explicativa:

P1) Los qualia como matices fenoménicos son propiedades básicas, simples; no poseen estructura[114].

Por tanto,

C2) Cualquier conexión entre los qualia y la estructura organizacional de sus substratos neurales debe ser arbitraria.

Por tanto,

C3) Las conexiones entre los qualia y sus bases neuronales no son inteligibles y se nos muestran como una brecha explicativa imposible de rellenar.

VanGulick se apoya en Hardin (1988) para rechazar el argumento. P1 es falsa. La idea de que los qualia son algo singular, excepcional, único, sin igual e inclasificable, la idea de tratar el quale relevante del color como una propiedad simple y sui generis del mundo, no es correcta, más bien, son elementos dentro de un espacio de color altamente organizado y estructurado (VanGulick, 1993: 144), y lo que es más importante, dicha organización y estructura es lo que proporciona la base para el establecimiento de conexiones

[114] La idea es que los qualia, vistos como las cualidades primarias de nuestra experiencia consciente, carecen de estructura, de ahí que su conexión con los procesos físicos subyacentes y estructurados sea completamente arbitraria.

explicativas entre los qualia y sus sustratos neuronales. Hardin (1988) proporciona algunos buenos ejemplos que nos permiten ver como el espacio de color fenomenológico posee una estructura organizacional compleja con múltiples dimensiones y relaciones de semejanza sistemáticas (por ejemplo, muchos colores tienen una calidez o frialdad asociados, y estas cualidades parecen estar asociadas con diferentes papeles funcionales). Dicha estructura relacional puede proporcionarnos no solo una mayor comprensión de la fenomenología del color, sino también tender los puentes necesarios para cerrar en un futuro la brecha. Si seguimos esta estrategia reduccionista no es difícil ver que a las múltiples experiencias de color y sus relaciones les corresponde múltiples representaciones de color y sus relaciones. Si esta es la estrategia correcta aún queda, sin duda, mucho camino por recorrer, el asunto de los qualia no se acaba aquí, el materialismo reduccionista defendido por los partidarios de la opción 2) es un programa de investigación sujeto a una hipótesis sumamente arriesgada, aunque plausible.

Los defensores de la posición 3) reconocen la existencia de una brecha conceptual, pero no ven en dicha brecha nada que haga de las tesis fisicalistas un proyecto estéril. Estos siguen la estrategia de los conceptos fenoménicos, que es, para algunos, quizás la opción más atractiva que puede tomar un fisicalista para responder al problema de la conciencia (Chalmers, 2006: 1). La estrategia de los conceptos fenoménicos ya ha sido analizada en este trabajo (especialmente en nuestro análisis del experimento mental de Jackson), pero ahora profundizaré un poco más en ella, además dejé para este apartado el análisis de algunas críticas muy recientes a dicha estrategia, por lo que serán ahora brevemente desarrolladas. Presentaré la estrategia y me cuestionaré sobre la naturaleza de dichos conceptos.

Los defensores de la estrategia de los conceptos fenoménicos coinciden en que dichos conceptos poseen una naturaleza especial, pero también coinciden en que dicha naturaleza no afecta a las posiciones fisicalistas. Según Loar (1997), Carruthers (2000), y Tye (2000), los conceptos fenoménicos son conceptos recognoscitivos de la experiencia. Un concepto recognoscitivo, a diferencia de un concepto teórico, se aplica directamente sobre la base de un conocimiento perceptual, es decir, si un concepto teórico es aquel que utiliza la ciencia para referirse a lo que es una referencia común, esto es, ciertas propiedades físico-funcionales del cerebro, un concepto fenoménico es aquel que aplicamos directamente sobre la base del conocimiento de experiencias, como cuando digo que esta es mi experiencia de una manzana roja. Perry (2001) y O'Dea (2002), por otro lado, argumentan que los conceptos fenoménicos son de algún modo indexicales, similares a los conceptos de yo, aquí o ahora. Son conceptos que escogen su referente bajo un modo indexical de presentación. Por ejemplo, el concepto *ésta* en "*ésta es mi experiencia de una manzana roja"* es un concepto indexical. Papineau (2002), por su parte, propone que los conceptos fenoménicos refieren no–descriptivamente por medio de una especie de *simulación* de las experiencias a las que se refieren, en dicha simulación lo que se produce es una activación de las mismas regiones del cerebro que se activaron en la experiencia original (algo similar es defendido en Balog, 2009). Estas son quizás las

versiones más potentes de la estrategia de los conceptos fenoménicos. No entraré en las discusiones internas de las diferentes versiones de la estrategia, ni en las críticas recibidas por las posiciones dualistas (véase especialmente la de Chalmers, 2006[115]), pero sí que trataré de analizar como desde esta estrategia se aborda el asunto de la brecha, y consideraré algunos problemas que creo deberían ser tenidos en cuenta por estos filósofos.

Ya vimos que lo atractivo de esta propuesta reside en la idea de que podemos aceptar un dualismo conceptual sin negar un monismo ontológico, ya vimos que Loar (quizás el máximo exponente de esta postura) detectó una confusión en los análisis anti-fisicalistas, estos no diferenciaban lo suficiente entre propiedades fenoménicas y conceptos fenoménicos, lo que creaba la ilusión de que estábamos ante un mundo con dos tipos de propiedades, las físicas y las mentales, pero al trasladar la fuerza de los argumentos de las propiedades a los conceptos tal ilusión desaparecía. La razón por la que no podemos dar una explicación materialista de porqué los cerebros producen propiedades fenoménicas no es que dichas propiedades no sean físicas, sino que los conceptos fenoménicos no se encuentran asociados con descripciones causales del mismo modo a como otros términos pre-teoréticos si lo están en otras áreas de la ciencia. Según los defensores de esta estrategia, lo que genera la brecha es la suposición de que todos los enunciados de identidad verdaderos deberían ser explicativos, el problema es que las identidades entre fenómenos físico-funcionales y fenomenológicos no lo son. Pero ahí está el error, de lo anterior no se sigue que los enunciados de identidad no explicativos sean falsos, podrían haber enunciados de identidad no explicativos pero verdaderos, por lo que lo que tenemos que explicar es más bien la suposición en sí misma, esto es, la intuición de que todos los enunciados de identidad deban ser explicativos[116]. En resumen estos autores no niegan la existencia de la brecha, el carácter especial de los conceptos fenoménicos nos impide detectar puentes de unión con los conceptos físico-funcionales, pero no encuentran necesario explicar dicha identidad, porque está completamente fuera de lugar exigir que una identidad sea explicativa. Si A es igual a B, A es igual a B, no procede cuestionarse sobre el porqué de que sean idénticas, explicar porque las experiencias conscientes son idénticas a ciertos estados cerebrales es algo innecesario para admitir el fisicalismo, demostrar que son idénticas ya es suficiente[117].

---

[115] Sobre este asunto Chalmers (2006) ofrece una crítica para muchos definitiva de la estrategia, acepta la existencia de los conceptos fenoménicos, pero no acepta que la estrategia de los fisicalistas denominados de tipo-B sea definitiva en favor de las posiciones fisicalistas. Para Chalmers la estrategia falla, no es lo suficientemente poderosa como para explicar nuestra situación epistémica con respecto a la conciencia, y menos aún para que esta sea explicada en términos físicos. Y esto es porque si las propiedades relevantes de los conceptos fenoménicos pueden ser explicadas en términos físicos, esas propiedades no pueden explicar la brecha explicativa, y si las propiedades pueden explicar la brecha explicativa, estas no podrán ser explicadas en términos físicos. (Chalmers, 2006: 2)

[116] Como señala Loar: "*el problema de la brecha explicativa proviene entonces de una ilusión. Lo que genera el problema es el no apreciar que puede haber dos aprehensiones directas independientes conceptualmente de una misma esencia, es decir, aprehensión demostrativa al experimentarla, y aprehensión en términos teóricos. La ilusión es de una transparencia esperada: una aprehensión directa de una propiedad debería revelar cómo está constituida internamente, y si no se revela como físicamente constituida entonces no lo es. El error es la idea de que una aprehensión directa de esencia debería ser una aprehensión transparente (…)"* (Loar, 1997: 608).

[117] Ahora bien, lo que si habría que explicar es el hecho de que exista tal identidad, o el hecho de que en dicha identidad encontremos ciertas asimetrías que no encontramos en otras identidades, quizás sea aquí donde se debería situar la brecha, aunque parece claro que esto sigue sin hacer ningún daño a la posición fisicalista, (no olvidemos que esta es una tesis ontológica).

Podemos decir que la estrategia de los conceptos fenoménicos es la opción más atractiva para la nueva ola materialista que trata de reconciliar las intuiciones sobre nuestras experiencias conscientes con el fisicalismo, recordemos que esta estrategia encuentra de un modo u otro una salida a los dilemas presentados por los dualistas: era un concepto fenoménico lo que aprendía Mary al salir de su habitación, eran los conceptos fenoménicos los que se invertían en el experimento del espectro invertido, era la carencia de conceptos fenoménicos lo que diferenciaba a nuestro gemelo zombi de nosotros, y son los conceptos fenoménicos los que generan la ilusión de que se abre una brecha insalvable entre nuestras descripciones del mundo físico y mental. Sin embargo, recientemente, una serie de autores han sugerido alternativas a esta estrategia, la idea general es que el asunto que tenemos entre manos no es meramente conceptual, después de todo, es posible defender el fisicalismo sin el apoyo de los conceptos fenoménicos (véase especialmente Prinz (2007); Ball (2009); Tye, (2009) o Fazekas (2011)). Veremos que sucede con la brecha cuando prescindimos de dichos conceptos como aquellos conceptos que debido a su naturaleza especial nos hacen ser como somos. Michael Tye es uno de esos filósofos que siempre se había mostrado fiel a la estrategia de los conceptos fenoménicos, pero en Tye (2009), parece encontrar razones para abandonar dicha estrategia, primero porque todas las propuestas sugeridas en esta línea acaban según él, siendo problemáticas, y segundo porque finalmente la noción misma de concepto fenoménico tal y como ha sido definida por los defensores de la estrategia acaba por crear más problemas de los que soluciona. Veamos su alternativa. Tye apuesta aquí por la noción de familiaridad (acquaintance[118]), pone especial énfasis en la distinción entre *conocimiento directo de una cosa* (o por familiaridad) y *conocimiento por descripción* (siguiendo a Russell, (1991) y su monismo de doble aspecto). Mientras el conocimiento por descripción es conceptual y proposicional, el conocimiento por familiaridad es no-conceptual y no proposicional, este es proporcionado por la propia toma de conciencia de la cosa, solo se requiere un encuentro entre la cosa y la experiencia de ella (Tye, 2009: 101). Desde este uso, se pierde la distinción entre conocimiento (en el sentido de *acquaintance*) y conciencia fenoménica, parece que hablemos de lo mismo[119]. Tye conecta la conciencia de una cosa con la posibilidad de cuestionarse o preguntarse acerca de esa cosa. Dice: *"Si un estado consciente es tal que, como mínimo, permite que se pregunte ¿qué es eso? con respecto a alguna entidad, y lo hace directamente en base únicamente a su carácter fenoménico, entonces soy consciente de esa entidad. Pero si un estado consciente no permite tal pregunta, entonces no soy consciente de la entidad pertinente"* (p. 14). La propuesta de Tye es sin duda arriesgada, nadie esperaba un giro tan brusco en un asunto tan central, pero, ¿qué consecuencias tiene todo esto para la brecha explicativa? La respuesta de Tye es que la ilusión de la brecha surge al ignorar la distinción entre conocimiento por descripción y conocimiento por familiaridad, "*el conocimiento que obtenemos por familiaridad del color rojo es lógicamente independiente de nuestro conocimiento acerca de otras*

[118] Traduzco aquí por familiaridad o conocimiento directo la noción de acquaintance.

[119] Quizás en este punto debiéramos exigirle a Tye una mayor precisión en la definición de la noción de acquaintance.

*verdades. Es físicamente posible… conocer todos los hechos físicos pertenecientes a la experiencia de rojo y no conocer el rojo (en ese sentido relevante de conocer)"* (p. 139). Esa independencia es la que genera la ilusión de que hay una brecha entre los hechos físicos y el fenómeno experimentado, hay una brecha entre el tipo de conocimiento que tenemos pero no en el fenómeno conocido en sí, finalmente el que haya una brecha epistémica no implica que debamos prescindir de la historia fisicalista después de todo. Pienso que lo que hace Tye es trasladar la fuerza de su argumentación de la noción de concepto fenoménico a la de conocimiento por familiaridad, pero el asunto sigue siendo el mismo, si la noción de concepto fenoménico pedía una aclaración, la de conocimiento por familiaridad también, si la idea de que hay una brecha epistémica entre nuestros conceptos fenoménicos y nuestros conceptos físicos generaba ciertos problemas, la idea de que hay una brecha epistémica entre nuestro conocimiento por familiaridad y nuestro conocimiento por descripción genera los mismos. Que no hay una brecha ontológica cada vez parece más claro, que no haya una brecha epistémica no esta tan claro, pero esta es precisamente la brecha que se pretende cerrar si queremos desarrollar una teoría fisicalista completa de nuestros estados fenoménicos[120].

Por último, según los partidarios de la opción 4), la brecha explicativa es una ilusión. La brecha que parece abrirse entre la explicación de nuestros estados fenoménicos y nuestros estados físicos no debe cerrarse sino disolverse. Una de las caracterizaciones que abre la supuesta brecha es como poco, incompleta, debería ser disuelta. El materialista de tipo A afirma que solo hay un tipo de propiedades en el mundo y que por tanto solo hay un tipo de conceptos que capten esas propiedades. El tipo A suele definirse como aquel que niega que haya una concepción especial de la conciencia fenoménica, una concepción que implique la existencia de propiedades fenoménicas especiales por encima de los procesos físico-funcionales. El tipo A parece negar lo que parece ser notoriamente manifiesto, que tenemos experiencias conscientes, pero esto son solo apariencias, el tipo A no se cuestiona sobre la existencia de la conciencia fenoménica sino sobre su naturaleza. El tipo A no cree que se abra una brecha explicativa entre las funciones cerebrales y la experiencia fenoménica porque no cree que haya tal disociación, toda la riqueza de nuestra fenomenología puede ser capturada a través de mecanismos cognitivos, más aún, sin una disolución del problema que se ha venido llamando el problema difícil de la conciencia (tratado ya en el anterior capítulo y fuertemente relacionado con este) difícilmente encontraremos los tan ansiados

[120] Prinz (2007) sustituye los conceptos fenoménicos por la idea de representaciones de nivel intermedio; Ball (2009) niega la existencia de conceptos fenoménicos pero no ofrece ninguna alternativa; Fazekas (2011) niega que los conceptos fenoménicos posean un carácter especial, explica las características peculiares de la experiencia consciente en términos de arquitectura cognitiva, dice que el lugar adecuado a tener en cuenta son las representaciones sensoriales/perceptuales y su interacción con las estructuras conceptuales generales. El problema que veo en todo este tipo de estrategias que tratan de introducir nociones puente entre nuestros estados mentales y nuestros estados cerebrales es que dichas nociones acaban resultando igualmente problemáticas. Tye no es muy preciso con su noción de acquaintance, al igual que Prinz con sus representaciones de nivel intermedio, y Fazekas nos debe una explicación más detallada de cómo es esa interacción entre las representaciones sensoriales/perceptuales y las estructuras conceptuales dentro de lo que llama la arquitectura cognitiva. Sin embargo, pienso que debemos a estos autores la osadía de escapar de la estrategia de los conceptos fenoménicos, al desmarcarse de esa línea de pensamiento y defender el fisicalismo se acercan a una posición como la que defienden los partidarios de la opción 4), negando así que haya alguna brecha conceptual o epistémica entre nuestros estados mentales y nuestros estados perceptuales.

correlatos neuronales de la conciencia[121], ya que si lo que buscamos en nuestro cerebro es algo que se encuentra al margen de las funciones que este desempeña, dichos correlatos se nos esconderán para siempre (Cohen y Dennett, 2011: 358). La brecha se abre porque pensamos que tras la experiencia de dolor subyace algo cuya naturaleza es de un modo u otro tan especial y tan íntima que difícilmente la aprehenderemos con sencillos trucos neuro-funcionales, pero desaparece en el mismo momento en el que observamos que tras la experiencia de dolor solo hay un cerebro funcionando. Obsérvese que los defensores de la opción 4) al igual que los de la opción 2) no creen que haya más en nuestras experiencias conscientes que ciertos mecanismos neuro-funcionales, pero así como los defensores de 2) creían que había una brecha o unos ciertos huecos que rellenar entre nuestra comprensión física y nuestra comprensión fenomenológica del mundo, los partidarios de la opción 4) nos recuerdan que nuestras intuiciones acerca del mundo en muchos casos necesitan pasar por un proceso de refinación, ser reemplazadas, revisadas y en última instancia incluso eliminadas. Para ilustrar lo que defienden los partidarios de la opción 4) emplearé de nuevo la analogía entre las experiencias conscientes y el vitalismo. Recordemos que el vitalismo es la idea de que los organismos vivos son fundamentalmente diferentes de los no vivos porque contienen un elemento no físico o porque están gobernados por diferentes principios. Del mismo modo podemos decir que un ser es consciente porque contiene algún elemento no físico o porque está gobernado por diferentes principios que aquel que no es consciente. En el momento en el que la biología comenzó a comprender el funcionamiento de la vida, al concepto de elan vital le quedaron dos opciones: ser reducido a aquellas propiedades que definen la vida o ser eliminado por ser un concepto impreciso. Lo más sensato resulto ser la eliminación, cuando un concepto no aclara sino que confunde el estudio de un cierto fenómeno lo más coherente es deshacerse de él. ¿Cómo podemos estar tan seguros de que no sucede lo mismo con la experiencia consciente? Muy probablemente cuando el desarrollo en neurociencia y ciencia cognitiva llegue a un nivel de desarrollo concreto, a la noción de conciencia, tal y como es entendida por los teóricos anti-fisicalistas, le quedarán las mimas opciones que a la de elan vital: ser reducida a procesos neuro-funcionales o ser eliminada por ser imprecisa.

Uno de los más férreos defensores de esta última opción es Daniel Dennett, en Dennett (1988, 1995, 2006) aboga por una eliminación de una buena parte de esa parafernalia conceptual que llevan empleando los filósofos desde hace siglos. Ve venir el día en que filósofos, científicos y legos se rían de los restos fósiles de nuestra confusión sobre la conciencia: *"Todavía parece que las teorías mecanicistas de la conciencia dejan algo sin explicar pero es solo una ilusión. De hecho, explican todos los aspectos que necesitan explicación"* (Dennett, 2006: 38). Dennett ofrece un método neutral, al que llama heterofenomenología, con el que

[121] En esta búsqueda se hallan inmersos entre muchos otros: Crick, F. and Koch, C. (1990, 1995, 2003); Lamme, V.A. (2003, 2006); Block, N. (2005, 2007); Koch, C. (2004); Koch, C. y Tsuchiya, N. (2007); Zeki, S. (2001, 2003); Zeki, S. y Bartels, A. (1999); Baars, B.J. (1989); Dehaene, S. et al. (2006) o Dehaene, S. y Naccache, L. (2001).

pretende llegar de los datos crudos a una explicación de la conciencia compatible con el método de la ciencia estándar desde la tercera persona (pp. 60-62). Tenemos:

a) "las experiencias conscientes en sí",
b) las creencias sobre esas experiencias,
c) "los juicios verbales" que expresan las creencias,
d) los enunciados de distinto tipo.

La idea es la siguiente: para el heterofenomenólogo los datos primarios son los sonidos del sujeto al mover la boca, los diversos enunciados, los datos en crudo, sin interpretar, que al interpretarse se transforman en (c), los juicios verbales, y a través de los juicios verbales o actos de habla llegamos a las creencias sobre esas experiencias, esto es a (b). La cuestión es, ¿es posible y necesario llegar hasta las experiencias conscientes en sí, hasta (a), para obtener una teoría sobre la conciencia? Dennett responde lo siguiente: *"Si (a) prevalece sobre (b) –si tenemos experiencias conscientes que no creemos tener-, esas experiencias son tan inaccesibles para el sujeto que las protagoniza como para el observador externo. Por consiguiente, el método de abordaje en primera persona no proporciona más datos útiles que la heterofenomenología. En cambio si (b) prevalece sobre (a) –si creemos que tenemos experiencias que de hecho no hemos tenido-, lo que hay que explicar son las creencias, y no las experiencias inexistentes. Entonces, ajustarse a las pautas de la heterofenomenología y considerar (b) como el conjunto máximo de datos primarios es una buena forma de evitar la necesidad de explicar datos espurios.*" (p. 61). Conclusión: la subjetividad de la experiencia esta de algún modo conectada con el acceso que el sujeto tiene de ella, las experiencias conscientes en sí no se diferencian en absoluto de las creencias que el sujeto tiene sobre esas experiencias, y por tanto no hay tal brecha explicativa, todo lo que necesita ser explicado ya es explicado.[122]

Otro defensor de esta postura es Michael Pauen, en Pauen (2011) encontramos hasta cuatro ataques diferentes a la idea de la existencia de la brecha. Primero, advierte que la brecha explicativa se sostiene sobre la plausibilidad de los experimentos mentales que he venido tratando en este trabajo, pero si finalmente Mary no aprende nada nuevo al salir de su cautiverio, si admitimos que el experimento del espectro invertido carece de sentido y negamos la posibilidad de zombis, entonces la historia de la brecha pierde todo su valor. Segundo, no hay un acceso privilegiado desde la primera persona a los qualia. Es obvio que no podemos ver las cosas desde la perspectiva en que las ven los demás, pero la brecha explicativa implica además que hay hechos acerca de nuestra experiencia en primera persona que nunca podremos conocer desde la perspectiva de la tercera persona. Pauen argumenta que si los sujetos son capaces de reconocer sus propios qualia esto debe suponer alguna diferencia en su disposición funcional, si no hay ninguna diferencia en su disposición funcional entonces los sujetos serán incapaces de reconocer

[122] Más sobre el método sugerido por Dennett en Dennett (2003, 2007), y para críticas sobre el método véase Gallagher y Zahavi (2013); Zahavi (2007); Varela y Shear (1999); Levine (1994); Soldati (2006) o Dreyfus y Kelly (2007).

sus propios qualia. Y si hay un cambio funcional en el sujeto de la experiencia entonces dicho cambio debe ser detectable desde la tercera persona. Tercero, el argumento de la brecha es circular, si necesitamos los experimentos mentales anti-fisicalistas para encontrarnos con la brecha y necesitamos la brecha para que los experimentos mentales tengan validez, entonces hay algo ahí que no funciona como debería, unos se sostienen sobre los otros, y todos finalmente se sostienen sobre la fuerte intuición de que hay algo que es como ser algo, una intuición que bien pudiera ser una ilusión. Por último Pauen ofrece un argumento histórico. Muchas de nuestras intuiciones en el pasado han podido ser explicadas y en muchos casos reducidas a elementos más simples, y dichas explicaciones son hoy en día satisfactorias, explican el fenómeno perfectamente, ¿qué nos hace pensar que con el asunto de los qualia y la conciencia es y será diferente?

¿Cuál de las cuatro posibilidades con respecto a la brecha deberíamos aceptar? Ya hemos visto que los defensores de (1) reconocen la existencia de la brecha epistémica, (algunos incluso detectan la presencia de una brecha ontológica), pero no tienen ni la menor idea de cómo unir ambas explicaciones, de hecho muchos piensan que estamos cognitivamente cerrados a dicha explicación, lo único de lo que parecen estar seguros es de que las explicaciones fenoménicas no son reducibles a explicaciones físicas, por lo que concluyen que el fisicalismo tiene aquí poco que decir, no nos proporcionan un panorama muy esperanzador. Los que se comprometen con la estrategia (2) reconocen la existencia de la brecha, pero tienen claro que la brecha existe debido a nuestra ignorancia acerca de cómo funciona nuestro cerebro. El futuro cerrará la brecha. Estos, al menos, mantienen la esperanza de descubrir las leyes puente que unan ambos mundos conceptuales, pero dan por sentado que ambos mundos, en la actualidad, existen. Los defensores de (3) no ven ningún problema en aceptar la existencia de la brecha, aunque no por ello se deshacen del fisicalismo, estos siguen la estrategia de los conceptos fenoménicos, se dedican a buscar la manera de que esos dos mundos conceptuales, el físico y el fenoménico, tengan cabida dentro de un marco enteramente físico, aceptan el dualismo conceptual, pero sin negar un monismo ontológico. El problema aquí es que al introducir nuevas nociones para tratar de encajar ambos mundos en uno, al final siempre resulta insatisfactoria la explicación de la noción introducida, por lo que solucionamos un problema con el alto coste de crear otro. Me inclino a pensar que (4), a pesar de ser la opción con menor consideración y la más contra-intuitiva de las expuestas, es la que más se acerca a una explicación satisfactoria ¿Por qué? Supongamos lo siguiente. Imaginemos que conseguimos fabricar un dispositivo que, conectado adecuadamente a la cabeza de un sujeto nos permite saber qué cambios se están produciendo en su cerebro. Imaginemos también que descubrimos que cuando nuestro sujeto tiene un estado mental con carácter fenoménico, la percepción de algo rojo o la presencia de dolor, las neuronas identificadas para tal fin se excitan y que cuando no existe tal estado mental, se mantienen en letargo. Diríamos que existe un mecanismo que permite explicar cómo la excitación del sistema cerebro-nervioso

de nuestro sujeto implementa un determinado estado con carácter fenoménico y consciente. Esto parece suficiente como para anular la posibilidad de que se abra una brecha entre la explicación física y la generación de contenido fenoménico consciente. La explicación científica basada en la ciencia neurofisiológica bastaría para rendir cuentas de la explicación de cómo se generan los estados fenoménicos, tan sólo deberíamos estudiar el modo en que la activación de ciertas neuronas implementan estados mentales con carácter fenoménico diferente según el caso. Pero si es cierto que cierta sustancia física (en nuestro caso neuronal) es capaz de implementar ciertos estados mentales en base a una serie de complejos algoritmos de entrada y salida, no será menos cierto que cualquier otra sustancia (pensemos en chips de silicio) será capaz, con el programa adecuado, de obtener idénticos resultados. Pensar lo contrario sería caer en un extraño biocentrismo que resulta difícil de defender. Habremos creado un dispositivo físicamente diferente pero funcionalmente idéntico a un ser con estados fenomenológicos conscientes ¿Es esto posible? ¿Es posible crear máquinas conscientes? Y si no lo es, ¿Qué nos impide hacerlo? La importancia de estas cuestiones con relación a nuestro tema reside en que si respondemos afirmativamente a estas preguntas, de algún modo u otro habremos demostrado que todo lo que acontece en nuestras mentes puede ser explicado si conseguimos explicar nuestros cerebros y su funcionamiento, si creamos un cerebro, sea este del material que sea, similar al de un ser humano y dicho cerebro, muestra una conducta inteligente y consciente entonces deberemos concluir que todo lo que hay por explicar sobre nuestras experiencias conscientes e inteligentes queda explicado por medio de una teoría físico-funcional de nuestros sistemas nerviosos. Sobre esto hablaré en el último apartado de este trabajo. Vamos a ello.

## La habitación china

Aunque durante décadas la idea de robots con características humanas ha estado presente en la literatura y en el cine, y aunque desde Türing[123] y el surgimiento de la Inteligencia Artificial se haya tenido muy presente, solo hace unas décadas que comienza a concederse cierta relevancia filosófica a la posible existencia de robots que funcional y psicológicamente sean idénticos a un individuo consciente (pero que ya no serían productos evolutivamente idénticos). Esta idea, un tanto desconcertante, pero sumamente atractiva para las discusiones filosóficas (y no filosóficas) actuales, ha sido criticada por diversos autores y desde diversos puntos de vista[124]. Muchos piensan que a pesar de que los estados internos de un robot

[123] En 1950, Alan Turing propuso una prueba conocida como el Test de Turing. La idea nuclear era que si una persona se comunicaba a través de un terminal de ordenador con otra persona y con un ordenador que estuvieran ocultos, y no pudiera discriminar a través de una serie de preguntas cuál era la persona y cuál el ordenador, entonces tendríamos una prueba de que el ordenador muestra la cualidad que conocemos como «inteligencia». Este Test es también llamado «Juego de Imitación». Para una aproximación histórica a la Inteligencia Artificial véase Trillas, 1998.

[124] Algunos cuestionan la tesis de la suficiencia computacional argumentando que hay ciertas habilidades humanas que nunca serán duplicadas computacionalmente (Dreyfus, 1974; Penrose, 1989), otros cuestionan la explicación computacional afirmando que el computacionalismo no es una línea de trabajo apropiada para la explicación de los procesos cognitivos (Edelman, 1989; Gibson, 1979) y otros están convencidos de que las descripciones computacionales de una sistema están vacías de contenido (Searle, 1980, 1984, 1990, 1996).

sean capaces de producir salidas electro-químicas en respuesta a entradas sensoriales, simplemente carecen de la experiencia consciente de todo aquello que son capaces de hacer, y que la biología del sistema nervioso tiene un estatus preferente y en algún sentido especial como el sustrato necesario para la mente. Un argumento que refuerza esta intuición es el de *la habitación china* (Searle 1980, 1984: 33-48 y 1992: 77-94)[125]. El escenario es el siguiente. Imaginemos que tenemos una habitación cerrada en la que se encuentra un individuo que tan sólo dispone de acceso al mundo exterior a través de dos rendijas, una de entrada, por la cual se le introducen láminas con símbolos impresos en un idioma que desconoce (pongamos por caso, el chino) y una de salida por la cual debe deslizar también láminas que lleven impresos símbolos también en chino que, sin embargo, respondan de manera adecuada a las órdenes que aparecían impresas en las láminas de entrada. Para poder dar correcta cuenta de dicha tarea, el individuo dispone de un enorme manual de instrucciones de manejo del chino que indica qué símbolos serían las salidas adecuadas a las entradas escritas en chino, aunque él desconoce por completo aquello que está haciendo. Según muestra este experimento, y según concluye Searle, en ningún caso podemos afirmar que el individuo encerrado en la habitación es conocedor del chino, porque aunque es capaz de aportar las respuestas de salida correctas a las entradas escritas en chino que se le introducen por la primera rendija, en ningún caso comprende aquello que dichas láminas en chino le están diciendo. Del mismo modo, nos dice Searle, actúa un instrumento mecánico como un computador.

Searle asegura que el individuo dentro de la habitación (que en el argumento original es el propio Searle), sólo ha manipulado símbolos sintácticamente, que es lo que en última instancia hace un computador digital, y que eso sólo no basta. No podemos llegar a afirmar que comprendemos el chino porque la sola sintaxis es insuficiente para comprender un lenguaje. Dicho de otra forma: la mera manipulación de símbolos, el hecho de instanciar un programa de ordenador, no hace que el computador comprenda un lenguaje en el sentido en que lo podría hacer un hablante nativo de dicho lenguaje.

La pretensión de Searle (1980) es elaborar un argumento en contra de lo que denomina Inteligencia Artificial Fuerte[126]. La clave del argumento es impedir la reducción de semántica a sintaxis en beneficio de la mente humana, lo que impediría a su vez una explicación funcionalista y computacional de la mente humana. En última instancia, lo que el experimento de la habitación china nos aportaría es una concepción de la mente humana como privilegiada respecto de la (posible) mente de las máquinas, ya que hay una diferencia esencial entre los seres realmente pensantes (o conscientes de lo que piensan, en el sentido en que tan sólo aquellos seres cuyos pensamientos estén dotados de semántica tienen realmente experiencia

[125] El blanco directo del argumento de Searle es el computacionalismo clásico, pero como veremos, sus críticas se extienden del mismo modo al computacionalismo de tipo conexionista.

[126] Lo que Searle llama Inteligencia Artificial Fuerte es el punto de vista según el cual un ordenador digital adecuadamente programado con las entradas y salidas de datos correctas tendría una mente en el mismo sentido en que la tenemos los seres humanos. O lo que es lo mismo, que las mentes humanas no son otra cosa que programas con entradas y salidas de datos. Por otro lado, la Inteligencia Artificial Débil es la postura que defiende que el computador es una herramienta útil para el estudio de la mente, ya que permite formular y comprobar hipótesis de un modo riguroso. Searle aceptará la segunda, y negará la primera.

consciente sobre ellos) de todos aquellos seres que sólo da la impresión de que piensan. Para Searle, los sistemas computacionales son meras versiones de silicio de lo que entendíamos por un zombi, son huecos por dentro, no hay nada en el interior de un computador que le permita cuestionarse sobre sí mismo ni comprender la semántica de las instrucciones para las que ha sido programado. Aunque el argumento original es dirigido contra la intencionalidad de las máquinas, considero que en la raíz del problema se encuentra el hecho de que el computador carece de estados conscientes, y es aquí donde este apartado adquiere su sentido dentro de este trabajo, el tipo de intencionalidad de la que habla Searle es fenomenológica, y ese es precisamente el tipo de intencionalidad inherente a los estados conscientes. No entraré aquí en la conexión entre intencionalidad y conciencia, pero parece claro, y Searle estaría de acuerdo, que si los programas fueran suficientes para la conciencia también lo serían para la intencionalidad, y que si esto finalmente fuera cierto deberíamos aceptar la Inteligencia Artificial Fuerte y por tanto rechazar el argumento de la Habitación China. El argumento de Searle se formularía así (Searle, 1984):

*(1) Los programas son puramente formales (es decir, sintácticos).*

*(2) Las mentes tienen contenidos mentales (es decir, contenidos semánticos).*

*(3) La sintaxis no es equivalente a la semántica, ni es suficiente por sí misma.*

*Por tanto,*

*(4) tener un programa, cualquier programa por sí mismo, no es suficiente ni equivalente a tener una mente.*

Varias han sido las réplicas a este argumento[127], algunas llevan a conclusiones diferentes al introducir ligeras modificaciones o añadidos al escenario original, entre ellas están: la réplica de los sistemas (la comprensión no es atribuible al individuo pero sí al sistema completo), la réplica del robot (si colocamos la computadora dentro de un robot y le permitimos interaccionar con el mundo, sería capaz de comprender incluso otros estados mentales), la réplica del simulador de cerebros (elaborando un programa que simule la secuencia real de emisiones neuronales y conexiones sinápticas de un cerebro humano), la réplica de la combinación (las tres réplicas anteriores consideradas en conjunto), la réplica de las otras mentes (si no podemos atribuir comprensión a la computadora, ¿cómo es posible atribuírsela a otra persona?) y la réplica de las mansiones múltiples (el argumento de Searle trata sobre la IA en su estado actual, pero no sabemos lo que nos deparará el futuro), todas ellas comentadas por Searle en su artículo (Searle, 1980)[128].

[127] Encontramos réplicas en: Dennett (1980, 1987), Boden (1988), Churchland, P. and Churchland, P. S. (1990), Hauser (1997), Thagard (1986), Rappaport (1986), Cole (1984), Chalmers (1999), Shaffer (2009) y muchos más.

[128] Las réplicas de Searle a la avalancha de críticas recibidas desde la publicación de su artículo han dado mucho de qué hablar. Pienso que no todas las réplicas de Searle son convincentes. Por ejemplo, su respuesta a la réplica de los sistemas es que si permitirnos que el individuo absorba todos los elementos del sistema, que memorice las reglas del libro y los bancos de datos de los símbolos chinos y que haga todos los

Aunque también podemos desmontar el argumento atacando a sus premisas. Este tipo de críticas se centran principalmente en la premisa (3), que es sobre la que se sostiene todo el argumento, aunque es posible atacar a todas ellas, como hace Dennett (Dennett, 1980, 1987). A mi parecer, las objeciones que presentan Dennett, Rapaport, y los Churchland en conjunto son definitivas. Ahí van una serie de reflexiones que pienso hacen del argumento de Searle un argumento insostenible.

1º En primer lugar, trataré de decir algo sobre la que considero que es la premisa clave para que el argumento se sostenga, la (3). A simple vista no parece una premisa problemática, es para Searle una verdad indudable e incuestionable que la sintaxis no es suficiente para la semántica. Me gustaría profundizar en este asunto y demostrar que lo que parece claro no es tan claro como parece. Se entiende por sintaxis las "*relaciones entre unidades formales sin interpretación, especialmente con respecto a cómo pueden y cómo no pueden ser concatenadas en agregados lineales*" (Moural, 2003: 180). Por otra parte, se entiende por semántica, "*la propiedad de un objeto de significar o simbolizar algo más allá de él mismo en una forma que sea entendida (al menos por algunos) usuarios del objeto*" (Moural, 2003: 181). Partiendo de estas dos definiciones, Searle asegura que es una verdad conceptual la idea de que la sintaxis no es suficiente para la semántica. Propondré, pues, el siguiente argumento:

(1´) No puede haber semántica sin sintaxis o la semántica es suficiente para la sintaxis.

(2´) No puede haber sintaxis sin semántica o la sintaxis es suficiente para la semántica.

(3´) Los computadores poseen sintaxis.

Luego,

(4´) Los computadores poseen semántica.

Con (1´) y (3`) no hay problemas, Searle estaría de acuerdo con ambas. Sin embargo, probar que la sintaxis es suficiente para la semántica, (2´), probaría que los computadores son capaces de auténtica comprensión de los símbolos que manipulan, lo que desmontaría el argumento de Searle. Para Searle la relación entre semántica y sintaxis no funciona en ambas direcciones. La cuestión es que si la semántica es suficiente, pero no necesaria para la sintaxis (esto es, no puede haber semántica sin sintaxis), y la sintaxis no es suficiente pero si necesaria para la semántica (esto es, puede haber sintaxis sin semántica) entonces la sintaxis debería ser previa a la semántica. La sintaxis debió de ser primero, pero ¿para qué una sintaxis inoperante semánticamente? Yo no encuentro ninguna razón ni mínimamente lógica de que esto pudiera ser así. Sin embargo, si admitimos que la sintaxis es necesaria y suficiente para la semántica y que la semántica es necesaria y suficiente para la sintaxis, es decir, que son de algún modo inseparables, la

cálculos mentalmente, el sistema en su conjunto seguirá sin entender chino, pero quizás ese y no otro sea el modo en que adquirimos intencionalidad, consciencia y entendimiento los seres humanos.

situación adquiere un sentido mucho más amplio y claro. Aquí es donde observo el problema, la sintaxis y la semántica son conceptualmente separables pero una separación conceptual no implica una separación real, no existen lenguajes naturales ni artificiales que tengan sintaxis sin semántica ni semántica sin sintaxis. La idea de una sintaxis aislada es absurda, no es concebible una sintaxis encerrada en un cajón sin algo más que la acompañe, es un error hablar de una sintaxis no puesta en práctica, inactiva, inoperante, aislada de algo para lo que ha sido creada. Decir que la sintaxis sola no es suficiente para la semántica carece entonces de sentido sin una puesta en marcha de esta, es como decir que una partitura es incapaz de hacer música por sí misma o que una receta de cocina no tiene ningún sabor. Es evidente que los programas tienen que ser ejecutados. Si bien el programa que se le proporciona a la persona dentro de la habitación es puramente sintáctico, la implementación de dicho programa permite la construcción de contenido semántico[129]. Esto es, no es necesario buscar el contenido semántico en el nivel de definición del programa, sino en el nivel de su implementación. Searle está confundiendo *programas* con *implementaciones* de los programas. Como bien apunta Chalmers:

*"Mientras los programas por sí mismos son objetos sintácticos, las implementaciones no lo son: estos son sistemas físicos reales con organización causal compleja, con causación física real dentro de ellos. En un computador electrónico, por ejemplo, los circuitos y los voltajes chocan entre sí de manera análoga a como las neuronas y las activaciones chocan entre sí. Es precisamente en virtud de esta causación que las implementaciones pueden tener propiedades cognitivas y, por lo tanto, semánticas"* (Chalmers, 1994: 16).

Debemos entender un programa como un algoritmo siendo ejecutado, pues solo un proceso en funcionamiento puede ser capaz de comprensión. Por consiguiente, los dominios semántico y sintáctico a pesar de no ser idénticos, pues podemos separarlos conceptualmente, deben ser tratados a la par. Pero véase que la cuestión esencial sigue en pie: ¿Cómo es posible que un computador pase del procesamiento sintáctico al contenido semántico? O lo que es lo mismo ¿Cómo consigue esto mismo, un cerebro? Rapaport trata de responder a esta cuestión en base a lo que llama el entendimiento sintáctico. Primero, debemos partir de que uno no puede decir que se halla en un dominio sintáctico si no es relativo a otro dominio que se toma como semántico, y viceversa (Rapaport, 1995, 60). Por otro lado, el *entendimiento semántico es una correspondencia entre dos dominios* (p. 49), un dominio A y un dominio B, por ejemplo. Así, un agente cognitivo entiende uno de los dominios, pongamos por caso el dominio A, en términos del otro dominio, el dominio B. Ahora bien, si el dominio A se entiende en términos del dominio B, entonces, ¿en qué términos se entiende el dominio B? Se entiende *recursivamente*, en términos de otro dominio, por

[129] La noción de implementación es la clave en este asunto. Según Chalmers (1999: 398) mientras la teoría de la computación se ocupa de objetos abstractos (como las máquinas de Turing), los sistemas cognitivos, en el mundo real, son objetos concretos, corporizados físicamente y que interactúan con otros objetos en el mundo físico. Si queremos utilizar la teoría de la computación para extraer conclusiones acerca de objetos concretos en el mundo real necesitamos un puente entre el dominio abstracto y concreto. Este puente nos lo proporciona la noción de implementación: la relación entre objetos computacionales abstractos y sistemas físicos que ocurre cuando un sistema físico *realiza* una computación o cuando una computación *describe* un sistema físico.

ejemplo el dominio C. Pero, ¿dónde acaba todo esto si es que acaba en algún lugar? Rapaport dice que todo esto termina en un dominio base que no se entiende en términos de otro dominio sino en términos de sí mismo, y para que este entendimiento en términos de sí mismo sea posible, tiene que haber un *entendimiento sintáctico*[130], esto es, debe haber un dominio que se entienda en términos de sí mismo, sin remisión, y por lo tanto, no relativo a ningún otro dominio. Dice: "*Podemos tener una "cadena" de dominios, cada uno de los cuales, excepto el primero, es un dominio semántico para su predecesor, y cada uno de los cuales, excepto el último, es un dominio sintáctico para su sucesor ¿Cómo entendemos el último? Sintácticamente"* (p. 57). Aclararé un poco todo esto, pues no parece satisfactoria la idea de que la regresión infinita por constante remisión entre dominios se rompa con una remisión circular en un dominio sintáctico inicial. La idea de Rappaport es que entender X significa o bien que X se entiende relativamente a algo (o por correspondencia, esto es, se establecen correspondencias entre un dominio desconocido y otro conocido a través de asociaciones sintácticas) o bien por la costumbre de usar X, (esto es, sin remisión entre dominios, pues un solo dominio se entiende en términos de sus partes. Este es el caso del entendimiento puramente sintáctico). Pues bien, el entendimiento sintáctico permite el entendimiento del primer dominio en términos de sus propias partes a través de algún tipo de auto-referencialidad. Esto significa que la circularidad del entendimiento sintáctico puede interpretarse como perteneciendo a un sistema auto-referencial y, por lo tanto, consecuente con la naturaleza de la mente humana[131]. Si con todo esto volvemos a la habitación china nos damos cuenta de que hay un caso de entendimiento semántico por correspondencia porque la persona que se encuentra dentro de la habitación, a la que se le exige entendimiento, cuenta con dos dominios: un dominio sintáctico, que son los símbolos chinos, y un dominio semántico, que es su lengua nativa. Si distinguimos como hace Rapaport entre actuar según las reglas (como cuando ejecutamos una acción que es nueva para nosotros y, por lo tanto, debemos actuar *según* los procedimientos con que contamos) y actuar de acuerdo a las reglas (como cuando nos hemos acostumbrado a la ejecución de una determinada acción que inicialmente pudo ser nueva y desconocida) (Rapaport, 1986: 272), vemos que el sujeto dentro de la habitación actúa del primer modo, según las reglas, pues nunca ha usado el chino, y por tanto sus primeras representaciones del chino consistirán únicamente en las reglas para el uso de los caracteres chinos. Esto significa que los símbolos chinos juegan el papel de dominio sintáctico y las instrucciones, en el idioma nativo, el del dominio semántico, y que probablemente, Searle, dentro de su habitación acabará formando representaciones mapeando ambos dominios en base a los símbolos desconocidos del chino (dominio sintáctico) y las reglas de manipulación

[130] Otra respuesta a la cuestión del dominio inicial puede encontrarse en Fodor, sin embargo, la naturaleza de lo que en Fodor es el lenguaje del pensamiento, guardará cierta distancia con lo que para Rapaport es el dominio inicial. La forma en que Fodor consigue salvar el regreso infinito es sosteniendo que *"el lenguaje del pensamiento es conocido pero no aprendido. Esto es, es innato"* (Fodor, 1975: 65). Según Fodor, lo verdaderamente importante del lenguaje del pensamiento es que sea entendido y que, por tanto, permita los procesos cognitivos del agente.

[131] A fin de cuentas, como dice Rapaport, *un agente C, con competencias de lenguaje natural entiende las emisiones de lenguaje natural de otro agente O, construyendo y manipulando los símbolos de un modelo interno (una interpretación) de la emisión de O, considerada como un sistema formal. El modelo interno de C será una representación de conocimiento y un sistema de razonamiento que manipula símbolos. Por lo tanto, el entendimiento semántico que C hace de O es una empresa sintáctica* (p. 51).

(dominio semántico). Con el paso del tiempo y el uso de las reglas, Searle establecerá asociaciones mentales entre los símbolos chinos y las reglas. Lo que sucederá es que Searle dejará de actuar según las reglas y comenzará a actuar de acuerdo a las reglas, se acostumbrará a ejecutar las acciones que para él eran inicialmente nuevas y desconocidas. Finalmente, aunque encerrado en una habitación no es el mejor modo de aprender un idioma, Searle aprenderá el chino.

La conclusión de todo esto es que si bien el programa que ejecuta la persona dentro de la habitación es inicialmente sintáctico, la implementación de dicho programa permite la construcción de contenido semántico. Esto quiere decir que no es necesario buscar un contenido semántico en el nivel de definición del programa, sino en el nivel de la implementación. A mí parecer esto es una explicación satisfactoria de como la sintaxis y la semántica guardan una estrecha relación, me parece suficiente como para aceptar (2), que no puede haber sintaxis sin semántica o que la sintaxis es suficiente para la semántica. Pero aunque no aceptáramos las explicaciones de Rapaport, aún le quedaría a Searle proporcionar una demostración de porqué solo el sustrato neurológico es capaz de semántica. A veces da la sensación de que al abrir un cerebro Searle ve significados mientras al abrir un computador solo ve simples cómputos que nada significan.

2º El escenario que presenta Searle es poco realista, nada tiene que ver con un computador. Si el sujeto que introducimos en la habitación posee una semántica previa, aunque esta no sea la del chino, el argumento se viene abajo, pues se hace uso constante de dicha semántica, y no puede ser neutral respecto de ella. Para que el argumento tuviera validez Searle debería abstenerse de introducirse a sí mismo en el interior de la habitación, o al menos de introducirse con semántica. Si según el propio Searle un ordenador digital no comprende como los seres humanos comprenden ya que únicamente manipula símbolos de acuerdo a determinadas reglas, y si no hay ningún ser con semántica e intencionalidad consciente en el interior de ningún computador serial, entonces ¿qué sentido tiene crear un ejemplo introduciendo en una habitación un ser consciente con intencionalidad y semántica? Pienso que ninguno. La analogía no es correcta. Y suponiendo que Searle al entrar en la habitación fuera incapaz de comprender su propio idioma, entonces sería incapaz de comprender el manual de instrucciones (el programa) que el argumento requiere para que todo esto tenga sentido, sería incapaz de emitir respuesta alguna. Si introducimos a un Searle a-semántico, entonces lo que sucedería es que al introducir las tarjetas por la ventanilla de entrada, esperaríamos y esperaríamos pero nunca obtendríamos respuesta, nos encontraríamos a un Searle aburrido, ausente e incapaz de comprender absolutamente nada de lo que tiene a su alrededor. Por tanto, el computador no debería entender absolutamente nada, pues si aceptamos su segunda premisa, esto es, que *la sintaxis no es equivalente a la semántica, ni es suficiente por sí misma*, ¿cómo es posible que un computador *comprenda* instrucciones del tipo "*cuando le introduzcamos plantillas de tal tipo deberá remitir plantillas de este otro tipo*" o "*las plantillas de un determinado tipo no contienen una respuesta*

*semántica adecuada, en este caso el sistema deberá informar de un error en el sistema",* etc.? Lo que sucede es que el argumento de Searle no depende tan solo de una imposible separación entre sintaxis y semántica sino del ya ampliamente comentado *punto de vista de la primera persona*[132]*.* Nuevamente sale a relucir ese misterioso lugar del que no nos desprendemos pero tampoco conseguimos aprehender. Con su experimento mental Searle cae una vez más en la trampa cartesiana del fantasma en la maquina detectada hace ya más de sesenta años por Ryle. Pensar en ese hombrecillo en nuestro interior es el gran error. Si lo que pretende Searle es crear un ejemplo plausible de cómo funciona un computador y compararlo con el funcionamiento de un cerebro humano el ejemplo es desafortunado, los cables, chips y conexiones no equivalen a un ser consciente que comprende el inglés, sino a un conjunto de neuronas, axones y conexiones que no comprenden nada por sí mismos. Lo que hace que un ser humano comprenda un idioma es, en todo caso, el sistema en su conjunto, desde las entradas (los estímulos), pasando por los cómputos realizados en el cerebro (procesos de aprendizaje, memoria, etc.), hasta las salidas (respuestas). La cuestión es que sustituir un conjunto de cables, chips y conexiones, que es lo que constituye un computador, por un ser humano consciente y con semántica previa no es una buena idea, no es la manera más adecuada para refutar los supuestos de la Inteligencia Artificial Fuerte. Llegar a la conclusión de que un computador no puede entender nada en virtud de que la persona dentro de la habitación tampoco entiende no es correcto, porque resulta de una asimetría entre los términos de la analogía. Por tanto, para que su argumento fuera válido Searle en el interior de la habitación no debería de entender nada, ¿o acaso pretende decirnos que esos cables, chips y conexiones hacen funcionar un programa que permite al sistema comprender unas determinadas instrucciones pero no comprender el chino? ¿Y si resulta que aprender el chino no es otra cosa que aprender ciertas instrucciones con las que manejar ciertos símbolos? ¿Y si fuéramos capaces de poner en marcha un programa con las instrucciones necesarias como para que de forma recurrente este fuera capaz de aprender, comprender y memorizar cualquier idioma? ¿Y si resultara que ese es precisamente el modo de funcionar de nuestras neuronas y sus conexiones? Ninguna de nuestras neuronas por sí solas es capaz de comprender nuestro idioma nativo, sin embargo, programadas de la manera adecuada, y en conjunto, consiguen realizar los cómputos necesarios como para aprender los significados de cualquier idioma. Searle no nació comprendiendo el inglés, pero sí con el programa que le permitió aprenderlo y comprenderlo, solo necesitó alguien que se lo enseñara, ejecutar el programa y tiempo para aprenderlo, ¿acaso no podemos pensar que sucedería lo mismo con un computador programado del mismo modo?

[132] En sus escritos posteriores a la elaboración del argumento de la habitación china Searle ha pretendido ir más allá afirmando que la sintaxis no es intrínseca a la física del sistema sino que es relativa al observador (Searle, 1996, 2000), lo que según él, dejaría al funcionalismo en una situación todavía más dramática. Esto quiere decir que no es posible *descubrir procesos computacionales en la naturaleza, independientemente de la interpretación humana porque cualquier proceso físico que se pueda encontrar es computacional solamente relativo a alguna interpretación*, por lo que el funcionalismo no es *lo suficientemente materialista* como dice que es, la *computación es un proceso matemático abstracto que existe sólo de manera relativa a la conciencia de los observadores y los intérpretes*. Aquello que dice ser tan materialista, el funcionalismo, se fundamenta, dice Searle, en última instancia en una abstracción matemática. Pienso que todo esto nos lleva a una situación absurda, al solipsismo, a la idea de que el mundo existe solo porque yo existo.

3º Cuando se pregunta "¿Puede pensar una máquina?" Searle responde que es obvio que sí, que nosotros somos precisamente esas máquinas. Cuando se pregunta "¿podría pensar un artefacto, una máquina hecha por el hombre?" Searle responde que si suponemos que puede producirse artificialmente una máquina que posea un sistema nervioso, neuronas con axones y dendritas, y todo lo demás, lo suficientemente semejantes a los nuestros, otra vez la respuesta a la pregunta sería obviamente que sí, que si se pueden duplicar exactamente las causas, podrían duplicarse los efectos, y que de hecho sería posible producir conciencia, intencionalidad, y todo lo demás utilizando algunos otros tipos de principios químicos que no sean los que utilizan los seres humanos ya que se trata, dice, de una cuestión empírica. Cuando se pregunta ¿podría pensar una computadora digital? responde que si por computadora digital nos referimos a cualquier cosa que tenga un nivel de descripción mediante el cual pueda describirse correctamente como la ejemplificación concreta de un programa de computadora, entonces otra vez la respuesta es que sí, que nosotros somos las ejemplificaciones concretas de cualquier número de programas de cómputo y podemos pensar. Sin embargo, cuando la cuestión es: ¿podría algo pensar, comprender, etc., exclusivamente en virtud de ser una computadora con el tipo correcto de programa? ¿Podría la ejemplificación concreta de un programa (del programa correcto, por supuesto) ser por sí misma condición suficiente para la comprensión? Entonces su respuesta es que no, que las manipulaciones de símbolos formales por sí mismas carecen de intencionalidad, que son carentes de sentido, y que ni siquiera son manipulaciones de símbolos, ya que los símbolos no simbolizan nada, sólo tienen sintaxis pero no semántica (Searle, 1980; 98). Searle no parece cansarse de repetir su mantra, "*la sintaxis no es suficiente para la semántica".* Pero el problema aquí es más bien que en el fondo, Searle piensa que solo el cerebro es capaz de producir consciencia, intencionalidad y entendimiento, ya que es incapaz de concebir que estas propiedades sean el producto de la implementación de un simple programa. Para él son algo más, ¿el qué? eso nadie lo sabe, él tampoco. Pero quizás por eso su experimento exige introducir un ser humano consciente en el interior de la habitación para producir en el lector la impresión de que hay algo en nuestra biología y solo en ella que permite el don de significar. Pero obsérvese que Searle supone que es posible responder en chino sin saber chino solo usando unas instrucciones mecánicas, aunque luego afirma que saber chino no es algo mecanizable y de ahí concluye que la intencionalidad, la consciencia y el entendimiento tampoco lo son. Pero para aceptar esto deberíamos mostrar que alguien sabe chino tan bien como un chino y además estar seguros de que no es un ser consciente, mientras tanto, el experimento mental no dice absolutamente nada. No podemos suponer que se puede conversar en chino sin saber chino siguiendo unas instrucciones que no implican saberlo, para luego concluir que entonces no se sabe chino. El problema aquí es que Searle interpreta el cerebro como *portador* de inteligencia consciente y no como su *fundamento causal* (Boden, 1990: 113).

4º El experimento mental de Searle depende, entera e ilícitamente, de que nos imaginemos un caso demasiado simple e irrelevante, para posteriormente extraer de él una conclusión obvia. Lo que sucede es que la simplicidad del experimento no debería permitirnos extraer conclusiones sobre algo tan complejo como nuestras mentes y nuestros cerebros. Dennett (2013: 320) astutamente señala que, para que un sistema como el que Searle postula realmente funcione, tiene que ser muchísimo más complejo y rápido. Searle presenta un sistema en el que le pasan una tarjeta por una rendija, el mismo consulta un libro de instrucciones y emite una respuesta que hace deslizar por una rendija de salida ¡Pero hay trillones de combinaciones posibles en la lengua china, además de trillones de contextos en que se pueden usar las palabras! En un caso como éste, la complejidad y la velocidad importan. Ya no se trataría meramente de una persona que consulta un pequeño manual, sino de un sistema que opera sobre una gigantesca base de datos y que calcula combinaciones casi infinitas. Las limitaciones del experimento mental de Searle deberían colocarnos en alerta, y tener presente que, como muchas otras situaciones imaginarias que se manejan en filosofía también son proclives a ser usadas defectuosamente. Finalmente, el hecho de que podamos llegar a crear programas con mentes similares a las nuestras podría ser sencillamente una cuestión de complejidad, y llegar a esta, solo una cuestión de tiempo[133].

5º La táctica empleada por Searle consiste en apelar una y otra vez al sentido común, a la intuición. Es como si la respuesta más convincente que Searle lanzara a sus críticos fuera algo así como: *"Pero, ¿Cómo puede alguien afirmar que el hombre dentro de la habitación entiende chino? ¡Vamos hombre, no tiene ni idea!"*[134]. Sin embargo, al examinar la historia de la ciencia vemos que ésta no ha sido muy comprensiva con las intuiciones que proponía el sentido común. Los filósofos Patricia y Paul Churchland (1990), por ejemplo, proponen imaginar el modo en que este mismo argumento de Searle hubiera podido utilizarse para refutar la teoría de Maxwell que afirma que la luz está formada por ondas electromagnéticas. El argumento por analogía dice así. Imaginemos un individuo que sostiene un imán en su mano y lo hace oscilar hacia arriba y hacia abajo, el individuo genera una radiación electromagnética, pero no produce luz, por lo tanto, deduciría Searle, la luz no es una oscilación electromagnética. La cuestión es que en el experimento mental se desaceleran las ondas hasta un umbral en el cual los seres humanos no las percibimos como luz. Al confiar en nuestras intuiciones durante el experimento mental, concluimos falsamente que tampoco las ondas rápidas producirán luz[135]. Algo similar sucede en el experimento de Searle, al desacelerar la computación mental, esta no supera el umbral en el que los humanos la

[133] Sobre este asunto: Kurzweil (1999, 2002), Moravec (1988, 1999) o Minsky (1986).

[134] Una cosa es que a Searle le parezca bastante obvio que el sujeto dentro de la habitación no entenderá chino y otra muy diferente que realmente no lo entienda.

[135] El argumento de la sala luminosa, con el que los Churchland pretenden echar por tierra el argumento de la habitación china dice así:
Axioma 1: La electricidad y el magnetismo son fuerzas.
Axioma 2: La propiedad esencial de la luz es la luminancia.
Axioma 3: Las fuerzas, de suyo, no son constitutivas ni suficientes para la luminancia.
Conclusión 1: La electricidad y el magnetismo no son constitutivas ni suficientes para la luminancia.
Hoy sabemos que esta conclusión no se sigue.

consideramos entendimiento (entender es un proceso mucho más rápido). Y al confiar en nuestras intuiciones durante el experimento, concluimos falsamente que la computación rápida tampoco puede ser entendimiento. Pero si imaginamos una versión más acelerada de la historia en la que una persona parece conversar de forma inteligente en chino desarrollando millones de reglas memorizadas en fracciones de segundo, entonces no quedaría tan claro que negásemos que dicha persona entiende chino[136].

6º Podríamos llegar a aceptar que el argumento fuera correcto para un sistema de procesamiento lineal de la información, (el funcionamiento del cerebro es radicalmente diferente al de una maquina digital y serial), pero hoy día tenemos ante nosotros una versión mejorada y más realista de cómo funcionan los cerebros, un nuevo candidato: el conexionismo. Las nuevas redes conexionistas son redes neurales que a diferencia de los programas de software convencionales carecen de un código rígido preestablecido, estas van generando su propio software a medida que interactúan con el entorno mediante un proceso prolongado en el tiempo. Durante este periodo de interacción y aprendizaje el computador va creando su propia semántica liberándose así de las instrucciones formales preestablecidas. Se trata de un lenguaje generado entre la máquina y el entorno, un lenguaje flexible y adaptable, la maquina aprende por sí misma y comprende por sí misma las relaciones entre ella y su entorno. La respuesta de Searle a las redes conexionistas es que estas tienen el mismo problema, carecen de contenido semántico. Dice: "*Cualquier función computable en una maquina paralela puede también computarse en una maquina serial (…) El procesamiento en paralelo no proporciona, pues, una forma de eludir el argumento de la sala china"* (Searle en I y C, 1990: 86). Para ilustrar el fracaso de las redes conexionistas Searle ha concebido un segundo experimento mental, el del gimnasio chino. Consiste en un gimnasio lleno de personas organizadas en una red en paralelo que solo hablan inglés y que se encargan de efectuar las mismas funciones que los nodos y las sinapsis de una arquitectura conexionista. El resultado es el mismo que si hubiera una sola persona manipulando símbolos, nadie en el gimnasio habla ni una palabra de chino, y sin embargo, el gimnasio en su conjunto puede proporcionar respuestas correctas a las preguntas realizadas en chino. Pero este argumento adolece de los mismos problemas que el anterior, no es importante que ninguna de las personas comprenda chino, ninguna de nuestras neuronas por si solas entiende ningún idioma, aunque nuestro cerebro en conjunto sí. También cae de nuevo en la cuestión de la simplicidad, necesitaríamos un gimnasio que albergara $10^{14}$ personas (el cerebro tiene aproximadamente $10^{11}$ neuronas cada una de las cuales realizan algo así como $10^{3}$ conexiones) lo que resulta casi inconcebible, pero además es una imprudencia extraer conclusiones a partir de un ejemplo tan simple, Searle está suponiendo gratuitamente que un gimnasio como el de su experimento mental pero que albergara $10^{14}$ personas carecería de entendimiento, intencionalidad y conciencia, pero carece de bases empíricas para mostrar

[136] Los artículos en los que Searle y los Churchland debaten sobre este asunto, y de donde extraigo estas ideas, fueron publicados en castellano en el nº 162 de la revista Investigación y Ciencia (I y C) en una sección titulada "Un debate sobre inteligencia artificial".

que esto pudiera ser así, se basa única y exclusivamente en una intuición cuyos límites se encuentran en su propia imaginación (Churchlands en I y C, 1990: 95).

Pienso que estas seis reflexiones son suficientes como para abandonar el experimento de la habitación china, demasiados cabos por atar, pero aún me gustaría hacer una última reflexión sobre la posibilidad de que los humanos seamos capaces en un futuro de crear máquinas conscientes[137]. En un principio, atribuimos experiencias conscientes a otros seres humanos y a otros animales porque compartimos una buena parte de nuestra biología, algo que no sucede con las máquinas, por lo que parece que deberemos buscar otra manera diferente de decidir si las máquinas son capaces como nosotros de tener experiencias conscientes. Una línea de investigación prometedora es estudiar que sucede en nuestra biología que nos hace conscientes (sean proteínas, neuronas, funciones o representaciones) y entonces mirar en el interior de un robot para ver si esas propiedades que producen consciencia también están presentes, se trata de una estrategia de inspiración biológica[138]. La principal razón por la que podríamos creer que algún día los robots serán conscientes, es precisamente el hecho de que los seres humanos somos un tipo de robot que es consciente (Dennett, 1994: 1), pero, ¿realmente podría un robot creado por el ser humano poseer experiencias cualitativas? ¿podrá algún día un robot tener experiencias conscientes y ser capaz de comunicarlas? Dennett analiza las posibles razones por las que podemos pensar que un robot nunca podrá llegar a ser consciente, son las siguientes (Dennett, 1994: 2-4):

(1) Los robots son objetos puramente materiales y la conciencia requiere de algún aspecto mental inmaterial (Dualismo anticuado).

Si ese aspecto inmaterial existe la ciencia nunca será capaz de captarlo. Ya he hablado ampliamente de las posiciones dualistas y los problemas con los que se encuentran, ahora solo diré que debemos evitar hasta el extremo postular entidades sobrenaturales. No hay más.

(2) Los robots son inorgánicos (por definición), y la conciencia puede existir solo en cerebros orgánicos.

No hay razones de peso para pensar que esto pudiera ser así, esta razón está impregnada de un vitalismo ya desfasado. No hay tesis científicas ni filosóficas lo suficientemente fuertes como para pensar que la materia orgánica sea un ingrediente necesario para la consciencia.

[137] Aquí analizaré brevemente el artículo de Dennett (1994). Para profundizar en los modelos desarrollados sobre el estudio científico de la conciencia y la posibilidad de máquinas conscientes, véase: Aleksander (2007), Anderson & Oates, (2007); Baars, (1988); Baars y Franklin, (2009), Blackmore, (2002); Chrisley, (2003); Clowes*, et al.*, (2007); Chella, (2009); Dennett, (1995, 2001); Densmore & Dennett, (1999); Gamez, (2008); Haikonen, (2007); Reggia, (2013); Rolls, (2007); Rosenthal, (1991); Sloman & Chrisley, (2003); Sun, (1999); Taylor, (2007); Velmans, (2002, 2009); Xue-Yan Zhang & Chang-Le Zhou (2013)

[138] Aleksander (2003) sugiere hasta cinco axiomas para determinar si un agente es consciente, son: presencia del agente en el mundo físico**,** imaginación y recuerdo de experiencias pasadas y presentes, atención como un determinante del contenido de la experiencia**,** volición (deseos, motivaciones, capacidad de crearse objetivos) y emoción. Si una máquina es capaz de todo ello podemos considerar que es consciente. .

(3) Los robots son artefactos, y la conciencia *aborrece* los artefactos. Sólo algo natural, no fabricado, podría exhibir conciencia genuina.

Ante esto uno podría objetar que un duplicado átomo a átomo de un ser humano, no sería ese ser humano, pero sería absurdo no reconocer que tal artefacto tendría sentimientos, qualia y conciencia, aunque no fuera un ser humano natural y genuino.

(4) Los robots serán siempre demasiado simples como para adquirir conciencia.

Pero esto es decir demasiado. Ciertamente los seres humanos estamos compuestos por billones de células de diferentes tipos y cada célula posee una maquinaria sumamente compleja. Todo hace pensar que tal complejidad es necesaria para la existencia de la conciencia, pero, ¿apostaríamos sobre seguro si afirmáramos que nunca llegaremos a crear robots tan complejos como nosotros? Según Dennett si no encontramos más razones que estas, entonces carece de sentido que nos sigamos manteniendo escépticos ante la posibilidad de robots conscientes en un futuro.

Aunque parece que en esta sección nos hayamos salido del tema que nos interesa, no es así. El problema que encuentra Searle en su argumento está enteramente relacionado con el problema de los qualia de la conciencia, pues Searle parte en todo momento que la IA fuerte nunca distinguirá entre un zombi y un ser con intencionalidad real, intrínseca, inefable, privada y directamente accesible a la conciencia. Esta es la idea que Searle tiene en mente, él lo ve todo desde el punto de vista de la primera persona de la persona que está en el interior de la habitación, atribuyéndole esas intuiciones tan profundamente arraigadas que tanto él como mucha gente tiene acerca de la conciencia y la imposibilidad de que esta sea implementada en una máquina. Sin duda, debemos agradecer a Searle el haber desarrollado su experimento, las cuestiones que ha suscitado son de suma importancia y de una profundidad sin igual, pero creo haber mostrado que de nuevo las fuertes intuiciones que sitúan los rasgos más maravillosos de nuestra existencia fuera del alcance de nuestra comprensión científica están francamente equivocadas, continúan teñidas de cartesianismo, salpicadas por fuertes dosis de antropocentrismo y aderezadas con demasiado biologicismo.

**Conclusión: El fin de una intuición**

---

Hemos visto que la cuestión que se plantea Nagel sobre la imposibilidad de un acercamiento a lo subjetivo desde lo objetivo acaba partiendo el mundo en dos, y esto acaba resultando sumamente problemático para las tesis fisicalistas. Pero también hemos visto que las consecuencias que se extraen de la primera premisa de su argumento, *que los seres humanos no somos capaces siquiera de imaginar lo que se siente al ser un murciélago debido al carácter subjetivo, privado e inefable de la experiencia*, son insostenibles.

Además vimos que un análisis más profundo de la expresión *como que es ser algo* nos sugiere que dicha expresión pudiera contener ciertos elementos reiterativos, como un hábil truco lingüístico que nos hace ver en ella algo significativo, pero que finalmente acaba por no referir claramente a nada. Hemos visto que muy probablemente Mary no aprenderá nada nuevo acerca del mundo cuando salga de su cautiverio monocromático, que no nos hacemos ni la menor idea de lo que significa conocerlo todo acerca de la visión del color, y que además aun en el caso de que Mary aprendiera algo, esto bien pudiera ser una nueva destreza, una nueva habilidad susceptible de ser estudiada física y funcionalmente. Hemos visto que el experimento del espectro invertido no nos proporciona razones suficientes como para pensar que se podría dar inversión fenomenológica sin inversión funcional, y por tanto que no es un experimento que demuestre, de ningún modo, que hay algo así como ciertas propiedades cualitativas que pertenecen a un ámbito en el que la ciencia tiene poco que decir. Hemos visto que el éxito del argumento de los zombis pasa por asumir la dudosa premisa de que lo concebible es posible, y de que los zombis, esos seres físicamente igual a nosotros pero vacíos por dentro, son algo concebible. Hay argumentos sólidos contra ambas premisas. Además, el argumento de los zombis debe superar otros obstáculos, como el sugerido por el análisis condicional y el contra-argumento de los anti-zombis. En esta sección también hemos visto que el problema difícil de la conciencia detectado por Chalmers bien pudiera ser una ilusión teórica fruto de nuestra ignorancia acerca de los problemas supuestamente fáciles, y que bien podría surgir de un error categorial. Hemos visto también que tanto los experimentos basados en el conocimiento como los basados en la concebibilidad pertenecen al ámbito epistemológico, no al ontológico, y que por tanto no afectan al fisicalismo, pues esta es una tesis ontológica. Aunque también hemos visto que esto trae consecuencias inesperadas, ya que parece abrirse una brecha explicativa insalvable entre nuestra descripción física del mundo y el mundo de las sensaciones subjetivas. Algunos no creen que sea posible cerrar tal brecha, otros piensan que podrá cerrarse en un futuro, otros, que aunque esa brecha exista no afecta a los tesis fisicalistas, y otros, que en realidad tal brecha es de nuevo una ilusión teórica, y que la explicación científica bastará para rendir cuentas de cómo se generan los estados fenoménicos (recordemos el método heterofenomenológico sugerido por Dennett). Por último, hemos visto que no hay ningún argumento filosófico lo suficientemente potente como para negar la posibilidad de que en un futuro podamos crear máquinas conscientes. El argumento de la habitación china da por buenas ciertas premisas que difícilmente se sostienen, es un argumento con demasiadas lagunas como para afirmar que la inteligencia artificial fuerte (la idea de que las mentes humanas no son otra cosa que complejos programas con entradas y salidas de datos) no es posible.

Todos estos experimentos fueron diseñados para convencernos de que debe haber una explicación alternativa, una explicación que vaya más allá de la de una mente puramente funcional y material, ahora bien, son los que ven la mente como algo puramente funcional y material los que esperan una explicación

alternativa, que por alguna misteriosa razón, nunca acaba de llegar. Para muchos, el funcionalismo materialista, como tantas otras teorías, está solo de paso por el mundo, piensan que llegará el momento en el que nos percatemos de su insuficiencia y que finalmente tendremos que desecharla como una teoría que aspira a explicar en toda su amplitud la naturaleza y funcionamiento de nuestros estados mentales. Pero harían falta argumentos de mayor calado para proclamar su definitiva defunción, y como hemos podido comprobar, esos argumentos todavía no han llegado, mientras tanto, las ciencias cognitivas y las neurociencias con el funcionalismo materialista como punto de partida, siguen recogiendo sus frutos. No hay ni una prueba concluyente de que esa apuesta de explicación de la mente en base a elementos físicos y con la mirada puesta en los modelos computacionales sea un cadáver filosófico. Aquellos que se apresuran a anunciar la defunción de tan ambiciosa propuesta deberían no solo darnos más razones y más pruebas, deberían también ofrecer alternativas viables de una explicación coherente y científica, pero eso sí, sin postular entidades inefables e indescriptibles, sin postular entidades fantasma, ni moverse en base a corazonadas, presentimientos o intuiciones.

Algunos dirán, "*pero si un fenómeno no entra dentro de los dominios de la ciencia no deberíamos negar la existencia del fenómeno sino ampliar o modificar los dominios en los que la ciencia se mueve para acomodar la ciencia al fenómeno y no el fenómeno a la ciencia*". Pero esto sería cierto si este fuera el caso. Nadie duda de la existencia del fenómeno y de que tenga cabida en una ciencia más madura, nadie duda de que seamos seres conscientes, de lo que se duda es de que ese fenómeno tenga las propiedades que algunos dicen que tiene. La obstinación de algunos científicos y filósofos a lo largo de la historia ha hecho que hoy sepamos que la tierra no es el centro del universo, que la vida no es nada parecido a poseer un élan vital, que la combustión de los cuerpos no se debe a sustancias intrínsecas a estos como el flogisto, o que las brujas, los dioses o las almas inmateriales no existen, por poner solo unos ejemplos. Ha sido mucho el tiempo y las energías perdidas en estos asuntos como para seguir postulando ideas basadas en experimentos mentales que son imposibles desde un punto de vista metafísico, imposibles desde un punto de vista lógico e imposibles desde un punto de vista físico.

Quizás el mayor problema sea que en lo más profundo de nuestra naturaleza sentimos más atracción por los problemas que por las soluciones, nos gusta resolver problemas, pero no queremos que estén todos resueltos, la vida resultaría excesivamente aburrida. Sin embargo, me inclino a pensar que muchas de las grandes cuestiones filosóficas, y esta es una de ellas, ya tienen solución, ya la hemos encontrado, solo nos queda refinar la búsqueda. La idea de un mundo enteramente mecanizable es demasiado arriesgada. Abandonar toda explicación que incluya elementos que habitan fuera del mundo físico exige para muchos dejar de lado lo más preciado de lo que creen que son, les cuesta reconocer que hemos estado inmersos en una profundidad equivocada, pero hasta que no abandonen tal convicción, hasta que no asuman que nuestras más claras y arraigadas intuiciones entran también dentro de aquello de lo que podemos dudar, y

sin duda estudiar, estas cuestiones están abocadas a no salir del espacio ocupado por el misterio, serán y tendrán cabida para siempre en el debate filosófico y científico, ¡esa y no otra es la brecha!

El principal objetivo de este trabajo ha sido colocar a la filosofía de la mente en el lugar que le corresponde, que no es otro que el de una disciplina que no puede ni debe esconderse tras conceptos trasnochados ni obstáculos filosóficos que no hacen otra cosa que impedir un acercamiento claro y coherente al objeto de estudio, que lejos de aclarar, más bien, endurecen el discurso. Su sitio debe ser más bien el de una disciplina cuyo objetivo sea deshacer la opacidad y disolver los misterios que impiden establecer y desarrollar uno de los proyectos más ambiciosos en los que se pueda embarcar el ser humano: desarrollar una ciencia completa y coherente de la mente humana.

**BIBLIOGRAFÍA**

**Aleksander, I. y Barry Dunmall, B. (2003)** Axioms and Tests for the Presence of Minimal Consciousness in Agents. Journal of Consciousness Studies, Vol 10, No. 4-5**.**

**Aleksander, I. (2007)** Modeling Consciousness in Virtual Computation Machines. Synthesis Philosophica, 44: 447-454.

**Alter, T. (1998)** A Limited Defense of the Knowledge Argument. Philosophical Studies 90, pp. 35-56.

**Alter, T. (2006)** Does representationalism undermine the knowledge argument? En T. Alter & S. Walter. Oxford University Press: 65–76.

**Alter, T. (2007)** On the Conditional Analysis of Phenomenal Concepts. Philosophical Studies, 134: 235-253.

**Anderson, M. L., y Oates, T. (2007)** A review of recent research in metareasoning and metalearning. *AI Magazine, 28*(1), 7-16.

**Armstrong, D. M. (1968)** A materialist theory of the mind. Londres, Routledge and Kegan Paul.

**Baars, B. J. (1988)** A cognitive theory of consciousness. Cambridge University Press.

**Baars, B. J. (1997)** In the Theater of Consciousness: Global Workspace Theory, a Rigorous Scientific Theory of Consciousness. Journal of Consciousness Studies, 4 (4): 292-309.

**Baars, B. J., and Franklin, S. (2009)** Consciousness is computational: The LIDA model of global worskspace theory. International Journal of Machine Consciousness, 1(1), 23-32.

**Bailey, A. (2007)** The Unsoundness of Arguments From Conceivability. University of Guelph. Ontario, Canada.

**Ball, D. (2009)** There are not phenomenal concepts. Mind, 118, No. 472: 935-962.

**Balog, K. (1999)** Conceivability, possibility, and the Mind-Body problem. The Philosophical review, Vol. 18, No. 4, pp. 497-528.

**Balog, K. (2009)** Phenomenal concepts. En Brian McLaughin, Ansgar Beckerman y Sven Walter (eds.). Oxford Handbook in the Philosophy of Mind. Oxford University Press, pp. 292-312.

**Balog, K. (2012a)** In defense of the phenomenal concept strategy. Philosophy and Phenomenological research 84.1: 1-23.

**Balog, K. (2012b)** Acquaintance and the Mind-Body Problem. In Identity Theory, Christopher Hill and Simone Gozzano (eds.), Cambridge University Press.

**Bechtel, W. (1991)** Filosofía de la mente. Madrid. Tecnos.

**Bickle, J. (2003)** Philosophy and neuroscience: A ruthlessly reductive account. Dordrecht: Kluwer Academic.

**Blackmore, S. (2002)** There is no stream of consciousness. Journal of Consciousness Studies, 9(5-6), 17-28.

**Block, N. (1978)** Troubles with functionalism. Reimpreso en Block (1980).

**Block, N. (1980)** Readings in the Philosophy of Psychology. Vol. 1, Harvard University Press, Cambridge.

**Block, N. (1990)** Inverted Earth. Philosophical Perspectives, 4**.**

**Block, N. (1994)** Qualia. In S. Guttenplan (Ed.). A Companion to Philosophy of Mind. Oxford: Oxford University Press.

**Block, N. (2002)** The Harder Problem of Consciousness. Journal of Philosophy, 99 (8): 391-425.

**Block, N. (2005)** Two neural correlates of consciousness. Trends Cogn. Sci. 9, 46–52.

**Block, N. (2007a)** Consciousness, accessibility, and the mesh between psychology and neuroscience. Behav. Brain Sci. 30, 481–499.

**Block, N (2007b)** Wittgenstein and Qualia. Philosophical Perspectives 21 (1):73-115.

**Block, N. y Stalnaker, R. (1999)** Conceptual Analysis, Dualism and Explanatory Gap. Philosophical Review, 108 (1): 1-46.

**Boden, M. A. (1988)** Escaping from the Chinese Room. En Margaret Boden, ed., The Philosophy of Artificial Intelligence, New York: Oxford University Press, pp. 89-104. [Trad. Cast. Escape de la habitación china. En Filosofía de la inteligencia Artificial. Margaret Boden, compiladora. México: Fondo de cultura económica. pp. 105 – 121].

**Boterell, A. (2001)** Conceiving What Is Not There. Journal of Consciousness Studies 8: 21-42.

**Braddon-Mitchell, D. (2003)** Qualia and Analytic Conditionals, *The Journal of Philosophy*, 100: 111-135.

**Broad, C. (1925)** The Mind and its Place in Nature. Rutledge and Kegan Paul.

**Brueckner, A. (2001)** Chalmers´s Conceivability Argument for Dualism, Analysis 61: 187- 193.

**Bunge, M. (2004)** Emergencia y convergencia. Novedad cualitativa y unidad del conocimiento. Barcelona, Gedisa.

**Carnap, R. (1932/33)** Psychology in Physical Language. *Erkenntnis*, 3: 107–42.

**Carruthers, P. (2000)** Phenomenal Consciousness. (Cambridge University Press).

**Chalmers, D. (1994)** *A Computational Foundation for the Study of Cognition*. URL= <http://jamaica.u.arizona.edu/~chalmers/papers/computation.html>.

**Chalmers, D. (1995)** Facing up to the problem of consciousness. Journal of Consciousness Studies, 2: 200-19.

**Chalmers, D. (1999)** La mente consciente. En busca de una teoría fundamental. Barcelona. Gedisa.

**Chalmers, D. (2002a)** Consciousness and its Place in Nature. En D.J. Chalmers (ed.), Philosophy of Mind: Classical and Contemporary Readings (Oxford UP, 2002), pp. 247-72.

**Chalmers, D. (2002b)** Does Conceivability Entail Possibility? En T.S. Gendler and J. Hawthorne (eds), Conceivability and Possibility (Oxford UP, 2002), pp. 145-200**.**

**Chalmers, D. (2006)** Phenomenal Concepts and the Explanatory Gap. En Torin Alter y Sven Walter (eds.). Phenomenal Concepts and Phenomenal Knouledge: New Essays on Conciousness and Physicalism. Oxford University Press.

**Chalmers, D. (2010)** The Two-Dimensional Argument against Materialism. En Chalmers, The Character of Consciousness. Oxford University Press.

**Chella, A. (2009)** Machine consciousness: a manifesto for robotics. *International Journal of Machine Consciousness, 1*(1), 33-51.

**Chisholm, R. M. (1957)** Perceiving: A Philosophical Study. Cornell University Press, Ithaca**.**

**Chrisley, R. (2003)** Embodied artificial intelligence. Artificial Intelligence, 149(1), 131-150.

**Churchland, P. (1981)** Eliminative Materialism and the Propositional Attitudes. Journal of Philosophy, 78: 67-90.

**Churchland, P. (1985)** Reduction, Qualia, and the Direct Introspection of Brain States. The Journal of Philosophy, Vol. 82, No. 1 (Jan., 1985), 8-28.

**Churchland, P. (1989)** Knowing Qualia: A Reply to Jackson. En A Neurocomputational Perspective, MIT Press.

**Churchland, P. (1999)** Materia y conciencia. Introducción contemporánea a la filosofía de la mente. Gedisa. Barcelona.

**Churchland, P. (2011)** Consciousness and the introspection of 'qualitative simples. University of California, San Diego, Department of Philosophy. Eidos n° 15 (2011), págs. 12-47.

**Churchland, P. and Churchland, P. S. (1981)** Functionalism, Qualia and Intentionality. Philosophical Topics, 12, 121-132.

**Churchland, P. and Churchland, P. S. (1990)** Could a Machine Think? *Scientific American* 262, pp. 32-39**.**

**Churchland, P. S. (1986)** Neurophilosophy. Cambridge, Mass. M.I.T. Press.

**Churchland, P. S. (1994)** Can neurobiology teach us anything about consciousness? Presidential Address to the American Philosophical Association, Pacific Division. In: Proceedings and Addresses of the American Philosophical Association. Lancaster, PA: Lancaster Press. 67-4: 23-40. Hay versión en castellano.

**Churchland, P. S. (2008)** The Impact of Neuroscience on Philosophy. Neuron 60, November 6, 2008.

**Clark, A. (1993)** Sensory Qualities. Oxford: Oxford University Press.

**Clark, A. (1994)** I am Joe´s Explanatory Gap. The American Philosophical Association symposium on Color Vision and the Explanatory Gap. Pacific Division Meeting. LA. http://selfpace.uccnn.edu/paper/PGAP.HTM

**Clowes, R., Torrance, S., and Chrisley, R. (2007)** Machine consciousness-Embodiment and imagination. Journal of Consciousness Studies, 14(7), 7-14.

**Cohen, M. A. and Dennett, D. (2011)** Consciousness cannot be separated from function. Trends in Cognitive Sciences August 2011, Vol. 15, Nº 8.

**Cole, D. (1984)** Thought and Thought Experiments. Philosophical Studies, 45: 431-44.

**Conee, E. (1994)** Phenomenal Knowledge. Australasian Journal of Philosophy 72: 136-50.

**Cornman, J. W. (1978)** A Nonreductive Identity Thesis about Mind and Body. En Reason and Responsibility, ed. J. Feinberg, Encimo: Dickenson Publishing.

**Crick, F. (1994)** The Astonishing Hypothesis: The Scientific Search for Soul. New York, Scribner. [Trad. Esp.: La búsqueda científica del alma: una hipótesis revolucionaria. Trad. F. Páez de la Cadena, Barcelona, Debate, 1994].

**Crick, F. y Koch, C. (1995)** Are we aware of neural activity in primary visual cortex. Nature 375, 121–123.

**Crick, F. y Koch, C. (1990)** Towards a neurobiological theory of consciousness. Semin. Neurosci. 2, 263–275.

**Crick, F. y Koch, C. (2003)** A framework for consciousness. Nat. Neurosci. 6, 119–126

**Davidson (1980)** Essays on Actions and Events, Oxford University Press, Oxford. [Versión en castellano: Ensayos sobre acciones y sucesos. Trads. Olbeth Hansberg, Jose Antonio Robles y Margarita Valdes. Instituto de Investigaciones Filosóficas. UNAM, Mexico, 1995.

**Dehaene, S. et al. (2006)** Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends Cogn. Sci. 10, 204–211.

**Dehaene, S. y Naccache, L. (2001)** Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework. Cognition, 79 (1): 1-37.

**Dennett, D. (1978)** Brainstorms. Cambridge, Mass.: MIT Press/A Bradford Book.

**Dennett, D. (1980)** The Milk of Human Intentionality. Behavioral and Brain Sciences 3, pp. 429-430.

**Dennett, D. (1987)** The Intentional Stance. Cambridge, MA, MIT Press (Bradford Books). [Trad. cast, de Daniel Zadunisky: La actitud intencional. Barcelona, Gedisa, 1991.]

**Dennett, D. (1988)** Quining Qualia. En Marcel, A. y E. Bisiach (comps.), Consciousness in Contemporary Science, Nueva York, Oxford University Press, pp. 42-77.

**Dennett, D. (1994)** Consciousness in Human and Robot Minds. IIAS Symposium on Cognition, Computation and Consciousness.

**Dennett, D (1995)** La Conciencia Explicada. Paidós. Barcelona.

**Dennett, D. (2001)** Are we explaining consciousness yet? Cognition, 79(1-2), 221-237.

**Dennett, D. (2003)** Who's on first? Heterophenomenology explained. Journal of Consciousness Studies, 10(9–10), 19–30.

**Dennett, D. (2006)** Dulces sueños. Obstáculos filosóficos para una ciencia de la conciencia. Katz. Buenos Aires.

**Dennett, D. (2007)** Heterophenomenology Reconsidered. Phenomenology and Cognitive Science, vol. 6, nos. 1 and 2 (special issue on heterophenomenology, Alva Noë, ed.), pp. 247–270.

**Dennett, D. (2009)** Heterophenomenology. In T. Bayne, A. Cleeremans, and P. Wilken, eds., *The Oxford Companion to Consciousness*. Oxford: Oxford University Press, pp. 345‑346.

**Dennett, D. (2013)** Intuition pumps and other tools for thinking. W.W. Norton and Company. New York.

**Densmore, S., and Dennett, D. (1999)** The virtues of virtual machines. Philosophy and Phenomenological Research, 59(3), 747-761.

**Descartes, R. (2005)** *Meditaciones metafísicas*. Madrid, Alianza.

**Díaz-León, E. (2010)** Can Phenomenal Concepts Explain the Epistemic Gap? Mind 119, 476: 933–951.

**Dretske, F. (1995)** Naturalizing the Mind. Cambridge, Mass.: MIT Press.

**Dreyfus, H. (1972)** What computers can´t do. New York, Basic Books**.**

**Dreyfus, H. y Kelly, D. S. (2007)** Heterophenomenology: Heavy-handed sleight-of-hand. Phenom Cogn Sci 6:45–55.

**Edelman, G. (1989)** The Remembered Present: A Biological Theory of Consciousness. New York, Basic Books.

**Evnine, S. J. (2008)** Kinds and conscious experience: is there anything that it is like to be something? Metaphilosophy. Vol. 39, No. 2, 185-202.

**Farrell, B. (1950)** Experience. Mind, 59, pp. 170- 198.

**Fazekas, P. (2011)** Cognitive Architecture and the Epistemic Gap: Defending Physicalism without Phenomenal Concepts. Published in Philosophia, 39: 21-29.

**Feigl, H. (1967)** The Mental and the Physical. The Essay and a Postscript. University of Minnesota Press. 135-169.

**Feyerabend, P. (1963)** Materialism and the mind-body problem. Review of Metaphysics, 17, 49-67.

**Flanagan, O. (1992)** Conciousness reconsiderd. Mitt Press, Cambridge, Mass.

**Fodor, J. (1975)** The Language of Thought. Cambridge: Harvard University Press. [Trad. cast, de Jesús Fernández Zulaira y revisión de José E. García-Albea: El lenguaje del pensamiento. Madrid, Alianza, 1985.]

**Foster, J. (1989)** A defense of dualism. En J. Smythies and J. Beloff. (eds.). The case for Dualism, University of Virginia Press.

**Foster, J. (1991)** The Immaterial Self: A Defense of the Cartesian Dualism Conception of Mind. Londres. Routledge.

**Foss, J. (1989)** On the Logic of What It Is Like to be a Conscious Subject. Australasian Journal of Philosophy 67, pp. 305-20.

**Frankish, K. (2007)** The anti-zombie argument. Philosophical Quarterly, 57 (299), pp. 650-666.

**Freeman, A. (2006)** Consciousness and Its Place in Nature: Does Physicalism Entail Panpsychism? Exeter, UK, Imprint Academic.

**Frege, G. (1879)** Sobre sentido y referencia. Trad. español U. Moulines, en Valdes Villanueva (comp.), *La búsqueda del significado,* Tecnos, Madrid, 1991.

**Gallager, S. y Zahavi, D. (2013)** La mente fenomenológica. Alianza. Madrid.

**Gamez, D. (2008)** Progress in machine consciousness. *Consciousness and Cognition, 17*(3), 887-910.

**García Suarez, A. (1995)** Qualia: propiedades fenomenológicas. En La mente humana. Ed. Fernando Broncano. Trotta.

**Gibson, J. J. (1979)** The Ecological Approach to Visual Perception. Boston: Houghton Mifflin.

**Gorman, M. (2005)** Nagasawa vs Nagel: Omnipotence, pseudo-tasks, and a recent discussion of Nagel´s doubts about physicalism. Inquiry 48 (5): 436-447.

**Graham, G. y T. Horgan (2000)** Mary Mary, quite contrary. Philosophical Studies, vol. 99, Nº1, 59–87.

**Hacker, P.M.S. (2002)** Is there anything it is like to be a bat? Philosophy 77 (300) pp. 157-174.

**Haikonen, P. O. A. (2007)** Essential issues of conscious machines. Journal of Consciousness Studies, 14(7), 72-84.

**Hampshire, S. (1950)** Critical Notice of Ryle, The Concept of Mind. Mind, LIX, 234, pp. 237-255.

**Hardin, C. (1987)** Qualia and Materialism: Closing the Explanatory Gap. Philosophy and Phenomenological Research, 48: 281–98.

**Hardin, C. (1988)** Color for Philosophers. Indianapolis: Hackett Publishers.

**Hardin, C. (1990)** Color and Illusion. Manuscrito. Presentado en el congreso: The Phenomenal Mind- How is it possible and Why is it Necessary? Bielefeld, Alemania, 14-17 de Mayo.

**Hardin, C. (1997)** Reinverting the Spectrum. En Readings on Color, Volume 1: The Philosophy of Color, A. Byrne and D. R. Hilbert (eds.), Cambridge, MA: MIT Press.

**Harman, G. (1990)** The Intrinsic Quality of Experience. En Toimberlin, J.E. (ed.), 3, pp. 1-53.

**Harman, G. (1996)** Explaining Objective Color in Terms of Subjective Reactions. Philosophical Issues, Vol. 7, pp. 1-17.

**Haukioja, J (2008)** A defense of the conditional analysis of phenomenal concepts. Philosophical Studies, 139: 145-151.

**Hauser, L. (1997)** Searle's Chinese Box: Debunking The Chinese Room Argument. Minds and Machines, Vol. 7, pp. 199-226.

**Hawthorne, J. (2002)** Advice for Physicalists, Philosophical Studies, 108: 17-52.

**Hempel, C. G. (1949)** The Logical Analysis of Psychology. En Block (ed.) (1980), Readings in Philosophy of Psychology, Harvard U.P., Cambridge, vol. 1.

**Hill, C. (1997)** Imaginability, Conceivability, Possibility, and the Mind-Body problem. Philosophical Studies, 87, 61-85.

**Hill, C., y McLaughlin, B. (1999)** There are fewer things in reality that are dreamt of in Chalmers´s philosophy. Philosophy and Phenomenological Research, 59: 445-454.

**Hoffman, D.D. (2006)** The scrambling theorem: A simple proof of the logical possibility of spectrum inversion. Consciousness and Cognition, 15, 31‑45.

**Horgan, T. (1984)** Jackson on Physical Information. Philosophical Quarterly, 34, 147-83.

**Hume, D. (2001)** Tratado de la naturaleza humana. Biblioteca de autores clásicos. Libros en la red. www. Dipualba.es/publicaciones.

**Jackson, F. (1982)** Epiphenomenal Qualia. The Philosophical Quarterly, 32, 127–36.

**Jackson, F. (1986)** What Mary Didn't Know. The Journal of Philosophy, vol. 83, pp. 291–295.

**Jackson, F. (1998)** Postscript on qualia. In Frank Jackson Mind, Method and Conditionals. London: Routledge.

**Jackson, F. (2003)** Mind and illusion. In Anthony O´Hear (ed), Royal Institute of Philosophy Supplement. Cambridge University Press. 421-442.

**Jackson, F. (2006)** The Knowledge Argument, Diaphanousness, Representationalism. In Torin Alter & Sven Walter (eds.), Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism. Oxford University Press. 52--64.

**Janzen, G. (2011)** In defense of the what-it-is-likeness of experience. The Southtern Journal of Philosophy, 49, pp. 271-293.

**Kim, J. (1982)** Psychophysical Supervenience. Philosophical Studies, 41, 51-70.

**Kim, J. (1989)** The Myth of Nonreductive Materialism. Proceedings and Addresses of the American Philosophical Association, Vol. 63, No. 3. pp. 31-47.

**Kim, J. (1990)** Supervenience as a Philosophical Concept. Metaphilosophy Vol. 21, Nos. 1 & 2, January/April 1990.

**Kim, J. (1992)** The nonreductivist's trouble with mental causation. In John Heil & Alfred R. Mele (eds.). Oxford University Press. [En castellano: **Kim, J. (2014)** El fisicalismo no reduccionista y su problema con la causalidad mental. Ideas y Valores, vol. LXIII, núm. 155, pp. 235-259. Universidad Nacional de Colombia. Bogotá, Colombia. http://www.redalyc.org/articulo.oa?id=80931523011]

**Kim, J. (2002)** El problema mente-cuerpo tras cincuenta años. Azafea. Rev. filos. 4, pp. 45-63.

**Kirk, R. (1974)** Zombies vs Materialists. Aristotelian Society Proceedings Supplementary, 48, 135-152.

**Kirk, R. (1996)** Why There Couldn´t Be Zombies. Supplement to the Proceedings of the Aristotelian Society 73: 1-16.

**Kirk, R. (2005)** Zombies and Consciousness, Oxford: Clarendon Press.

**Kirk, R. (2008)** The Inconceivability of Zombies, Philosophical Studies, 139: 73-89.

**Koch, C. (2004)** The Quest for Consciousness. Roberts and Company.

**Koch, C. y Tsuchiya, N. (2007)** Phenomenology without conscious access is a form of consciousness without top-down attention. Behav. Brain Sci. 30, 509–510.

**Kripke, S. (1978)** Identidad y Necesidad. Trad. Español M. Valdes, UNAM, Cuadernos de Crítica, México.

**Kripke, S. (1995)** EI nombrar y la necesidad. Trad. español M. Valdes, UNAM, Mexico.

**Kuna, M. (2004)** The Knowledge Argument and the Refutation of Physicalism. Organon F, 11, Nº 2. Págs. 128-142.

**Kurzweil, R. (1999)** La era de las máquinas espirituales. Ed. Planeta. Barcelona.

**Kurzweil, R. (2002)** Locked in his Chinese Room. In Richards, 128–171.

**Lamme, V.A. (2003)** Why visual attention and awareness are different. Trends Cogn. Sci. 7, 12–18.

**Lamme, V.A. (2006)** Towards a true neural stance on consciousness. Trends Cogn. Sci. 10, 494–501.

**Levin, J. (2008)** Taking Type-B Materialism Seriously. Mind and Language 23 (4): 402-425.

**Levine, J. (1983)** Materialism and Qualia: The Explanatory Gap. Pacific Philosophical Quarterly, vol. 64, pp. 354-361.

**Levine, J. (1993)** On leaving out what it´s like. En M. Davies y G. Humphreys (comps.), Consciousness: Psychological and Philosophical Essays, Oxford, Blackwell.

**Levine, J. (1994)** Out of the closet: A qualophile confronts qualophobia. Philosophical Topics, 22, 107– 126.

**Levine, J. (2001)** Purple Haze. NY: Oxford University Press.

**Lewis, D. (1966)** An argument for the identity theory. Journal of Philosophy, 63, 17-25.

**Lewis, D. (1983)** Mad Pain and Martian Pain. Philosophical Papers, vol. 1. (New York: Oxford; 1983).

**Lewis, D. (1990)** What Experience Teaches. In William G. Lycan (ed.), Mind and Cognition. Blackwell. 29-57.

**Loar, B. (1990-1997)** Phenomenal states. In N. Block, O. Flanagan, and G. Guzeldere eds. The Nature of Consciousness. Cambridge, MA: MIT Press. Pp. 597-616.

**Locke, J. (1690)** Ensayo sobre el entendimiento humano. México, D. F. Fondo de Cultura Económica, 2005.

**Lycan, W. (1973)** Inverted Spectrum. Ratio, vol. XV.

**Lycan, W. (1990)** What is the "Subjectivity" of the Mental? Philosophical Perspectives 11 (2):229-238.

**Lycan, W. (1995)** A Limited Defense of Phenomenal Information**.** In Thomas Metzinger (ed.). Conscious Experience. Imprint Academic. 243-58.

**Lycan, W. (1996)** Consciousness and Experience. Cambridge Mass: The MIT Press, Bradford Books.

**Lycan, W. (2003)** Perspectival Representation and the Knowledge Argument. In Quentin Smith & Aleksandar Jokic (eds.), Consciousness: New Philosophical Perspectives. Oxford University Press. 384.

**Lycan, W. (2009)** Giving Dualism its Due, Australasian Journal of Philosophy. Vol. 89, No 4, pp. 551-563.

**Majeed, R. (2014)** A Priori Conditionals and the Conceivability of Zombies. Philosophical Papers 43 (2):227-253.

**Marcus, E. (2004)** Why Zombies are Inconceivable, Australasian Journal of Philosophy. Vol. 82, No. 3: pp 477-490.

**Marton, P. (1998)** Zombies vs Materialists: The battle for conceivability. Southwest Philosophy Review, 14: 131-138.

**McConnell, J. (1994)** In Defense of the Knowledge Argument. Philosophical Topics 22 (1&2): 157–187.

**McGinn, C. (1989)** Can we solve the mind-body problem? Mind. Vol. 98: 349-366.

**Meyer, U. (2001)** The Knowledge Argument, Abilities, and Metalinguistic Beliefs. Erkenntnis 55 (3):325-347.

**Minsky, M. (1986)** The Society of Mind. Simon and Schuster.

**Mizrahi y Morrow (2015)** Does Conceivability Entail Metaphysical Possibility? Ratio, 28 (1) 1-13.

**Moravec, H. (1988)** Mind Children. Harvard University Press.

**Moravec, H. (1999)** Robot: Mere Machine to Transcendent Mind. New York: Oxford University Press.

**Moural, J. (2003)** The Chinese Room Argument. En J. Searle. Barry Smith, ed. New York: Cambridge University Press. pp. 214 – 260.

**Nagasawa, Y. (2002)** The Knowledge Argument Against Dualism. Theoria Vol. LXIII, pp. 205-223.

**Nagasawa, Y. (2003)** Thomas vs Thomas: a new approach to Nagel´s bat argument. Inquiry, 46, pp. 377-394.

**Nagel, E. (1979)** *The Structure of Science. Problems in the Logic of Scientific Explanation. Hackett Publishing Company. Cambridge.*

**Nagel, T. (1974)** What Is It Like To Be a Bat? Philosophical Review, no. 83: 435-450.

**Nagel, T. (1979)** Panpsychism. En Mortal Questions. Cambridge, UK, Cambridge University Press.

**Nagel, T. (1986)** The view from nowhere. NY. Oxford University Press. [En Cast. Una visión desde ningun lugar. FCE. Mexico].

**Nagel, T. (2014)** La mente y el cosmos. Ed. Biblioteca Nueva. Madrid.

**Nida-Rümelin, M. (1996)** Pseudonormal Vision. An actual case of Qualia Inversion? Philosophical Studies 82: 145-157.

**Nida-Rümelin, M. (1998)** On Belief about Experiences. An Epistemological Distinction Applied to the Knowledge Argument. Philosophy and Phenomenological Research 58 (1998): 51-73.

**Nemirow, L. (1980)** Review of Nagel´s Mortal Questions. Philosophical Review 89: 473-477.

**Nemirow, L. (1990)** Physicalism and the Cognitive Role of Acquaintance. En Lycan, W.G. (ed.), 490-499.

**O'Dea, J. (2002)** The Indexical Nature of Sensory Concepts. En Philosophical Papers, 31:2, 169-181.

**Orlando, E. (1997)** Contenido y conciencia: el debate en torno a los qualia. Dianoia, XLIII.

**Papineau, D. (1993)** Philosophical Naturalism. Oxford: Blackwell.

**Papineau, D. (2002**) Thinking about Consciousness. Oxford: Oxford University Press.

**Papineau, D. (2007)** Phenomenal concepts and perceptual concepts. En A. Torin and S. Walter (eds.). Phenomenal concepts and Phenomenal Knowledge. Oxford: Oxford University Press, 111-144.

**Pauen, M. (2011)** Materialism, Metaphysics and the intuition of Distinctness. Journal of consciousness, Studies 18 (7-8) pp. 71-98.

**Peacocke, C. (1983)** Sense and Content. Oxford: Oxford University Press.

**Penrose, R. (1989)** The Emperor´s New Mind. Oxford, Oxford University Press [Trad. Cast. La nueva mente del emperador. Barcelona, Grijalbo, 1996].

**Perry, J. (2001)** Knowledge, Possibility and Conciousness. MIT Press**.**

**Place, U. T. (1956)** Is Consciousness a Brain Process? British Journal of Psychology, 47.

**Prinz, J. (2007)** Mental Pointing: Phenomenal Knowledge Without Concepts. Journal of Consciousness Studies, 14.

**Putnam, H. (1960)** Minds and machines. En: Hook S, ed. Dimensions of Mind. Albany, NY: New York University Press. Pp. 138-164.

**Putnam (1963)** Mind and behavior. En Philosophical Papers, vol. 2. Cambridge University Press, 1975.

**Putnam, H. (1967a)** The Mental Life of some Machines. Mind, Language and Reality. Philosophical papers, pp. 408-428.

**Putnam, H. (1967b)** The Nature of Mental States. Mind, Language and Reality. Philosophical papers, pp. 429-441.

**Putnam H. (1967c)** Psychological predicates. In: Capitan WH, Merrill DD eds. Art, Mind and Religion. Pittsburgh, PA: University of Pittsburgh Press. 37-48.

**Putnam, H. (1975)** EI significado de significado. Trad. español J.J. Acero, en L. Valdes Villanueva (comp.), La búsqueda del significado, Tecnos, Madrid, 1991.

**Putnam, H. (2001)** La trenza de tres cabos. La mente, el cuerpo y el mundo. Siglo XXI España editores. Madrid.

**Quine, W.O. (1960)** Word and Object. Cambridge. Mass. The MIT Press. [Ed. Cast. Palabra y objeto, Barcelona, Herder, 2001]

**Ramsey, W., S. Stich, and J. Garon (1991)** Connectionism, Eliminativism, and the Future of Folk Psychology. En Greenwood, 93-119.

**Rapaport, W. (1986)** Searle´s Experiments with Tought. Philosophy of Science, Vol. 53, No. 2, 271-279.

**Rapaport, W. (1995)** Understanding Understanding: Syntactic Semantics and Computational Cognition. Philosophical Perspectives, Vol 9, Issue AI, Connectionism and Philosophical Psychology. Pp. 49 – 88.

**Reggia, J. (2013)** The rise of machine consciousness: Studying consciousness with computational models. Neural Networks, 44: 112-131.

**Rey, G. (1983)** A Reason for doubting the Existence of Consciousness. En R. Davidson, G. Schwartz and D. Shapiro (eds.) Consciousness and self-regulation Vol. 3. New York, Plenum: 1-39.

**Rey, G. (1995)** Toward a Projectivist Account of Conscious Experience. En Thomas Metzinger (ed.). Conscious Experience. Ferdinand Schoningh. Pps. 123-142.

**Robinson, H. (1982)** Matter and Sense: A Critique of Contemporary Materialism. Cambridge University Press.

**Robinson, H. (1993)** Dennett on the Knowledge Argument. Analysis 53, pp. 174-77.

**Rolls, E. T. (2007)** A computational neuroscience approach to consciousness. Neural Networks, 20(9), 962-982.

**Rorty, R. (1970)** In defense of Eliminative Materialism. The review of Metaphysics XXIV, 1, pp. 329-359.

**Rosenthal, D. M. (1991)** The nature of Mind. Oxford University Press.

**Rosenthal, D. M. (1996)** A theory of consciousness. En N. Block, O. Flanagan y G. Güzeldere (comps.). The nature of consciousness. Cambridge. Mass., MIT Press.

**Russell, B. (1991)** Los problemas de la filosofía. Ed. Labor, Barcelona.

**Ryle, G. (2005)** El concepto de lo mental. Barcelona. Paidós.

**Searle, J. (1980)** Minds, Brains, and Programs. En M. Boden (1990), *The philosophy of Artificial Intelligence*, Oxford University Press, New York, 1990, 67-88. Searle, J. Mentes, Cerebros y Programas. En *Filosofía de la inteligencia Artificial.* Margaret Boden, compiladora. México: Fondo de cultura económica, 1990. pp. 82 – 104.

**Searle, J. (1984)** Minds, Brains and Science. Cambridge. Mass., Harvard University Press.

**Searle, J. (1990)** Is the brain a digital computer? Proceedings and addresses of the American Philosophical Association 64: 21-37.

**Searle, J. (1992)** Intencionalidad. Un ensayo en la filosofía de la mente. Tecnos, Madrid.

**Searle, J. (1996)** El redescubrimiento de la mente. Madrid. Ed. Crítica.

**Searle, J. (2000)** El misterio de la conciencia. Barcelona, Paidós.

**Searle, J. (2002)** Why I Am Not a Property Dualist. Journal of Consciousness Studies, 9, No. 12, pp. 57–64.

**Searle, J. (2004)** Libertad y neurobiología: reflexiones sobre el libre albedrio, el lenguaje y el poder político. Ed Paidos Ibérica. Barcelona.

**Searle, J. y Churchland, P. y Churchland, P. S. (1990)** Un debate sobre inteligencia artificial. Investigación y Ciencia, nº 162. Pp. 82-96.

**Sellars, W. (1956)** Empiricism and the Philosophy of Mind. En Feigl H. y Scriven M. (eds) The Foundations of Science and the Concepts of Psychology and Psychoanalysis: Minnesota Studies in the Philosophy of Science, Vol. 1. Minneapolis: University of Minnesota Press: 253–329.

**Shaffer, M. J. (2009)** A Logical Hole in the Chinese Room. Minds and Machines, 19 (2): 229-235.

**Schiffer, S. (1996)** Language-created Language-independent Entities. Philosophical Topics 24: 149-167.

**Shoemaker, S. (1975)** Functionalism and Qualia. Philosophical Studies. Vol 27, Issue 5, pp. 291-315.

**Shoemaker, S. (1982)** The Inverted Spectrum. The Journal of Philosophy, Vol. 79, No. 7. pp. 357-381.

**Shoemaker, S. (1981)** Absent qualia are impossible—A replay to Block. The Philosophical Review, 90, 581-599.

**Shoemaker, S. (1991)** Qualia and Consciousness. Mind, vol. 4.

**Shoemaker, S. (1994a)** Phenomenal Character. Noûs 28 (1): 21-38.

**Shoemaker, S. (1994b)** Lecture III: The phenomenal character of experience. Self-knowledge and inner sense. Philosophy and Phenomenological Research 54 (2):291-314.

**Shoemaker, S. (1996)** Colors, Subjective Reactions, and Qualia. En Enrique Villanueva (ed.), Philosophical Issues. Atascadero: Ridgeview. 55-66.

**Shoemaker, S. (1999)** On David Chalmers´s 'The Conscious Mind', Philosophy and Phenomenological Research 59: 439-444.

**Sloman, A. y Chrisley, R. (2003**) Virtual machines and consciousness. Journal of Consciousness Studies, 10: 133-172.

**Smart, J. J. C. (1958)** Sensations and Brain Processes. Philosophical review. Vol. 68.

**Snowdon, P. (2010)** On the what-it-is-like-ness of experience. The southern Journal of Philosophy. Vol. 48. Issue 1. 8-27.

**Soldati, G. (2007)** Subjectivity in heterophenomenology. Phenom Cogn Sci, 6: 89-98.

**Stalnaker, R (2002)** What is it like to be a zombie? In T. S. Gendler & J. Hawthorne (eds.), Conceivability and Possibility. Oxford: Oxford University Press.

**Stoljar, D. (2001)** Two Conceptions of the Physical, Philosophy and Phenomenological Research, 62: 253-281.

**Stoljar, D. (2005)** Physicalism and Phenomenal Concepts. Mind and Language, 20 (2): 296-302.

**Stoljar, D. (2014)** The Semantics of 'What it's like' and the Nature of Consciousness. Mind, forthcoming.

**Stoljar, D. & Nagasawa, J. (2003)** 'Introduction'; There's Something About Mary. Cambridge, MIT Press.

**Sturgeon, S. (2000)** Matters of Mind: Consciousness, reason and nature. London and New York: Routledge.

**Strawson, G. (2006)** Panpsychism? Reply to commentators with a celebration of Descartes. *Journal of Consciousness Studies* 13 (10-11):184-280.

**Sun, R. (1999)** Accounting for the computational basis of consciousness. Consciousness and Cognition, 8: 529-565.

**Sundström, P. (2002)** An argument against spectrum inversion. In: Physicalism, Consciousness, and Modality: Essays in the Philosophy of Mind, Sten Lindström and Pär Sundström (eds.), (Umeå: Department of Philosophy and Linguistics), 65-94.

**Taylor, J. (2007)** CODAM: a neural network model of consciousness. Neural Networks, 20: 983-992.

**Thagard, P. (1986)** The Emergence of Meaning: How to Escape Searle's Chinese Room. Behaviorism, Vol. 14, No. 2, pp. 139-146.

**Thomasson, A. (2001)** Ontological minimalism. American Philosophical Quarterly 38: 319-331.

**Trillas, E. (1998)** La Inteligencia Artificial. Máquinas y personas. Ed. Debate.

**Turing, A. (1950)** Computing Machinery and Intelligence. Mind, 59, pp. 433-460.

**Tye, M. (1986)** The Subjective Qualities of Experience. Mind, New Series, Vol. 95, No. 377 (Jan., 1986), pp. 1-17.

**Tye, M. (1995)** Ten problems of consciousness. Cambridge (Ma.): MIT Press.

**Tye, M. (1999)** Phenomenal Consciousness: The Explanatory Gap as a Cognitive Illusion. En *Mind*, 108: 432, 705-725.

**Tye, M. (2000)** Consciousness, color, and content. Cambridge (Ma.): MIT Press.

**Tye, M. (2004)** Knowing What It Is Like: The Ability Hypothesis and the Knowledge Argument. En Ludlow, P., Nagasawa, Y., and Stoljar, D. (eds.), 2004, There's Something About Mary, MIT Press.

**Tye, M. (2005)** Qualia. Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/qualia.

**Tye, M. (2006)** Absent Qualia and the Mind-Body Problem. Philosophical Review, Vol. 115, No. 2. 139-168.

**Tye, M. (2009)** Consciousness Revisited: Materialism Without Phenomenal Concepts. MIT Press.

**Van Gulick, R. (1985)** Physicalism and the subjectivity of mental. Philosophical Topics. 51-70.

**Van Gulick, R. (1993)** Understanding the Phenomenal Mind: Are We All just Armadillos? In M. Davies and G. Humphreys, eds., *Consciousness: Psychological and Philosophical Essays*. Oxford: Blackwell, 137-54.

**Van Gulick, R. (1999)** Toward a Science of Consciousness III. The Third Tucson Discussions and Debates. Edited by Stuart R. Hameroff, Alfred W. Kaszniak, and David J. Chalmers. A Bradford Book. The MIT Press. Pgs. 13-23.

**Van Gulick, R. (2004)** So Many Ways of Saying No to Mary. Published in P. Ludlow, Y. Nagasawa & D. Stoljar (eds.) There´s something about Mary: Essays on Frank Jackson's Knowledge Argument, Cambridge, MA: MIT Press, pp. 365-405.

**Van Gulick, R. (2009)** Jackson's change of mind: representationalism, apriorism and the knowledge argument. En Ian Ravenscroft (ed.), Minds, Ethics, and Conditionals: Themes from the Philosophy of Frank Jackson. Oup Oxford.

**Varela, F y Shear, J. (1999)** First-person methodologies: what, why, how? En Journal of Consciousness Studies, vol. 6, Nº 2-3, pp. 1-14.

**Velmans, M. (2002)** Making sense of causal interactions between consciousness and brain. Journal of Consciousness Studies, 9 (11): 69-95.

**Velmans, M. (2009)** How to define consciousness: And how not to define consciousness. Journal of Consciousness Studies, 16 (5): 139-156.

**Wilkes, K. (1988)** Yishi, Duh, Um and Consciousness. En Marcel, A. and Bisiach, E. (eds.), Consciousness in Contemporary Science. Oxford: Oxford University Press.

**Wittgenstein, L. (1958/1988**) Investigaciones filosóficas. Barcelona: UNAM/Crítica.

**Wue-Yan Zhang y Chang-Le Zhou (2013)** From Biological Consciousness to Machine Consciousness: An Approach to Make Smarter Machines. International Journal of Automation and Computing, 10 (6): 498-505.

**Yablo, S. (1993)** Is Conceivability a Guide to Possibility? Philosophy and Phenomenological Research. Vol. LIII, No. 1.

**Yablo, S. (1999)** Concepts and Consciousness. Philosophy and Phenomenological Research 59: 455-463.

**Yetter-Chappell, Y. (2013)** Circularity in the Conditional Analysis of Phenomenal Concepts. Philosophical Studies 165 (2): 553-572.

Brain Sci. 30, 509–510

**Zahavi, D. (2007)** Killing the strawman: Dennett and Phenomenology. Phenomenology and the cognitive science 6/1-2, 21-43.

**Zeki, S. (2001)** Localization and globalization in conscious vision. Annu. Rev. Neurosci. 24, 57–86

**Zeki, S. (2003)** The disunity of consciousness. Trends Cogn. Sci. 7, 214–218

**Zeki, S. y Bartels, A. (1999)** Toward a theory of visual consciousness. Conscious. Cogn. 8, 225–259.

**Zoglauer, T. (1999)** Qualiaphobia: Paul Churchland's Critique of the Knowledge Argument. In Julian Nida-Rümelin (ed.), Rationality, Realism and Revision. 536--542.